Summer 2010

The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups

Sarah Lynn Hagge
University of Iowa

THE IMPACT OF EQUATING METHOD AND FORMAT REPRESENTATION OF COMMON ITEMS ON THE ADEQUACY OF MIXED-FORMAT TEST EQUATING USING NONEQUIVALENT GROUPS

by

Sarah Lynn Hagge

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

July 2010

Thesis Supervisor: Professor Michael J. Kolen

ABSTRACT

Mixed-format tests containing both multiple-choice and constructed-response items are widely used on educational tests. Such tests combine the broad content coverage and efficient scoring of multiple-choice items with the assessment of higher-order thinking skills thought to be provided by constructed-response items. However, the combination of both item formats on a single test complicates the use of psychometric procedures. The purpose of this dissertation was to examine how characteristics of mixed-format tests and composition of the common-item set impact the accuracy of equating results in the common-item nonequivalent groups design.

Operational examinee item responses for two classes of data were considered in this dissertation: (1) operational test forms and (2) pseudo-test forms that were assembled from portions of operational test forms. Analyses were conducted on three mixed-format tests from the Advanced Placement Examination program: English Language, Spanish Language, and Chemistry.

For the operational test form analyses, two factors of investigation were considered as follows: (1) difference in proficiency between old and new form groups of examinees and (2) relative difficulty of multiple-choice and constructed-response items. For the pseudo-test form analyses, two additional factors of investigation were considered: (1) format representativeness of the common-item set and (2) statistical representativeness of the common-item set. For each study condition, two traditional equating methods, frequency estimation and chained equipercentile equating, and two item response theory (IRT) equating methods, IRT true score and IRT observed score methods, were considered.

There were five main findings from the operational and pseudo-test form analyses. (1) As the difference in proficiency between old and new form groups of examinees increased, bias also tended to increase. (2) Relative to the criterion equating

relationship for a given equating method, increases in bias were typically largest for frequency estimation and smallest for the IRT equating methods. However, it is important to note that the criterion equating relationship was different for each equating method. Additionally, only one smoothing value was analyzed for the traditional equating methods. (3) Standard errors of equating tended to be smallest for IRT observed score equating and largest for chained equipercentile equating. (4) Results for the operational and pseudo-test analyses were similar when the pseudo-tests were constructed to be similar to the operational test forms. (5) Results were mixed regarding which common-item set composition resulted in the least bias.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

THE IMPACT OF EQUATING METHOD AND FORMAT REPRESENTATION OF COMMON ITEMS ON THE ADEQUACY OF MIXED-FORMAT TEST EQUATING USING NONEQUIVALENT GROUPS

by

Sarah Lynn Hagge

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

July 2010

Thesis Supervisor: Professor Michael J. Kolen

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

Sarah Lynn Hagge

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 July 2010 graduation.

Thesis Committee: _____
Michael J. Kolen, Thesis Supervisor

_____
Donald B. Yarbrough

_____
Robert L. Brennan

_____
Won-Chan Lee

_____
Mary Kathryn Cowles

ACKNOWLEDGMENTS

As I wrote my acknowledgements, I felt so blessed to think of all the people who have supported, loved, encouraged, and mentored me. Although I cannot individually thank everyone who has had an influence in my life, I want to acknowledge that I would not be where I am or who I am without the people who have been beside me.

I want to first thank my committee for their guidance and encouragement through the dissertation process. Thank you to my chair, Dr. Michael Kolen, for inspiring my interest in equating and for being a remarkable mentor who was always willing to answer questions, provide feedback, or offer encouragement. My academic advisor throughout graduate school, Dr. Don Yarbrough, has always expressed a genuine interest in and concern for both my academic and non-academic endeavors. It is because of his encouragement to expand my research and career interests rather than limit them that I have explored many opportunities and found the place that was right for me. Through his insightful comments and attention to detail, Dr. Robert Brennan modeled how to think as a researcher and taught me how to more fully develop my research ideas. Dr. Won-Chan Lee's office was always open to answer my questions or discuss ideas. His thorough and thoughtful reading of my dissertation helped me to clarify ideas and develop stronger research. Last, I am continually inspired by Dr. Kate Cowles's knowledge about and authentic excitement for statistics. Not only was she enthusiastic about what she taught, but she also taught her students how to apply lessons learned in the classroom for present research and future jobs.

I am thankful to Dr. Kris Waltman, my research advisor for the first three years of graduate school, for taking a chance on me as a new graduate student. Through countless hours of mentoring, she taught me how to think about and understand data, develop research ideas and questions, write technically, attend to details, and always think

critically. Most importantly, she listened, offered advice when needed, encouraged, and celebrated.

Without Anna Marie Guengerich, our department librarian, much of the research I have done in graduate school would not have been possible. She searched out any test, article, or book I needed, happily going far beyond expectations. Her sincere excitement for and interest in both academic and personal successes made even the small accomplishments important.

I appreciate the many friends who have supported me through graduate school. Their phone calls, cards, emails, and words of wisdom often carried me through the most challenging days. In particular, thank you to Messus for staying up with me long nights to finish our papers in college and our books in graduate school. I am grateful to Katherine Furgol and Karoline Jarr for studying with me countless hours, being sweet companions when there was no studying to be done, and getting me out of the office to go on Iowa adventures when I needed a break. April Gonzalez, Scott Wood, and Brad Brossman, have been there from the beginning of this graduate school adventure, studying together, helping me understand the tough concepts, and providing comic relief.

I especially want to thank the friends and colleagues I have shared 204 Lindquist Center with throughout my time in graduate school. They have never tired of answering questions, teaching me new software, providing encouragement, celebrating milestones, studying together, or assuring me I was not just an imposter. Also, thank you to Chunyan Liu, Sonya Powers, Benjamin Andrews, and Katherine Furgol for helping me develop the ideas and software code for my dissertation.

Finally, my family deserves so much more than can be expressed through a simple “thank you”. Laurel and Kirsten, my dear sisters, have been incredibly patient with me as I have pursued degree after degree. They have listened attentively as I relayed my fears and stressors, offered words of encouragement, and opened their homes to care for me as I traveled home to visit. I could not have been blessed with two better sisters. I

also thank my grandparents, Grandma Smith, Grandma Helen, and Grandpa, for loving me unconditionally and caring for me tirelessly. No matter how big or small the accomplishment, they have always been proud. And, when the task seemed insurmountable, a little friendly bet gave me the drive to finish what I had started. (I'll be waiting for my homemade apple pie, Grandpa.)

My parents, Jim and Susan Hagge, have been patient through many years of education. They have always encouraged me to be whatever I wanted to be and have helped make my dreams reality. Dad, thank you for stopping to answer all of the questions of my inquisitive young mind, even though there was work to be done. Some of the most important things I have learned in life I learned while working with you on the farm or riding with you in the grain truck. Mom, thank you for spending late nights helping me to finish school projects in elementary and high school. Through college and graduate school, your emails, cards, and phone calls with prayers and words of encouragement got me through many tests, papers, presentations, and important meetings. I cannot thank you both enough for instilling in me humility, conscientiousness, a strong work ethic, and the persistence to complete the tasks I have started. I know it is your love and prayers that have brought me to where I am today.

Last, thank you to my fiancé, Burgess Smith, for taking on this graduate school journey of mine as though it were his own. He has cared for me, lovingly supported me through the hard times, excitedly celebrated the accomplishments, and patiently waited for the journey to be complete.

ABSTRACT

Mixed-format tests containing both multiple-choice and constructed-response items are widely used on educational tests. Such tests combine the broad content coverage and efficient scoring of multiple-choice items with the assessment of higher-order thinking skills thought to be provided by constructed-response items. However, the combination of both item formats on a single test complicates the use of psychometric procedures. The purpose of this dissertation was to examine how characteristics of mixed-format tests and composition of the common-item set impact the accuracy of equating results in the common-item nonequivalent groups design.

Operational examinee item responses for two classes of data were considered in this dissertation: (1) operational test forms and (2) pseudo-test forms that were assembled from portions of operational test forms. Analyses were conducted on three mixed-format tests from the Advanced Placement Examination program: English Language, Spanish Language, and Chemistry.

For the operational test form analyses, two factors of investigation were considered as follows: (1) difference in proficiency between old and new form groups of examinees and (2) relative difficulty of multiple-choice and constructed-response items. For the pseudo-test form analyses, two additional factors of investigation were considered: (1) format representativeness of the common-item set and (2) statistical representativeness of the common-item set. For each study condition, two traditional equating methods, frequency estimation and chained equipercentile equating, and two item response theory (IRT) equating methods, IRT true score and IRT observed score methods, were considered.

There were five main findings from the operational and pseudo-test form analyses. (1) As the difference in proficiency between old and new form groups of examinees increased, bias also tended to increase. (2) Relative to the criterion equating

relationship for a given equating method, increases in bias were typically largest for frequency estimation and smallest for the IRT equating methods. However, it is important to note that the criterion equating relationship was different for each equating method. Additionally, only one smoothing value was analyzed for the traditional equating methods. (3) Standard errors of equating tended to be smallest for IRT observed score equating and largest for chained equipercentile equating. (4) Results for the operational and pseudo-test analyses were similar when the pseudo-tests were constructed to be similar to the operational test forms. (5) Results were mixed regarding which common-item set composition resulted in the least bias.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

3PL	Three-parameter logistic model
AP	Advanced placement
CE	Chained equipercentile equating method
CI	Common item(s)
CI 0.00	Common item effect size equals 0.00
CI 0.20	Common item effect size equals 0.20
CI 0.40	Common item effect size equals 0.40
CI 0.60	Common item effect size equals 0.60
CI ES	Common item effect size
CINEG	Common-item nonequivalent groups design
CO	Composite score(s)
CR	Constructed response
CSE	Conditional standard error of equating
DTM	Difference that matters
ES	Effect size(s)
FCR	Common item set contains MC and CR items (**F**ull **C**onstructed **R**esponse)
FE	Frequency estimation equating method
GPCM	Generalized partial credit model
IRT	Item response theory
MC	Multiple choice
MC-CR 0.00	Difference between the MC and CR effect sizes for the study conditions are similar to the difference for the criterion
MC-CR 0.25	Difference between the MC and CR effect sizes for the study conditions are different from the difference for the criterion
MSE	Mean squared error
NCR	Common item set contains only MC items (**N**o **C**onstructed **R**esponse)
NCR MT	NCR mini-test (CI set is constructed to be a mini version of the total test)
NCR SM	NCR semi-miditest (CI set is constructed to be a semi-miditest)

NCR DS	NCR difficulty shift (CI set is constructed to be harder than the total test)
NEAT	Nonequivalent groups with anchor test design
OS	IRT observed score equating method
PCA	Principal components analysis
REMSD	Root expected mean square difference
RMSD	Root mean square difference
RMSE	Root mean squared error
TS	IRT true score equating method
WARMSB	Weighted average root mean squared bias
WARMSE	Weighted average root mean squared error
WASE	Weighted average standard error of equating

CHAPTER ONE:
INTRODUCTION

For most of the 20th century, multiple-choice (MC) items were the item format most commonly used for standardized tests (Koretz & Hamilton, 2006), such as the *Army Alpha* used in World War I, college admissions tests like the *Scholastic Aptitude Test*, and tests designed to measure K-12 student learning, such as *the Iowa Tests of Basic Skills*. The inclusion of constructed-response (CR) items on standardized tests appears at first glance to be a relatively recent trend fueled by the current educational accountability climate. However, for the field of educational measurement, CR items are nothing new. In the third edition of Educational Measurement, Whitney (1989) briefly discussed that prior to around 1920 and the development of the *Scholastic Aptitude Test*, college admissions tests were created locally and comprised of essays. CR items were the item format primarily used on these assessments. Although MC items were widely used for the rest of the 20th century, the conversation about CR items continued among educational measurement professionals. Lindquist (1951) argued in the first edition of *Educational Measurement* that "the most important consideration is that the test questions require the examinee to do the *same* things, *however complex*, that he is required to do in the criterion situations (p. 154). Ebel (1980) contended that "While each type [of examination] has its own special values and limitations, they are largely interchangeable. The quality of an examination depends less on the particular form used than on the skill with which it is used" (p. 124).

In the late 1980s and 1990s, MC items became less favorable, and CR items made their way back onto the testing scene. According to Koretz and Hamilton (2006), there were three primary motivations for the reintroduction of assessments comprised of CR item formats. Assessments with CR item formats were believed to assess higher-order thinking skills, shape classroom instruction better than MC, and reduce test score

inflation presumed to result from test-preparation practices. Many testing programs now widely use both MC and CR item formats in mixed-format tests. By using both item formats in the same test, tests are believed to combine the high reliability and broad content coverage afforded by MC items while using CR items to closely mimic real-life situations and invoke complex levels of thought. Although mixed-format tests combine the positive aspects of both item formats, they also combine the challenges of both formats. Arguing for the merits of MC or CR items is beyond the scope of this dissertation. Rather, this dissertation acknowledges that mixed-format tests are, and will likely continue to be, widely used. The goal of this dissertation is to understand the impact that the inclusion of CR items has on current psychometric methods. Specifically, this dissertation examines how the characteristics of mixed-format tests might adversely impact equating and aims to explore test characteristics that may lead to satisfactory equating with mixed-format tests.

Definitions and Characteristics of MC and CR Items

MC items are a type of selected response item with a question or statement stem followed by possible answer choices (Ferrara & DeMauro, 2006). The possible answer choices include one correct answer and a number of incorrect answer choices (Ebel & Frisbie, 1991). Many of the strengths of MC items lie in efficiency of administration and scoring. However, MC items have been criticized for their perceived inability to assess higher-order thinking skills and curricular overemphasis on MC specific test-taking strategies (Ferrara & DeMauro, 2006).

CR items require examinees to produce an answer or product without existing answer choices. CR items can be classified into many different categories, such as short answer, products, performances, completion, or construction (Bennett, 1993; Ferrara & DeMauro, 2006). Some of the strengths of CR items include that they are easy to create relative to MC items and place a curricular emphasis on writing and away from

memorization and recall (Ebel & Frisbie, 1991; Ferrara & DeMauro, 2006; Clauser, Margolis, & Case, 2006). However, CR items are more expensive to score, easier to memorize, and typically have lower reliability compared to MC items. Fewer tasks can be administered during an administration period, resulting in limited sampling of the content domain (Ferrara & DeMauro, 2006; Ebel & Frisbie, 1991).

A balance between MC and CR items appears to have been found with mixed-format tests that contain both MC and CR items. By including each item format on a test, some of the weaknesses of each item format are mitigated by the strengths of the other. However, the combination of both item formats on the same test introduces other potential problems for existing psychometric methodologies, such as multidimensionality or how to appropriately weight MC and CR items to create composite scores. Of particular interest for this dissertation is the way in which characteristics of mixed-format tests impact equating.

Definition of Equating

"*Equating* is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content" (Kolen & Brennan, 2004, p. 2). Different forms of the same test are rarely, if ever, created perfectly parallel. Consequently, test forms almost always differ in difficulty, necessitating the use of equating. Without equating, if test forms differ in difficulty, examinees from one administration may be unfairly advantaged or disadvantaged relative to examinees taking a test form from a different test administration.

A variety of data collection designs and methodologies for equating have been developed. Three data collection designs are commonly used for equating: single group design (with and without counterbalancing), random groups design, and the common-item nonequivalent groups (CINEG) design, which is sometimes referred to as the

nonequivalent groups with anchor test (NEAT) design. These three equating designs are described in detail in Chapter Two. The focus of the current dissertation is on the CINEG design. In the CINEG design, some items are selected to be administered on two test forms. These items are referred to as common items, or the items are also referred to as an anchor test. Two groups of examinees that are not assumed to be equivalent in overall proficiency take one of two test forms. The common items allow for separation of differences in examinee score distributions across the two test forms into form difficulty and examinee proficiency. The CINEG design is very flexible and widely used, but it requires strong statistical assumptions and careful development of a set of common items. A number of equating methodologies have been developed for use with the CINEG equating design. This dissertation considers two traditional and two item response theory (IRT) equating methods: frequency estimation (FE), chained equipercentile (CE), IRT true score (TS), and IRT observed score (OS). A detailed description of these methods is presented in Chapter Two.

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

One of the most crucial components of the CINEG design is the careful development of common-item sets that accurately reflect the total test. For MC tests, Kolen and Brennan (2004) discussed desirable characteristics of common-item sets. They indicated that the common items should be a "mini version" of the total test. The common items should be "proportionally representative of the total test forms in content and statistical characteristics" (Kolen & Brennan, 2004, p. 19). Additionally, the common items should remain unchanged from the old form to the new form and be placed in the same position on both forms. Recent research suggests statistical guidelines for common items could be relaxed for MC-only tests (e.g., Sinharay and Holland, 2007).

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

The challenges in using CR items as common items for mixed-format tests have resulted in many testing programs electing to use MC-only common items for mixed-format tests. Therefore, it is important to understand which characteristics of a mixed-format test might impact equating results with MC-only common items. One important test characteristic to consider is multidimensionality. Multidimensionality occurs when a test measures more than one latent construct. Although the possibility for multidimensionality exists in any test, concerns are heightened for mixed-format tests where item formats are intended to measure different content and/or processes. If multidimensionality is present in a mixed-format test, MC-only common items may not accurately reflect the total test characteristics. Therefore, the extent to which multidimensionality exists in mixed-format tests and the extent to which multidimensionality affects equated scores are two pertinent problems to consider. Other

test or examinee characteristics, such as the ratio of MC to CR points, test length, or examinee proficiency on MC items may interact with multidimensionality. A more in-depth overview of the guidelines for selecting common items and potential problems associated with equating mixed-format tests in the CINEG design is presented in Chapter Two.

Equating Research Data Classes

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

A third class of data is pseudo-test forms. Pseudo-test forms are created by splitting the items from one test form from an operational administration in half to create two test forms. A pseudo-test form uses an operational test form and creates a reasonable criterion, because data exist for the same examinees on two pseudo-test forms. Additionally, pseudo-test forms allow the researcher to manipulate the composition of items on the test forms. Of particular importance for this dissertation is that the common items on the original operational test forms contain only MC items. Pseudo-test forms

could be created so that the common items contain both MC and CR items. However, pseudo-test forms may not accurately represent operational test forms for a number of reasons. Pseudo-test forms are shortened forms of the original test. Further, pseudo-test forms are intended to be parallel in content and statistical specifications, but it is plausible that this parallelism may not be achieved in practice. Also, pseudo-test forms are typically created in a way to address a particular problem, which may or may not reflect the way operational test forms exist in practice.

Purpose and Research Questions

The potential problems associated with equating mixed-format tests, and the potential benefit to operational testing programs of knowing which test characteristics are likely to be conducive to adequate equating motivated this dissertation. The first goal of this dissertation is to gain increased understanding about the extent to which characteristics of operational mixed-format tests affect equating results. Further, this dissertation seeks to investigate when use of MC-only common items yield plausible equating results and whether the addition of CR items increases the accuracy of equating relationships. A second purpose of this dissertation is to understand whether similar analyses on two different classes of data lead to the same results and conclusions. Specifically, this dissertation addresses the following questions as they pertain to mixed-format tests and the CINEG design:

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

4. How do the content and statistical specifications of a test (e.g., subject area, correlation between MC and CR scores, and composition of common items) impact equated scores?
5. To what extent do analyses with two different classes of data, operational test forms and pseudo-test forms, result in the same findings?

Overview of the Dissertation

Chapter One addressed the background of mixed-format tests, common items, equating methodologies, and potential problems for equating mixed-format tests. It further provided a rationale for the specific research purposes and questions of the dissertation. Chapter Two contains an in-depth review of relevant literature in the areas of mixed-format tests, equating, and classes of data. Chapter Three provides a detailed explanation of the methodology used to address the specific research questions in this dissertation. Chapter Four contains a summary of the results, and Chapter Five provides a discussion of the findings and implications of the results of this dissertation.

CHAPTER TWO:
LITERATURE REVIEW

Chapter Two consists of four main sections which each address the specific topics in this dissertation. The first section provides a basic overview of the definitions, strengths, and weaknesses of MC items, CR items, and mixed-format tests. The second section provides an overview of equating designs, and the third section consists of an overview of commonly used equating methods. The fourth section provides a summary of past research related to the topics in this dissertation.

MC Items, CR Items, and Mixed-Format Tests

This section is divided into three subsections, which provide an overview of MC items, CR items, and mixed-format tests, respectively.

MC Items

According to Mislevy (1993), MC items have been the staple of educational assessment since the Army Alpha Intelligence Test of World War I. MC items are one type of selected-response item that can be administered in a small amount of time, allowing for broad coverage of content domains and higher reliability (Ferrara & DeMauro, 2006). MC items are also inexpensive to score. Further, a great deal of attention has focused on developing psychometric procedures that allow for efficient evaluation of MC items (Ferrara & De Mauro, 2006). However, Haladyna (1992) discussed several conditions that impacted the status of selected-response items. Some of these conditions included a criticism that MC items emphasize only memorization and recall, the desire to measure performance more "authentically", and a lack of scientific attention to item writing. Ferrara and DeMauro (2006) also noted that MC items have been criticized because they encourage test preparation focusing primarily on test-taking strategies at the expense of important curriculum. Linn, Baker, and Dunbar (1991) stated,

"In judging results from traditional standardized tests, we should demand evidence regarding the degree to which the skills and knowledge that lead to successful performance on multiple-choice test questions transfer to other tasks" (p. 19).

CR Items

CR items are a family of items for which examinees are expected to construct an answer or produce a product (Bennett, 1993). They include a broad range of tasks extending from word fill-ins to complex portfolios or performances. Bennett (1993) categorized CR items into a number of different categories, including substitution/correction, completion, construction, and presentation. Ferrara and DeMauro (2006) classified CR items into three categories: short CR items, products, and performances. Short CR items included tasks such as short answer and word fill-ins. Products involved longer responses, such as essays, papers, science projects, and artwork. Oral presentations or music and dance performances were examples of performance items.

Although MC items were the primary assessment item on educational tests since World War I, Bennett claimed in 1993 that CR items were receiving more attention than at any time in history. Presumably, the emphasis on CR items has grown even more since 1993, given the No Child Left Behind context. CR items are believed to assess higher-order thinking skills and emphasize a curricular focus on writing rather than test-preparation practices. They do not allow examinees the opportunity to guess the correct answer (Clauser, Margolis, & Case, 2006). Although CR items are relatively inexpensive to develop (Ebel & Frisbie, 1991), they are more costly and inefficient to score than MC items (Ferrara & DeMauro, 2006). Also, CR items require longer administration times; consequently, the content coverage of CR items is much less than MC items. Limited sampling of the content domain and subjectivity of tasks and scoring leads to lower

reliability (Ebel & Frisbie, 1991). CR items may also be easier to memorize than MC items, causing concerns about the security of the items.

Mixed-Format Tests

Neither CR nor MC assessments perfectly fulfill all desirable characteristics of assessments. Mixed-format tests incorporate both item formats, and they typically contain a large number of MC items with a few short CR items (Koretz & Hamilton, 2006). By including each item format on a test, presumably, some of the weaknesses of each item format may be mitigated by the strengths of the other. However, because MC and CR items are believed to measure different skills and processes, the inclusion of both item formats presents challenges for existing psychometric procedures. The equivalence of MC and CR items and the impact of using these two items in combination on a single test are discussed in detail later in the review of relevant literature.

Equating Designs

Equating designs are implemented for the purpose of collecting test data for use in equating. In practice, three equating designs are commonly used: single group, random groups, and common-item nonequivalent groups (CINEG) design. The choice of equating design is determined based on a number of considerations, such as the examinee population available, concerns about test security, and the degree to which statistical assumptions are expected to hold.

Single Group Design

In the single group design, the same examinees take two forms of a test (Kolen & Brennan, 2004). Data exists for the same examinees for both tests, making it relatively straightforward to determine which form is more difficult. One concern about the single group design is that examinee fatigue and practice effects may bias scores. Therefore, it is necessary to control for order effects of forms through counterbalancing, where half of

the examinees take one form first, and the other half of the examinees take the other form first. A second consideration of the single group design is that testing time for a given examinee is increased, because each examinee must take two test forms.

Random Groups Design

In the random groups design, groups of examinees are randomly equivalent (Kolen & Brennan, 2004). The only difference in ability between the two groups is assumed to be the result of random error. Therefore, the difference in scores between the two groups of examinees is considered to be a direct reflection of the difference in difficulty between the two forms. Unlike the single group design, examinees are required to take only one test form. However, all of the test forms must be administered in the same test administration. The administration of multiple forms on the same test date requires that all forms are constructed for a given test date and increases concerns about test security. Large samples of examinees are also needed, because only a portion of the total sample of examinees takes each test form (Kolen and Brennan, 2004).

CINEG Design

In the CINEG design, only one form is administered per test date, and the groups of examinees taking test forms on different administration dates are not assumed to be equivalent in proficiency. Differences in the distribution of examinee total scores between the two test forms are a combination of form differences and differences in examinee proficiency. In order to determine to what the differences in total scores are attributable, test forms are constructed with items in common between test forms. For example, consider Form Y and Form X. Form Y is administered to examinees on one test date, and Form X is administered to examinees on a later date. Both Forms Y and X contain items unique to the form and share a set of common items. If the two groups of examinees taking Form X and Form Y were the same in proficiency, it would be expected for the groups of examinees to have the same distribution of common-item

scores on Form X as compared to Form Y. Therefore, differences in the distribution of scores on common items are attributed to examinee proficiency, and the remaining differences are attributed to form difficulty.

Equating Methods

Both linear and nonlinear equating methods can be used with the CINEG design. This dissertation focuses on nonlinear equating methods, so only nonlinear methods are discussed in detail. The two traditional equating methods and two IRT equating methods being studied in this dissertation are considered in this section.

Traditional Equating Methods

This section is divided into two subsections. The first subsection provides an overview of the frequency estimation method of equating. The second subsection provides a description of the chained equipercentile equating method.

Frequency Estimation (FE)

As the name implies, the FE method of equating (Angoff, 1971; Braun & Holland, 1982; Kolen & Brennan, 2004) involves estimating distributions of scores on test forms. In the CINEG equating design, two test forms are administered: Form X and Form Y. Form X is considered the new test form, and it is administered to Population 1. Form Y is the old test form, and it is administered to Population 2. The distribution of total scores is available on the new form for Population 1 only and on the old form for Population 2 only. The distribution of common-item scores is available for both Population 1 and 2. Common items are assumed to be representative of the total test; therefore, the distribution of common-item scores is used to estimate distributions of total scores for a synthetic Population 1 on Form Y and a synthetic Population 2 on Form X. The synthetic distributions are considered weighted combinations of the distributions of

both populations. As described by Kolen and Brennan (2004), the equations for estimating scores for the synthetic populations are

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x) \tag{2.1}$$

and

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y), \tag{2.2}$$

where an s subscript refers to a synthetic population, the subscript 1 refers to Population 1 taking Form X, the subscript 2 refers to Population 2 taking Form Y, and w_1 and w_2 refer to weights for Populations 1 and 2, respectively. x and y refer to the old and new test forms, respectively. f and g refer to distributions of total scores for Form X and Form Y, respectively. $f_1$and g_2 are observed, but f_2 and g_1 must be estimated because data are not available for Population 2 on Form X or Population 1 on Form Y. Estimating these two quantities requires the assumption that the conditional distribution of total score given common-item score (v) is the same for both populations. Equations 2.3 and 2.4 are used to estimate score distributions for synthetic populations on Form X and Form Y, respectively,

$$f_s(x) = w_1 f_1(x) + w_2 \sum_v f_1(x|v) h_2(v) \tag{2.3}$$

$$g_s(y) = w_1 \sum_v g_2(y|v) h_1(v) + w_2 g_2(y), \tag{2.4}$$

where $w_1$and w_2 are the weights for Population 1 and Population 2, respectively. $f_1(x)$ is the probability of earning a score of x on Form X in Population 1, $f_1(x|v)h_2(v)$ is the product of the probability of earning a score of x given a score of v in Population 1 and the probability of earning a score of v in Population 2. The product of these two quantities is summed over all v. $f_s(x)$ is the probability of earning a score of x in the synthetic population. Values of $f_s(x)$ can be cumulated over all values of x to produce

$F_s(x)$, which is the cumulative distribution of Form X scores for the synthetic population. Similarly, for $g_s(y)$, $g_2(y)$ is the probability of earning a score of y on Form Y in Population 2, $g_2(y|v)h_1(v)$ is the product of the probability of earning a score of y given a score of v in Population 2 and the probability of earning a score of v in Population 1. The product of these two quantities is summed over all v. $g_s(y)$ is the probability of earning a score of y in the synthetic population. Values of $g_s(y)$ can be cumulated over all values of y to produce $G_s(y)$, which is the cumulative distribution of Form Y scores for the synthetic population.

After the distributions of scores for the synthetic populations on Form X and Form Y have been found, equating is conducted using Equation 2.5

$$e_{Ys}(x) = Q_s^{-1}[P_s(x)], \tag{2.5}$$

where $e_{Ys}(x)$ is the Form Y equivalent score of the Form X score x, $P_s(x)$ is the percentile rank function for Form X, and $Q_s^{-1}$is the percentile function for Form Y. One of the primary assumptions of FE is that the conditional distribution of total score given common-item score is the same for both populations. The more similar the two populations are taking the test forms, the more likely it is that this assumption holds.

Chained Equipercentile Equating (CE)

CE equating (Angoff, 1971; Dorans, 1990; Kolen & Brennan, 2004) is another equating method used with the CINEG design. Kolen and Brennan (2004) provide a step-by-step description of the steps in CE equating. First, using equipercentile equating methods, Form X scores are converted to common-item scores in Population 1 ($e_{V1}(x)$). Then, common-item scores are converted to Form Y scores in Population 2, also using equipercentile methods ($e_{Y2}(v)$). The two functions are chained together to produce the Form Y equivalent score of a Form X score, using Equation 2.6:

$$e_{Y(chain)} = e_{Y2}[e_{V1}(x)]. \tag{2.6}$$

CE does not require the same assumptions as FE. That is, there is no explicit assumption that the conditional distributions are the same in both populations. Two of the shortcomings of this method are that it equates a long test to a short test, and because there is no synthetic population, it is unclear to what population the equating relationship applies (Kolen & Brennan, 2004). However, von Davier, Holland, and Thayer (2004) showed that CE is expected to be population invariant when two assumptions hold: the link between scores on X and V is population invariant and the link between scores on V and Y is also population invariant.

Item Response Theory (IRT)

This section is divided into four subsections. The first subsection provides an overview of key IRT assumptions and commonly used models. The second subsection provides a description of IRT scale linking procedures. The last two subsections provide details on IRT true and observed score equating methods, respectively.

IRT Assumptions and Models

Before considering IRT equating, it is necessary to first consider IRT assumptions and models. Two assumptions of IRT are unidimensionality and local dependence. The assumption of unidimensionality means that the proficiency of an examinee on a given test is measured by a single latent variable, or construct. A related assumption is local independence, which means that items are statistically independent after accounting for examinee proficiency. In a CINEG equating context, there is also an assumption that the same unidimensional proficiency is being modeled for the examinee groups taking the two forms.

A number of unidimensional models exist for the estimation of item parameters in IRT. Dichotomous items (two score categories) are commonly modeled by the one-parameter logistic (also often referred to as the Rasch model), two-parameter logistic, or three-parameter logistic (3PL) models. Polytomous items (more than two score

categories) are commonly modeled by Samejima's graded response model, Bock's nominal model, and Muraki's generalized partial credit model (GPCM). In this dissertation, dichotomous items were modeled using the 3PL model and polytomous items using the GPCM.

3PL

Under the 3PL model, three item parameters are considered: discrimination (a), difficulty (b), and pseudo-chance (c). Equation 2.7 provides the probability that an examinee of a given proficiency would correctly answer an item given certain item parameters:

$$p_{ij} = p_{ij}(\theta_i; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\left[Da_j(\theta_i - b_j)\right]}{1 + \exp\left[Da_j(\theta_i - b_j)\right]}. \quad (2.7)$$

In Equation 2.7, p_{ij} is the probability that a person i would correctly answer item j. θ_i is the proficiency parameter for person i, and a_j, b_j, and c_j are the discrimination, difficulty, and pseudo-chance parameters, respectively, for item j. D is a constant typically set to 1.7. c_j is referred to as a pseudo-chance parameter, because it represents the probability of an examinee of low proficiency correctly answer the item.

GPCM

The GPCM can be used to model polytomous items with ordered or unordered categories. There is a discrimination and difficulty parameter for each item, and a difficulty parameter for each item category. Equation 2.8 gives the parameterization of the GPCM:

$$p_{ijk} = p_{ijk}\left(\theta_i; a_j, b_j, d_{j1}, d_{j2}, \ldots d_{jm_j}\right) = \frac{\exp\left[\sum_{h=1}^{k} Da_j(\theta_i - b_j + d_{jh})\right]}{\sum_{g=1}^{m_j}\exp\left[\sum_{h=1}^{g} Da_j(\theta_i - b_j + d_{jh})\right]}. \quad (2.8)$$

In Equation 2.8, p_{ijk} is the probability that a person i would correctly answer item j with category parameters k. θ_i is the proficiency parameter for person i, and a_j, b_j, and d_j are

the discrimination, difficulty, and category difficulty parameters, respectively, for item j. The number of d_j are equal to the number of item categories. D is a constant typically set to 1.7. For items with only two categories, the GPCM model reduces to the 2PL.

IRT Scale Linking

When using the CINEG design, IRT item and proficiency parameters are typically estimated separately for old and new forms. Linking procedures such as mean/mean (Loyd & Hoover, 1980), mean/sigma (Marco, 1977), Haebara (1980), and Stocking-Lord (1983) are necessary to place new form parameter estimates on the same scale as the old form estimates. Essentially, these methods estimate a slope and intercept which are then used to transform new form item parameters onto the old form scale. The Stocking-Lord and Haebara methods (characteristic curve methods) find a slope and intercept that minimize the sum of squared differences between two characteristic curves. For the Stocking-Lord method, a criterion is found that minimizes the sum of the squared differences between test characteristic curves. For the Haebara method, a criterion is found that minimizes the sum of the squared differences between item characteristic curves. Additional details about these methods can be found in Kolen and Brennan (2004).

IRT True Score (TS) Equating

After item parameter estimates have been transformed onto the same scale, equating can be conducted. IRT TS equating is one of two IRT equating methods typically used in practice. In TS equating, the true number correct score on one form associated with a given θ is considered to be equivalent to the true number correct score on another form associated with the same θ. Kolen and Brennan (2004) describe TS as a three-step process.

1. Choose a true score on Form X (new form).

2. Find the θ_i that corresponds to the true score selected in Step 1. This step requires an iterative process.
3. Find the true score on Form Y (old form) that corresponds to the same θ_i.

The three-step process is repeated for all true score values on Form X. In practice, true scores are not known; thus, observed scores are used in place of true scores. When observed scores are used in place of true scores, an ad hoc procedure (Kolen & Brennan, 2004) is used to convert scores less than the sum of the c_j. Additional details on TS equating can be found in Kolen and Brennan (2004).

IRT Observed Score (OS) Equating

A second IRT equating method is IRT OS equating. The IRT model(s) used to estimate item parameters are used to estimate a distribution of observed number-correct scores for both Form X and Form Y (Kolen & Brennan, 2004). Using a procedure described by Lord and Wingersky (1984), the distribution of observed number-correct scores is generated for examinees of a given θ_i. Distributions are estimated for all values of θ_i, and the distributions are cumulated to produce a distribution of number-correct observed scores for one form, such as Form X. This process is also repeated for Form Y. Similar to FE, the distributions of observed scores are estimated for synthetic populations on Form X and Form Y. Four distributions of number-correct observed scores are found: Form X for Population 1, Form X for Population 2, Form Y for Population 1, and Form Y for Population 2. The estimated number-correct observed-score distributions are equated using equipercentile methods. Additional details on OS equating can be found in Kolen and Brennan (2004).

Classes of Data and Criteria for Evaluation

Equating studies commonly use many different classes of data and criteria for evaluation. As discussed in Chapter One, data classes commonly include one of three approaches: operational test forms, pseudo-test forms, or simulated test forms. For

operational test forms, one of the most common processes in equating studies is to conduct various equatings and compare results across the equating methods (Harris & Crouse, 1993). For this method, there is generally no criterion to judge the source of the different results across equating methods. Pseudo-test forms appear to be a relatively recent method of assessing equating results. Commonly, a single-group equating relationship is used as the criterion relationship for pseudo-test forms, because data exist on both pseudo-test forms for all examinees. For simulated test forms, equating results are typically evaluated by how well the true population relationship is recovered (Harris & Crouse, 1993). However, if the model used to generate the data is also implemented as a study condition, results may be biased more favorably towards the generating model (Harris & Crouse, 1993).

In addition to the use of many different classes of data in equating studies, studies also implement various types of indices for evaluation. Harris and Crouse (1993) provided an extensive review of equating criteria used in equating studies. Many of the indices they discussed in their article were used in the studies reviewed in this dissertation, including root mean square differences (RMSD) (Fitzpatrick & Yen, 2001; Tan, Kim, Paek, & Xian, 2009; Walker & Kim, 2009), mean squared error (MSE) (Kim & Lee, 2006; Kim & Kolen, 2006), bias, standard error of equating, and root mean squared error (RMSE) (Cao, 2008; Kim & Lee, 2006; Sinharay & Holland, 2007; Cao, 2008; Walker & Kim, 2009; Wu, Huang, Huh, & Harris, 2009). Harris and Crouse (1993) can be consulted for an in-depth discussion of the ways in which these indices have been calculated across various studies.

Review of Relevant Literature

The review of relevant literature contains four sections corresponding to three areas of literature that were reviewed for this dissertation. These sections include equivalence of MC and CR items, equating methods, and test, common-item, and

examinee characteristics. The literature on equating is extensive; consequently, reviewing all literature pertaining to these four areas is beyond the scope of this review. Studies focusing on mixed-format tests and the CINEG design, specifically as related to the topics listed above, were reviewed. However, research on mixed-format tests is limited in many areas, such as comparison of nonlinear equating methods. Therefore, for some topics, research from MC-only tests or the random groups design was considered as a logical starting point for examining the impact on equating mixed-format tests. For each of the equating studies reviewed in this section, the specific class of data and criteria for evaluation used in the study is provided. It is also important to note that many equating studies incorporate a number of test or examinee characteristics as study conditions. Therefore, there is some overlap in the studies discussed in each of the sections.

Equivalence of Constructs Measured by MC and CR Items

CR items are often presumed to measure different constructs than those measured by MC items. This is potentially problematic for equating, especially when using MC-only common items. In this instance, the common items may not be representative of the total test, and equating results may be biased. However, as noted in the introduction of this dissertation, many item formats fall under the broad classification of CR item formats. In essence, in terms of the cognitive processes required to answer items, a continuum of CR items exists ranging from very similar to MC items to very different from MC items. Traub (1993) reviewed nine studies to examine the trait equivalence of MC and CR items. Each of the nine studies used both MC and CR items to measure a given construct. Constructs from the nine studies were grouped into either Quantitative or Language (Writing, Word Knowledge, and Reading Comprehension) domains. Traub (1993) suggested that in Writing and possibly Word Knowledge, different formats might measure different characteristics. However, the studies from the Word Knowledge domain provided contradictory evidence as to whether or not MC and CR items measured

different constructs. In the Reading Comprehension and Quantitative domains, tests of different formats did not appear to measure different constructs.

Rodriguez (2003) conducted a meta-analysis of studies examining construct equivalence of MC and CR items. One commonly used method of investigating construct equivalence is to create MC and CR items with equivalent stems. Stem-equivalent means the stem for the MC item is the same as the stem for the CR item. The only difference between the items is that answer choices are not provided for the CR items. Content equivalence is another way of examining the equivalence of MC and CR items. Content-equivalent items can be created by writing MC and CR items with different stems to measure the same content area. Content-nonequivalent items are specifically designed to measure different content domains. Rodriguez included only those studies that reported the correlation between the total score for all items of MC format and the total score for all items of CR format, indicated whether or not stem-equivalent methods had been used, incorporated MC items with more than one answer choice, and reported the subject area of the test. In general, Rodriguez found that when MC and CR items were constructed to be stem equivalent, correlations were higher than when items were not stem equivalent. The mean disattenuated correlation between scores for stem-equivalent MC and CR items was 0.92, and the mean disattenuated correlation between scores for stem-nonequivalent MC and CR items was 0.85. Additionally, content-equivalent items tended to have higher MC and CR correlations than content-nonequivalent items. Although Rodriguez did not find construct equivalence based on Traub's (1993) definition (true-score correlations of 1.00), he suggested that how the items were designed moderated the correlations. "When the items are designed to measure the same aspects of the content domain or cognitive ability, the resulting correlations are higher." (p. 179). He further suggested that when MC and CR items are carefully designed to measure the same content and cognitive aspects, it is more appropriate to combine MC and CR scores.

Bennett, Rock, and Wang (1991) used a two-factor confirmatory factor analysis model for an Advanced Placement (AP) Computer Science test for determining the equivalence of MC and CR items. For the first factor, they constructed five ten-item MC sets similar in content and mean difficulty. The second factor was marked by five nine-point CR items. A one-factor model was the most parsimonious model, suggesting MC and CR items measured similar constructs. An informal analysis of the items indicated that some of the processes necessary for solving MC and CR items overlapped. Thissen, Wainer, and Wang (1994) proposed an alternative three-factor model for the AP Computer Science test. Their model included a general factor and two CR specific factors. CR items loaded on both the general and specific factors, although loadings were typically higher on the general factor. Consequently, they concluded CR items were mostly related to the construct measured by the general factor, but some CR specific dependence existed among CR items. Further, because loadings were small on the CR factors, Thissen, Wainer, and Wang concluded that many CR items would be required to reliably measure the CR specific factors. Similar results were found for an AP Chemistry test. Sykes, Hou, Hanson, and Wang (2002) also used factor analysis to assess dimensionality on an operational mixed-format mathematics test. They found that both CR and MC items loaded heavily on the first factor, while only MC items loaded heavily on a second factor.

In addition to focusing on the equivalence of MC and CR items, research has also examined the value of these two item types. Lukhele, Thissen, and Wainer (1993) examined the relative value of CR items as compared to MC items on AP U.S. History and Chemistry tests. They concluded that "constructed-response items provide less information in more time at a greater cost than do multiple-choice items" (p. 15). For the Chemistry test, the inclusion of two CR items reduced error. However, in the amount of time it took to answer only one CR item, 16 MC items could have been administered. The 16 MC items would have also resulted in greater information across the entire score

scale. A 75-minute MC test would have had approximately the same reliability as a 185-minute CR test. Similar results were found for U.S. History.

Wainer and Thissen (1993) investigated the combination of MC and CR scores on mixed-format tests. One example in their study was from an AP European History test. The CR section was found to have an internal consistency reliability estimate of 0.46, while the MC section had a reliability estimate of 0.90. The two sections were equally weighted, resulting in a composite reliability of 0.80, which was lower than the reliability of the MC section alone. On the use of CR items, Wainer and Thissen (1993) stated, "Measuring something that is not quite right accurately may yield far better measurement than measuring the right thing poorly" (p. 115).

Equating Methods

Many studies in the current literature comparing equating methods have been conducted on tests containing only MC items; although, some studies have compared equating methods for mixed-format tests. Additionally, recent research has included alternative equating methods, such as trend scoring. The literature reviewed in the equating methods section is summarized in two sections. The first section provides a review of some relevant literature comparing equating methods for tests containing only MC items. The second section contains a summary of research on equating methods for mixed-format tests.

Comparison of Equating Methods for MC-Only Tests

Harris and Kolen (1990) compared FE and CE using operational test forms for a 200 MC item certification test. They conducted two equatings using three different test forms. The new test form was the same for both equatings. For one equating, examinees were similar in proficiency across old and new test forms. However, for the other equating, examinees on the old form were lower in proficiency than examinees on the new test form by approximately 0.30 standard deviation units. Three different statistics

were used to compare equivalents for the two equating methods: RMSE, mean absolute difference, and mean signed difference. FE was considered the criterion equating. In some places in the score scale, differences between CE and FE exceeded more than two standard errors of equating based on the FE equating method, especially when old and new form groups differed in proficiency. However, because the samples of examinees were large, standard errors of equating were small. Standard errors of equating were also calculated based on a sample size of 1,250. When the differences in equated scores for FE and CE for examinee groups differing in proficiency were compared to these standard errors, the differences were still larger than the standard errors.

Holland, Sinharay, von Davier, and Han (2008) examined missing data assumptions for the chain (CE) and post-stratification (or, FE) nonlinear equating methods in the CINEG design for an MC-only test. Pseudo-test forms were the class of data used in their study. They began with one form of 120 MC items that had been administered to two populations. The means for these two populations differed by approximately 0.25 standard deviation units. The form was split into two pseudo-test forms of 44 MC items. The new test form was intentionally created to be easier than the old test form. Three external and three internal common-item lengths were also created. The mean difference for the two populations on the common items was similar to the mean difference for the two populations before the test was split. Raw score distributions were smoothed using presmoothing, and marginal distributions for CE were continuized using both the traditional linear interpolation method and a kernel smoothing method. The criterion for evaluation was how closely predicted score distributions approximated the observed score distributions. This criterion was evaluated using plots of differences, goodness-of-fit measures, and a comparison of the first four moments of observed and predicted distributions. Both CE and FE methods performed similarly in how well predicted score distributions approximated the observed score distributions, although CE predictions were slightly more accurate.

Ricker and von Davier (2007) examined the impact of common-item length for four different equating methods: FE, CE, kernel poststratification equating with optimal bandwidths, and linear kernel poststratification equating (large bandwidths). Their study also used the pseudo-test form class of data and the same operational test form used by Holland, Sinharay, von Davier, and Han (2008). Their criterion equating was an equivalent groups equipercentile equating, and they evaluated the magnitude of the differences between the criterion and study condition equatings using RMSD and a "difference that matters" (DTM). For all methods except kernel with large bandwidth, the conversions were very similar to the criterion, although some differences were observed across the common-item set length.

Sinharay and Holland (2007) conducted a study examining the distribution of item difficulty in a common-item set. They used simulation studies under unidimensional and multidimensional IRT models as well as pseudo-test forms in their study. They simulated a number of conditions related to test and examinee characteristics, which are discussed in greater detail in a later section. The criterion equating for the simulations was a population equipercentile equating function, which was essentially the population IRT observed score equating function. For the pseudo-test forms, the criterion was a single-group equipercentile equating. They compared FE and CE with respect to bias, standard error, and RMSE. They found that CE performed better than FE with respect to bias, but FE was better than CE with respect to SE. Similar results were found for the multidimensional simulation and pseudo-test forms.

Wang, Lee, Brennan, and Kolen (2008) compared FE and CE through a simulation study. Two test lengths (60 and 120 MC items) for a mathematics test were simulated. Additionally, two ratios of common-item test length were simulated. Their evaluation criteria were bias, standard error of equating, and RMSE, and the criterion equating relationship was an IRT observed score equating relationship. When there was no group difference in proficiency across examinees taking the old and new test forms,

both CE and FE were unbiased, except at low scores. However, when the group difference was large, both methods were more biased, and the difference between FE and CE was also greater. CE was less sensitive to differences in group means than was FE. Additionally, they found that FE resulted in smaller standard errors than CE.

Research studies comparing IRT and traditional equating methods appear to be somewhat less prevalent in the literature. Therefore, some of the studies reviewed that compared IRT and traditional equating methods were conducted using the random groups design rather than the CINEG design. Using operational test forms, Harris and Kolen (1986) compared equipercentile and TS with five *American College Testing* forms of Mathematics Usage. They used the 3PL model and a random groups equating design. They conducted equatings with both high and low proficiency subgroups; although, it appears that equating was not conducted *between* high and low ability subgroups. That is, test forms with samples of high proficiency subgroups were equated to each other, and test forms with samples of low proficiency subgroups were equated to each other. Results were evaluated using RMSE, a mean absolute difference criterion, and a mean difference criterion between equatings for the high and low ability groups. They found that the equipercentile and TS equating methods performed similarly, concluding that both methods were relatively invariant to the examinee group proficiency level.

Another study by Han, Kolen, and Pohlmann (1997) also used operational test forms and the random groups equating design. They compared TS and OS to equipercentile equating for *American College Testing* test forms. Results were evaluated using bias, mean absolute loss, root mean square loss, mean signed difference, mean absolute difference, and root mean square difference. The criterion for their study was an equating a test to itself design. They found that none of the three equating methods consistently produced more equating loss relative to the other methods. However, TS tended to produce equating conversions that were more stable than those for equipercentile or OS. Additionally, OS tended to be more stable than equipercentile

equating. However, among the three methods, mean differences in equating stability were generally small. Larger mean differences in equating stability among the equating methods appeared to be related to larger differences in test form difficulty.

Tsai, Hanson, Kolen, and Forsyth (2001) compared bootstrap standard errors of equating for IRT equating methods for MC-only tests using operational test forms and two sample sizes: 500 and approximately 1,500 to 1,800. In their study, they examined various methods of placing item parameters on the same scale (e.g., Stocking-Lord, concurrent calibration). They found that OS resulted in smaller standard errors than TS for both sample sizes.

Comparison of Equating Methods for Mixed-Format Tests

Although research on mixed-format tests is limited, some studies have compared equating methods with mixed-format tests. Lee, Hagge, He, Kolen, and Wang (2010) conducted a study in which a mixed-format test based on an AP World History exam was simulated. The purpose of their study was to examine the impact on equating results for various levels of examinee group differences and correlations between MC and CR scores. They compared three smoothing approaches for FE and CE, and their criterion equating was a simple-structure multidimensional IRT observed-score equating. Squared bias, variance, and MSE were used to evaluate the results. They found that with small group differences (0.05 standard deviation units), FE and CE both resulted in adequate equating relationships as evaluated by bias compared to a DTM. However, with large group differences, CE resulted in less bias and MSE than FE. Variance was generally larger for CE as compared to FE.

Von Davier and Wilson (2008) examined the population sensitivity for male and female examinees of TS and CE using an operational AP Calculus AB exam containing both MC and CR items. The correlation between MC and CR sections was 0.86 and 0.87 for the old and new test forms, respectively. They used the CINEG design with an

internal set of MC-only common items. They conducted two studies. For the first study, equating was conducted for only the MC portion of the test. For the second study, equating was conducted using both the MC and CR items. The same MC-only common-item set was used for both studies. RMSD and root expected mean square difference (REMSD) were used to evaluate the results. In the first study, when only MC items were included on the test forms, they determined that the TS equating results for the male and female subgroups did not differ substantially from the results based on the combined male and female group. Results for CE were similar to those for TS. For the study containing both MC and CR items, they found that the differences between the equating results for the male and female subgroups and the combined group were larger relative to the study containing only MC items. However, the overall value of REMSD was still smaller than the standardized DTM, indicating that the equating results were still relatively group invariant.

Tate (1999) cautioned that use of standard linking or equating procedures without taking into account changes in rater severity across years could result in incorrect equating results. He proposed a linking method for mixed-format tests that incorporates a second rating of CR items. Rerating CR responses with raters from the new test form administration has become known as trend scoring. Tate and some of his colleagues conducted studies investigating the proposed linking method (2000, 2003, & 2005). Tate (2000) used simulated data to create 11 study conditions that varied multidimensionality, examinee proficiency, sample size, and characteristics of the total test and common items. The criterion for evaluation used in the study was the extent to which linking coefficients were accurately recovered across the various common-item compositions. Tate found that the linking error for the proposed linking method was reasonable across changes in examinee ability. The accuracy of the recovery of the linking coefficients improved with increased common-item length or sample size.

Tan, Kim, Paek, and Xiang (2009) compared an MC-only common-item design to the trend scoring design. Their study created pseudo-test forms by shortening operational test forms. Using a 90 item MC and three item CR test form, they created test forms with varying MC to CR point ratios and MC and CR correlations. The criterion equating relationship in this study was equating using trend scoring. The differences between the equated raw scores for the study conditions and the equated raw scores for the trend-scored equating relationship were evaluated using average weighted differences, RMSD, and conditional RMSD. These statistics were compared to a DTM. For CE, RMSD was only lower than the DTM with an MC to CR ratio of 2 to 1 and a moderately high correlation. The MC-only common-item design was comparable to the trend scoring design when the MC to CR point ratio and correlation between MC and CR were moderately high to high.

Kim, Walker, and McHale (2008) also examined trend scoring in equating. They created two pseudo-test forms with eight unique MC and four unique CR, and an internal common-item set with eight MC and four CR. The criterion in their study was a linear single-group equating. They compared a NEAT (CINEG) design and a hybrid no-anchor common-item design. For the CINEG design, equating was conducted using three different compositions of common items: MC plus no-trend CR items, MC plus trend CR items, and only MC items. The second equating design was a hybrid no-anchor design, which incorporated a combination of single-group and equivalent groups equating designs. Chained linear equating was used for the three common-item compositions for the CINEG design. Only the chained linear method was used for the hybrid no-anchor design. The hybrid no-anchor design resulted in the smallest RMSD and bias compared to the other three common-item sets, although equating error was slightly larger. Of the three common-item sets, MC plus trend-scored CR items resulted in the smallest RMSD and bias.

Paek and Kim (2007) investigated alternatives to the trend-scoring method for detecting shifts in the scoring of CR items using a differential bundle function. In their analysis of operational data, they compared differential bundle functioning to trend-scoring methods, and they found both methods performed similarly in detecting scoring shifts. However, the extent to which an MC-only common-item set would adjust shifts in scoring as well as trend-scoring adjusts for shifts was unknown. Some factors thought to influence the adequacy of the MC-only common-item set included the MC and CR correlation and the ratio of MC to CR points.

Test, Common-Item, and Examinee Characteristics

Test, common-item, and examinee characteristics are discussed together in the literature review, because many studies investigated a combination of these characteristics. Two studies focusing specifically on CR test length and rater consistency are discussed first. The remainder of the section is organized according to the common-item characteristics studied in this dissertation as follows: statistical and format representativeness.

CR Test Length and Rater Consistency

Fitzpatrick and Yen (2001) examined the impact of test length and sample size on random groups equating for CR tests. They simulated data for tests containing 2, 4, 8, 12, and 20 CR items with 2, 4, or 6 categories. They also investigated sample sizes of 200, 500, and 1,000 examinees. Equating was conducted using equipercentile methods to equate observed θ estimates. The criterion in their study was the difference between examinees' true θ and the equated estimates of θ. Two of the indices used to measure differences were RMSD and classification consistency. In general, RMSD decreased as the numbers of items or score points increased. Sample size did not appear to result in systematic effects. Error rates in classification consistency were less than 10% with 12 or 20 items. Tests with eight, four- or six-point items also had an error rate approximately

equal to or less than 10%. Greater gains in accuracy were found by increasing items than were found by increasing the number of score points.

Fitzpatrick, Ercikan, Yen, and Ferrara (1998) examined the consistency of raters from 1991-1993 for Grades 3, 5, and 8 on the Maryland School Performance Assessment Program. They found that ratings across test years were inconsistent, with scores typically varying on average one- to two-tenths of a standard deviation. For all grades except Grade 8, mathematics content and process had the most consistent ratings, while the three language areas (writing, reading, and language arts) had the least consistent ratings. One important note about this study is that 1991 was the first year of administration. Consequently, it might be expected for ratings to be more inconsistent in the first years of the testing program. Further, the scoring contractor changed between 1991 and 1992, which may have resulted in different rater training across these two years. Lukhele, Thissen, and Wainer (1993) found that analytically scored items, often found in quantitative domains, appeared to reduce inter-rater variability more than holistically scored items.

Common Items

Although evidence has varied somewhat according to test specifications or examinee characteristics, research has generally suggested that for MC-only tests, common items should be representative of the content and statistical specifications on the entire test (Klein & Jarjoura, 1985; Cook & Petersen, 1987; Harris, 1991; Petersen, Marco, & Stewart, 1992). Additionally, Kolen and Brennan (2004) suggested a set of common items should be at least 20% of the length of a test unless the test is very long. The inclusion of more items tends to lead to fewer problems in equating (Budescu, 1985; Petersen, Cook, & Stocking, 1983). Cook and Petersen (1987) suggested items should be kept in the same positions on both test forms. Answer choices for MC items should also be kept in the same order (Cizek, 1994). Some recent literature suggests that statistical

specifications for the common-item set could be relaxed (Sinharay & Holland, 2006; Sinharay & Holland, 2007). Although these studies have been conducted for MC-only tests, it is plausible that relaxing some of the statistical specifications may allow for the creation of common-item sets that represent mixed-format tests without the inclusion of CR items in the common-item set.

Statistical Representativeness

Although it is usually recommended for common items to be representative of the total test in terms of mean and variability of difficulty and discrimination, research by Sinharay and Holland (2006) suggests it may be possible to relax some of the statistical guidelines. Their study was conducted on MC-only tests; however, relaxing restrictions on the representativeness of statistical characteristics may be useful when constructing common-item sets for mixed-format tests. Sinharay and Holland (2006) examined statistical properties of common-item sets by evaluating content representative and content non-representative minitests, miditests, and semi-miditests. They defined a minitest as representative of the statistical characteristics of the total test. A miditest was defined as containing items of only medium difficulty, and a semi-miditest contained a spread of item difficulties somewhere between a minitest and miditest. Sinharay and Holland (2006) investigated the correlation between the anchor and total test, citing Angoff's (1971, p. 577) comment that "the usefulness of an anchor test depends on the extent to which it is correlated with the test being equated." Although they conducted analyses for both internal and external common-item sets, results were reported only for external common-item sets. It is plausible that when the common-item set is internal, it may not be difficult to find and select items that represent a mini version of the total test.

Sinharay and Holland (2006) simulated univariate and bivariate tests with 40 items and 20 common items. (They examined both internal and external common-item sets, but they reported results only for external common-item sets.) The conditions they

varied included difficulty of the total test, difficulty of the common items, and equivalence of examinee proficiency. Across all study conditions for the simulated univariate test forms, the miditests had the highest average correlations and correlations for the semi-miditests were nearly as high. Correlations were always lowest for the minitests. For the bivariate test forms, common-item sets were created to be content representative and content non-representative. Correlations for the content non-representative common items were lower than correlations for the content representative common items, and correlations were lower for the minitests than for the semi-miditests. Additional simulations were conducted, and similar results were found. In addition, Sinharay and Holland (2006) found that a miditest or semi-miditest that was not centered on the mean difficulty of the total test had higher correlations than a centered minitest. They also found similar results for an operational basic skills test.

Sinharay and Holland (2007) went a step further to investigate the impact of miditests, semi-miditests, and minitests on equating. Using the simulated data, they created three unidimensional tests that differed in subject area and test length. Additional factors that were investigated included sample size, difference in mean proficiency of examinees, difference in difficulty of the test forms, and CE and FE equating methods. As described earlier in this review, their criterion equating relationship was the population equipercentile equating function. They compared the difference between estimated equated scores for the various study conditions and the population equipercentile equating function using bias, RMSE, and standard errors. Miditests and semi-miditests were usually less biased than minitests, and standard errors and RMSE were also usually smaller for miditests and semi-miditests, although the effect was small. The effect of common-item set had a much smaller impact on equating than the effect of CE versus FE, sample size, group differences, or test length. They also simulated test forms using a multidimensional model. Although some of the study conditions slightly favored a particular common-item set, they concluded that there was “no practically

significant difference in equating performance of the three anchor tests" (p. 267). One interesting finding was that when the difference in examinee proficiency differed across dimensions, FE performed better than CE. Sinharay and Holland (2007) also considered pseudo-test forms with tests of 44 items and external anchor tests of 16, 20, and 24 items. Their criterion was a single-group equipercentile equating based on all examinees. Differences between the equated scores based on the study conditions and the criterion equating were evaluated using bias and SE. They found that there was relatively little difference between the minitest and semi-miditest in terms of bias or standard errors. The minitest performed slightly better with FE, and the semi-miditest performed slightly better with CE.

Sykes, Hou, Hanson, and Wang (2002) examined the impact on equating of common-item statistical representativeness in terms of dimensionality. Dimensionality was assessed using Poly-Dimtest and a factor analysis of item responses. Two significant dimensions were found. Four content and difficulty representative anchor sets were created. Two of the common-item sets were criterion sets, meaning that the loadings across the two significant dimensions were balanced. Another common-item set included items loading more heavily on the first factor (F1), and the last anchor set included items loading more heavily on the second factor (F2). The operational administration of a test containing 35 MC and 10 SA was linked to the field test administration of the test using Stocking-Lord. Item parameter estimates for the field test were linearly transformed to a scale score metric with a mean of 500 and standard deviation of 50. Operational item parameter estimates were linked to this scale. Using one of the criterion common-item sets as the criterion, F1 resulted in five times more total error, and F2 resulted in over twice as much error. 95% and 37% of the total error for F1 and F2, respectively, was attributable to bias. As a result, the mean scale score for F2 was two points larger than that for the criterion, and the mean scale score for F1 was five points smaller. In

comparison, the difference in scale scores based on results from the two criterion anchor sets was less than one point.

Format Representativeness

Use of MC-only common items may bias results, because examinee proficiency and item difficulty on CR items may not be represented by the MC common items. However, inclusion of CR items in the common-item set raises concerns about the security of the CR items. Additionally, CR items contain an additional source of rater error. Further, CR items may not be representative of the content domain because of limited sampling resulting from administration time constraints for CR items. Therefore, it is of practical interest to measurement professionals to determine when MC-only common-item sets lead to adequate equating results. A number of studies have investigated both MC-only common items and inclusion of CR items in the common-item set.

Kim and Kolen (2006) examined the robustness of four IRT linking methods and concurrent calibration in the context of format effects. They used simulated test forms and a common-item set containing only MC items. Format effects were simulated by varying the degree of correlation among constructs measured by MC and CR items. Three levels of format effects, two types of mixed-format tests (wide and narrow range), and three levels of equivalence of examinee proficiency were simulated. The wide range test represented a standardized achievement test, and the narrow-range test represented a professional certification test. To evaluate the results of their study, they used an observed score difference criterion, which was based on the difference between the estimated and true observed score distributions. MSE for the squared difference was reported. For the wide range test, they found that as the correlation between MC and CR items decreased, MSE increased, although increases were markedly higher for moment scale linking methods as compared to Stocking-Lord and Haebara. Similar results were

seen for the narrow-range test; however, the pattern of MSE across the score scale differed for the two test types.

Wu, Huang, Huh, and Harris (2009) examined the effectiveness of using MC items as an external common-item set for a CR test. They used simulated data as the class of data to create external common-item sets of eight MC items for a test containing eight five-category CR items. Three levels of form differences were created for the CR test, as well as three levels of format effects, which were determined by the correlation between MC and CR latent proficiencies. Further, five levels of group differences, three sample sizes, and two linear equating methods were investigated. The criterion equating relationship in this study was the difference between the estimated sample equivalents and the population equivalents. Differences were evaluated using root weighted mean squared error, weighted squared bias, and weighted standard error of equating. They found that when the correlation between MC and CR items was lower, greater bias occurred in the sample equating relationship. Additionally, as the mean difference between groups increased, the amount of bias in the sample equating relationship also increased. Sample size did not impact bias, although it did contribute to the standard error of equating. Difference in form difficulty had little effect on bias or the standard error of equating. The authors stated that when a 30-item common-item set was used, bias and the standard error of equating decreased substantially, although results were not reported.

Walker and Kim (2009) considered the use of MC-only common items for a mixed-format test. They created two pseudo-test forms, each with 16 MC and eight CR items. The two test forms had internal common items with eight MC and four CR items in common, but only the MC items were treated as common items. The observed correlation between the composite score and MC common-item score was approximately 0.56, and the observed correlation between the MC section and MC common-item score was approximately 0.83. Disattenuated correlations were not reported. They conducted chained linear and chained equipercentile equating using direct and two-stage linking. In

direct linking, linking was conducted by linking a new form to a reference form using MC-only common items. In two-stage linking, the MC section on the new form was equated to the MC section on the reference (old) form using MC-only common items. Then, the reference MC score was scaled to the composite score on the reference form, and the same procedure was used for MC and composite scores on the new form. The criterion equating in this study was a single-group equating, and separate linear and equipercentile criteria were established. Equated raw scores for each study condition were compared to the appropriate linear or equipercentile criterion. Results were evaluated using RMSD, bias, standard errors, and RMSE. The direct and two-stage linking strategies produced very similar results for the MC-only common items, but bias was approximately 1.5 points for both strategies. They suggested that MC-only common items may result in an adequate equating relationship if the relationship between MC and composite scores was consistent across reference and new form examinees.

As described previously, Tan, Kim, Paek, and Xiang (2009) used pseudo-test forms to compare trend scoring to equating with an MC-only common-item set. Their results were evaluated based on how similar the equating with the MC-only common-item set was to equating based on trend scoring. RMSD and a DTM were used to judge the difference between these two methods. They found that as the MC to CR point ratio and MC and CR correlation increased, the MC-only common items produced results comparable to the trend scoring method. As discussed previously, using pseudo-test forms, Kim, Walker, and McHale (2008) also compared equating with an MC-only common set, a common-item set containing MC plus trend-scored CR, and a common-item set containing MC plus no-trend CR. Based on a single-group equipercentile equating criterion, across the three common-item sets, RMSD was lowest for the common-item set containing MC and trend-scored CR items. Equating bias was largest for the common-item set containing MC plus no-trend CR. However, bias was only slightly larger for this common-item set as compared to the MC-only common-item set.

Through a simulation study, Tate (2000) found that when a test was unidimensional, MC-only common-item sets resulted in reasonable linking bias. When unidimensionality was violated, common-item sets containing proportional numbers of MC and CR items resulted in reasonable linking bias with the proposed trend scoring equating method.

Kirkpatrick (2005) used both operational data and simulated data to investigate the impact of the inclusion of a CR item in the common-item set. Two of the equating methods Kirkpatrick considered were IRT TS and OS. For the operational data analyses, equating results based on two sets of common items were compared. One common-item set contained a two-point CR item, and the other set replaced the CR item with two MC items from the same content area as the CR item. Kirkpatrick conducted analyses for a reading and a mathematics test for one grade level each in elementary, middle school, and high school. For the operational data analyses, Kirkpatrick found that the impact of the inclusion of the CR item in the common-item set was small for all equatings except elementary mathematics. In the simulated data investigation, three levels of dimensionality, as measured by the correlation between MC and CR items were considered (0.5, 0.8, and 1.0). One reading and one mathematics test were simulated, and seven levels of MC and CR means were examined. Some of the levels of mean differences represented differential examinee performance on the MC and CR items. Kirkpatrick found that within a dimensionality level, as the mean difference between the two dimensions increased, the magnitude of the differences between equating results also increased. Additionally, when the MC and CR correlation was low (0.5), there was a noticeable difference in equating results for the common-item set that included the CR item as compared to the common-item set that excluded the CR item. When the correlation was moderate (0.8), the differences in equating results between the two common-item sets were small enough that they were likely not of practical significance. However, the correlation between MC and CR items impacted equated scores less than

the difference in means across MC and CR scores. The impact on equating results of the inclusion or exclusion of the CR item was most noticeable when the difference between MC and CR means was large.

Cao (2008) simulated a number of characteristics of a mixed-format test, including differences in examinee proficiency distributions, and content, format, and statistical representativeness of the common-item set for various test dimensionality structures. The class of data used in the study was simulated test forms. Statistical representativeness meant the common items were created to be similar to the total test. For the statistical non-representativeness condition, average item difficulty in the common-item set was 0.30 different from the average item difficulty in the total test. Format non-representative common items contained only MC items, and format representative common items contained both MC and CR items proportionally representative to the total test. The evaluation criterion in this study was the difference between the true expected total scores and the estimated expected total scores. Bias, RMSE, and classification consistency were used to evaluate the magnitude of the differences. Bias and RMSE were always smaller and classification consistency was always higher for the equivalent groups condition as compared to the nonequivalent groups condition. Bias and RMSE for the statistical representativeness condition were usually smaller relative to the statistical non-representativeness condition. Additionally, statistical representativeness led to higher classification consistency rates in the nonequivalent groups condition. When the degree of multidimensionality increased, format representativeness became more salient. Specifically, for nonequivalent groups, when the degree of multidimensionality increased, format representativeness resulted in significantly smaller bias and RMSE and higher classification consistency as compared to format non-representativeness with nonequivalent groups.

Kim and Lee (2006) examined extensions of four IRT linking methods (mean/mean, mean/sigma, Stocking-Lord, and Haebara) for mixed-format tests. For each

of four IRT linking methods, they investigated degree of examinee equivalence, sample size, type of item format on the total test, and type of item format used as linking items. They used simulated test forms as the class of data in their study, but the simulated test forms were based on statistical distributions rather than operational test forms. The criterion for evaluation was a characteristic curve criterion, which was based on the difference between estimated and true characteristic curves. The conditions they considered for the common-item set included linking through both dichotomous and polytomous item formats, linking through dichotomous items only, and linking through polytomous items only. Using the category characteristic curve criterion, they found larger bias and MSE for nonequivalent groups as compared to equivalent groups. Additionally, using both dichotomous and polytomous items as linking items typically resulted in smaller MSE than using only one item format. When only dichotomous or polytomous items were used as linking items, linking through the dominant item type on the total test resulted in lower MSE. For example, for the test containing 10 dichotomous items and 10 five-category polytomous items, using only polytomous items resulted in smaller MSE than using only dichotomous items. It is important to note that all items on the test of a given format were used as linking items. Consequently, item format and number of linking items may be somewhat confounded.

Summary

Research on equating has taken into account many factors, including equating methods and test characteristics, examinee proficiency, and composition of the common-item set. Although much of the research has been conducted on tests containing only MC items, a growing body of literature has focused on CR-only tests and mixed-format tests. A number of consistent results have been found across the equating studies that have been conducted.

Equivalence of Constructs Measured by MC and CR Items

One factor that may impact equating is the extent to which MC and CR items are measuring equivalent constructs on a given test. The extent to which MC and CR items measure equivalent constructs may differ according to the subject area as well as format of the CR item. Some evidence has suggested that in the writing domain, MC and CR items may not be measuring equivalent constructs (Traub, 1993). In the quantitative domain, there is evidence to suggest that MC and CR items may measure similar constructs (Bennett, Rock, & Wang, 1991; Traub, 1993). However, factor analyses have indicated the presence of a possible weak format factor, even in the quantitative domain (Thissen, Wainer, & Wang, 1994; Sykes, Hou, Hanson, & Wang, 2002). Other studies have suggested that when MC and CR items are designed to measure equivalent constructs, they perform similarly (Rodriguez, 2003).

Equating Methods and Designs

For MC tests, FE and CE equating methods typically performed similarly when examinee groups were similar in proficiency (Wang, Lee, Brennan, & Kolen, 2008), but when examinee groups differed in proficiency, FE and CE led to different results (Harris & Kolen, 1990). Specifically, research has found that when groups differed substantially, CE produced somewhat more accurate equating results than FE (Holland, Sinharay, von Davier, & Han, 2008; Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008). Similar results were found for a mixed-format test study (Lee, et al., 2010). In research for both MC and mixed-format tests, FE has been found to result in smaller standard errors than CE (Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008; Lee, Hagge, He, Kolen, & Wang, 2010).

Although many of the comparisons of IRT and traditional equating methods have been conducted for random group designs, there is evidence to suggest IRT equating results might lead to more stable or accurate equating results than equipercentile methods

(Han, Kolen, & Pohlmann, 1997). However, there is also evidence to suggest that IRT and traditional equating methods lead to similar results (Harris & Kolen, 1986; von Davier & Wilson, 2008). One study also found that standard errors of equating were smaller for OS as compared to TS (Tsai, Hanson, Kolen, & Forsyth, 2001). Newly proposed equating methods, such as the hybrid no-anchor design and trend scoring, have appeared to perform better for mixed-format tests than some of the current equating methods (Kim, Walker, & McHale, 2008; Tate, 2000; Tan, Kim, Paek, & Xiang, 2009).

Test, Common Item, and Examinee Characteristics

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

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

differs across item formats (Cao, 2008; Kirkpatrick, 2005). However, use of trend scoring may be recommended when CR items are included in the common-item set (Kim, Walker, & McHale, 2008; Tate, 2005). Using only the dominant item type included on the test may also lead to reasonable equating results (Kim & Lee, 2006).

Operational Test Forms and Pseudo-Test Forms

The extent to which findings and conclusions varied as a result of the class of data used in the study was difficult to determine from the literature reviewed. The primary classes of data in the studies reviewed were simulated test forms and pseudo-test forms, and for the most part, studies resulted in similar conclusions. However, because the majority of the studies did not investigate multiple classes of data within a study, it is impossible to tell whether differences in findings were the result of the class of data or the study characteristics. Sinharay and Holland (2006, 2007) were one exception. They used multiple classes of data in their studies and found similar results. However, their research was conducted on MC-only tests.

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

CHAPTER THREE:
METHODOLOGY

This dissertation used operational test forms and pseudo-test forms to investigate how the inclusion of both MC and CR item formats on the same test impacts equated scores in the CINEG design. Chapter Three is divided into four sections. The first section provides a general overview of the original operational test forms used in this dissertation. The second section describes the methodology implemented for the operational test form analyses, and the third section describes the methodology implemented for the pseudo-test form analyses. The final section describes how results were compared across the two classes of data. A number of abbreviations are used throughout the methodology and results. A list of abbreviations and acronyms was provided at the beginning of this dissertation.

Original Operational Test Forms

This section is divided into three subsections. First, the process for selecting the tests used in this dissertation is described. Second, steps taken to prepare the data for analyses are outlined. Last, the methods used to assess dimensionality for the selected tests are discussed.

Selection of Tests

Data for this dissertation were from College Board Advanced Placement (AP) tests. It is important to note that, although Advanced Placement (AP) tests were used for analyses in this dissertation, the exams were manipulated in order to investigate how equating methods, test characteristics, and differences in examinee group proficiency affect equating results for mixed-format tests. The tests were modified in such a way that the characteristics of the tests and groups of examinees no longer represented the AP tests

as administered operationally. Consequently, generalizations of the results and findings from this dissertation should not be made to the AP tests.

Three factors were considered when selecting the specific AP tests for this dissertation as follows: subject area, observed and disattenuated MC and CR correlations, and number of examinees. The first factor considered was the test subject area. A variety of subject area tests were desired in order to determine whether similar patterns of findings occurred across different subject areas. Three tests were selected spanning two broad subject areas as follows: science and language. Specifically, this dissertation investigated tests of English Language, Spanish Language, and Chemistry. For this dissertation, one operational equating relationship and corresponding datasets were selected for each of the three tests.

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

Table 3-1 contains a summary of the number of MC and CR items for each test by year of administration. The first column in Table 3-1 lists the subject area test, and the second column provides the year of administration. The third column provides the

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

The second factor considered in selecting tests was observed and disattenuated correlations between MC and CR scores. A range of correlations was desired in order to investigate how the MC and CR correlation impacts equated scores. Observed MC and CR correlations were calculated using Pearson correlations and disattenuated MC and CR correlations were calculated using the observed Pearson correlations and coefficient α estimates of reliability. Three levels of observed correlations were selected: less than 0.70, 0.70 to 0.80, and greater than 0.80. These correlations were calculated using formula scored data in order to select tests for use in this dissertation. After the tests were selected, correlations were recalculated using imputed number correct data, as described in the next section. The correlations based on the two types of scoring were similar in magnitude.

English Language was the test selected with an observed MC and CR correlation less than 0.70. English Language also had the lowest disattenuated correlations (approximately 0.75). Spanish Language was the test selected with an observed MC and CR correlation between 0.70 and 0.80. Chemistry was the third test selected, with an observed MC and CR correlation greater than 0.80. Chemistry had the highest disattenuated correlations of all the tests (approximately 0.95). It is important to note that because a different subject area was selected for each level of correlation, subject area and MC and CR correlation were completely confounded. However, the operational AP science exams tended to have higher MC and CR correlations, and the foreign language exams tended to have moderate correlations. A third factor considered in selecting tests for this dissertation was the number of examinees who had taken each test. As described later in this chapter, samples of examinees were selected from the full sample of test takers in order to address the first and second research questions. In order to ensure adequate sample sizes of examinees, a minimum of 10,000 examinees for each test form was required.

It is important to note that, operationally, for the foreign language exams, examinees are included in equating analyses if they do not regularly speak the foreign language in their home and have not lived for more than one month in a country where the foreign language was primarily spoken. However, in order to ensure adequate sample sizes for Spanish Language analyses, all examinees were used in this dissertation.

Data Preparation

For this dissertation, number-correct scoring was used. However, formula scoring was used for the original operational test forms. Examinees were advised that they would be penalized for incorrect answers on the MC items. Depending on the number of MC answer choices, -0.25 or -0.33 points were subtracted from examinee scores for incorrect MC answers. Consequently, some examinees had missing responses for a large number

of MC items. In order to use number-correct scoring, examinees completing fewer than 80% of the MC items were removed. Then, formula scores were transformed to number-correct scores. Incorrect responses were coded as 0, correct responses were coded as 1, and imputation was used for the missing responses.

The two-way imputation procedure described by Sijtsma and van der Ark (2003) was used in this dissertation. The person mean across all available items, item mean for each item, and overall mean were calculated. The two-way imputed value was the sum of the person-mean and item mean minus overall mean. If the two-way imputed value was greater than 0.5, a value of 1 (correct) was assigned to the examinee on that particular item. If the two-way imputed value was less than 0.5, a value of 0 (incorrect) was assigned. Descriptive statistics for the imputed data for the three tests are provided in Table 3-2. In Table 3-2, MC, CR, common-item (CI), and composite (CO) scores are provided. For each set of scores, means, standard deviations, skewness, and kurtosis are provided. For MC, CR, and CO, α is also given. For CI, the correlation between composite and common-item scores is given (CO Corr.). Correlations between MC and CR scores are provided in Chapter 4. It is plausible that the specific imputation procedure selected for this dissertation may have impacted equating results. However, the imputation method was not a factor of investigation in this dissertation; therefore, an imputation procedure used in previous research on AP exams was selected.

Operationally, weights used for AP exams are very complex. Weights differ by item format in order to ensure that each item format is given the intended proportion of points according to test specifications. Additionally, because test forms do not contain the same number of items across test forms, weights ensure the number of score points is the same across years. Although the choice of weighting may impact equating results, weighting was not considered as a factor of investigation in this dissertation. Summed scores were used. That is, each MC or CR point was worth one point and was not differentially weighted. Number-correct scoring and summed score weighting was chosen

for this dissertation, because current psychometric software cannot easily handle non-integer scores. Additionally, AP exams may use number-correct scoring rather than formula scoring in the future. As discussed previously, because of the use of imputation and summed score weighting, as well as the additional manipulations to the data that are discussed later, results from this dissertation should not be generalized to the AP exams.

Dimensionality Assessment

In order to evaluate whether the IRT assumption of unidimensionality held for the tests, a dimensionality assessment was conducted. For this dissertation, disattenuated correlations between MC and CR scores were considered, and a principal components analysis (PCA) was conducted using tetrachoric and polychoric correlations among all individual items to assess the assumption of unidimensionality.

In order to evaluate the results, disattenuated MC and CR correlations near 1.00 indicated that the MC and CR sections of the test were measuring essentially the same dimensions. For the PCA, four rules of thumb were used to determine whether a test was sufficiently unidimensional for IRT analyses. The first guideline was to retain only those components with eigenvalues greater than one (Orlando, 2004; Rencher, 2002). However, Orlando indicated that despite many eigenvalues being larger than one, a test may still be sufficiently unidimensional for IRT analyses. A second guideline was to examine the scree plots of eigenvalues for a break between what might be considered large and small eigenvalues (Orlando, 2004; Rencher, 2002). A third guideline used to select the number of principal components was to compare the ratio of the first and second eigenvalue to the ratio of the second and third eigenvalues (Lord, 1980; Divgi, 1980; Hattie, 1985; Jiao, 2004). The difference between the first and second principal components was divided by the difference between the second and third principal components. If this ratio was larger than 3, then the first principal component was considered strong relative to the other principal components, indicating a unidimensional tendency. Lastly, Reckase (1979) recommended that 20% or more of the total variance should be explained by the first

principal component in order for a test to be considered unidimensional. Based on these four criteria, decisions about the unidimensionality of each test were made.

Operational Test Form Analyses

The first class of data considered in this dissertation was based on operational test forms. Operational test form analyses were conducted with the test forms intact. That is, no changes were made to which items were included on each test form, and equating was conducted using the operational items and operational MC-only common items. However, as described later in this section, different samples of examinees were selected for the operational test form analyses. This section on operational test form analyses is divided into two additional sections: factors of investigation and evaluation. Each of these two sections contains a number of subsections that outline the specific procedures for the operational test form analyses. The procedures described for the operational test form analyses were repeated for each of the three tests selected for the dissertation. The operational test form analyses primarily addressed research questions one through three:

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

Factors of Investigation

Corresponding to the first three research questions, three factors of investigation were studied for the operational test form analyses as follows: examinee common-item effect sizes, MC and CR relative difficulty, and equating methods.

Examinee Common-Item Effect Sizes

The examinees taking the operational old and new test forms were very similar in terms of common-item means and standard deviations. Consequently, in order to investigate research Question One, it was necessary to create differences in examinee score means across test forms for common-item scores. Differences in common-item score means were created using a sampling process based on demographic variables in order to create various effect sizes across old and new test forms for common-item scores. Common-item effect sizes (CI ES) indicated the standardized amount mean common-item scores of the examinees differed across old and new test forms. Demographic variables were used to sample examinees rather than sampling based on common-item scores, because sampling based on common-item scores could introduce correlated error. The sampling process used to create levels of CI ES is described in the following steps.

Step 1.1: Select Demographic Variables

Information was available for examinees for six demographic variables: gender, ethnicity (Ethnicity), grade in high school, reduction of test fee, region of the United States where the examinee lived, and highest level of education attained by either parent (Ed Parents). Only the demographic variables of Ed Parents and Ethnicity were considered in this dissertation. There were five levels of the variable Ed Parents and eight levels of the variable Ethnicity. One of the levels for each of the variables was for missing responses. Ed Parents and Ethnicity were selected because of the large number of levels as well as the extensive range of CI ES across old and new test forms. Table 3-3 contains a description of the levels for both of these variables. It is important to note that categories 5 and 6 were collapsed into other levels; consequently, there are only eight levels.

Step 1.2. Categorize Examinees on Demographic Variables

Figure 3-1 illustrates the process of creating categories based on the two demographic variables. There are two contingency tables in Figure 3-1, each containing the five levels of Ed Parents across the top and the eight levels of Ethnicity along the left. The table on the left of Figure 3-1 represents the old form, and the table on the right represents the new form. The intersection between the two variables resulted in a total of 40 possible categories into which examinees could be classified. The forty possible categories are illustrated by the cells in the middle of the tables. The first cell in each table, "00", indicates the category where examinees were classified if they had a "0" for Ed Parents and a "0" for Ethnicity. Similarly, the second cell in the first row of each table contains "10". Examinees were classified into this cell if they had a "1" for Ed Parents and a "0" for Ethnicity. For each form separately, examinees were classified into only one of the 40 cells in the table.

Figure 3-2 provides an example of the classification process for Spanish Language. As in Figure 3-1, the table on the left of the figure is for the old form (Spanish Language 2004), and the table on the right is for the new form (Spanish Language 2006). The value in the first cell of the Spanish Language 2004 table on the left of Figure 3-2 (125) is the number of examinees classified into the "00" category. Similarly, the value in the first cell of the Spanish Language 2006 table on the right of Figure 3-2 (333) is the number of examinees classified into the "00" category. When all of the cells of the Spanish Language 2004 table are summed, the total is 19,010, which is the number of examinees taking Spanish Language 2004 after imputation (Table 3-2). Similarly, 18,022 is the number of examinees taking Spanish Language 2006 after imputation (Table 3-2).

Step 1.3: Calculate Effect Sizes

The third step in the creation of differences in common-item means was to calculate effect sizes based on common-item scores. Effect sizes were calculated for all

possible combinations of the old form and new form categories shown in Figure 3-1. Figure 3-3 illustrates how this process was conducted. Across the top of the table in Figure 3-3 are the 40 new form categories. Along the left side of the table are the 40 old form categories. Each of the 40 new form categories was paired with each of the 40 old form categories. The first cell in the table in Figure 3-3 illustrates that one CI ES across the old and new forms was obtained by calculating the CI ES for only those examinees in the old form category of "00" and the new form category of "00". The last cell in the first row of the table illustrates that another CI ES was calculated for only those examinees in the new form category of "49" and the old form category of "00". For the old form category of "00", there were 40 possible CI ES. Similarly, for the old form category of "01", there were also 40 possible CI ES. The entire process of calculating CI ES resulted in a possible total of 1,600 CI ES across the old and new test forms. ES were calculated for new form minus old form common-item scores, as shown in Equation (3.1)

$$ES = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}}, \tag{3.1}$$

where

$\bar{x}_1 =$ Mean for new form

$\bar{x}_2 =$ Mean for old form

$n_1 =$ Number of examinees for new form

$n_2 =$ Number of examinees for old form

$s_1^2 =$ Variance of scores for new form

$s_2^2 =$ Variance of scores for old form

Step 1.4: Select Target Levels of CI ES

Four levels of CI ES were chosen for this dissertation: 0.00, 0.20, 0.40, and 0.60. For the operational test forms, only 0.00, 0.20, and 0.40 could be obtained. The lower two levels were selected to represent differences in mean proficiency that might be seen in practice. The highest levels were selected to represent a large CI ES. Kolen and Brennan (2004) stated that differences of 0.30 standard deviation units or more could result in large differences across equating methods. The sampling process used to create the selected levels of CI ES is described after discussing the MC and CR relative difficulty factor of investigation, because these conditions were jointly considered in sampling examinees.

MC and CR Relative Difficulty

If MC and CR items measure different constructs, it is plausible that examinee proficiency may differ across item formats. When the common items contain only items representative of the MC items, there is an implied assumption that examinees will perform the same on both MC and CR items. If examinee proficiency differs across formats, having only MC items in the common-item set may be potentially problematic for equating. Research Question Two addressed this possibility. In order to assess the impact differences in the relative difficulty of MC and CR items have on equated scores, a procedure similar to the procedure described for producing group differences in the common-item mean scores was used to create differences in the relative difficulty of MC and CR items.

Steps 2.1-2.2: Select Demographic Variables and
Categorize Examinees

Steps 2.1 and 2.2 were described previously in Steps 1.1 and 1.2 in the section entitled “Examinee Common-Item Effect Sizes”. These two steps were completed only

once, but the categories of examinees were also used in the creation of differences in MC and CR relative difficulty.

Step 2.3: Calculate ES for MC and CR

Step 2.3 is similar to Step 1.3. The same formula was used to calculate effect sizes, and the same procedure was used to determine all possible combinations of categories across old and new forms. However, effect sizes were calculated for new form scores minus old form scores for both MC and CR scores, separately, on the entire form (not just on the common items) for all possible combinations of demographic variables. The process is described further with an example in the description of sampling examinees in the following section.

For the operational test forms selected for this dissertation, the number of MC or CR items differed across old and new forms. When the number of items differed across forms, a weighted MC or CR score was calculated in order to place scores from the two forms onto a similar scale. For example, Spanish Language 2004 contained 90 MC items and Spanish Language 2006 contained 75 MC items. Examinees would appear to have higher scores on Spanish Language 2004 because there were fifteen additional items on this form. To make MC scores comparable across forms, MC items on Spanish Language 2006 were multiplied by a weight of 90/75 (1.20). This weighted score was used only for calculating effect sizes. Additionally, the weighted score was used only for effect sizes for operational test form analyses, because pseudo-test forms were constructed to have the same number of score points. After conducting analyses for English Language and Spanish Language, it was determined that the magnitude of the unweighted effect sizes may have impacted equating results rather than the magnitude of the weighted effect sizes. Therefore, sampling for Chemistry was based on the unweighted effect sizes in order to determine whether this was the case. The Chemistry 2005 and 2007 operational

test forms differed by only two points. Therefore, the unweighted and weighted effect sizes were very similar.

Table 3-4 compares descriptive statistics for the unweighted and weighted scores. The first block of data in Table 3-4 contains descriptive statistics for the unweighted scores, and the second block of data in Table 3-4 contains descriptive statistics for the weighted scores. The first column corresponds to English Language 2007 MC scores, the second column corresponds to Spanish Language 2006 MC scores, and the last column corresponds to Chemistry 2007 CR scores. Weighting was conducted only for these three test forms.

Step 2.4. Calculate Effect Sizes for MC and CR Relative Difficulty

In order to determine the difference in relative difficulty across old and new forms for MC and CR, the difference in effect sizes for MC and CR was calculated. The MC and CR effect sizes calculated in Step 2.3 were used in this step. For all possible combinations of categories (i.e., 1,600 combinations), the CR effect size was subtracted from the MC effect size. This effect size is referred to as the MC-CR ES.

Step 2.5: Select Target Levels of MC-CR ES

Two target levels of MC-CR ES were selected: 0.00 and 0.25. These two levels were combined with the levels of CI ES to create six combinations of CI ES and MC-CR ES for the operational test forms and eight combinations for the pseudo-test forms. The target effect size patterns are shown in Table 3-5. It is important to note that although these were the target effect size patterns, because of limitations of sampling real data, the exact effect sizes could not always be achieved.

Sampling Examinees

The sampling process for creating effect sizes involved the consideration of both MC-CR ES and CI ES. The process is described in the following steps.

Step 3.1: Select a Target Effect Size Pattern

The first step in creating samples of examinees was to select a target effect size pattern from Table 3-5 for which to create samples of examinees. As an example of the sampling process, CI 0.00 MC-CR 0.00 was selected. This pattern represents a CI ES of approximately 0.00 and an MC-CR ES of approximately 0.00. (It is important to note that in practice, the actual effect size was only approximately equal to the target effect size patterns in Table 3-5. This is the result of using existing data to create the effect size patterns.)

Step 3-2: Select an Initial Sample of Examinees

Based on the target effect size pattern, an initial sample of examinees was selected by choosing combinations of categories (Figure 3-3) with effect sizes near the values in the target effect size pattern. In the current example, the target CI ES was 0.00, and the target MC-CR ES was 0.00. Consequently, all combinations of categories with CI ES from 0.00 to 0.05 and MC-CR ES from 0.00 to 0.05 were selected from the 1,600 total combinations. This step of the process is illustrated in Table 3-6 for Spanish Language 2004-2006. Five combinations of categories fit the initial requirement. Starting with the first row in Table 3-6, examinees on the old form (Spanish Language 2004) were eligible to be sampled if they had a "1" for Ed Parents and a "7" for Ethnicity. Examinees were eligible to be sampled on the new form (Spanish Language 2006) if they had a "2" for Ed Parents and an "8" for Ethnicity. After all combinations were selected, one combination was randomly drawn from all combinations fitting the criteria. The number of examinees sampled from each row was determined by the number of examinees in the category. In order to ensure that the old and new forms had the same number of examinees, the same

number of examinees was sampled from each form every time a sample was drawn. Consequently, the number sampled was based on the smallest number of examinees across the old and new forms. For example, in the first row of Table 3-6, 288 examinees were eligible to be sampled for Spanish Language 2004; however, only 62 examinees were eligible for sampling on Spanish Language 2006. Consequently, only 62 examinees were eligible to be sampled in the first round of sampling. For all rounds of sampling, examinees were randomly sampled without replacement. After the first sample of examinees was drawn, CI, MC, CR, and MC-CR ES were calculated. The effect sizes were evaluated against stopping rules, which are discussed in the next step. The same process was repeated for each of the remaining four rows in Table 3-6.

Step 3.3: Iterate the Sampling Process

After each sample of examinees was drawn, it was necessary to determine whether further sampling iterations were needed. Additional iterations were necessary if any of the following five outcomes occurred:

1. The CI ES was lower than the range of acceptable values. (Acceptable values were ± 0.03 of the target effect size.)
2. The CI ES was higher than the range of acceptable values.
3. The MC-CR ES was lower than the range of acceptable values.
4. The MC-CR ES was higher than the range of acceptable values.
5. The sample size was not within the range of 1,900 to 2,100. (For Chemistry, sample sizes of only 1,500 could be obtained.)

To illustrate how the next step in the sampling process was carried out, consider the first outcome: CI ES *lower* than the range of acceptable values. If CI ES was *lower* than the range of acceptable values, then all category combinations with CI ES *higher* than the current CI ES were eligible for sampling. Additionally, one of three subconditions occurred in combination, so a second restriction was placed on which

category combinations were eligible for sampling. The three subconditions are listed below as well as the sampling process that resulted from each subcondition (for a CI ES *lower* than the range of acceptable values).

1. The MC-CR ES was lower than the range of acceptable values.

   Category combinations with CI ES *higher* than the current CI ES and MC-CR ES *higher* than the current MC-CR ES were eligible for sampling.

2. The MC-CR ES was higher than the range of acceptable values.

   Category combinations with CI ES *higher* than the current CI ES and MC-CR ES *lower* than the current MC-CR ES were eligible for sampling.

3. The MC-CR ES was within the range of acceptable values.

   Category combinations with CI ES *higher* than the current CI ES and MC-CR ES within $\pm$ 0.10 of the current MC-CR ES were eligible for sampling.

As described for the initial sample, examinees were sampled from one category combination at a time. After a category combination was randomly sampled, effect sizes were calculated, and the outcomes were evaluated. If one of the other five outcomes occurred, different category combinations were selected. For example, if after evaluation it was determined that the CI ES was now *higher* than the acceptable range of ES, category combinations with CI ES *lower* than the current CI ES were eligible for sampling. A function was created in R (R Development Core Team, 2009) to carry out the sampling process described above. This sampling process iterated until the CI ES and MC-CR ES were within the range of acceptable values, and the sample size was between 1,900 and 2,100, or 1,500 for Chemistry. One important note is that the target effect size patterns could not always be created. When this occurred, the CI ES was kept within $\pm$0.03 of the target effect size, and the closest possible MC-CR ES was obtained. Descriptive statistics and ES are provided in Chapter Four for all operational test form and pseudo-test form samples for English Language, Spanish Language, and Chemistry.

Equating Methods

For each of the CI ES and MC-CR ES conditions, equating was conducted for the CINEG design using the four equating methods described in Chapter Two: FE, CE, TS, and OS. For the traditional equating methods, FE and CE, equating was conducted using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009). Cubic spline postsmoothing was conducted using eight *S*-values ranging from 0 to 1.00. The *S*-value that resulted in the smoothest distribution within the 0.5 and 99.5 percentiles while remaining within ± 1 standard error of the unsmoothed equating relationship was selected to smooth the equated results. For FE, the new form population received a weight of one to create synthetic populations. A new form weight of one was chosen, because this weight was used in similar research for the same tests. It is plausible that different weights may have led to different equating results, especially when the old and new form groups differed substantially. However, choice of new and old form weights was not a factor of investigation in this dissertation.

The equating process was replicated using 500 bootstrap replications. For the bootstrap replications, a sample of the same size was selected with replacement from the sample selected for a given effect size pattern. For example, for Spanish Language, 1,900 examinees were sampled with replacement from the existing sample of 1,900 examinees for each form. The equating process was then repeated on this bootstrap sample of examinees. It is important to note that the sample of examinees was a bootstrap sample, meaning that examinees were resampled from the sample of 1,900 examinees. A new sample was not selected from the original operational data. The effect size combinations were often difficult to obtain, requiring that most or all of the examinees with similar effect size patterns be selected in order to obtain the effect sizes. Consequently, if multiple samples were selected from the same original operational data, it was likely that the examinees would be the same. Therefore, it was determined that bootstrap samples

would likely result in samples that were less similar than what could be obtained from selecting multiple samples from the original operational data.

Before conducting equating for the IRT methods, it was necessary to first obtain item parameter estimates and estimates of the distribution of proficiency. Estimates were obtained using PARSCALE (Muraki & Bock, 2001). PARSCALE was used because of its capacity to estimate item parameters for both dichotomous and polytomous item types. Additionally, PARSCALE was chosen because of prior use of the program in previous research studies. The 3PL model was used to estimate parameters for the MC items, and the GPCM was used to estimate parameters for the CR items. The 3PL and GPCM were chosen because of previous work with these two models on another research study involving mixed-format tests. The new form item and proficiency parameter estimates were placed onto the scale of the old form using STUIRT (Kim & Kolen, 2004) and the Haebara method of scale linking. TS and OS equating were conducted for raw scores using POLYEQUATE (Kolen, 2003). STUIRT and POLYEQUATE were chosen for their compatibility with both dichotomous and polytomous item formats. In order to maintain consistency with FE, the new form population received a weight of one for OS. The entire IRT equating process was also replicated using 500 bootstrap replications, as described for the traditional equating methods.

The process for conducting bootstrap replications for IRT equating methods was complex and involved a series of computer software packages. It was impossible to check all of the output for accuracy. Therefore, a number of steps were taken in an attempt to ensure that the IRT analyses were accurate. When PARSCALE could not estimate a pseudo-chance item parameter, as evidenced by 0 for the item parameter and standard error estimates, that item parameter was fixed to 0 in the command file. However, for a given test form, the same control card was used across all bootstrap replications. Additional checks were also built into the IRT standard error of equating computer code.

These checks included removing replications when PARSCALE did not converge and removing replications for which the equating results seemed unreasonable.

Evaluation

The two factors of investigation for the operational test forms resulted in 6 operational test form analysis conditions for four different equating methods, as summarized in Table 3-7. A separate criterion equating was established for each of the four equating methods. This criterion equating was for CI 0.00 MC-CR 0.00 (shown in the first row of Table 3-7). This pattern was selected as the criterion because group differences were closest to zero; consequently, assumptions were least likely to be violated. Additionally, this pattern essentially represented no difference in proficiency between old and new forms and no difference in the relative difficulty of MC and CR items for old and new form examinees. Although this criterion is reasonable, it is not an absolute criterion. It is important to note that the criterion equating relationship differed for each equating method. That is, the criterion equating relationship for FE was calculated using the FE equating method for the CI 0.00 MC-CR 0.00 sampling condition. The criterion equating relationship for CE was calculated using the CE equating method for the CI 0.00 MC-CR 0.00 sampling condition. The criterion equating relationship for TS was calculated using the TS equating method for the CI 0.00 MC-CR 0.00 sampling condition. The criterion equating relationship for OS was calculated using the OS equating method for the CI 0.00 MC-CR 0.00 sampling condition. To evaluate the results from the operational test form analyses at each raw score point, conditional bias, root mean squared error (RMSE), conditional standard error of equating (CSE), and conditional Difference were calculated and plotted for each score point. Classification consistency, weighted average root mean squared bias (WARMSB), weighted average RMSE (WARMSE), the weighted average standard error of equating (WASE), and Difference were calculated as overall summary measures.

Bias, CSE, RMSE, and Difference

Bias, CSE, RMSE, and Difference were used to evaluate equating results at each score point. Equations 3.2, 3.3, 3.4 and 3.5 represent these statistics.

$$Bias_i = \frac{\sum_{j=1}^{J}[\hat{e}_j(x_i) - e^*(x_i)]}{J} \tag{3.2}$$

$$CSE_i = \sqrt{Var_i\left[\hat{e}_j(x_i) - \bar{\hat{e}}(x_i)\right]} \tag{3.3}$$

$$RMSE_i = \sqrt{Bias_i^2 + CSE_i^2} \tag{3.4}$$

$$Difference_i = \hat{e}(x_i) - e^*(x_i) \tag{3.5}$$

In these three equations, *i* is a score point, *j* is a replication, *J* is the total number of replications, $e^*(x_i)$ is the old form equivalent of a new form raw score for the criterion equating relationship, and $\hat{e}_j(x_i)$ is the old form equivalent of a new form raw score for a study condition equating relationship. Values of Difference are similar to WARMSB, but they are based on only the original sample of data for each study condition equating relationship. That is, Difference was not based on bootstrap replications. 500 replications of the criterion equating relationship were conducted in order to compare the magnitude of the standard errors for the criterion to the standard errors for the study condition equatings. The study conditions were evaluated against the mean criterion equating over 500 replications. Plots of conditional bias, RMSE, CSE, and Difference were created to graphically represent differences between the criterion and study condition equating relationships for a given equating method.

WARMSB, WASE, WARMSE, and Difference

Statistics were also calculated in order to summarize the amount of error over the entire score scale. Equations 3.6, 3.7, 3.8, and 3.9 contain formulas for WARMSB, WASE, WARMSE, and Difference.

$$WARMSB = \sqrt{\sum_i w_i Bias_i^2} \tag{3.6}$$

$$WASE = \sqrt{\sum_i w_i CSE_i^2} \tag{3.7}$$

$$WARMSE = \sqrt{\sum_i w_i RMSE_i^2} \tag{3.8}$$

$$Difference = \sqrt{\sum_i w_i Difference_i^2} \tag{3.9}$$

In equations 3.6 through 3.9, bias, CSE, RMSE, and Difference are from Equations 3.2, 3.3, 3.4, and 3.5. As described previously, *i* is a score point. w_i is the proportion of examinees scoring at each new form score. Weighted statistics were used because the number of examinees scoring at each score point was not the same. Typically, there were a number of score points where no examinees scored. That is, there were no data at some of the score points. Consequently, it was desirable to weight those score points so that they received less weight relative to scores where a large number of examinees scored. It is important to note that the choice of weighting selected for this dissertation is one of many weighting schemes that could have been considered. Consequently, generalization of results should be limited to this type of weighting. Standardized WARMSB and WASE were also calculated. The values were standardized by the standard deviation of the old form for the study condition equating relationship.

Classification Consistency

For this dissertation, number-correct scoring and summed score weighting were used; however formula scoring and non-integer weights were used operationally. Therefore, it was not desirable to use the operational cut scores. New cut scores were obtained by setting cut scores so that similar percentages of examinees were classified

into a given grade level for both the original operational test forms and the operational test forms in this dissertation. For example, if 15% of examinees received a grade of 1 on the original operational forms, the cut score was set for the operational test form analyses in this dissertation so that 15% of examinees received a 1. This was carried out by finding the score near the 15^{th} percentile and adding 0.5, as was done operationally. This score was the cut score between the 1 and 2 grade levels. Four cut scores were set, resulting in five proficiency levels. Cut scores were first set for the old form for the criterion sample (i.e., CI 0.00 MC-CR 0.00). Then, the new form for the criterion sample was equated to the old form for the criterion sample to find the new form cut scores. Each equating method potentially resulted in different new form cut scores. For each of the four equating methods separately, all examinees taking the original operational test form (i.e., not just the 1,900 examinees in the sampling condition) were assigned a grade based on the cut scores for the criterion equating relationship.

Cut scores on the new form were also obtained for each of the sampling conditions (e.g., CI 0.20 MC-CR 0.00) in the same manner. For each equating method separately, classification consistency was determined by calculating the percentage of examinees receiving the same grade based on the cut scores for the criterion equating relationship and for a given study condition equating relationship. Cut scores on the new form for the criterion equating relationships are provided for all operational test form and pseudo-test forms following discussion of the pseudo-test form methodology.

Pseudo-Test Form Analyses

Pseudo-test forms were created by dividing the items from one operational test form into two similar test forms. In this dissertation, data on pseudo-test forms were used as a class of data to compare results obtained from operational test form analyses and to examine the impact of test and common-item composition on equating, which could not be investigated through operational test form analyses. The section on pseudo-test forms

is divided into five additional sections. The first section addresses the general procedures used to construct the pseudo-test forms. Sections two through four summarize the factors investigated in this dissertation, and the last section addresses the procedures used for evaluating the equating results. Each section contains smaller subsections pertaining to the details of this dissertation. Pseudo-test forms were used to address research questions one through four. Research questions one through three were investigated using pseudo-test forms in the same manner as described for the operational test form analyses. Additionally, pseudo-test form analyses addressed research question four:

4. How do the content and statistical specifications of a test (e.g., subject area, correlation between MC and CR scores, and composition of common items) impact equated scores?

Construction of Pseudo-Test Forms

The three tests used for operational test form analyses were also used for the pseudo-test form analyses. Only one year of administration was selected for each of the three tests.

Pseudo-Test Forms

Pseudo-test forms were created by splitting one test form into two new test forms. An example of this process is shown in Figure 3-4. On the left of the figure is the Spanish Language 2004 operational test form containing 90 MC items and 27 CR items. On the right of the figure are two pseudo-test forms, created from the operational Spanish 2004 test form on the left. The items on the operational Spanish 2004 test form were split in order to create two separate test forms. In the example in Figure 3-4, the new pseudo-test form and old pseudo-test form each contain 21 unique MC items and seven unique CR items. In addition, the old pseudo-test and new pseudo-test form contain 46 MC items and 13 CR items in common with each other. All of the items contained in the new pseudo-test form and old pseudo-test form were originally on the operational Spanish

Language 2004 form. It is important to note that although the old and new Spanish Language pseudo-test forms contain 46 of the same MC items and 13 of the same CR items, only a subset of these items were actually used as common items for analyses, which is discussed in a later section.

Pseudo-test forms were created to be as similar as possible in terms of content and statistical specifications based on the sample of examinees completing at least 80% of the items. That is, the pseudo-test form datasets were constructed before sampling examinees based on demographic variables. The amount of content information available for some tests or item formats was limited. When information was available, items were assigned to the pseudo-test forms in such a way as to create forms with similar proportions of items from each content category. Pseudo-test forms were also created to be similar in terms of the mean and standard deviation of difficulty and discrimination. Difficulty was calculated for each item as the mean item score over all examinees. For polytomous items, the mean item score was divided by the total number of points possible for the item to put the difficulty of the items on the same scale as that of the MC items. Pearson correlations between each MC or CR item and the total scores were calculated as discrimination.

Factors of Investigation

Four factors of investigation were considered for the pseudo-test form analyses. Three of the factors considered for the operational test form analyses were also considered for the pseudo-test form analyses: CI ES, MC-CR ES, and equating method. The same methodology and conditions used to investigate these factors for the operational test form analyses were also used for the pseudo-test form analyses. Two additional factors of investigation were considered for the pseudo-test form analyses: format representativeness of common items and statistical representativeness of common items.

Examinee Characteristics

Factor 1: Examinee Common-Item Effect Sizes

Examinees were separated into subgroups using the same sampling procedure described for the operational test form analyses. The same conditions for the operational test form analyses were also used for the pseudo-test form analyses, with the addition of CI 0.60.

Factor 2: MC and CR Relative Difficulty

The same sampling process used for the operational test form analyses was also used for the pseudo-test form analyses to create differences in the relative difficulty of MC and CR items. The same conditions for the operational test form analyses were also used for the pseudo-test form analyses.

Equating Methods

Factor 3: Equating Methods

The four equating methods described for the operational test form analyses were also implemented for the pseudo-test form analyses.

Test Composition

Test length was not a factor of investigation in this dissertation. The length of each pseudo-test form was determined by the subject area and number of items available on the original operational form. Additionally, although the ratio of MC to CR points was not considered as a factor of investigation, the ratio of MC to CR points was similar to the ratio of MC to CR points for the operational test forms.

Common-Item Composition

Content representation of the common-item set was not considered in this dissertation, because a number of factors limited the feasibility of developing content

representative common items. For some of the tests used for the current dissertation, only limited content information was available. An additional complication to varying the degree of content balance was that many of the items, both MC and single-point CR items, were part of a testlet. For these items, it was not reasonable to place items from the same testlet on different test forms. A third challenge in creating differences in content representativeness was that to some extent, different formats also represented different content areas. It appeared that content and format were often confounded and difficult to separate. However, in order to ensure reasonable content representation, common-items from the operational test administrations were used whenever possible.

One common-item set length was considered for this dissertation. A length of approximately 30% of the total points on the test was chosen for the length. One common recommendation for common items is that the common items should be at least 20% of the length of the total test (Kolen & Brennan, 2004). Many of the CR items on the tests in this dissertation had large point values. Consequently, restricting the common-item set to only 20% of the length of the total test limited the extent to which MC and CR items could be used in combination as common items. Therefore, a longer common-item length was selected in order to be able to include an adequate number of points allocated to MC and CR items.

Two factors of investigation were considered in creating the common items as follows: format representativeness and statistical representativeness. As stated earlier, one of the primary goals of this dissertation was to determine when an MC only common-item set might result in an adequate equating and when the inclusion of CR items may improve the equating relationship. In order to hold test characteristics constant, the items on each test form were the same across common-item conditions. The only aspect that varied was which items were treated as common items. Consequently, both the old and new pseudo-test forms contained some items in common that were not treated as common for all of the common-item set conditions. Additionally, it is

important to note that the majority of the MC common items were common items on the original operational test forms.

Factor 4: Format Representativeness

Two levels of format representativeness were used in this dissertation: **N**o **CR** representation (NCR) and **F**ull **CR** representation (FCR). No CR representation means that the common items contained only MC items. Full CR representation means that the items in the common-item set were proportionally representative of the formats on the total test. That is, the common-item set contained 30% of the total points allocated to MC and 30% of the total points allocated to CR. This factor was investigated for all three of the subject area tests. However, for English Language and Chemistry, because CR point values were large for all of the items, the actual percentage of CR points was 50% of the CR points on the total test for English Language and approximately 20% of the CR points on the total test for Chemistry.

Factor 5: Statistical Representativeness

Statistical representativeness of common items in this dissertation meant the degree to which the common-item set was representative of the difficulty and discrimination in the total test. Given that MC and CR items are often intended to measure different constructs, it is plausible that one item format may appear more difficult than another. Therefore, in order to create content and format representative sets of items in practice, it may be necessary to relax some of the restrictions on statistical representation. Three levels of statistical representativeness of the common items were explored in this dissertation for the NCR common items only: minitest (NCR MT), semi-miditest (NCR SM), and difficulty shift (NCR DS). NCR MT means that the common-item set was a mini version of the test. The common items were representative of the total test in terms of both mean and standard deviation of difficulty and discrimination. The semi-miditest (NCR SM) was based on Sinharay and Holland's (2006, 2007) idea of a set

of common items with a spread of difficulty less than that of the total test. For NCR SM, the mean difficulty and discrimination were similar to the total test, but the standard deviation of difficulty was approximately 0.04 smaller than for the total test. The third common-item condition, NCR DS, represented a shift in the mean difficulty of the common-item set as compared to the total test. The target shift in mean difficulty considered for this dissertation was 0.10 for both test forms. However, a shift in difficulty of only 0.08 was possible. The NCR SM and NCR DS statistical representativeness conditions were only considered for Spanish Language.

Table 3-8 contains information regarding the composition of the pseudo-test forms and common-item sets. The first column indicates old or new pseudo-test forms, the second column contains the test, and the third column contains the item type. For example, total test refers to the items on the total test, NCR refers to MC-only common items, FCR refers to combined MC and CR common items, and Full CI refers to all items in common between the old and new pseudo-test forms. The next two columns contain the number of MC items (number of MC points in parentheses) and number of CR items (number of CR points in parentheses), respectively. The last four columns of data in Table 3-8 contain the mean and standard deviation of difficulty and discrimination.

Using English Language as an example, as shown in the first row of Table 3-8, the old total pseudo-test form has 36 MC items worth 36 points and two CR items worth a total of 18 points (or, nine points each). As shown in the last row of the English Language block, there were 19 MC items and one CR item worth nine points in common between the old and new pseudo-test forms. The NCR set (second row) contained 17 MC items and no CR items, and the FCR set (third row) contained eight MC items and one CR item. The new pseudo-test forms contained the same numbers of items as the old pseudo-test forms. For English Language, mean difficulty for FCR was 0.721 (third row and sixth column), and mean difficulty for NCR was 0.756 (second row and sixth column). The mean difficulty for the total test was 0.713 (first row and sixth column).

Therefore, the mean difficulty for FCR was more similar than NCR to the mean difficulty for the total test. Conversely, mean discrimination for NCR (second to last column and second row) was more similar than FCR (second to last column and third row) to the mean discrimination for the total test (second to last column and first row). As can be seen in the third to last column, the standard deviation of difficulty for NCR was more similar than FCR to the standard deviation of difficulty for the total test. The opposite was true for the standard deviation of discrimination, shown in the last column. Similar trends were seen for the new pseudo-test form, although the pattern of results for discrimination was slightly different.

As another example, for Spanish Language (second block of data), the three NCR sets represent the three levels of statistical representativeness for common-item sets. All three of the NCR sets contained 33 MC items worth 33 points. The FCR set contained 21 MC items worth 21 points and six CR items worth a total of 12 points. As intended, mean difficulty for NCR MT, NCR SM, and FCR were quite similar to each other and to the mean difficulty for the total test, as shown in the sixth column. Also, as intended, mean difficulty was lower for NCR DS. Standard deviations of difficulty were also similar to the standard deviation of difficulty for the total test (within approximately 0.03), with the exception of NCR SM. As intended, NCR SM had the lowest standard deviation of difficulty, which was approximately 0.06 less than the total test. Mean discrimination for all of the common-item sets were lower than for the total test.

For Chemistry, the total test contained 54 MC items worth 54 points and four CR items worth a total of 42 points. NCR contained 28 MC items, and FCR contained 19 MC items and one CR item worth nine points. Mean difficulty for NCR and FCR was higher than for the total test, but mean difficulty for NCR and FCR was very similar. The standard deviation of difficulty and mean discrimination for the common-item sets were also very similar to the total test. However, the standard deviation of discrimination was lower for FCR as compared to NCR or the total test.

Tables 3-9 and 3-10 contain descriptive statistics for the English Language, Spanish Language, and Chemistry old and new pseudo-test forms. Table 3-9 contains descriptive statistics for the MC, CR, and CO scores. Table 3-10 contains descriptive statistics for the various compositions of common-item sets. For each set of scores, means, standard deviations, skewness, and kurtosis are provided. For MC, CR, and CO, α is also given. For the common-item compositions (Table 3-10), the correlation between composite and common-item scores is given (CO Corr.). It is important to note that descriptive statistics for the common-item sets were the same for the old and new pseudo-test forms, because the same examinees took both the old and new pseudo-test forms.

Tables 3-11 and 3-12 contain summaries of the factors of investigation for the pseudo-test form analyses. Table 3-11 contains the factors of investigation for English Language and Chemistry. For these two tests, only examinee CI ES, MC-CR ES, common-item format representation, and equating methods were considered. Table 3-12 contains the factors of investigation for Spanish Language. In addition to the factors investigated for English Language and Chemistry, common-item statistical representation was also considered for the NCR common items for Spanish Language.

Evaluation

Two pseudo-test forms were created using a single test form; consequently, the same examinees took both pseudo-test forms. Therefore, the criterion equating relationship was established for each subject area test for the pseudo-tests using the single group equating design. After the pseudo-test forms were created, single group equating was conducted prior to sampling subgroups of examinees. A single group equipercentile equating relationship was used as the criterion for FE and CE. TS and OS were used as the criterion equating relationships for TS and OS, respectively. Concurrent calibration was used for estimating item and proficiency parameters for the single group. Using the single-group equating relationship as the criterion equating relationship rather than CI

0.00 MC-CR 0.00, the same methods of evaluation used for the operational test form analyses were also used for the pseudo-test form analyses. These included bias, RMSE, CSE, WARMSB, WARMSE, WASE, Difference and classification consistency. Cut scores for the criterion samples are shown in Table 3-13. For the operational test form analyses, the criterion sample was the CI 0.00 MC-CR 0.00 sample of examinees. For the pseudo-test analyses, the criterion sample was the original sample of examinees on the old pseudo-test form.

Comparison of Results across Operational Test Forms and Pseudo-Test Forms

The final research question addressed the extent to which operational test form analyses and pseudo-test form analyses yielded results that led to the same conclusions. The primary focus of comparison across operational test form and pseudo-test form analyses was at the general findings level. That is, the similarity of the two classes of data was determined based on whether conclusions were the same across classes of data.

Table 3-1. Description of Selected AP Tests

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

Table 3-2. Descriptive Statistics for English Language, Spanish Language, and Chemistry after Imputation

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

[a] Before imputation

[b] After imputation

Table 3-3. Levels of Ed Parents and Ethnicity

Variable	Level	Description
Ed Parents	0	Not stated
	1	High school degree or less
	2	Trade school, some college, associates
	3	Bachelors, some graduate/professional school
	4	Graduate/professional degree
Ethnicity	0	Not stated
	1	American Indian, Alaska Native
	2	African American
	3	Mexican American
	4	Asian
	7	White
	8	Other
	9	Puerto Rican, Hispanic, Latino

Table 3-4. Comparison of Unweighted and Weighted Scores

Score Type	Descriptive Statistics	English Language 2007 MC	Spanish Language 2006 MC	Chemistry 2007 CR
Unweighted	Mean	35.201	50.656	29.179
	SD	8.774	11.908	15.601
	Skewness	-0.540	-0.425	-0.198
	Kurtosis	2.773	2.627	2.001
Weighted	Mean	35.878	60.787	28.222
	SD	8.942	14.290	15.089
	Skewness	-0.540	-0.425	-0.198
	Kurtosis	2.773	2.627	2.001

Note: English Language weight equals 53/52 (MC on old form)/(MC on new form)

Note: Spanish Language weight equals 90/75 (MC on old form)/(MC on new form)

Note: Chemistry weight equals 59/61 (CR points on old form)/(CR points on new form)

Table 3-5. Target Effect Size Patterns

Pattern	ES	
	CI	MC-CR
CI 0.00 MC-CR 0.00	0.00	0.00
CI 0.00 MC-CR 0.25	0.00	0.25
CI 0.20 MC-CR 0.00	0.20	0.00
CI 0.20 MC-CR 0.25	0.20	0.25
CI 0.40 MC-CR 0.00	0.40	0.00
CI 0.40 MC-CR 0.25	0.40	0.25
CI 0.60 MC-CR 0.00	0.60	0.00
CI 0.60 MC-CR 0.25	0.60	0.25

Note: A CI ES of 0.60 was not considered for operational test form analyses.

Table 3-6. Example of ES Sampling Process for Spanish Language 2004-2006

2004			2006			CI ES	MC-CR ES
Ed Parents	Ethnicity	N	Ed Parents	Ethnicity	N		
1	7	288	2	8	62	0.049	0.024
2	0	81	0	9	444	0.040	0.023
3	2	96	0	2	21	0.028	0.011
3	7	1998	2	9	949	0.009	0.037
4	2	100	2	9	949	0.030	0.009

Table 3-7. Factors of Investigation for Operational Test Form Analyses

Conditions	ES CI	MC-CR ES	Equating Methods
01	0.00	0.00	FE, CE, TS, OS
02	0.00	0.25	FE, CE, TS, OS
03	0.20	0.00	FE, CE, TS, OS
04	0.20	0.25	FE, CE, TS, OS
05	0.40	0.00	FE, CE, TS, OS
06	0.40	0.25	FE, CE, TS, OS

Notes: Each condition was analyzed using four equating methods.

Table 3-8. Composition of Pseudo-Test Forms

Form	Test	Item Type	Number of MC (Points)	Number of CR (Points)	Difficulty		Discrimination	
					Mean	SD	Mean	SD
Old	English Language	Total Test	36 (36)	2 (18)	0.713	0.131	0.375	0.097
		NCR	17 (17)	0 (0)	0.756	0.124	0.358	0.124
		FCR	8 (8)	1 (9)	0.721	0.085	0.416	0.090
		Full CI	19 (19)	1 (9)	0.756	0.103	0.375	0.102
	Spanish Language	Total Test	68 (68)	20 (40)	0.648	0.177	0.376	0.124
		NCR MT	33 (33)	0 (0)	0.649	0.151	0.358	0.097
		NCR SM	33 (33)	0 (0)	0.655	0.119	0.365	0.101
		NCR DS	33 (33)	0 (0)	0.570	0.151	0.348	0.088
		FCR	21 (21)	6 (12)	0.650	0.165	0.364	0.113
		Full CI	46 (46)	13 (22)	0.631	0.169	0.370	0.118
	Chemistry	Total Test	54 (54)	4 (42)	0.588	0.155	0.464	0.139
		NCR	28 (28)	0 (0)	0.641	0.155	0.460	0.113
		FCR	19 (19)	1 (9)	0.636	0.146	0.453	0.083
		Full CI	33 (33)	2 (23)	0.617	0.151	0.475	0.121
New	English Language	Total Test	36 (36)	2 (18)	0.706	0.121	0.402	0.101
		NCR	17 (17)	0 (0)	0.756	0.124	0.364	0.119
		FCR	8 (8)	1 (9)	0.721	0.085	0.420	0.091
		Full CI	19 (19)	1 (9)	0.756	0.103	0.380	0.099
	Spanish Language	Total Test	68 (68)	20 (40)	0.641	0.167	0.382	0.136
		NCR MT	33 (33)	0 (0)	0.649	0.151	0.350	0.107
		NCR SM	33 (33)	0 (0)	0.655	0.119	0.353	0.115
		NCR DS	33 (33)	0 (0)	0.570	0.151	0.333	0.097
		FCR	21 (21)	6 (12)	0.650	0.165	0.355	0.129
		Full CI	46 (46)	13 (22)	0.631	0.169	0.360	0.135
	Chemistry	Total Test	54 (54)	4 (42)	0.601	0.152	0.465	0.128
		NCR	28 (28)	0 (0)	0.641	0.155	0.459	0.115
		FCR	19 (19)	1 (9)	0.636	0.146	0.453	0.084
		Full CI	33 (33)	2 (23)	0.617	0.151	0.475	0.122

Table 3-9. Descriptive Statistics for English Language, Spanish Language, and Chemistry Pseudo-Test Forms

Item Format	Descriptive Statistics	English Language		Spanish Language		Chemistry	
		Old	New	Old	New	Old	New
	N	15,820	15,820	19,010	19,010	12,328	12,328
MC	Mean	26.023	25.735	45.029	44.486	32.159	32.898
	SD	6.042	6.602	11.286	11.795	11.341	11.361
	Skewness	-0.713	-0.774	-0.401	-0.398	-0.332	-0.353
	Kurtosis	3.003	3.036	2.624	2.541	2.094	2.086
	α	0.841	0.867	0.909	0.914	0.926	0.927
CR	Mean	9.564	9.760	25.438	27.110	20.765	20.894
	SD	2.842	2.805	7.030	7.930	10.951	11.269
	Skewness	-0.208	-0.122	-0.614	-0.732	-0.255	-0.233
	Kurtosis	3.075	2.977	2.897	2.662	2.027	1.986
	α	0.575	0.527	0.821	0.818	0.864	0.868
CO	Mean	35.587	35.496	70.467	71.596	52.923	53.792
	SD	7.848	8.378	17.245	18.536	21.498	21.799
	Skewness	-0.628	-0.654	-0.517	-0.575	-0.316	-0.318
	Kurtosis	3.069	3.031	2.809	2.677	2.064	2.047
	α	0.820	0.842	0.929	0.930	0.899	0.898

Table 3-10. Descriptive Statistics for Common Items for English Language, Spanish Language, and Chemistry Pseudo-Test Forms

Item Format	Descriptive Statistics	English Language		Spanish Language		Chemistry	
		Old	New	Old	New	Old	New
	N	15,820	15,820	19,010	19,010	12,328	12,328
NCR MT (NCR)	Mean	12.853	12.853	21.423	21.423	17.941	17.941
	SD	3.042	3.042	5.855	5.855	6.300	6.300
	Skewness	-0.991	-0.991	-0.321	-0.321	-0.521	-0.521
	Kurtosis	3.726	3.726	2.485	2.485	2.229	2.229
	CO Corr.	0.846	0.859	0.913	0.891	0.936	0.936
NCR SM	Mean	--	--	21.625	21.625	--	--
	SD	--	--	6.077	6.077	--	--
	Skewness	--	--	-0.380	-0.380	--	--
	Kurtosis	--	--	2.496	2.496	--	--
	CO Corr.	--	--	0.912	0.883	--	--
NCR DS	Mean	--	--	18.812	18.812	--	--
	SD	--	--	6.159	6.159	--	--
	Skewness	--	--	-0.032	-0.032	--	--
	Kurtosis	--	--	2.333	2.333	--	--
	CO Corr.	--	--	0.883	0.845	--	--
FCR	Mean	10.609	10.609	22.139	22.139	17.107	17.107
	SD	2.783	2.783	5.705	5.705	6.464	6.464
	Skewness	-0.455	-0.455	-0.490	-0.490	-0.461	-0.461
	Kurtosis	2.998	2.998	2.752	2.752	2.253	2.253
	CO Corr.	0.853	0.850	0.934	0.926	0.943	0.944
Full CI	Mean	19.236	19.236	44.194	44.194	32.997	32.997
	SD	4.216	4.216	11.359	11.359	13.242	13.242
	Skewness	-0.841	-0.841	-0.455	-0.455	-0.431	-0.431
	Kurtosis	3.656	3.656	2.700	2.700	2.117	2.117
	CO Corr.	0.918	0.922	0.983	0.969	0.984	0.983

Table 3-11. Factors of Investigation for English Language and Chemistry Pseudo-Test Form Analyses

Conditions	CI ES	MC-CR ES	CI Format	Equating Methods
01-02	0.00	0.00	NCR, FCR	FE, CE, TS, OS
03-04	0.00	0.25	NCR, FCR	FE, CE, TS, OS
05-06	0.20	0.00	NCR, FCR	FE, CE, TS, OS
07-08	0.20	0.25	NCR, FCR	FE, CE, TS, OS
09-10	0.40	0.00	NCR, FCR	FE, CE, TS, OS
11-12	0.40	0.25	NCR, FCR	FE, CE, TS, OS
13-14	0.60	0.00	NCR, FCR	FE, CE, TS, OS
15-16	0.60	0.25	NCR, FCR	FE, CE, TS, OS

Notes: Each condition was analyzed using four equating methods.

Table 3-12. Factors of Investigation for Spanish Language Pseudo-Test Form Analyses

Conditions	CI ES	MC-CR ES	CI Format	CI Statistical (NCR Only)	Equating Methods
01-04	0.00	0.00	NCR, FCR	MT, SM, DS	FE, CE, TS, OS
05-08	0.00	0.25	NCR, FCR	MT, SM, DS	FE, CE, TS, OS
09-12	0.20	0.00	NCR, FCR	MT, SM, DS	FE, CE, TS, OS
13-16	0.20	0.25	NCR, FCR	MT, SM, DS	FE, CE, TS, OS
17-20	0.40	0.00	NCR, FCR	MT, SM, DS	FE, CE, TS, OS
21-24	0.40	0.25	NCR, FCR	MT, SM, DS	FE, CE, TS, OS
25-28	0.60	0.00	NCR, FCR	MT, SM, DS	FE, CE,TS, OS
29-32	0.60	0.25	NCR, FCR	MT, SM, DS	FE, CE, TS, OS

Notes: Each condition was analyzed using four equating methods.

Table 3-13. New Form Cut Scores for the Criterion Equating

Cut	Equating Method	Operational			Pseudo-Tests		
		English Language	Spanish Language	Chemistry	English Language	Spanish Language	Chemistry
1/2	FE	27.5	68.5	34.5	22.5	51.5	36.5
	CE	27.5	69.5	35.5	22.5	51.5	36.5
	TS	27.5	72.5	34.5	22.5	50.5	37.5
	OS	27.5	72.5	34.5	22.5	51.5	37.5
2/3	FE	44.5	88.5	52.5	34.5	68.5	52.5
	CE	44.5	88.5	51.5	34.5	68.5	52.5
	TS	44.5	90.5	51.5	34.5	68.5	52.5
	OS	44.5	90.5	51.5	34.5	68.5	52.5
3/4	FE	55.5	102.5	75.5	40.5	83.5	66.5
	CE	55.5	101.5	75.5	40.5	83.5	66.5
	TS	55.5	102.5	74.5	40.5	83.5	66.5
	OS	55.5	102.5	74.5	40.5	83.5	66.5
4/5	FE	62.5	114.5	95.5	44.5	93.5	76.5
	CE	62.5	114.5	94.5	44.5	93.5	76.5
	TS	62.5	113.5	95.5	45.5	94.5	76.5
	OS	62.5	113.5	95.5	45.5	94.5	76.5

Old Form Categories

Ed Parents

Ethnicity	0	1	2	3	4
0	00	10	20	30	40
1	01	11	21	31	41
2	02	12	22	32	42
3	03	13	23	33	43
4	04	14	24	34	44
7	07	17	27	37	47
8	08	18	28	38	48
9	09	19	29	39	49

New Form Categories

Ed Parents

Ethnicity	0	1	2	3	4
0	00	10	20	30	40
1	01	11	21	31	41
2	02	12	22	32	42
3	03	13	23	33	43
4	04	14	24	34	44
7	07	17	27	37	47
8	08	18	28	38	48
9	09	19	29	39	49

Figure 3-1. Creating demographic variable categories.

Spanish Language 2004
Number of Examinees

Ed Parents

Ethnicity	0	1	2	3	4	
0	125	121	81	106	132	
1	0	7	10	3	12	
2	19	43	86	96	100	
3	404	3479	902	3301	339	
4	34	124	157	356	536	
7	140	288	774	1998	2774	
8	36	86	81	119	205	
9	475	1975	931	649	877	
Total						19,010

Spanish Language 2006
Number of Examinees

Ed Parents

Ethnicity	0	1	2	3	4	
0	333	157	109	129	130	
1	0	3	8	14	6	
2	21	37	86	113	97	
3	402	3394	861	368	340	
4	31	114	159	301	461	
7	153	245	663	1875	2434	
8	17	49	62	114	185	
9	444	1782	949	617	759	
Total						18,022

Figure 3-2. Classification of examinees for Spanish Language 2004-2006.

		New Form Categories								
		1	**2**	**3**	**4**	**5**		**38**	**39**	**40**
		00	**01**	**02**	**03**	**04**	**…**	**47**	**48**	**49**
1	**00**	00-00	01-00	02-00	03-00	04-00		47-00	48-00	49-00
2	**01**	00-01	01-01	02-01	03-01	04-01		47-01	48-01	49-01
3	**02**	00-02	01-02	02-02	03-02	04-02		47-02	48-02	49-02
4	**03**	00-03	01-03	02-03	03-03	04-03		47-03	48-03	49-03
5	**04**	00-04	01-04	02-04	03-04	04-04	…	47-04	48-04	49-04
	…	…						…		
38	**47**	00-07	01-07	02-07	03-07	04-07		47-47	48-47	49-47
39	**48**	00-08	01-08	02-08	03-08	04-08		47-48	48-48	49-48
40	**49**	00-09	01-09	02-09	03-09	04-09		47-49	48-49	49-49

Old Form Categories

Figure 3-3. Calculation of effect sizes for demographic variable categories.

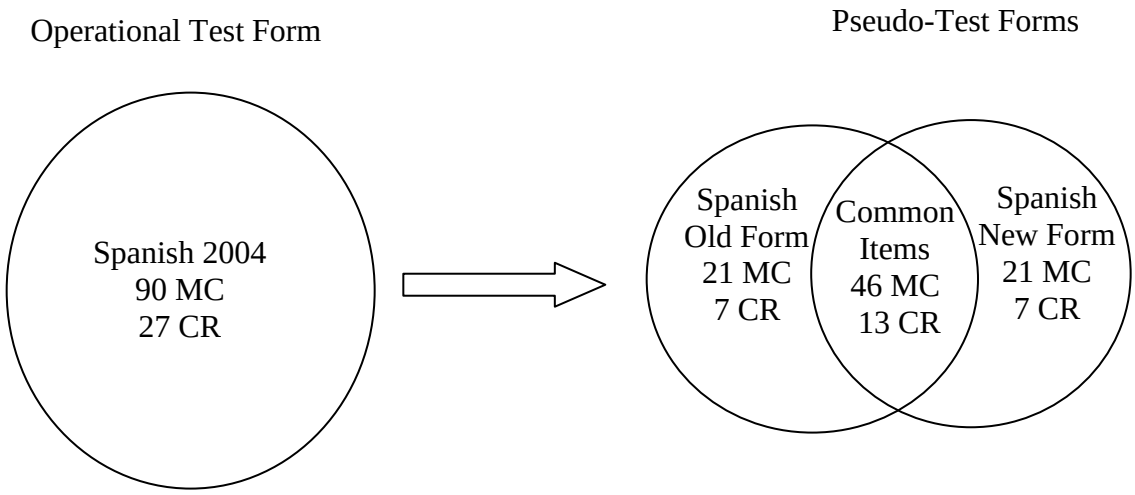

Figure 3-4. Example of creation of pseudo-test forms.

CHAPTER FOUR:
RESULTS

Chapter Four is arranged into four sections of results: dimensionality assessment, cubic spline postsmoothing, operational test forms, and pseudo-test forms. The first section provides results for the dimensionality analyses. The second section illustrates the process of selecting cubic spline postsmoothing values for the equating results. The third section contains results for the operational test forms, and results for the pseudo-test analyses are provided in the last section of Chapter Four. It is important to note that all operational and pseudo-test form samples for English Language and Spanish Language contained 1,900 examinees. The Chemistry operational and pseudo-test form samples contained 1,500 examinees.

Dimensionality Assessment

The dimensionality assessment section is divided into three subsections. Results are presented first for the English Language operational test forms, second for the Spanish Language operational test forms, and last for the Chemistry operational test forms. Two methods were considered for assessing the dimensionality of the tests analyzed in this dissertation: disattenuated MC and CR correlations and principal components analysis (PCA). For each of the three tests, results are presented for the MC and CR observed and disattenuated correlations as well as for the PCA.

English Language

Disattenuated correlations, using Cronbach's α as the estimate of reliability, between MC and CR scores were the first pieces of data considered for assessing dimensionality. Disattenuated correlations near one would indicate the MC and CR items were measuring the same constructs. Table 4-1 contains observed and disattenuated MC and CR correlations for the English Language 2004 and 2007 operational test forms. The

first column in Table 4-1 indicates the English Language operational test form. The second column indicates whether the observed or disattenuated MC and CR correlation is being considered. The first column of data, labeled "Original Operational Form" contains the correlation based on the full sample of examinees for the operational test forms. The next three columns of data contain correlations for the MC-CR 0.00 sampling conditions, and the last three columns of data contain correlations for the MC-CR 0.25 sampling conditions. For example, the column labeled CI 0.00 under MC-CR 0.00 contains MC and CR correlations based on the CI 0.00 MC-CR 0.00 sample of 1,900 examinees. For both test forms and across all sampling conditions, observed correlations were in the range of 0.55 to 0.62. Disattenuated correlations ranged from approximately 0.75 to 0.80. Correlations for the English Language pseudo-test forms are presented in Table 4-2. Table 4-2 is similar to Table 4-1 with two differences. The column labeled "Single Group" contains the correlation based on the single group of examinees. There are also two additional columns for the CI 0.60 sampling conditions. The correlations for the pseudo-test forms tended to be somewhat smaller in magnitude than the correlations for the operational test forms.

PCA was the second method used to assess the dimensionality of the tests. Four criteria, as described in Chapter Three, were used to determine whether the PCA results indicated a test was unidimensional. The first criterion was the number of eigenvalues greater than one. Table 4-3 contains eigenvalues for the first 10 principal components for the 2004 and 2007 English Language operational test forms. Seven eigenvalues were one or greater than one for the English Language 2004 operational test form, and eight were one or greater than one for the English Language 2007 operational test form. The second criterion was based on examining the scree plots for a logical break between eigenvalues. Although there were a number of eigenvalues greater than one, it is evident by examining the scree plots in Figure 4-1 that the first eigenvalue was much larger than the other eigenvalues for both English Language 2004 and English Language 2007. The

eigenvalues did not level off until after the third eigenvalue, suggesting that there may have been two weak components in addition to the first strong component. The third criterion took into consideration the total variance accounted for by the first principal component. For English Language 2004, the first principal component accounted for approximately 28% of the total variance, and the first principal component for English Language 2007 also accounted for approximately 28% of the total variance. The fourth criterion used to evaluate the PCA results was the ratio of the difference between the first and second eigenvalues and the difference between the second and third eigenvalues. As discussed previously, a ratio less than three is one commonly used guideline that indicates possible multidimensionality in a test. For English Language 2004, the ratio of the differences was approximately 39 [(15.61-1.78)/(1.78-1.43)], and for English Language 2007, the ratio was approximately 106 [(14.37-1.60)/(1.60-1.48)]. The conclusions from the four criteria for evaluating the PCA results, along with the disattenuated MC and CR correlations suggest that, although more than one dimension may have existed for the English Language tests, additional dimensions were weak. The same analyses were conducted for the English Language sampling conditions as well as for the pseudo-test forms, and similar results were found.

Spanish Language

Table 4-4 contains observed and disattenuated MC and CR correlations for the Spanish Language 2004 and 2006 operational test forms. For both test forms, and across all sampling conditions, observed correlations were in the range of 0.71 to 0.83. Disattenuated correlations ranged from approximately 0.80 to 0.91. It is important to note that correlations for Spanish Language 2006 were lower than those for Spanish Language 2004, because Spanish Language 2006 contained 15 fewer MC items. Correlations for the Spanish Language pseudo-test forms are presented in Table 4-5. The correlations for the pseudo-test forms were similar in magnitude to the correlations for the operational test

forms. Correlations for the pseudo-test forms were not consistently higher or lower than those for the operational test forms.

Table 4-6 contains eigenvalues for the first 15 principal components for the Spanish Language operational test forms. For Spanish Language 2004, 13 eigenvalues were one or greater than one, and 12 eigenvalues were one or greater than one for Spanish Language 2006. By examining the scree plots in Figure 4-2, it is evident that the first and second eigenvalues were substantially larger than the other eigenvalues for both Spanish Language 2004 and 2006. The two large eigenvalues relative to the other eigenvalues suggest potentially two stronger dimensions. Additionally, the eigenvalues did not level off until after the fourth eigenvalue, suggesting the possibility of two weaker dimensions. However, the first principal component accounted for 27% of the total variance for Spanish Language 2004 and 25% of the total variance for Spanish Language 2006. Additionally, the ratio of the difference between the first and second eigenvalues and the difference between the second and third eigenvalues was approximately 4.5 for 2004 and 4.8 for 2006 test forms. Although this ratio was much closer to three than was found for the English Language operational test forms, it was still larger than three. The conclusions from the four criteria for evaluating the PCA results along with the disattenuated MC and CR correlations suggest it is plausible that more than one dimension may have existed for Spanish Language, although three of the four criteria indicated the test was satisfactorily unidimensional for IRT analyses. The same analyses were conducted for the Spanish Language sampling conditions and pseudo-test forms, and similar results were found.

Chemistry

Table 4-7 contains observed and disattenuated correlations for the Chemistry operational 2005 and 2007 test forms. For both test forms, and across all sampling conditions, observed correlations were in the range of 0.85 to 0.91. Disattenuated

correlations ranged from approximately 0.94 to 0.97, indicating MC and CR items were measuring essentially the same content and/or processes. Correlations for the Chemistry pseudo-test forms are presented in Table 4-8. Correlations for the pseudo-test forms were similar in magnitude to the correlations for the operational test forms. Correlations for the pseudo-test forms were not consistently higher or lower than those for the operational test forms.

Table 4-9 contains eigenvalues for the first 10 principal components for the Chemistry operational test forms. For both Chemistry 2005 and 2007, seven eigenvalues were one or greater than one. By examining the scree plots in Figure 4-3, it is evident that the first eigenvalue was large in magnitude relative to the other eigenvalues for both Chemistry 2005 and 2007. However, the eigenvalues did not level off until after the second or third eigenvalue, suggesting one or two weak dimensions in addition to the first strong dimension. Two additional criteria were also considered. The first principal component accounted for approximately 36% of the total variance for Chemistry 2005 and 34% of the total variance for Chemistry 2007. Additionally, the ratio of the difference between the first and second eigenvalues and the difference between the second and third eigenvalues was approximately 39 for Chemistry 2005 and 21 for Chemistry 2007. The conclusions from the four criteria for evaluating the PCA results, along with the disattenuated MC and CR correlations suggest that, although there may be more than one dimension for the Chemistry operational test forms, additional dimensions are weak. The same analyses were conducted for the Chemistry sampling conditions as well as for the pseudo-test forms, and similar results were found.

Cubic Spline Postsmoothing

Figures 4-4 and 4-5 contain illustrations of the process used for determining smoothing values for English Language for FE. Although only plots for FE are shown in Figures 4-4 and 4-5, the same process was used for CE. The plots in Figure 4-4 illustrate

the process of selecting a smoothing value for the CI 0.00 MC-CR 0.25 sampling condition for FE. In the top plot of Figure 4-4, the equating relationship for a value of S=0.1 is illustrated, and in the bottom plot of Figure 4-4, the equating relationship for a value of S=0.2 is illustrated. The solid line in each plot represents the unsmoothed equating relationship, the dashed line represents the smoothed equating relationship, and the dotted lines represent +/- one standard error of equating for the unsmoothed equating relationship. The dark boxes in the plots indicate an area where the unsmoothed equating relationship falls on or outside the standard error band for S=0.2. This area is also magnified in the box below the equating relationships. In the bottom plot, for S=0.2, the smoothed equating relationship was outside the standard error band inside the box; however, in the S=0.1 plot on the top, the smoothed equating relationship was just inside the standard error band. Figure 4-5 contains the same plots as Figure 4-4, but for the CI 0.40 MC-CR 0.25 sampling condition. For S=0.2, the unsmoothed equating relationship just touches the standard error band. Although S=0.2 could have been chosen for this sampling condition, it was desired to consistently use the same smoothing value across all sampling conditions. The same process was conducted for each of the equating relationships for the different operational test forms, pseudo-test forms, and sampling conditions. In order to maintain consistency across analyses, S=0.1 was selected as the smoothing value for all of the equating analyses.

Operational Test Forms

Results for the operational test forms are divided into three subsections corresponding to the three operational tests analyzed: English Language, Spanish Language, and Chemistry. Results are presented first for the English Language operational test forms, second for the Spanish Language operational test forms, and last for the Chemistry operational test forms. Additionally, for each test, results are presented for descriptive statistics, equated moments, equating relationships, conditional bias,

conditional standard error of equating (CSE), overall summary statistics, and classification consistency. The formats of tables and figures are similar across all three tests; therefore, details about formats are only provided in the English Language subsection. In order to maintain consistency across the tests, MC-CR 0.00 is consistently used to indicate the MC-CR ES that is most similar to the criterion sample. MC-CR 0.25 is consistently used to indicate the MC-CR ES that is most different from the criterion sample. It is important to note that the actual effect sizes differ by test. Also, it is important to remember that the criterion equating relationship differed according to equating method.

English Language

Results for the English Language operational test forms are presented in the following subsections: descriptive statistics, equated moments, equating relationships, conditional bias, conditional standard error of equating (CSE), overall summary statistics, and classification consistency.

Descriptive Statistics

Descriptive statistics were provided for the original operational English Language 2004 and 2007 test forms in Chapter Three in Table 3-2. Tables 4-10 and 4-11 contain descriptive statistics for the six sampling conditions for 2004 and 2007, respectively. Descriptive statistics include number of examinees, mean, standard deviation, skewness, and kurtosis for MC, CI, CR, and CO. Cronbach's α is also provided as an estimate of internal consistency reliability for MC, CR, and CO scores. Additionally, weighted descriptive statistics are provided for MC (Wt. MC) and CO (Wt. CO) scores for English Language 2007 (Table 4-11). Recall that the weight for English Language 2007 MC and CO scores was 53/52, or approximately 1.019. Tables 4-10 and 4-11 contain a large number of values; however, there are a few important things to note:

1. Weighting new form MC scores increased MC and CO scores for the 2007 data by approximately 0.60 to 0.70 points.
2. By comparing score means in Table 4-10 to weighted means in Table 4-11, it is evident that means were generally lower on the new form as compared to the old form.
3. In general, means on the old form increased as the CI ES increased, although this trend did not strictly hold. Conversely, means on the new form tended to decrease as the CI ES increased.
4. Samples with larger CI ES tended to be more negatively skewed on the old form and less negatively skewed on the new form, although this trend did not strictly hold.
5. Values of Cronbach's α were similar across sampling conditions, although larger fluctuations were seen for CR scores.

Effect sizes for English Language are shown in Table 4-12. The effect sizes provided in Table 4-12 were calculated as new form mean minus old form mean; consequently, a negative effect size indicates that means on the new form were lower than means on the old form. Each row in Table 4-12 contains effect sizes for the scores listed in the "Score Type" column. For example, the row labeled Wt. CO contains effect sizes for weighted composite scores. Each column contains effect sizes for a given sampling condition. Although sampling was based on weighted effect sizes, effect sizes based on unweighted scores were also calculated. Recall that the target CI ES were 0.00, 0.20, and 0.40. The actual CI ES are shown in the third row of data in Table 4-12. The target MC-CR ES were 0.00 and 0.25, and the actual MC-CR ES are shown in the seventh and eighth rows of data in Table 4-12.

There are a number of important things to note about the effect sizes. The actual CI ES were within 0.02 of the target CI ES. The actual *weighted* MC-CR ES (second to last row) were also within approximately 0.02 of the target MC-CR ES. Although the

actual *unweighted* MC-CR ES (last row) were larger in absolute value than the target MC-CR ES, the three effect sizes for the MC-CR 0.00 sampling conditions were still within 0.02 of each other. The three unweighted effect sizes for the MC-CR 0.25 sampling conditions, shown in the last three columns of data, differed by less than 0.03. Another interesting finding to note is the difference between the CI ES and CO ES. The CI ES indicates difference in examinee proficiency, and the CO ES incorporates both differences in examinee proficiency and form difficulty. As the CI ES increased, the CO ES also increased. Interestingly, the difference between the CI ES and CO ES also increased. For example, for CI 0.00 MC-CR 0.00 (first column of data), the unweighted CO ES was -0.172 and the CI ES was -0.014, resulting in a difference of 0.158. For CI 0.40 MC-CR 0.00, the difference was 0.30. Consequently, it appeared that as the CI ES increased, the common items were less representative of the items on the total test.

In general, for a given target CI ES, the differences between the CI ES and CO ES were smaller for the MC-CR 0.25 sampling conditions than for the MC-CR 0.00 sampling conditions. For example, as noted earlier in this paragraph, for CI 0.00 MC-CR 0.00, the difference between the unweighted CO ES and CI ES was 0.158. For CI 0.00 MC-CR 0.25, the difference was 0.107.

Equated Moments

Moments for new form scores equated to the old form scale are provided in Table 4-13 for the six sampling conditions and four equating methods. For CI 0.00 MC-CR 0.00 (first column of data), means and standard deviations were similar across the four equating methods. As CI ES increased, means for FE and CE were less similar to means for TS and OS. Specifically, for both CI 0.40 MC-CR 0.00 and CI 0.40 MC-CR 0.25 sampling conditions, means for TS and OS were approximately 1 point smaller than the mean for CE and 2 points smaller than the mean for FE. Standard deviations for the four equating methods also fluctuated somewhat across sampling conditions. The largest

differences among standard deviations across the four methods was evident for the CI 0.00 MC-CR 0.25 sampling condition.

Equating Relationships

Figure 4-6 contains a comparison of the equating relationships for English Language 2004-2007. There are four plots in Figure 4-6, one for each equating method. The two plots in the top row are for the two traditional equating methods (FE and CE, on the left and right, respectively), and the two plots in the bottom row are for the two IRT equating methods (TS and OS, on the left and right, respectively). The six lines in each plot represent the six sampling conditions. In each plot, the three MC-CR 0.00 sampling conditions are illustrated by bold lines: CI 0.00 is solid, CI 0.20 is dashed, and CI 0.40 is dotted. The three MC-CR 0.25 samples are illustrated by non-bold lines. The same line types used to represent the CI ES for the three MC-CR 0.00 samples were also used for the three MC-CR 0.25 samples (e.g., CI 0.00 is solid). Additionally, only the score range from the 1st to the 99th percentile of examinees is plotted. Recall that for the operational test forms, the criterion equating relationship was the equating relationship for the CI 0.00 MC-CR 0.00 sampling condition, which is represented by the solid bold line.

Consider the top left plot in Figure 4-6 for the FE equating method. First, examine the equating relationships for the bold lines, which represent the MC-CR 0.00 equating relationships. It is evident that as the CI ES increased (dashed and dotted bold lines), the equating relationships became increasingly different from the criterion equating relationship (solid bold line). Equated scores were as much as approximately 2 points higher for CI 0.20 and 3 points higher for CI 0.40. Second, examine the non-bold equating relationships, which represent the MC-CR 0.25 equating relationships. Consider the non-bold dashed (CI 0.20) and non-bold dotted (CI 0.40) lines in relation to the non-bold solid line (CI 0.00). It is again evident that as the CI ES increased, equating relationships became increasingly different from the equating relationship for CI 0.00.

For both MC-CR 0.00 and MC-CR 0.25, the difference between the study condition and criterion equating relationships increased as the CI ES increased. The same trend was also evident in the plots for CE, TS, and OS. However, the trend occurred to a lesser extent for CE, TS, and OS.

Now, compare the bold lines in relation to the non-bold counterparts in Figure 4-6 for the FE equating method. That is, compare the solid bold line to the non-bold solid line, the bold dashed line to the non-bold dashed line, and the bold dotted line to the non-bold dotted line. By examining these three pairs of lines, it is evident that the MC-CR 0.25 (non-bold lines) resulted in lower equated scores than MC-CR 0.00. One plausible explanation for this result is that the differences between the CI ES and CO ES were smaller for the MC-CR 0.25 sampling conditions as compared to the MC-CR 0.00 sampling conditions, as described previously in the descriptive statistics section. Because of the tendency for equated scores to be lower for MC-CR 0.25, an unexpected result occurred: The CI 0.20 MC-CR 0.25 equating relationship appeared most similar to the criterion (CI 0.00 MC-CR 0.00) equating relationship. Effect sizes were examined in order to determine whether the difference between the CI ES and CO ES were influencing the patterns of results. As mentioned previously, for the criterion equating relationship, the unweighted CO ES was -0.172 and the CI ES was -0.014, resulting in a difference of 0.158. For the CI 0.20 MC-CR 0.25 equating relationship, the unweighted CO ES was -0.361 and the CI ES was -0.199, resulting in a difference of 0.162. The differences between the CI ES and CO ES were very similar for the two equating relationships, which may have resulted in the CI 0.20 MC-CR 0.25 equating relationship appearing most similar to the criterion. In the top right plot of Figure 4-6, it is evident that for CE, the MC-CR 0.00 and MC-CR 0.25 equating relationship patterns were similar to those for FE. To some extent, patterns for TS and OS (bottom row of Figure 4-6) were similar to those for FE and CE. However, for TS and OS, CI 0.40 MC-CR 0.25 appeared to result in the equating relationship most similar to the criterion equating

relationship. This result does not appear to be influenced by the difference between the CI ES and CO ES.

Figure 4-7 contains six plots of equating relationships, one for each of the six sampling conditions. There are four lines in each plot, one for each equating method. The two traditional equating methods (FE and CE) are illustrated by solid and dashed lines, respectively. The two IRT methods (TS and OS) are illustrated by dotted and dotted-dashed lines, respectively. The plots in the left column of Figure 4-7 illustrate equating relationships for the MC-CR 0.00 sampling conditions, and the plots in the right column illustrate equating relationships for the MC-CR 0.25 sampling conditions. The first row contains plots for CI 0.00, the second row contains plots for CI 0.20, and the last row contains plots for CI 0.40. Consider the first plot in the top left corner for the criterion equating relationships. FE and CE resulted in nearly identical equating relationships, and TS and OS also resulted in nearly identical equating relationships. In addition, the traditional and IRT equating relationships differed by 0.50 score points or less throughout the score scale. Next, consider the plots for CI 0.20 MC-CR 0.00 and CI 0.40 MC-CR 0.00, shown in the second and third rows of the left column. As the CI ES increased, the differences among the equating methods also increased. For CI 0.40, equivalents for FE were approximately 2 points higher than those for TS or OS. Equivalents for CE were approximately 1.50 points higher than those for TS or OS. However, equating relationships for TS and OS remained very similar to each other. Similar trends can be seen in the right column of Figure 4-7 for the MC-CR 0.25 sampling conditions.

Conditional Bias

Plots of conditional bias are shown in Figure 4-8. Recall that the criterion equating relationship was CI 0.00 MC-CR 0.00. Figure 4-8 contains five plots of conditional bias, one for each of the other five sampling conditions. There are four lines in each plot, one for each of the four equating methods. The two traditional methods (FE

and CE) are illustrated by the solid and dashed lines, respectively. The two IRT methods (TS and OS) are illustrated by the dotted and dotted-dashed lines, respectively. The plots in the left column of Figure 4-8 illustrate conditional bias for the MC-CR 0.00 sampling conditions, and the plots in the right column illustrate conditional bias for the MC-CR 0.25 sampling conditions.

First, consider the plots in the left column of Figure 4-8 for the MC-CR 0.00 sampling conditions. As discussed previously, it is evident that as the CI ES increased, conditional bias also increased for FE and CE. For CI 0.20, bias was approximately 1 to 1.50 score points for FE and CE. For CI 0.40, conditional bias for FE was approximately 2.50 to 3 score points for new form raw scores greater than 40. For CE, conditional bias was less than two score points across most of the score scale. For both CI 0.20 and CI 0.40, conditional bias for TS and OS was less than 1 score point across the entire score scale.

In the right column of Figure 4-8 for the MC-CR 0.25 sampling conditions, consistent with the lower equated scores that were seen in Figure 4-6, conditional bias was negative for CI 0.00. As the CI ES increased, bias became more positive. For CI 0.20, conditional bias was near zero for FE and CE. For TS and OS, conditional bias was approximately -1 across the score scale. For CI 0.40, conditional bias was approximately 0 for TS and OS for new form raw scores greater than 35. For FE and CE, conditional bias was approximately 1 to 2 score points for new form raw scores greater than 35.

Conditional Standard Error of Equating (CSE)

Figure 4-9 contains six plots of CSE, one for each of the six sampling conditions. There are four lines in each plot, one for each of the four equating methods. The two traditional methods (FE and CE) are illustrated by the solid and dashed lines, respectively. The two IRT methods (TS and OS) are illustrated by the dotted and dotted-dashed lines, respectively. The plots in the left column of Figure 4-9 illustrate CSE for

the MC-CR 0.00 sampling conditions, and the plots in the right column illustrate CSE for the MC-CR 0.25 sampling conditions. The first row contains plots for CI 0.00, the second row contains plots for CI 0.20, and the third row contains plots for CI 0.40.

Consider first the plot in the top left corner for CI 0.00 MC-CR 0.00. It is evident that CSE was smallest for the two IRT equating methods and largest for the two traditional equating methods. More specifically, CSE was smallest for OS, except at scores greater than approximately 55. CSE was largest for CE across most of the score scale. By examining the other five plots in Figure 4-9, it is evident that OS always results in the smallest CSE (except at scores greater than approximately 55), and CE always resulted in the largest CSE (except at scores greater than approximately 55). It is also important to consider the magnitude of standard errors of equating when interpreting the differences among equating relationships. Generally, CSE was approximately 0.3 to 0.5 score points across much of the score scale.

Overall Summary Statistics

WARMSB, WASE, WARMSE, Difference, standardized WARMSB, and standardized WASE, which are overall averages across the score scale, are contained in Table 4-14. (These statistics were described in Chapter Three, and definitions of the acronyms are provided in the list of acronyms and abbreviations at the beginning of this dissertation.) Each one of these six overall summary statistics are contained in a "block" of values in Table 4-14, and each block contains a row for each of the four equating methods. For example, the first block and column in Table 4-14 is labeled WARMSB. There are four rows adjacent to WARMSB, labeled FE, CE, TS, and OS. The first row, labeled FE, contains values of WARMSB for the FE equating method. There are six columns of data in Table 4-14, one for each of the six sampling conditions. Recall that the criterion equating relationship was CI 0.00 MC-CR 0.00. WARMSB, WARMSE, and Difference could not be calculated for this condition.

Consider the first block of values (WARMSB) for the MC-CR 0.00 sampling conditions, as shown in the first three columns of data. Values of WARMSB indicate a weighted average bias across the score scale. For CI 0.20, WARMSB was approximately 0.70 points for TS and OS, 1.20 points for CE, and 1.50 points for FE. For CI 0.40, WARMSB was still approximately 0.70 points for TS and OS, but WARMSB increased to 1.60 and 2.30 points for CE and FE, respectively. WARMSB was smallest for TS and OS and largest for FE for both sampling conditions. Additionally, of the four equating methods, the increase in WARMSB from CI 0.20 to CI 0.40 was largest for FE. For both sampling conditions and all equating methods, WARMSB was larger than 0.50, which is often used as a reasonable "difference that matters" (DTM). For TS and OS, values of standardized WARMSB (second to last block of data) were always less than 0.10 standard deviation units. Values of standardized WARMSB ranged from approximately 0.10 standard deviation units to approximately 0.20 standard deviation units for FE and CE.

By considering only the values of WARMSB in Table 4-14, it appears that a somewhat different pattern occurred for the MC-CR 0.25 sampling conditions (last three columns of data). For CI 0.00 MC-CR 0.25, WARMSB was approximately 1 point for FE and CE and 1.20 points for TS and OS. For FE and CE, values of WARMSB then *decreased* for the CI 0.20 MC-CR 0.25 sampling condition and *increased* for the CI 0.40 MC-CR 0.25 sampling condition. For TS and OS, values of WARMSB *decreased* for both the CI 0.20 and CI 0.40 sampling conditions. Although these results seem somewhat counterintuitive, they were consistent with the equating relationships shown in Figure 4-6. In Figure 4-6 to 4-8, it was evident that the CI 0.00 MC-CR 0.25 equating relationship resulted in *lower* equivalents than the CI 0.00 MC-CR 0.00 equating relationship, leading to negative bias for CI 0.00 MC-CR 0.25. As CI ES increased for the MC-CR 0.25 equating relationships, bias became less negative (or, more positive). The CI ES impacted the traditional methods to a greater extent that the IRT methods. Consequently, for FE

and CE, the CI 0.20 MC-CR 0.25 equating relationships were most similar to the criterion. For TS and OS, the CI 0.40 MC-CR 0.25 equating relationships were most similar to the criterion.

As evidenced by the figures examined earlier in this section, values of WARMSB based on the CI 0.00 MC-CR 0.00 equating relationship masked some of the patterns of results. Therefore, in order to better evaluate the influence of CI ES on WARMSB for the MC-CR 0.25 conditions, a different criterion equating relationship was considered. Table 4-15 contains values of WARMSB based on the different criterion equating relationships. The values in Table 4-15 are based on two criteria. The three columns of data on the left of the table contain the same values of WARMSB as shown in Table 4-14. These values of WARMSB were calculated using CI 0.00 MC-CR 0.00 as the criterion. The three columns of data on the right of the table for the MC-CR 0.25 sampling were calculated using CI 0.00 MC-CR 0.25 as the criterion. Based on the values of WARMSB in Table 4-14, it appeared that results for MC-CR 0.25 were irregular. However, it is now evident from the values in Table 4-15 that WARMSB increased as CI ES increased for both MC-CR 0.00 and MC-CR 0.25 sampling conditions.

Returning to Table 4-14, the second block of data contains values of WASE. Values of WASE indicate average standard errors of equating across the score scale. Consistent with plots in Figure 4-9, WASE was smallest for OS across all sampling conditions (approximately 0.34 to 0.37) and largest for CE across all sampling conditions (approximately 0.46 to 0.52). Values of WASE did not differ substantially across sampling conditions. The last block of data in Table 4-14 contains values of standardized WASE. Values of standardized WASE ranged from approximately 0.03 to 0.05 standard deviation units.

The third block of data in Table 4-14 contains values for WARMSE, which is an index of the overall average error across the score scale. The same patterns that were found for WARMSB were also found for WARMSE. Additionally, the magnitude of

WARMSE appeared to be primarily attributable to bias. The fourth block of data in Table 4-14 contains values of Difference. Overall, the results for Difference were very similar to the results for WARMSB.

Classification Consistency

Examinees are typically classified into one of five AP grade levels based on their composite scores. Although the tests used in this dissertation were modified and no longer representative of the AP tests, classification consistency was still used to investigate the practical significance of differences among sampling conditions and equating methods. New form cut scores were provided in Chapter Three for the criterion equating relationship (CI 0.00 MC-CR 0.00) for each of the tests. Table 4-16 contains percentages of classification consistency. The first column (in bold) contains classification consistency for the criterion equating relationships. In order to determine how similar the four equating methods were for the criterion (CI 0.00 MC-CR 0.00), the grades examinees received on CE, TS, and OS were compared to the grades received on FE. The percentages contained in the first column are the percentages of examinees receiving the same grade based on FE and the other equating method for the criterion. For example, in the row labeled CE, the first column contains the value 100.0. This value indicates that 100% of examinees received the same grades using the FE and CE equating methods. Similarly, for the row labeled TS, 100% of examinees received the same grades using the FE and TS methods. The same result was found for OS. For the CI 0.00 MC-CR 0.00 samples, all examinees received the same grade, regardless of the equating method.

The percentages contained in the remaining five columns of data were calculated differently. These columns of data contain the overall percentage of examinees receiving the same grade based on the criterion equating relationship and the study condition equating relationship for a given equating method. For example, in the row labeled FE,

the second column contains the value 87.5. This value indicates that the FE equating relationship for the CI 0.20 MC-CR 0.00 sample of examinees and the FE equating relationship for the criterion resulted in 87.5% of examinees receiving the same grade. Similarly, in the row labeled CE, the second column contains the value 86.8. This value indicates that the CE equating relationship for the CI 0.20 MC-CR 0.00 sample of examinees and the CE equating relationship for the criterion resulted in 86.8% of examinees receiving the same grade.

Across all sampling conditions for TS and OS, at least approximately 90% of examinees received the same grade as they received based on the criterion equating relationship. With the exception of CI 0.00 MC-CR 0.25 and CI 0.20 MC-CR 0.25, TS and OS resulted in larger percentages of classification consistency than FE and CE. TS and OS also tended to result in similar percentages. For CI 0.40 MC-CR 0.00 and CI 0.40 MC-CR 0.25, FE resulted in the smallest percentage of examinees receiving the same grade as the criterion equating relationship. Percentages for FE and CE also tended to be similar except for CI 0.40 MC-CR 0.00 and CI 0.40 MC-CR 0.25. For these two conditions, percentages for CE were 7% to 10% higher than for FE.

Summary

For English Language, for both the MC-CR 0.00 and MC-CR 0.25 sampling conditions, as the common-item effect size increased, the differences among the equating relationships also increased. Specifically, because scores on the new form were *lower* than scores on the old form, old form equivalents were *higher* for large common-item effect sizes. Equating relationships for the MC-CR 0.25 sampling conditions resulted in lower equivalents relative to the MC-CR 0.00 equating relationships. Consequently, equating relationships for CI 0.20 MC-CR 0.25 and CI 0.40 MC-CR 0.25 were more similar to the criterion than the MC-CR 0.00 counterparts. This result may have occurred because the difference between the common-item effect size and composite effect size

was more similar to the difference for the criterion for the MC-CR 0.25 sampling conditions. TS and OS did not differ substantially across common-item effect sizes. Standard errors of equating were smallest for OS and largest for CE across all sampling conditions.

Spanish Language

Results for the Spanish Language operational test forms are presented in the following subsections: descriptive statistics, equated moments, equating relationships, conditional bias, conditional standard error of equating (CSE), overall summary statistics, and classification consistency. The Spanish Language operational test forms differed by 15 items. That is, Spanish Language 2004 contained 15 more MC items than Spanish Language 2006. Consequently, results for Spanish Language reflect a situation in which equating (in the strictest sense) may not be advisable.

Descriptive Statistics

Descriptive statistics for the operational Spanish Language 2004 and 2006 test forms for the six sampling conditions are provided in Tables 4-17 and 4-18, respectively. It is important to note that the unweighted and weighted MC and CO scores differed substantially for Spanish Language 2006, because Spanish Language 2006 contained 15 fewer MC items than Spanish Language 2004. Recall that the weight for Spanish Language 2006 MC and CO scores was 90/75, or 1.20. Other important information to note about the descriptive statistics is as follows:

1. Weighting Spanish Language 2006 MC scores increased MC and CO means by approximately 10 to 11 points.
2. By comparing MC and CO means in Table 4-17 to weighted MC and CO means in Table 4-18, it is evident that means on the new form were generally higher than those on the old form.

3. In general, means on the old form for MC, CI, and CO scores decreased as the CI ES increased. Means on the new form did not consistently increase or decrease.
4. Samples with larger CI ES tended to be less negatively skewed on the old form. On the new form, skewness was not consistently impacted by the CI ES.
5. Values of α were similar across sampling conditions.

Effect sizes for the Spanish Language operational test forms are shown in Table 4-19. Recall that negative effect sizes indicate means on the new form were lower than means on the old form. In contrast to English Language, CI ES were positive for Spanish Language. Additionally, MC-CR target effect size patterns were calculated using weighted MC scores, but unweighted effect sizes are also provided for comparison. It is evident that the MC, CO, and MC-CR ES were much larger in absolute value for the unweighted scores as compared to the weighted scores, because of the large difference in the number of MC items on Spanish Language 2004 and 2006.

The CI ES, shown in the third row, were within 0.03 of the target effect sizes, and the *weighted* MC-CR ES (second to last row) were within 0.02 of the target effect sizes. The *unweighted* MC-CR ES were less similar to each other. For MC-CR 0.00, the three *unweighted* MC-CR ES differed by approximately 0.08 (see the first three columns and last row of data). Similarly, for MC-CR 0.25, the three *unweighted* MC-CR ES differed by approximately 0.10 (see the last three columns and last row of data). The difference between the unweighted CI ES and CO ES varied across sampling conditions as it did for English Language. However, in contrast to English Language, the difference between the CI ES and CO ES was larger for the MC-CR 0.25 sampling conditions as compared to the MC-CR 0.00 sampling conditions. For example, the difference between the CI ES and CO ES was 0.460 for CI 0.00 MC-CR 0.00. For CI 0.00 MC-CR 0.25, this difference was 0.635. Also in contrast to English Language, as the CI ES increased, the difference between the CI ES and CO ES *decreased*. For example, the difference between the CI ES and CO ES was 0.46 for CI 0.00 MC-CR 0.00 as compared to 0.20 for CI 0.40 MC-CR

0.00. Similarly, for CI 0.00 MC-CR 0.25, the difference was 0.635 as compared to 0.281 for CI 0.40 MC-CR 0.25.

Equated Moments

Equated moments are given in Table 4-20 for the six sampling conditions and four equating methods. For CI 0.00 MC-CR 0.00, means for TS and OS were smaller than means for FE and CE by approximately 1.30 points. For CI 0.20 MC-CR 0.00, means were similar across all four equating methods. For CI 0.40 MC-CR 0.00, means for TS and OS were approximately 4 points higher than the FE mean and 2.50 points higher than the CE mean. For the MC-CR 0.25 sampling conditions, similar trends occurred. For the MC-CR 0.00 sampling conditions, standard deviations were relatively similar across equating methods for CI 0.00 and CI 0.40. Standard deviations were substantially larger for TS and OS as compared to FE and CE for CI 0.20. For the MC-CR 0.25 sampling conditions, standard deviations were larger for TS and OS as compared to FE and CE for CI 0.00 and CI 0.20. Standard deviations were relatively similar across equating methods for CI 0.40.

Equating Relationships

Figure 4-10 contains plots of the equating relationships for Spanish Language 2004-2006. Recall that for the operational test forms, the criterion was the CI 0.00 MC-CR 0.00 sampling condition. Also recall that MC-CR 0.00 relationships are represented by bold lines, and MC-CR 0.25 relationships are represented by non-bold lines. Additionally, CI 0.00 is represented by a sold line, CI 0.20 by a dashed line, and CI 0.40 by a dotted line. Consider the first plot in Figure 4-10 for the FE equating method. First, examine the bold lines, which represent the MC-CR 0.00 equating relationships. The equating relationships for CI 0.00 and CI 0.20 appeared relatively similar except at scores greater than 90. The equating relationship for CI 0.40 resulted in substantially lower scores than the criterion equating relationship. Again, it is plausible that these differences

reflect the difference between the CI ES and CO ES. For CI 0.20, the difference between the CI ES and CO ES was 0.439, which was very similar to the difference for CI 0.00 (0.460). For CI 0.40, the difference was 0.20. Second, examine the non-bold equating relationships, which represent the MC-CR 0.25 equating relationships. Consider the non-bold dashed (CI 0.20) and non-bold dotted (CI 0.40) lines in relation to the non-bold solid line (CI 0.00). For these three equating relationships, it is evident that as the CI ES increased, the equating relationships become increasingly different from the equating relationship for CI 0.00. Similar patterns were seen for CE (top right), but the patterns occurred to a lesser extent.

A somewhat different trend occurred for the TS and OS equating relationships, shown in the bottom row of Figure 4-10. For the MC-CR 0.00 sampling conditions (bold lines), equating relationships for the three CI ES were similar throughout much of the score distribution. For the MC-CR 0.25 equating relationships (non-bold lines), CI 0.00 and CI 0.20 resulted in similar equating relationships, but the equating relationship for CI 0.40 was quite different.

Now, compare the bold lines in relation to the non-bold counterparts in Figure 4-10 for the FE equating method. That is, compare the solid bold line to the non-bold solid line, the bold dashed line to the non-bold dashed line, and the bold dotted line to the non-bold dotted line. By examining these three pairs of lines, it is evident that the MC-CR 0.25 (non-bold) sampling conditions resulted in higher equated scores than for the MC-CR 0.00 sampling conditions. This trend is also seen in the CE, TS, and OS plots. As described previously in this section, it is plausible that the difference between the CI ES and CO ES impacted how similar the FE and CE equating relationships were to each other. The CI 0.20 MC-CR 0.00 and CI 0.20 MC-CR 0.25 equating relationships (bold and non-bold dashed lines) were most similar to the criterion equating relationship. The difference between the CI ES and CO ES for the criterion was 0.460. For CI 0.00 MC-CR 0.00, this difference was approximately 0.440, and for CI 0.20 MC-CR 0.25, this

difference was approximately 0.480. TS and OS did not appear to be impacted by the CI ES and CO ES differences.

Interestingly, for both English Language and Spanish Language, the three MC-CR 0.00 equating relationships were similar, and the three MC-CR 0.25 equating relationships were similar for TS and OS. There was relatively little overlap of the MC-CR 0.00 and MC-CR 0.25 equating relationships. However, this was not the case for FE and CE.

Figure 4-11 also contains plots of equating relationships, one for each of the six sampling conditions. Consider the plot in the top left corner for the criterion equating relationships. FE (solid line) and CE (dashed line) resulted in similar equating relationships, and TS (dotted line) and OS (dotted-dashed line) also resulted in similar equating relationships. However, the traditional and IRT equating relationships were quite different between scores of approximately 50 and 90. Next, consider the plots for CI 0.20 MC-CR 0.00 and CI 0.40 MC-CR 0.00, shown in the left column and second and third rows of Figure 4-11, respectively. As the CI ES increased to 0.40, FE and CE became somewhat less similar to each other, yet TS and OS remained very similar to each other.

The equating relationships in the right column of Figure 4-11 for the MC-CR 0.25 sampling conditions appear quite different in shape from the MC-CR 0.00 equating relationships in the left column. However, some similar trends can be seen. The FE and CE equating relationships were similar to each other, as were the TS and OS equating relationships. The TS and OS equating relationships remained very similar across all three CI ES, while the equivalent scores for FE and CE were markedly different across the three CI ES.

Conditional Bias

Figure 4-12 contains plots of conditional bias. Recall that CI 0.00 MC-CR 0.00 was the criterion equating relationship. First, consider the plots in the left column of Figure 4-12. Bias for the CI 0.20 MC-CR 0.00 (second row) equating relationships was near zero across most of the score scale, especially for TS and OS. For CI 0.40 MC-CR 0.00 (third row), it is evident that as the CI ES increased, bias also increased for all methods. However, bias was much larger for FE and CE as compared to TS and OS. In the right column of Figure 4-12 for the MC-CR 0.25 sampling conditions, it is evident that the higher equated scores seen in Figure 4-10 resulted in large positive bias across most of the score scale for CI 0.00. As the CI ES increased, bias became less positive. Consequently, of the three CI ES for MC-CR 0.25, CI 0.00 appeared to result in the largest bias for all four equating methods. CI ES impacted FE and CE to a greater extent than TS and OS. Therefore, for CI 0.40 MC-CR 0.25, FE and CE resulted in large negative bias, and TS and OS appeared to result in smaller positive bias.

Conditional Standard Error of Equating (CSE)

Figure 4-13 contains plots of CSE for each of the six sampling conditions. Unlike the English Language operational test forms, the magnitude and patterns of CSE for Spanish Language appeared to vary by sample. First, CSE was not always smallest for the TS and OS equating conditions. Consider the plot in the top left corner for CI 0.00 MC-CR 0.00. It is evident in this plot that throughout most of the score scale, FE resulted in the smallest CSE, and CSE for TS, OS, and CE were similar. For CI 0.20 MC-CR 0.00 (second row, left column) and CI 0.00 MC-CR 0.25 (first row, right column), CSE for TS and OS appeared larger than both FE and CE across much of the score scale. For the remaining three samples, CSE for TS and OS appeared to be similar to or smaller than CSE for FE and CE. Further, the magnitude of the CSE appeared to vary substantially across the six sampling conditions. For example, CSE appeared much larger for CI 0.00

MC-CR 0.25 relative to some of the other conditions. However, it is important to remember that there were a larger number of score points for Spanish Language and only 1,900 examinees in each sample.

Overall Summary Statistics

Table 4-21 contains WARMSB, WASE, WARMSE, Difference, standardized WARMSB, and standardized WASE for Spanish Language 2004-2006. As was previously found for the English Language operational test forms, for the MC-CR 0.00 sampling conditions, WARMSB increased as CI ES increased. WARMSB was also substantially lower for TS and OS as compared to FE and CE. WARMSB was less than 0.50 score points for TS and OS for CI 0.20, but not for CI 0.40. For FE and CE, WARMSB was greater than 0.5 for both CI 0.20 and CI 0.40 samples. The pattern of results for MC-CR 0.25 was also somewhat similar to the results found for the English Language operational test forms. For CI 0.00 MC-CR 0.25, WARMSB was approximately 6.0 points for FE and CE and 5.8 points for TS and OS. These values indicate that the MC-CR 0.25 resulted in average bias across the score scale of approximately 6 score points for all equating methods. For FE, values of WARMSB then *decreased* for the CI 0.20 MC-CR 0.25 sampling condition and *increased* for the CI 0.40 MC-CR 0.25 sampling condition, which was also found for English Language. For CE, values of WARMSB *decreased* for both the CI 0.20 and CI 0.40 sampling conditions. For TS and OS, values of WARMSB *increased* slightly for CI 0.20 and then *decreased* for the CI 0.40 sampling conditions. WARMSB was not less than 0.50 score points for any of the MC-CR 0.25 samples or equating methods. These results are consistent with the conditional bias in Figure 4-12, although it was difficult to determine this pattern of results visually. As seen in Figure 4-12, the CI 0.00 MC-CR 0.25 equating relationship resulted in *higher* equated scores than the CI 0.00 MC-CR 0.00 equating relationship. As the CI ES increased, bias became less positive for the MC-CR 0.25 equating

relationships. Consequently, for FE, the CI 0.20 MC-CR 0.25 equating relationship was most similar to the criterion. For CE, TS, and OS, the CI 0.40 MC-CR 0.25 equating relationships were most similar to the criterion. The second to last block of data contains values of standardized WARMSB. For the MC-CR 0.00 sampling conditions, standardized WARMSB ranged from approximately 0.02 standard deviation units for OS (CI 0.20) to 0.25 standard deviation units for FE (CI 0.40). For the MC-CR 0.25 sampling conditions, standardized WARMSB ranged from approximately 0.10 to 0.27 standard deviation units.

Similar to the English Language operational test forms, values of WARMSB did not adequately illustrate the patterns seen in Figures 4-10 through 4-12 for the MC-CR 0.25 conditions. Table 4-22 contains values of WARMSB based on different criterion equating relationships to illustrate the influence of CI ES on WARMSB. Recall that the three columns of data on the left of the table contain the same values of WARMSB shown in Table 4-21. These values of WARMSB were calculated using CI 0.00 MC-CR 0.00 as the criterion. The three columns of data on the right of the table for the MC-CR 0.25 sampling conditions were calculated using CI 0.00 MC-CR 0.25 as the criterion. Based on the values of WARMSB in Table 4-21, the MC-CR 0.25 results appeared inconsistent. However, it is evident in Table 4-22 that for MC-CR 0.25, WARMSB also increased as CI ES increased.

Returning to Table 4-21, the second block of data in Table 4-21 contains values of WASE. Values of WASE were much larger for the Spanish Language operational test forms as compared to the English Language operational test forms. This likely resulted from the large number of score points and small number of examinees. Additionally, as seen in Figure 4-13, values of WASE appeared to vary substantially across sampling conditions, particularly for TS and OS. Across sampling conditions, values of WASE ranged from approximately 0.74 to 1.30 score points for TS and OS. Consistent with what was seen in the plots of CSE, values of WASE were smallest for FE for CI 0.00

MC-CR 0.00, CI 0.20 MC-CR 0.00, and CI 0.00 MC-CR 0.25. For the other three conditions, TS and OS had the smallest values of WASE. CE resulted in the largest values of WASE for all conditions except CI 0.00 MC-CR 0.25. Values of standardized WASE (last block of data in Table 4-21) were approximately 0.03 to 0.06 standard deviation units across all sampling conditions and equating methods.

The third block of data in Table 4-21 contains values for WARMSE. Patterns of WARMSE were similar to those for WARMSB. However, because values of WASE were large in magnitude, WASE contributed to the overall error to a greater extent than what was found for English Language. This was especially the case for TS and OS for CI 0.20 MC-CR 0.00. The fourth block of data in Table 4-21 contains values of Difference. For FE and CE, the results for Difference were similar to the results for WARMSB. However, for TS and OS, the two statistics differed by as much as 0.50 points for some of the sampling conditions.

Classification Consistency

Table 4-23 contains classification consistency results for the Spanish Language operational test forms. In Table 4-23, the first column (in bold) contains percentages of classification consistency based on the criterion equating relationships for CE, TS, and OS in comparison to the FE criterion equating relationship. CE resulted in 97.1% of examinees receiving the same grade as they received based on FE. For TS and OS, 91.5% of examinees received the same grade as they did based on FE. When interpreting the remaining percentages, it is important to consider that CE, TS, and OS did not result in the same percentages.

The remaining columns of data contain the percentage of examinees receiving the same grade for the criterion equating relationship and the study condition equating relationship for a given equating method. For the MC-CR 0.00 sampling conditions, TS and OS resulted in approximately 98% and 92% of examinees receiving the same grade

as the criterion equating relationship for CI 0.20 and CI 0.40, respectively. Approximately 90% of examinees received the same grade for the FE and CE equating methods for CI 0.20. For CI 0.40, only 65% and 71% of examinees received the same grade for FE and CE, respectively. For the MC-CR 0.25 sampling conditions, all four of the equating methods resulted in approximately 73% for CI 0.00. For CI 0.20 and CI 0.40, FE and CE resulted in higher percentages of classification consistency than TS and OS. Given the plots and results described in earlier sections for the MC-CR 0.25 sampling conditions, this result was not unexpected.

Summary

For Spanish Language, similar to English Language, for both the MC-CR 0.00 and MC-CR 0.25 sampling conditions, as the common-item effect size increased, the differences among the equating relationships also increased. Specifically, because scores on the new form were *higher* than scores on the old form, old form equivalents were *lower* for large common-item effect sizes. Equating relationships for the MC-CR 0.25 sampling conditions resulted in higher equivalents relative to the MC-CR 0.00 equating relationships. Consequently, for FE and CE, bias appeared larger for CI 0.40 MC-CR 0.00 relative to CI 0.40 MC-CR 0.25. The TS and OS equating relationships did not differ substantially across most of the common-item effect size levels. Standard errors of equating were smallest for OS for three of the sampling conditions and largest for CE for five of the sampling conditions. However, no pattern was evident regarding why the standard errors of equating were smallest for OS for only three of the sampling conditions.

Chemistry

Results for the Chemistry operational test forms are presented in the following subsections: descriptive statistics, equated moments, equating relationships, conditional

bias, conditional standard error of equating (CSE), overall summary statistics, and classification consistency.

Descriptive Statistics

Descriptive statistics for the Chemistry 2005 and 2007 operational old and new test forms for the six sampling conditions are provided in Tables 4-24 and 4-25, respectively. It is important to note that there were two more CR points on Chemistry 2007 than Chemistry 2005. The Chemistry 2007 CR scores were weighted to create comparable scores to Chemistry 2005 for sampling purposes. Recall that the weight for Chemistry 2007 MC and CO scores was 59/61, or approximately 0.967. There are a few important things to note about the Chemistry descriptive statistics:

1. Weighting new form CR scores decreased mean CR and CO scores by approximately one point.
2. By comparing score means in Table 4-24 to weighted score means in Table 4-25, it is evident that means on the new form were generally higher than those for the old form. Thought, this trend did not strictly hold, especially for CR scores.
3. In general, means on the old form decreased as the CI ES increased. Means on the new form generally increased as the CI ES increased.
4. Samples with larger CI ES tended to be less negatively skewed (or more positively skewed) on the old form. The reverse was true for the new form.
5. Values of α were fairly similar across sampling conditions.

Effect sizes for the Chemistry operational test forms are shown in Table 4-26. The effect sizes were calculated as new form minus old form scores; consequently, a negative effect size indicates that scores on the new form were lower than scores on the old form. Similar to Spanish Language, CI ES for Chemistry were positive. For Chemistry, unlike English Language and Spanish Language, the target MC-CR ES patterns were calculated using unweighted CR means, but weighted MC-CR ES are also provided. As noted in

Chapter Three, this decision was made after reviewing results for English Language and Spanish Language. For those two tests, especially Spanish Language, it appeared the unweighted effect sizes might have better explained the equating results than the weighted effect sizes. Further, bias appeared to be impacted by the similarity of the study condition effect size to the criterion effect size, rather than the magnitude of the effect size.

Recall that the target CI ES were 0.00, 0.20, and 0.40. The CI ES (third row) were within 0.025 of the target effect sizes. The target MC-CR ES were 0.00 and 0.25. The unweighted MC-CR 0.00 ES were within 0.01 of the target effect sizes. However, it was not possible to obtain the MC-CR 0.25 target effect size. The MC-CR 0.25 ES for Chemistry was actually an effect size of approximately only 0.10, but all three effect sizes were within 0.01 of this value. The weighted MC-CR ES were slightly larger than the unweighted MC-CR ES. However, they were also within 0.02 of each other.

As described previously for English Language and Spanish Language, it is also important to note that the difference between the CI ES and CO ES fluctuated across sampling conditions. Similarly to Spanish Language, the differences between the CI ES and CO ES were larger for the MC-CR 0.25 sampling conditions as compared to the MC-CR 0.00 sampling conditions. However, within the MC-CR 0.00 or MC-CR 0.25 sampling conditions, the magnitude of the differences was very similar. The difference between the unweighted CI ES and CO ES for the criterion (CI 0.00 MC-CR 0.00) was 0.132, which was the smallest difference between the CI ES and CO ES. The difference for CI 0.40 MC-CR 0.00 was only 0.150. The largest difference between the CI ES and CO ES was only 0.217 for CI 0.00 MC-CR 0.25.

Equated Moments

Equated moments are given in Table 4-27 for the six sampling conditions and four equating methods. For CI 0.00 MC-CR 0.00 and CI 0.00 MC-CR 0.25, means tended

to be similar across equating methods. As the CI ES increased, means for the four equating methods became less similar. Specifically, means for TS and OS were larger than means for FE and CE. For both CI 0.40 MC-CR 0.00 and CI 0.40 MC-CR 0.25, means for TS and OS were approximately 1.70 points higher than for FE and approximately 0.75 points higher than for CE. Standard deviations tended to be similar across equating methods; although, the standard deviations differed by as much as approximately 1.30 points across equating methods.

Equating Relationships

Figure 4-14 contains a comparison of the equating relationships for Chemistry, with one plot for each equating method. Consider the top left in Figure 4-14 for the FE equating method. Consider first the bold lines for the MC-CR 0.00 sampling conditions. It is evident that all three equating relationships were similar. For FE and CE, the largest difference was one score point around scores of 75 to 105. Results were very similar for CE, in the top right plot. For TS and OS (bottom row), large differences of approximately 5 to 6 score points occurred at low scores. However, at scores greater than approximately 60, equating relationships for the three CI ES were very similar. Now, consider the non-bold lines for the MC-CR 0.25 sampling conditions. In the first plot for FE, it is evident that the MC-CR 0.25 equating relationships were also similar to each other. Similar results can be seen for CE, TS, and OS.

Comparing the bold lines for MC-CR 0.00 to the non-bold lines for MC-CR 0.25, it is evident that the MC-CR 0.25 equating relationships resulted in slightly higher equated scores as compared to the MC-CR 0.00 equating relationships. Also, for all four equating methods the three MC-CR 0.00 equating relationships were similar to each other, and the three MC-CR 0.25 equating relationships were also similar to each other. There was relatively little overlap of the MC-CR 0.00 and MC-CR 0.25 equating relationships. It is plausible that this trend resulted because, as discussed previously, the

differences between the CI ES and CO ES were larger for the MC-CR 0.25 sampling conditions than for the MC-CR 0.00 sampling conditions.

One interesting finding about the Chemistry equating relationships was that across all CI ES levels, the equating relationships for FE and CE were similar. However, even though the same CI ES levels were used for English Language and Spanish Language, equating relationships for those tests were typically not similar across CI ES levels for FE and CE. As frequently noted throughout the results section, one plausible explanation for this finding is the difference between the CI ES and CO ES. These differences were very similar across the Chemistry sampling conditions, but for English Language and Spanish Language, there was substantial variability in the magnitude of differences between CI ES and CO ES. Perhaps, because the MC and CR correlations were high for Chemistry, the common items were representative of the items on the total test, even for large CI ES.

Figure 4-15 contains six plots, one for each of the six sampling conditions for Chemistry 2005-2007. Consider the first plot in the top left corner for the criterion equating relationships. Equating relationships for all four equating methods were similar, although small differences did exist. Next, consider the plots for CI 0.20 MC-CR 0.00 and CI 0.40 MC-CR 0.00, shown in the left column and second and third rows, respectively. As the CI ES increased to 0.40, differences in equating relationships increased among equating methods. Although, TS and OS remained very similar to each other, equated scores were substantially higher for TS and OS than for FE and CE. Further, equated scores were higher for CE as compared to FE. Not surprisingly, the equating relationships for MC-CR 0.25 were very similar to those for MC-CR 0.00.

Conditional Bias

Figure 4-16 contains plots of conditional bias. Recall that CI 0.00 MC-CR 0.00 was the criterion equating relationship. First, consider the plots in the left column of Figure 4-16 for the MC-CR 0.00 sampling conditions. For CI 0.20 MC-CR 0.00, bias was

around zero across the score scale, except at low scores for TS and OS and high scores for FE and CE. For CI 0.40 MC-CR 0.00 (bottom row, left column), conditional bias appeared to increase, except at low scores for TS and OS. In the right column of Figure 4-16 for the MC-CR 0.25 sampling conditions, it is evident that the slightly higher equated scores seen in Figure 4-14 for MC-CR 0.25 resulted in somewhat larger positive bias. As CI ES increased, bias became less positive for FE and CE, but slightly more positive for TS and OS. Consequently, of the three CI ES for MC-CR 0.25, CI 0.40 appeared to result in the least bias for FE and CE, but CI 0.00 appeared to result in the least bias for TS and OS. Additionally, for both MC-CR 0.00 and MC-CR 0.25, as the CI ES increased, the differences among the four equating methods became larger.

Conditional Standard Error of Equating (CSE)

Figure 4-17 contains plots of CSE for each of the six sampling conditions. Similar to Spanish Language, CSE for Chemistry varied somewhat in magnitude and pattern according to the sample. Also similar to Spanish Language, Chemistry contained a large number of score points and a relatively small number of examinees; therefore, CSE were similar in magnitude to CSE for Spanish Language. For four of the six sampling conditions, TS and OS resulted in the smallest CSE. For CI 0.20 MC-CR 0.00 and CI 0.40 MC-CR 0.25, TS and OS were similar to or slightly larger than CSE for FE and CE. Given the small differences among the Chemistry equating relationships and the large CSE, it is especially important to consider the magnitude of standard errors of equating when interpreting the differences among equating relationships. Generally, CSE ranged from approximately 0.50 to 1 score points.

Overall Summary Statistics

Table 4-28 contains WARMSB, WASE, WARMSE, Difference, standardized WARMSB, and standardized WASE for Chemistry 2005-2007. As previously found for the English Language and Spanish Language operational test forms, for the MC-CR 0.00

sampling conditions, as CI ES increased, WARMSB also increased slightly for FE and CE. For FE, WARMSB increased by approximately 0.5 points, and CE increased by approximately 0.1 points. Values of WARMSB for both CI 0.20 and CI 0.40 were larger than the DTM; however. However, for TS and OS, WARMSB actually decreased as the CI ES increased. By reexamining the plots of equating relationships bias in Figure 4-14, it is conceivable that this trend occurred because, at low scores, the CI 0.20 equating relationship resulted in equated scores that were approximately three points higher than the CI 0.40 equating relationship. Across all four equating methods, WARMSB was higher for TS and OS for CI 0.20, but WARMSB was similar across all equating methods for CI 0.40. For the MC-CR 0.00 sampling conditions, standardized WARMSB (second to last block of data) ranged from approximately 0.02 standard deviation units for FE to 0.06 standard deviation units across all sampling conditions.

The pattern of results for MC-CR 0.25 was somewhat different from the results found for the English Language and Spanish Language operational test forms. For CI 0.00 MC-CR 0.25, WARMSB was approximately 2.20 points for FE, 2.60 points for CE, and 1.50 points for TS and OS. These values indicate that the MC-CR 0.25 resulted in average bias across the score scale of approximately 1.50 to 2.50 score points. For FE and CE, values of WARMSB *decreased* for both the CI 0.20 and CI 0.40 sampling conditions. Further, values of WARMSB were larger for CE as compared to FE. For TS and OS, values of WARMSB *increased* for both the CI 0.20 and CI 0.40 sampling conditions. However, WARMSB increased by less than 0.5 score points between CI 0.00 and CI 0.40. For the MC-CR 0.25 sampling conditions, standardized WARMSB ranged from approximately 0.03 to 0.09 standard deviation units across equating methods and sampling conditions.

As described throughout Chapter Four, the values of WARMSB in Table 4-28 do not capture all of the trends seen in Figures 4-14 through 4-16. Table 4-29 contains values of WARMSB based on different criterion equating relationships. The three

columns of data on the left of Table 4-29 contain the same values of WARMSB as shown in Table 4-28. These values of WARMSB were calculated using the CI 0.00 MC-CR 0.00 as the criterion. The three columns of data on the right of the table for the MC-CR 0.25 sampling conditions were calculated using CI 0.00 MC-CR 0.25 as the criterion. Based on the values of WARMSB in Table 4-28, the pattern of results for MC-CR 0.25 seemed somewhat contradictory to those found for English Language and Spanish Language. However, based on the values in Table 4-29, it is evident that WARMSB increased as the CI ES increased for all methods except CE.

Returning to Table 4-28, the second block of data contains values of WASE. Values of WASE were larger for Chemistry as compared to English Language, and similar to or smaller than for Spanish Language. Additionally, values of WASE varied somewhat across sampling conditions, especially for TS and OS. Across sampling conditions, values of WASE ranged from approximately 0.60 to approximately 1 score point for TS and OS. Although plots in Figure 4-16 indicated CSE might be larger for TS and OS as compared to FE and CE for two of the six sampling conditions, OS resulted in the smallest values of WASE for all samples except CI 0.40 MC-CR 0.25. CE consistently resulted in the largest values of WASE. Values of standardized WASE (shown in the last block of Table 4-28) ranged from approximately 0.02 to 0.035 standard deviation units across all sampling conditions and equating methods.

The third block of data in Table 4-28 contains values for WARMSE. Patterns for WARMSE were similar to WARMSB. However, because values of WASE were somewhat large in magnitude, WASE contributed to the overall error to a greater extent than what was found for English Language. This was especially the case for CI 0.20 MC-CR 0.00 and CI 0.40 MC-CR 0.25 for FE. The fourth block of data in Table 4-28 contains values of Difference. Overall, the results for Difference were similar to the results for WARMSB, and the values differed by approximately 0.06 or less.

Classification Consistency

Table 4-30 contains classification consistency results for the Chemistry operational test forms. In Table 4-30, the first column (in bold) contains percentages of classification consistency based on the criterion equating relationships for a given equating method in comparison to the FE criterion equating relationship. CE resulted in 97.3% of examinees receiving the same grade as they received based on FE. For TS and OS, 97.9% of examinees received the same grade as they received based on FE. In interpreting the remaining percentages, it is important to remember that CE, TS, and OS did not result in the same percentages of examinees receiving the same grade as FE for the criterion equating relationships. The remaining columns of data contain the percentage of examinees receiving the same grade for the criterion equating relationship and the study condition equating relationship for a given equating method. Across all sampling conditions, all four equating methods resulted in approximately 90% or more of examinees receiving the same grade as the criterion equating relationship.

Summary

For Chemistry, similar to English Language and Spanish Language, for both the MC-CR 0.00 and MC-CR 0.25 sampling conditions, as the common-item effect size increased, the differences among the equating relationships also increased slightly. However, for the MC-CR 0.00 sampling conditions, the largest difference among the equating relationships was approximately only one point. In contrast to English Language or Spanish Language, bias typically appeared larger for the MC-CR 0.25 conditions relative to the MC-CR 0.00 conditions. However, by comparing standardized values of average bias (WARMSB) across the three tests, Chemistry appeared to typically result in smaller standardized bias than either English Language or Spanish Language for the MC-CR 0.00 sampling conditions. Because the differences among the equating methods were small in comparison to the standard errors of equating, it is important not to

overemphasize the differences among the equating relationships. Standard errors of equating were smallest for OS for five of the sampling conditions and largest for CE for all six of the sampling conditions.

Pseudo-Test Forms

The pseudo-test forms section is divided into three subsections corresponding to the three tests analyzed: English Language, Spanish Language, and Chemistry. Results are presented first for the English Language pseudo-test forms, second for the Spanish Language pseudo-test forms, and last for the Chemistry pseudo-test forms. Additionally, for each test, results are presented for descriptive statistics, equated moments, equating relationships, conditional bias, CSE, overall summary statistics, and classification consistency. The formats of tables and figures are similar across all three tests; therefore, details about formats are primarily provided in the English Language subsection. In order to maintain consistency across the tests, MC-CR 0.00 is consistently used to indicate the MC-CR ES that is most similar to the criterion sample. MC-CR 0.25 is consistently used to indicate the MC-CR ES that is most different from the criterion sample. It is important to note that the actual effect size differs by test.

Throughout the pseudo-test forms results section, when different compositions of common items are considered (i.e., NCR, FCR, NCR MT, NCR SM, NCR DS, or FCR), they are referred to as common-item sets. The abbreviation, CI, is used to indicate different CI ES only. This is being done in an effort to limit confusion between references to the common-item sets and the CI ES.

English Language

Results for the English Language pseudo-test forms are presented in the following subsections: descriptive statistics, equated moments, equating relationships, conditional bias, conditional standard error of equating (CSE), overall summary statistics, and classification consistency.

Descriptive Statistics

Descriptive statistics for the single group old and new English Language pseudo-test forms were provided in Chapter Three in Tables 3-9 and 3-10. Recall that a fourth CI ES (0.60) was considered for the pseudo-test forms, resulting in eight sampling conditions. For the eight sampling conditions, descriptive statistics for the old and new English Language pseudo-test forms are provided in Tables 4-31 and 4-32, respectively. Tables 4-31 and 4-32 contain descriptive statistics for MC, CR, CO, and CI scores. Recall that for the pseudo-test forms, two compositions of common-item sets were created: **N**o **CR** representation (NCR) and **F**ull **CR** representation (FCR). Therefore, there are three sets of descriptive statistics for the common items: NCR, FCR, and Full CI. NCR is the common-item set containing only MC items, FCR is the common-item set containing both MC and CR items, and Full CI is all items in common across the old and new pseudo-test forms. Examinees were sampled using Full CI to determine the target effect sizes. It is important to note that for English Language, means on the new form tended to be lower than means on the old form. This is important to note, because for Spanish Language, means on the new form tended to be higher than means on the old form.

Effect sizes for the English Language pseudo-test forms are shown in Table 4-33. The effect sizes were calculated as new form minus old form scores; consequently, a negative effect size indicates that scores on the new form were lower than scores on the old form. It is important to note that the single group MC-CR ES was -0.115. Therefore, for the MC-CR 0.00 ES, the target effect size was actually -0.115 in order to create samples with MC-CR ES similar to the single group MC-CR ES. The MC-CR ES for the four MC-CR 0.00 sampling conditions were within 0.01 of the target effect size. The target MC-CR 0.25 ES was still an effect size of -0.25, and the four MC-CR 0.25 sampling conditions were within 0.025 of the target effect size. It is important to note that the difference between the MC-CR 0.00 and MC-CR 0.25 ES was approximately 0.15.

The four target CI ES levels were CI 0.00, CI 0.20, CI 0.40, and CI 0.60. The actual Full CI ES were within 0.025 of the target CI ES.

It is important to note the magnitude of the NCR and FCR ES. For the MC-CR 0.00 sampling conditions, the FCR ES were larger in absolute value than the NCR ES. Consequently, FCR ES were more similar than NCR ES to CO ES. For the MC-CR 0.25 sampling conditions, results were mixed. As was previously discussed for the operational test forms, the difference between the CI ES and CO ES varied across sampling conditions. For the single group, this difference was -0.011. The difference based on FCR was generally more similar to the single group criterion than the difference based on NCR, except for CI 0.20 MC-CR 0.25 and CI 0.60 MC-CR 0.25.

Equated Moments

Equated moments are given in Tables 4-34 and 4-35 for the eight sampling conditions and four equating methods for NCR and FCR, respectively. In Table 4-34, for CI 0.00 MC-CR 0.00, means and standard deviations were similar across the four equating methods. As CI ES increased, means for TS and OS were lower than means for CE and FE. For CI 0.60 MC-CR 0.00, means for TS and OS were approximately 0.60 score points lower than the mean for CE and 1.20 score points lower than the mean for FE. Standard deviations were still relatively similar. Similar trends occurred for the MC-CR 0.25 sampling conditions. In Table 4-35, somewhat different results were found for FCR. For the MC-CR 0.00 sampling conditions, FE, CE, TS, and OS resulted in somewhat similar means for all four CI ES. For CI 0.40 and CI 0.60, the mean for FE was somewhat higher than means for CE, TS, and OS. For the MC-CR 0.25 sampling conditions, across all sampling conditions, means for TS and OS were approximately 0.50 to 0.90 score points lower than the mean for CE and 0.60 to 1.50 score points for FE. For each of the sampling conditions, standard deviations were similar for all four equating methods.

Equating Relationships

The criterion equating relationships were based on a single group equating design for the entire sample of examinees taking the pseudo-test forms. Three criterion equating relationships were calculated: equipercentile, TS, and OS. For both the FE and CE methods, the criterion equating relationship was the equipercentile equating relationship. For TS and OS, the criterion equating relationships were TS and OS equating relationships, respectively. Figure 4-18 contains a plot of the three criterion equating relationships. The solid line represents the equipercentile equating relationship, the dotted line represents TS, and the dotted-dashed line represents OS. It is evident in Figure 4-18 that the single-group equating relationships were very similar.

Figures 4-19 through 4-22 contain plots comparing NCR and FCR the equating relationships for each equating method. Each figure contains plots for a different equating method, and each figure contains four plots. The top row contains plots for NCR, and the bottom row contains plots for FCR. Plots for the MC-CR 0.00 sampling conditions are in the left column, and plots for the MC-CR 0.25 sampling conditions are in the right column. For example, in Figure 4-19, the top left plot contains equating relationships for NCR MC-CR 0.00. The bottom left plot contains equating relationships for FCR MC-CR 0.00. The solid bold line in each plot represents the criterion equating relationship for the given equating method, as described in the previous paragraph. There are four additional lines in each plot for the four CI ES levels. In each plot, the solid line represents CI 0.00, the dashed line represents CI 0.20, the dotted line represents CI 0.40, and the dotted-dashed line represents CI 0.60.

Consider Figure 4-19 for FE. In the top left plot for NCR MC-CR 0.00, it is evident that the equating relationship for CI 0.00 was similar to the criterion equating relationship. Further, as noted for the operational test forms, as CI ES increased, the differences between the study condition equating relationships and criterion equating relationship also increased. Now, consider the plot on the left bottom of Figure 4-19 for

FCR MC-CR 0.00. It is evident in this plot that the equating relationships were more similar to each other than they were for NCR. That is, CI ES appeared to impact equating relationships less for the FCR common-item set. As noted for the operational test forms, this result may have occurred because of the difference between the CI ES and CO ES. For NCR, the difference between the CI ES and CO ES was approximately 0.186 for CI 0.60 MC-CR 0.00. For FCR, this difference was only 0.091. It is important to note that, as discussed in Chapter Three, average difficulty for FCR was more similar to the total test than average difficulty for NCR.

Also interesting to note is the magnitude of the CI ES. Sampling was conducted based on Full CI, which was composed of all of the items in common between the two test forms. Consequently, there were no restrictions on the magnitude of the effect size for either NCR or FCR. For CI 0.60 MC-CR 0.00, the effect size for NCR was only approximately -0.50, which was one standard deviation unit less than the target effect size and one standard deviation unit less than the effect size for FCR. If differences among equating relationships resulted primarily from group differences in examinee proficiency, then it would be expected for the equating relationship for NCR to be more similar to the criterion than the equating relationship for FCR. However, equating relationships for NCR were not more similar to the criterion.

Now, consider the plot in the right of the top row in Figure 4-19 for NCR MC-CR 0.25. In this plot, it is evident that the equivalent scores were lower for MC-CR 0.25 than for MC-CR 0.00 (top left), especially for CI 0.00 and CI 0.20. Consequently, the CI 0.20 equating relationship appeared most similar to the criterion. Similar results were seen for the English Language operational test forms. It is interesting to note that for FCR MC-CR 0.25 (bottom right), the equating relationships for CI 0.00 and 0.20 appeared somewhat more similar to those for NCR MC-CR 0.00 (top left). Similar trends can be seen for CE in Figure 4-20. However, it is evident that the equating relationships for CE were more similar across CI ES levels than they were for FE.

Figure 4-21 contains the same plots for the TS equating method. The plot for NCR MC-CR 0.00 (top left) illustrates that all four equating relationships were very similar to the criterion. Equating relationships for FCR MC-CR 0.00 (bottom left) were also similar to the criterion, although they resulted in slightly lower equated scores as compared to NCR. For NCR MC-CR 0.25 (top right), the equating relationships resulted in lower equated scores as compared to NCR MC-CR 0.00. In contrast to the results found for FE and CE, FCR (bottom right) did not appear to result in equating relationships more similar to NCR MC-CR 0.00. Similar results were found for OS (Figure 4-22).

Conditional Bias

Plots of conditional bias are shown in Figures 4-23 and 4-24 for NCR and FCR, respectively. Each figure contains eight plots, one for each of the eight sampling conditions. In each plot, there is one line for each of the four equating methods. FE is represented by the solid line, CE is represented by the dashed line, TS is represented by the dotted line, and OS is represented by the dotted-dashed line. These plots illustrate how similar the four equating methods were for each of the sampling conditions. In Figure 4-23 for NCR, it is evident that for CI 0.00 MC-CR 0.00 (top left) and CI 0.00 MC-CR 0.25 (top right), all four equating methods resulted in similar bias across the score scale. Bias was slightly negative for CI 0.00 MC-CR 0.25. The second row contains plots of conditional bias for CI 0.20. TS and OS appeared to result in less bias than FE and CE for MC-CR 0.00, but more bias than FE and CE for MC-CR 0.25. The third row contains plots for CI 0.40, and the fourth row contains plots for CI 0.60. For all four plots, it is evident that bias was smallest for TS and OS, and largest for FE. Additionally, as CI ES increased, bias also increased; however, the increase in bias for TS and OS was minimal.

Figure 4-24 is identical to Figure 4-23, except that it contains plots for FCR. For CI 0.00 (top row), the four equating methods are less similar to each other than for the NCR equating relationships, especially for MC-CR 0.25 (top right). In the four MC-CR 0.00 plots (left column), it is evident that bias increased slightly as the CI ES increased. However, even for CI 0.60, bias was less than 1.50 score points across the score scale. For MC-CR 0.25 (right column), the magnitude of conditional bias was larger than for MC-CR 0.00. Additionally, patterns of bias were similar to NCR MC-CR 0.25.

Conditional Standard Error of Equating (CSE)

Figures 4-25 to 4-26 contain plots of CSE for MC-CR 0.00 and MC-CR 0.25, respectively. Each plot in these figures represents a different sampling condition. Within each plot, there is one line for each of the four equating methods. The plots in the left column of both figures are for NCR, and the plots in the right column of both figures are for FCR. For MC-CR 0.00 (Figure 4-25), it appears that NCR typically resulted in smaller CSE as compared to FCR, although it is difficult to determine from visual inspection alone. TS and OS typically resulted in smaller CSE than FE and CE for NCR. However, for some of the FCR conditions (CI 0.00 and CI 0.60), TS and OS resulted in larger CSE than FE or CE across most of the score scale. Similar results can be seen in Figure 4-26 for MC-CR 0.25. However, TS and OS appeared to result in larger CSE then FE or CE for all four FCR equating conditions. It is important to note that the same item parameter estimates were used for both NCR and FCR. The only difference was that a different set of common items was used to transform new form estimates onto the old form scale.

Overall Summary Statistics

Tables 4-36 and 4-37 contain overall summary statistics for the eight sampling conditions and four equating methods. Table 4-36 contains summary statistics based on the NCR common-item set, and Table 4-37 contains summary statistics based on the FCR

common-item set. The first four columns of data contain results for the four MC-CR 0.00 sampling conditions (i.e., CI 0.00, CI 0.20, CI 0.40, and CI 0.60), and the last four columns of data contain results for the four MC-CR 0.25 sampling conditions. Recall that for the pseudo-test forms, the criterion equating relationship was the single-group equating relationship.

Consider first Table 4-36 for the NCR common-item set. The first block of data contains values of WARMSB. In the first four columns of data for MC-CR 0.00, it is evident that for CI 0.00, all four equating methods resulted in similar values of WARMSB. Additionally, consistent with results found previously for the operational test forms, values of WARMSB tended to increase as the CI ES increased. As CI ES increased, TS and OS always resulted in the smallest values of WARMSB, and FE always resulted in the largest values of WARMSB. For all of the conditions except CI 0.60, values of WARMSB were less than 0.50 score points for TS and OS. For FE and CE, values of WARMSB were less than 0.50 score points only for CI 0.00. Standardized WARMSB (second to last block of data) was less than 0.10 standard deviation units for all conditions except CE for CI 0.60 and FE for CI 0.20, CI 0.40, and CI 0.60.

For the MC-CR 0.25 sampling conditions, shown in the last four columns of data in Table 4-36, a different trend occurred. Values of WARMSB did not consistently increase as the CI ES increased. This trend was similar to what was found for the English Language operational equating results. For FE and CE, values of WARMSB were smaller for CI 0.20 as compared to CI 0.00. For TS and OS, values of WARMSB decreased as CI ES increased. Standardized WARMSB was less than 0.10 standard deviation units for all conditions except FE for CI 0.40 and CI 0.60.

Now, consider Table 4-37, which contains summary statistics based on the FCR common-item set. Similar trends were seen for the FCR common-item set. Next, compare values of WARMSB in the first four columns of Table 4-36 to values of WARMSB in the first four columns of Table 4-37. It is evident that the magnitude of WARMSB for

larger CI ES was much lower for FCR (Table 4-37) as compared to NCR (Table 4-36) for FE and CE. For CI 0.60, based on FCR as compared to NCR, values of WARMSB were approximately 0.60 (1.819-1.249) points smaller for FE and 0.80 (1.327-0.543) points smaller for CE. For TS and OS, the magnitude of WARMSB was similar for both NCR and FCR. Unexpectedly, similar trends were not found for MC-CR 0.25. For FE and CE, values of WARMSB were only slightly lower for CI 0.00 and CI 0.40 when FCR was the common-item set. However, values of WARMSB were slightly larger for CI 0.20 and CI 0.60 when FCR was the common-item set as compared to NCR. As mentioned previously, for the MC-CR 0.00 sampling conditions, the difference between the CI ES and CO ES was more similar to the difference for the criterion for FCR than for NCR. However, for the MC-CR 0.25 sampling conditions, the same trend did not occur. The second to last block of data in Tables 4-36 and 4-37 contain standardized values of WARMSB. In general, similar conclusions were reached based on the standardized values. Additionally, standardized WARMSB ranged from approximately 0.02 to 0.24 standard deviation units for NCR. For FCR, standardized WARMSB ranged from approximately 0.03 to 0.20 standard deviation units.

As noted previously for the operational test forms, values of WARMSB often masked some of the patterns of results that were evident by examining the figures. Table 4-38 contains values of WARMSB for illustrating the influence of CI ES on WARMSB based on different criterion equating relationships. The columns of data on the left of the table were calculated using CI 0.00 MC-CR 0.00 as the criterion equating relationship. The columns of data on the right of the table were calculated using CI 0.00 MC-CR 0.25 as the criterion equating relationship. The first block of data contains values based on the NCR common-item set. It is evident that for both MC-CR 0.00 and MC-CR 0.25, as CI ES increased WARMSB also increased. The second block of data contains values based on the FCR common-item set. Similar results can be seen for FCR, although values of WARMSB did not increase for TS and OS for the MC-CR 0.25 conditions. However, it is

interesting to note that values of WARMSB based on FCR were generally lower for FE and CE as compared to values based on NCR. This result indicates that the equating relationships based on the FCR common-item set were more similar across CI ES. Values of WARMSB were typically larger for TS and OS based on FCR, however.

Returning to Tables 4-36 and 4-37, the second block of data contains values for WASE. Across all sampling conditions for NCR (Table 4-36), values of WASE were smallest for OS and largest for CE. Values of WASE were approximately 0.23 for OS, 0.25 for TS, 0.28 for FE, and 0.33 for CE. There was some variability across sampling conditions, but WASE typically varied by 0.05 or less. For FCR (Table 4-37), values of WASE were smallest for OS for only two sampling conditions. For the remainder of the conditions, FE resulted in the smallest values of WASE. However, OS still resulted in smaller values of WASE as compared to TS. CE also always resulted in larger values of WASE as compared to FE. Comparing values of WASE for NCR (Table 4-36) to those for FCR (Table 4-37), for TS and OS, values of WASE were consistently higher for FCR as compared to NCR. Similar results were not found for FE and CE. Standardized values of WASE, shown in the last block of data in Tables 4-36 and 4-37 were approximately 0.03 to 0.05 standard deviation units across all sampling conditions and equating methods.

The third block of data in Tables 4-36 and 4-37 contains values for WARMSE. Patterns of WARMSE were similar to patterns of WARMSB. The fourth block of data in Tables 4-36 and 4-37 contains values of Difference. Overall, the results for Difference were similar to the results for WARMSB and differed by only approximately 0.04 or less.

Classification Consistency

Table 4-38 contains classification consistency results for the English Language pseudo-test forms. Recall that there were three equating relationships for the criterion: equipercentile, TS, and OS. The first column of data (in bold) contains the percentages of

classification consistency based on the TS or OS criterion equating relationships in comparison to the equipercentile criterion equating relationship. The criterion for both FE and CE was the equipercentile equating relationship. The first block of values contains classification consistency percentages for NCR for the four equating methods, and the second block contains the same values for FCR. It is important to note that the same values are listed for the single group criterion for both NCR and FCR in order to more easily make comparisons. The single group criterion equating relationship was calculated without the use of common items. The value of 96.5 for TS in the first column and third row (or seventh row) indicates that 96.5 percent of examinees received the same grade based on the equipercentile criterion equating relationship as they did based on the TS criterion equating relationship. TS and OS resulted in the same percentage.

The remaining columns of data contain the percentage of examinees receiving the same grade for the criterion equating relationship and the study condition equating relationship for a given equating method. For example, in the second column (CI 0.00 MC-CR 0.00) and first row (FE) of data, 100.0 is the percentage of examinees receiving the same grade for the FE CI 0.00 MC-CR 0.00 (NCR) study condition equating relationship as they received for the equipercentile criterion equating relationship. Similarly, 100.0 is the percentage of examinees receiving the same grade for the CE CI 0.00 MC-CR 0.00 (NCR) study condition equating relationship as they received for the equipercentile criterion equating relationship. Across all sampling conditions and common-item compositions, TS and OS resulted in classification consistency percentages of 90% or more. However, it is important to remember that the criterion TS and OS equating relationships resulted in slightly different percentages of classification consistency from the equipercentile criterion. With the exception of three sampling conditions for NCR and one sampling condition for FCR, CE also resulted in classification consistency percentages of 90% or more. Across nearly all sampling conditions for NCR, FE resulted in classification consistency percentages less than 90%.

For FCR, FE resulted in classification consistency percentages greater than 90% for half of the sampling conditions. In general, for FE and CE, percentages of classification consistency were higher for FCR as compared to NCR. Percentages of classification consistency for TS and OS were similar across NCR and FCR, although percentages were higher by approximately 5% for FCR for CI 0.60 MC-CR 0.00. Percentages were approximately 5% lower for FCR for CI 0.40 MC-CR 0.25.

Summary

For the English Language pseudo-test forms, consistent with previous results, as the common-item effect size increased, the differences among the equating relationships also increased somewhat. However, the largest differences among the equating relationships were generally less than 2 score points. Of particular importance for the pseudo-test analyses is the comparison of the common-item sets containing only multiple-choice items and the common-item sets containing both multiple-choice and constructed-response items. For the MC-CR 0.00 conditions, the common-item set that included both multiple-choice and constructed-response items generally resulted in less bias than the common-item set that included only multiple choice items, especially for FE and CE. For the MC-CR 0.25 conditions, both common-item sets resulted in similar amounts of bias. For the common-item set containing only multiple-choice items, standard errors of equating were smallest for OS and largest for CE across all sampling conditions. However, for the common-item set containing both multiple-choice and constructed-response items, standard errors of equating were smallest for OS for only two of the eight sampling conditions. For the other six sampling conditions, FE resulted in the smallest standard errors. CE resulted in the largest standard errors for only two of the sampling conditions, and TS resulted in the largest standard errors for the other six sampling conditions.

Spanish Language

Results for the Spanish Language pseudo-test forms are presented in the following subsections: descriptive statistics, equated moments, equating relationships, conditional bias, conditional standard error of equating (CSE), overall summary statistics, and classification consistency.

Descriptive Statistics

Descriptive statistics for the old and new Spanish Language pseudo-test form sampling conditions are provided in Tables 4-40 through 4-43. Descriptive statistics for the single group pseudo-test forms were provided in Tables 3-9 and 3-10. It is also important to remember from Table 3-8 that mean difficulty was similar for NCR MT, NCR SM, and FCR. Table 4-40 contains descriptive statistics for MC, CR, and CO scores for the old form for the eight sampling conditions. Table 4-41 contains descriptive statistics for the common-item sets for the old form. Tables 4-42 and 4-43 contain the same information for the new pseudo-test form. For the pseudo-test form analyses, descriptive statistics for the common items are especially important to consider (Tables 4-41 and 4-43). In both tables, NCR MT refers to the minitest common-item set, NCR SM refers to the semi-miditest common-item set, NCR DS refers to the difficulty shift common-item set, and FCR refers to the common-item set containing both MC and CR items. Full CI refers to all items in common between the two test forms. Means for NCR SM and FCR were generally higher than for NCR MT. The mean for NCR DS was generally lower than for NCR MT. Additionally, standard deviations were generally larger for NCR SM and NCR DS than for NCR MT. Standard deviations were generally smallest for FCR. NCR DS was much less negatively skewed than the other common-item sets. However, it is important to note that the skewness varied substantially across sampling conditions.

The rows labeled “CO Corr.” in Tables 4-41 and 4-43 contain correlations between CI and CO scores. In Table 4-41 for the old pseudo-test form samples, the FCR common-item set always resulted in the largest correlation between CI and CO scores. NCR DS always resulted in the lowest correlation. For the MC-CR 0.00 sampling conditions (first four columns of data), NCR SM resulted in correlations that were slightly larger than those for NCR MT. For the MC-CR 0.25 sampling conditions, correlations between CI and CO scores were typically larger for NCR MT as compared to NCR SM. In Table 4-43 for the new form pseudo-test form sampling conditions, similar trends occurred.

Table 4-44 contains effect sizes based on new form minus old form scores; consequently, negative scores indicate that means on the new form were lower than means on the old form. The first row contains effect sizes for CO scores. The second block of values contains effect sizes for NCR MT, NCR SM, NCR DS, FCR, and Full CI. The MC, CR, and MC-CR effect sizes are contained in the last block of Table 4-44. It is important to note that the MC-CR effect size for the single group was -0.270. Therefore, for the MC-CR 0.00 sampling conditions, the MC-CR effect sizes were similar to this value. All of the MC-CR 0.00 effect sizes were within 0.02 standard deviation units of -0.270. The MC-CR effect size for the MC-CR 0.25 sampling conditions differed from -0.270 by approximately 0.25 standard deviation units. The effect sizes for the MC-CR 0.25 sampling conditions were approximately 0. Although the magnitude of the effect sizes and labels is counterintuitive, remember that MC-CR 0.00 represents the sampling conditions with the MC-CR effect size most similar to the criterion. MC-CR 0.25 represents the sampling conditions with the MC-CR effect sizes most different from the criterion.

As described previously for the English Language pseudo-test forms, the CI ES were created by sampling examinees using Full CI to obtain the target effect size. Across all sampling conditions, the effect sizes for Full CI were within 0.035 of the target effect

sizes. For the MC-CR 0.00 sampling conditions, effect sizes for the four common-item sets considered for Spanish Language were relatively similar. Across the compositions, effect sizes differed by only 0.05 standard deviation units. For example, for CI 0.60 MC-CR 0.00, the smallest effect size was -0.561 for NCR DS. The largest effect size was -0.612 for FCR. However, results were much different for the MC-CR 0.25 sampling conditions. For these conditions, effect sizes for the four common-item sets differed by as much as 0.25 standard deviation units. For CI 0.60 MC-CR 0.25, the smallest effect size was -0.356 for NCR DS, and the largest effect size was -0.590 for FCR. For CI 0.00 MC-CR 0.25; however, effect sizes differed by only approximately 0.06 standard deviation units.

Equated Moments

Equated moments are provided in Tables 4-45 through 4-48. Each table contains equated moments based on one of the four common-item sets (i.e., NCR MT, NCR SM, NCR DS, or FCR). The first column of data in each table contains equated moments for the single group. The next four columns of data contain equated moments for the MC-CR 0.00 sampling conditions, and the last four columns of data contain equated moments for the MC-CR 0.25 sampling conditions. For example, Table 4-45 contains equated moments for equating conducted using NCR MT as the set of common items.

Consider first Table 4-45 for NCR MT. For the MC-CR 0.00 sampling conditions, for CI 0.00, means for the four equating methods were within approximately 0.50 score points of each other. As the CI ES increased, means for the four equating methods became substantially different, with means for TS and OS being approximately 2 points smaller than the mean for FE and 1.20 points smaller than the mean for CE. For the MC-CR 0.25 sampling conditions, for CI 0.00, across the four equating methods, means were very similar. For CI 0.60, means for TS and OS were smaller by approximately 3 points for FE and 2 points for CE. With the exception of CI 0.60 MC-CR 0.25, standard

deviations were typically within 0.50 score points across the four equating methods. Similar trends occurred for the means for NCR SM, as shown in Table 4-46. However, differences in means among the four equating methods were generally larger than those for NCR MT. Additionally, standard deviations were not as similar across the four equating methods. Somewhat similar trends also occurred for NCR DS (Table 4-47). However, for the MC-CR 0.25 sampling conditions, means for the four equating methods differed by only approximately 1 score point or less. Lastly, for FCR (Table 4-48), means for the four equating methods typically differed by less than 1.50 score points for the MC-CR 0.00 sampling conditions. For the MC-CR 0.25 sampling conditions, means for the four equating methods differed by less than 1 score point. Standard deviations were typically within 0.50 score points across the four equating methods. Additionally, by comparing Tables 4-45 through 4-48, it is evident that, of the four common-item set compositions, FCR typically resulted in the smallest means while NCR DS typically resulted in the largest means.

Equating Relationships

As previously described, the criterion equating relationships for Spanish Language pseudo-test forms were based on a single-group equating. For both the FE and CE methods, the criterion was an equipercentile equating relationship. For TS and OS, the criterion equating relationships were TS and OS equating relationships, respectively. Figure 4-27 contains a plot of the three criterion equating relationships. It is evident in Figure 4-27 that the three single-group equating relationships were very similar across the score scale.

Figures 4-28 through 4-31 contain plots of the equating relationships based on the four different common-item sets. Each figure contains plots for a different equating method. For example, Figure 4-28 contains plots for FE; Figure 4-29 contains plots for CE, and so on. Each figure contains eight plots. The left column contains plots for the

MC-CR 0.00 equating relationships, and the right column contains plots for the MC-CR 0.25 equating relationships. Each row contains plots for a different common-item set composition. For example, the first row in Figure 4-28 contains plots for NCR MT for the FE equating method. The second row in Figure 4-28 contains plots for NCR SM for the FE equating method. In each plot, there are five lines, one for each CI ES sampling condition (i.e., CI 0.00, CI 0.20, CI 0.40, and CI 0.60) and one for the criterion equating relationship. The criterion equating relationship is illustrated by the solid dark black line. The CI ES equating relationships are illustrated by solid (CI 0.00), dashed (CI 0.20), dotted (CI 0.40), and dotted-dashed (CI 0.60) lines.

Consider Figure 4-28 for the FE equating method first. The plot on the left of the first row is for NCR MT for the MC-CR 0.00 sampling conditions. As previously noted for other tests, the difference between the study condition equating relationships and the criterion equating relationship increased as the CI ES increased. It is important to note that the equating relationship for CI 0.00 MC-CR 0.00 resulted in slightly lower equated scores than the criterion equating relationship. The plot on the right of the first row is for NCR MT for the MC-CR 0.25 sampling conditions. It is evident in this plot that equating relationships for MC-CR 0.25 resulted in substantially higher equated scores as compared to MC-CR 0.00. Next, consider the plots for NCR SM and NCR DS, in the second and third rows, respectively. Similar trends occurred; however, the differences among the equating relationships for the four CI ES were larger. Lastly, consider the plots for FCR in the last row of Figure 4-28. By comparing FCR (last row) to NCR MT (first row), it is evident that the equating relationships for MC-CR 0.00 (left column) were similar for both common-item sets. However, in the plot for MC-CR 0.25 (right column), the equating relationships for FCR were substantially different from the equating relationships for the other three common-item sets. First, the equating relationships were more similar to the criterion. Second, the equating relationships for the four CI ES were all very similar to each other, especially at scores greater than approximately 60. The

pattern of results described for FE in Figure 4-28 can also be seen in Figure 4-29 for CE. However, the differences among the equating relationships for different CI ES were smaller for CE as compared to FE.

Although the general pattern of equating results was similar for TS and OS as compared to FE and CE, there were notable differences. Figures 4-30 and 4-31 contain plots of the equating relationships for TS and OS, respectively. In contrast to FE and CE, for the MC-CR 0.00 sampling conditions (left column), equating relationships differed very little across the four CI ES levels, even for NCR DS. The largest differences among equating relationships was for the NCR DS common-item set, although equating relationships differed by only approximately 2 score points or less. Further, there was relatively little difference in the equating relationships for the four common-item sets. Similar to FE and CE, for the MC-CR 0.25 (right column) sampling conditions, the equating relationships generally resulted in higher equated scores relative to the MC-CR 0.00 sampling conditions. However, for all of the common-item sets, the four CI ES equating relationships still remained very similar to each other. Again, the largest differences among equating relationships occurred for NCR DS, but equated scores differed by approximately 3 points or less. Further, it is not clear whether FCR was the common-item set resulting in equating relationships most similar to the criterion.

Conditional Bias

Figures 4-32 through 4-35 contain plots of conditional bias. Each figure contains eight plots of bias. The left column contains plots for the MC-CR 0.00 equating relationships, and the right column contains plots for the MC-CR 0.25 equating relationships. Each row contains plots for a different common-item set composition. Additionally, each figure contains plots for a different CI ES (i.e., CI 0.00, CI 0.20, CI 0.40, and CI 0.60). In each plot, there are four lines, one for each equating method. The

equating methods are illustrated by solid (FE), dashed (CE), dotted (TS), and dotted-dashed (OS) lines.

Consider the plots of conditional bias in Figure 4-32 for CI 0.00. By comparing the MC-CR 0.00 sampling conditions (left column) to the MC-CR 0.25 sampling conditions (right column), it is evident that the patterns of conditional bias were quite different across the two MC-CR ES levels. For the three MC-only common-item sets (i.e., NCR MT, NCR SM, and NCR DS), FE and CE resulted in similar patterns of conditional bias. TS and OS also resulted in similar patterns of conditional bias. However, conditional bias for FE and CE differed from conditional bias for TS and OS by approximately 2 score points. For FCR, all four equating methods resulted in similar conditional bias. Similarly, for CI 0.20 in Figure 4-33, all four equating methods for FCR resulted in similar bias for both the MC-CR 0.00 and MC-CR 0.25 sampling conditions. For the other three common-item sets, it was evident that bias was larger for FE and CE as compared to TS and OS. Interestingly, for NCR DS, bias for TS and OS appeared to increase. For the other three common-item sets, bias did not appear to increase substantially for TS and OS. Similar trends can also be seen in Figures 4-34 and 4-35 for CI 0.40 and CI 0.60, respectively. Even for CI 0.60, for both the MC-CR 0.00 and MC-CR 0.25 sampling conditions, the four equating methods resulted in similar bias for FCR. For the three MC-only common-item sets, it is evident than FE and CE resulted in much larger bias than TS or OS.

Conditional Standard Error of Equating

Conditional standard errors of equating are shown in Figures 4-36 through 4-39. The figures of CSE are organized the same way as the figures of conditional bias. In Figure 4-36, for CI 0.00, it is evident that TS and OS resulted in smaller values of CSE across most of the score scale as compared to FE and CE. CE resulted in the largest values of CSE. In Figures 4-37 and 4-38 for CI 0.20 and CI 0.40, respectively, values of

CSE for TS and OS were typically smaller than CSE for FE and CE, but were similar to or larger than values of CSE for FE for one condition. (See the plot for the NCR DS common-item set and MC-CR 0.25 sampling condition in the third row and right column.) In Figure 4-39, TS and OS often appeared to result in values of CSE similar to or larger than those for FE and CE. This trend was most evident for the MC-CR 0.25 sampling conditions (right column) or the NCR DS common-item set (third row). For each of the four CI ES levels, the magnitude of standard errors did not appear to differ substantially across the four common-item sets.

Overall Summary Statistics

Tables 4-49 through 4-52 contain overall summary statistics for the eight sampling conditions and four equating methods. Each table contains results for a different common-item set. That is, Table 4-49 contains results for NCR MT; Table 4-50 contains results for NCR SM, and so on. Results for these tables are discussed collectively, because many of the same patterns occurred across the various common-item sets. As has generally been found for all of the tests in this dissertation, for MC-CR 0.00, for FE and CE, WARMSB generally increased as the CI ES increased. However, for NCR SM, NCR DS, and FCR, values of WARMSB for CI 0.20 MC-CR 0.00 were smaller than values for CI 0.00 MC-CR 0.00. This result was expected given the equating relationships shown in Figures 4-28 through 4-31. Additionally, values of WARMSB for CI 0.20 MC-CR 0.25 for FCR were also smaller than the values for CI 0.00 MC-CR 0.25. Further, for TS and OS, values of WARMSB did not exhibit a consistent increasing or decreasing trend, especially for NCR DS and FCR. For both the MC-CR 0.00 and MC-CR 0.25 sampling conditions, values of WARMSB were highest for FE and CE and lowest for TS and OS. The only conditions for which this trend did not occur were the FCR MC-CR 0.25 sampling conditions. Comparing values of WARMSB for MC-CR 0.25 to those for MC-

CR 0.00, values of WARMSB were substantially larger across all common-item sets for the MC-CR 0.25 sampling conditions.

Comparing the magnitude of WARMSB across the four common-item sets is of particular importance. Comparing the magnitude of the three MC-only common-item sets, values of WARMSB were generally largest for NCR DS and smallest for NCR MT. Values of WARMSB were always largest for NCR DS. However, there were a few conditions for which values of WARMSB for NCR SM were smaller than those for NCR MT. These values are shaded in grey in Table 4-50.

Table 4-52 contains overall summary statistics for FCR. It is especially important to note a few things about the results for FCR in comparison to those for NCR MT (Table 4-49). For the MC-CR 0.00 sampling conditions, values of WARMSB for FE and CE were typically smaller for FCR as compared to NCR MT. However, values of WARMSB for TS and OS were typically smaller for NCR MT. For the MC-CR 0.25 sampling conditions, FCR resulted in substantially lower values of WARMSB for FE and CE relative to NCR MT. For CI 0.60 MC-CR 0.25, values of WARMSB for FCR were approximately 2 score points lower for both FE and CE. However, for TS and OS, values of WARMSB were larger for FCR as compared to NCR MT by as much as approximately 0.30 score points. As noted previously in the descriptive statistics section, the effect sizes for FCR were much more similar to the composite score effect sizes, especially for large CI ES. Consequently, it appeared that when the relative difficulty of MC and CR items differed across test forms, common items consisting of both MC and CR items more accurately represented the total test.

The second to last block of data in Tables 4-49 to 4-52 contain standardized values of WARMSB. Standardized WARMSB generally ranged from approximately 0.025 standard deviation units to 0.38 standard deviation units, although standardized WARMSB was as large as approximately 0.45 standard deviation units for NCR DS. Values of standardized WARMSB were smaller for FCR as compared to NCR MT by as

much as 0.04 standard deviation units for MC-CR 0.00. For MC-CR 0.25, standardized WARMSB was smaller for FCR than NCR MT by as much as 0.11 standard deviation units.

Table 4-53 provides values of WARMSB to better illustrate the impact of CI ES on WARMSB using different criterion equating relationships. As for the other tests studied in this dissertation, as the CI ES increased from 0.00, WARMSB also typically increased for both MC-CR 0.00 and MC-CR 0.25 sampling conditions. For NCR DS, values of WARMSB did not steadily increase for TS and OS as the CI ES increased. Additionally, for FCR (MC-CR 0.00), values of WARMSB did not steadily increase for CE, TS, or OS. Further, for MC-CR 0.25, there was little change in the values of WARMSB based on FCR for CE, TS, or OS as CI ES increased.

Returning to Tables 4-49 to 4-52, the second block of data contains values for WASE. Across common-item sets, values of WASE ranged from 0.35 for OS using FCR to 0.80 for CE using NCR DS. Values of WASE were typically similar for NCR SM as compared to NCR MT. Values of WASE tended to be larger for NCR DS than for NCR MT across all equating methods and sampling conditions. Values of WASE also tended to be smaller for FCR than for NCR MT. Across all common-item sets and sampling conditions, OS typically resulted in the smallest values of WASE and CE typically resulted in the largest values of WASE. Values of standardized WASE (last block of data) were approximately 0.03 to 0.05 standard deviation units across all common-item sets, sampling conditions, and equating methods.

The third block of data in Tables 4-49 to 4-52 contains values for WARMSE. Similar patterns were found for both WARMSB and WARMSE. Overall, the results for Difference (fourth block of data) were similar to the results for WARMSB. However, for TS and OS, differences between WARMSB and Difference were as large as 0.20 points.

Classification Consistency

Table 4-54 contains classification consistency percentages for the Spanish Language pseudo-test forms. The first column of data (in bold) represents the criterion equating relationship based on the single group equating relationship, as described previously. Recall that there were three equating relationships for the single group: equipercentile, TS, and OS. The TS and OS criterion equating relationships resulted in slightly different percentages of examinees receiving the same grade as the equipercentile criterion equating relationship (97.5% and 98.4%, respectively).

The remaining columns of data contain the percentages of examinees receiving the same grades based on the study condition equating relationships and the criterion equating relationship for a given equating method. The next four columns of data are for MC-CR 0.00 sampling conditions, and the last four columns of data are for MC-CR 0.25 sampling conditions. Additionally, there are four blocks of data in Table 4-54. Each block contains classification consistency percentages for a different common-item set.

For the MC-CR 0.00 sampling conditions, results were mixed as to which common-item set resulted in the highest percentage of examinees receiving the same grade for the study condition equating relationship and the criterion equating relationship. In general, the percentages were approximately 90% or higher for all equating methods, sampling conditions, and common-item sets. The only exceptions were CE for NCR DS and FE. It is not surprising that the classification consistency percentages were similar, given that the differences between the CI ES and CO ES were all very similar across the common-item sets for MC-CR 0.00. For the MC-CR 0.25 sampling conditions, none of the common-item sets resulted in high percentages of classification consistency for any of the CI ES. For FE and CE, FCR typically resulted in higher percentages of classification consistency than the other common-item sets. For FE, classification percentages were as much as 12% higher for FCR as compared to NCR MT. For TS and OS, NCR MT

typically resulted in the highest percentages of classification consistency, although the percentages for NCR MT, NCR SM, and FCR were all relatively similar.

Summary

For the Spanish Language pseudo-test forms, consistent with the other tests, as the common-item effect size increased, bias also increased. Of particular importance for the Spanish Language pseudo-test forms is the impact of common-item composition on bias. Four common-item compositions were considered: three common-item sets containing only multiple-choice items and one common-item set containing both multiple-choice and constructed-response items. Of the three common-item sets containing only multiple-choice items, the minitest typically resulted in the smallest bias, but bias was similar for the semi-miditest. The difficulty shift common-item set resulted in the largest bias. For the MC-CR 0.00 sampling conditions, the common-item set containing both multiple-choice and constructed-response items resulted in similar, but typically somewhat smaller bias, than the minitest. However, for the MC-CR 0.25 sampling conditions, bias for the common-item set containing both multiple-choice and constructed-response items was substantially smaller than bias for the minitest for FE and CE. Similar to previous tests, bias did not vary substantially across common-item effect sizes for TS or OS.

Chemistry

Results for the Chemistry pseudo-test forms are presented in the following subsections: descriptive statistics, equated moments, equating relationships, conditional bias, conditional standard error of equating (CSE), overall summary statistics, and classification consistency.

Descriptive Statistics

Descriptive statistics for the old and new Chemistry pseudo-test forms are provided in Tables 4-55 and 4-56 for the old and new forms, respectively, for the eight

sampling conditions. Descriptive statistics for the single group were provided in Chapter Three in Tables 3-9 and 3-10. The descriptive statistics are given for MC, CR, CO, and CI items. There are also three sets of descriptive statistics for the common items: NCR, FCR, and Full CI. NCR is the common-item set containing only MC items, FCR is the common-item set containing both MC and CR items, and Full CI is all items in common across the old and new pseudo-test forms. Examinees were sampled using Full CI to determine the target effect sizes. There are a few important things to note about the descriptive statistics:

1. By comparing means in Table 4-55 to those in Table 4-56, it is evident that means on the old form were generally lower than means on the new form.
2. In general, old form scores tended to be less negatively skewed than new form scores. Additionally, samples with larger CI ES tended to be less negatively skewed on the old form and more negatively skewed on the new form.
3. Means were lower and standard deviations were larger for FCR as compared to NCR.

Effect sizes for the Chemistry pseudo-test forms are shown in Table 4-57. The effect sizes were calculated as new form minus old form scores; consequently, a negative effect size indicates that scores on the new form were lower than scores on the old form. It is important to note that the single group MC-CR ES was approximately 0.05. Consequently, for the MC-CR 0.00 ES, the target effect size was actually 0.05 in order to create samples with similar effect sizes. It was not possible to obtain a large MC-CR ES, so the target MC-CR 0.25 ES was actually an effect size of only approximately 0.11. This means that the difference between the two target MC-CR ES was only approximately 0.06. The MC-CR ES for the four MC-CR 0.00 sampling conditions were within 0.01 of the target effect size. The four MC-CR 0.25 sampling conditions were within 0.015 of the target effect size. The four target CI ES levels were CI 0.00, CI 0.20, CI 0.40, and CI 0.60. The actual Full CI ES were within approximately 0.025 of the target CI ES. It is

also important to note that, for MC-CR 0.00, effect sizes for FCR tended to be more similar to the CO ES. For MC-CR 0.25, effect sizes for NCR tended to be more similar to the CO ES. However, the differences between the NCR and FCR effect sizes were never larger than 0.03.

Equated Moments

Equated moments are given in Tables 4-58 and 4-59 for the eight sampling conditions and four equating methods for NCR and FCR, respectively. For both NCR and FCR, for CI 0.00 MC-CR 0.00, means and standard deviations were similar across the four equating methods. For the NCR common-item set, as the CI ES increased, the differences between means also increased. For CI 0.60 MC-CR 0.00, means for TS and OS were approximately 1.50 points higher than the mean for FE and 0.80 points higher than the mean for CE. Similar results were seen for the MC-CR 0.25 sampling conditions. Means for TS and OS were larger than means for FE and CE for large CI ES. Standard deviations remained relatively similar, even for large CI ES.

For the FCR common-item set, means for TS and OS were approximately 1 point higher than the mean for FE and 0.30 points higher than the mean for CE for CI 0.60 MC-CR 0.00. For CI 0.60 MC-CR 0.25, means for TS and OS were approximately 2 points higher than the mean for FE and 1.60 points higher than the mean for CE. Standard deviations were still relatively similar.

Equating Relationships

The criterion equating relationships were based on a single group equating for the entire sample of examinees taking the pseudo-test forms. For both the FE and CE methods, the criterion was an equipercentile equating relationship. For TS and OS, the criterion equating relationships were TS and OS equating relationships, respectively. Figure 4-40 contains a plot of the three criterion equating relationships. It is evident in

Figure 4-40 that the single-group equating relationships were very similar across most of the score scale.

Figures 4-41 through 4-44 contain plots comparing the equating relationships for NCR and FCR. In discussion of these plots, it is important not to overemphasize the differences among equating methods, because the differences among equating relationships were small, and the standard errors of equating were relatively large in comparison. Each figure contains four plots. The top row contains plots for NCR, and the bottom row contains plots for FCR. In the left column are plots for the MC-CR 0.00 sampling conditions, and in the right column are plots for the MC-CR 0.25 sampling conditions. For example, the top left plot contains equating relationships for NCR MC-CR 0.00. The bottom left plot contains equating relationships for FCR MC-CR 0.00. The solid bold line in each plot represents the criterion equating relationship for the specific equating method, as described in the previous paragraph. There are four additional lines in each plot for the four CI ES levels. In each plot, the solid line represents CI 0.00, the dashed line represents CI 0.20, the dotted line represents CI 0.40, and the dotted-dashed line represents CI 0.60. Each figure contains plots for a different equating method.

Consider the first plot in Figure 4-41 for the FE method and MC-CR 0.00. In this plot, it is evident that the equating relationship for CI 0.00 is similar to the criterion equating relationship at high and low scores. Throughout the middle of the score distribution, CI 0.20 appears to result in an equating relationship that is more similar to the criterion than the equating relationship for CI 0.00. The difference between the study condition equating relationships and the criterion equating relationship did not consistently increase as CI ES increased. Now, consider the plot on the left in the bottom row of Figure 4-41 for FCR MC-CR 0.00. The equating relationships for FCR are very similar to those for NCR; although, the equating relationships for the four CI ES appeared more similar to each other than the equating relationships for NCR.

Consider the plot in the right column and top row of Figure 4-41 for NCR MC-CR 0.25. In this plot, it is evident that the equating relationships generally resulted in slightly higher equated scores than the equating relationships for MC-CR 0.00. The equating relationships for the four CI ES also appeared to be somewhat more similar to each other. In the right column and bottom row, for FCR MC-CR 0.25, the equating relationships appeared very similar to the NCR MC-CR 0.25 equating relationships. Similar trends also occurred for CE (Figure 4-42).

Figure 4-43 contains the same plots, but for the TS equating method. The plot for NCR MC-CR 0.00 (top left) illustrates that the equating relationships were very similar across all four CI ES. For FCR MC-CR 0.00 (bottom left), the equating relationships were slightly less similar across all four CI ES. For MC-CR 0.25, the equating relationships resulted in somewhat higher scores as compared to MC-CR 0.00 for both NCR (top right) and FCR (bottom right). Similar results were found for OS, shown in Figure 4-44.

Conditional Bias

Plots of conditional bias are shown in Figures 4-45 and 4-46. Figure 4-45 contains plots for NCR, and Figure 4-46 contains plots for FCR. Each figure contains eight plots, one for each of the eight sampling conditions. The left column contains plots for MC-CR 0.00, and the right column contains plots for MC-CR 0.25. In each plot, there are four lines: one for each of the four equating methods. These plots illustrate how similar the four equating methods were to one another for each of the sampling conditions. Considering Figure 4-45 for NCR MC-CR 0.00 (left column), it is evident that for CI 0.00, all four equating methods resulted in similar amounts of bias across the score scale. In the second and third rows of plots, for CI 0.20 and CI 0.40, TS and OS appeared to result in less bias than FE and CE. Bias was near 0 across the score scale for TS and OS. For CI 0.60 in the last row of plots, TS and OS still appeared to result in less bias than FE

and CE; however, the pattern of bias differed between traditional and IRT equating methods.

In the second column of Figure 4-45 for MC-CR 0.25, as the CI ES increased, the difference in bias among the four equating relationships also increased. All four equating methods appeared to result in similar bias for CI 0.00 and CI 0.20. However, for CI 0.40 and CI 0.60, it was difficult to determine by visual inspection which method resulted in smaller bias. Figure 4-46 is identical to Figure 4-45, except that it contains plots for FCR. By comparing Figure 4-45 to Figure 4-46, it is evident that the patterns of bias for FCR were similar to those for NCR. It is difficult to determine by visual inspection whether FCR or NCR resulted in larger bias.

Conditional Standard Error of Equating (CSE)

Figures 4-47 and 4-48 contain plots of CSE. Figure 4-47 contains CSE for the MC-CR 0.00 equating relationships, and Figure 4-48 contains CSE for the MC-CR 0.25 equating relationships. The plots on the left of both of the figures are for NCR, and the plots on the right of both of the figures are for FCR. It is difficult to determine by looking only at the plots whether NCR or FCR resulted in smaller CSE. However, across all conditions, TS and OS resulted in smaller CSE throughout most of the score scale, and CE resulted in the largest.

Overall Summary Statistics

Tables 4-60 (NCR) and 4-61 (FCR) contain overall summary statistics for the eight sampling conditions and four equating methods. Consider first Table 4-60 for NCR. By looking at each equating method individually, for MC-CR 0.00, values of WARMSB tended to increase as the CI ES increased, although WARMSB was slightly lower for TS and OS for CI 0.20. By comparing results across the four equating methods in the WARMSB block of data, it is also evident that TS and OS always resulted in the smallest values of WARMSB. CE resulted in the largest values of WARMSB for CI 0.00, and FE

resulted in the largest values of WARMSB for the remaining three CI ES. For MC-CR 0.25, for all four equating methods, there was not a consistent increasing trend. Additionally, for FCR, in Table 4-61, FE was the only equating method for which WARMSB consistently increased as CI ES increased. This was the case for both MC-CR 0.00 and MC-CR 0.25. Standardized WARMSB was less than 0.10 standard deviation units across all equating methods, sampling conditions, and common-item sets.

It is also important to compare values of WARMSB for FCR to those for NCR. For MC-CR 0.00, values of WARMSB for FE and CE were smaller for FCR as compared to those for NCR. Additionally, it is interesting to note that values of WARMSB for CE did not appear to increase as the CI ES increased. For MC-CR 0.25, values of WARMSB for FCR tended to be larger than values of WARMSB for NCR. It is plausible that this occurred because the effect sizes for NCR were more similar than the effect sizes for FCR to the CO ES, as was shown in Table 4-57. However, it is also important to note that the differences in WARMSB between FCR and NCR were small and may have been the result of random equating error.

Table 4-62 contains values of WARMSB for illustrating the influence of CI ES on WARMSB based on different criterion equating relationships. In the first block of data for NCR, for both MC-CR 0.00 and MC-CR 0.25, as CI ES increased WARMSB also tended to increase, with the exception of CE. Similar results can be seen for FCR in the second block of data.

Returning to Tables 4-60 and 4-61, the second block of data contains values for WASE. For both NCR and FCR, values of WASE were smallest for OS and largest for CE across all sampling conditions. Values of WASE for TS and OS were approximately 0.45 to 0.55. For CE, values of WASE were in the range of 0.66 to 0.72. Standardized values of WASE were approximately 0.02 to 0.03 standard deviation units across all sampling conditions and equating methods.

The third block of data in Tables 4-60 and 4-61 contains values for WARMSE. Similar patterns were seen for WARMSB and WARMSE. The total error in the third block of data in WARMSE appeared to result from both WARMSB and WARMSE. Tables 4-60 and 4-61 contain values of Difference in the fourth block of data. Overall, the results for Difference were similar to the results for WARMSB, and for both FCR and NCR, the results differed by 0.05 or less.

Classification Consistency

Table 4-63 contains classification consistency results. The first column of data (in bold) represents the criterion equating relationship based on the single group equating relationship. The value of 99.0 in the first column and third row (TS) indicates that 99.0 percent of examinees received the same grade based on the equipercentile and TS equating relationships. TS and OS resulted in the same percentage of examinees receiving the same grade.

The remaining columns of data contain the percentages of examinees receiving the same grades for the criterion equating relationship and the study condition equating relationships for a given equating method. For example, in the second column (CI 0.00 MC-CR 0.00) and first row (FE) of data, 96.6 is the percentage of examinees receiving the same grade for the single group equipercentile equating and the FE equating method. Similarly, 96.6 (second row of data) is the percentage of examinees receiving the same grade for the single group equipercentile equating and the CE equating method. Across all conditions, classification consistency percentages were approximately 90% or higher.

Summary

As noted in the results, it is important not to overemphasize differences in the equating relationships for the Chemistry pseudo-test forms. Although values of WARMSB indicated that bias increased slightly as the common-item effect size increased, the equating relationships were all very similar across sampling conditions and

common-item compositions. Standard errors were smallest for OS and largest for CE for all of the sampling conditions.

Table 4-1. Observed and Disattenuated MC and CR Correlations for English Language Operational Test Forms

Form	Correlation	Original Operational Form	MC-CR 0.00			MC-CR 0.25		
			CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
2004	MC and CR	0.571	0.617	0.605	0.584	0.599	0.608	0.601
	Disattenuated MC and CR	0.755	0.769	0.763	0.737	0.783	0.791	0.765
2007	MC and CR	0.578	0.570	0.559	0.583	0.600	0.551	0.561
	Disattenuated MC and CR	0.750	0.754	0.749	0.790	0.776	0.752	0.754

Table 4-2. Observed and Disattenuated MC and CR Correlations for English Language Pseudo-Test Forms

Form	Correlation	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
Old	MC and CR	0.495	0.487	0.495	0.484	0.479	0.521	0.533	0.471	0.435
	Disattenuated MC and CR	0.712	0.687	0.708	0.690	0.694	0.708	0.741	0.715	0.654
New	MC and CR	0.506	0.505	0.542	0.519	0.521	0.526	0.530	0.519	0.519
	Disattenuated MC and CR	0.749	0.736	0.765	0.752	0.751	0.792	0.773	0.736	0.767

Table 4-3. Eigenvalues for English Language Operational Test Forms

Principal Component	Eigenvalues	
	2004	2007
1	15.61	14.37
2	1.78	1.60
3	1.43	1.48
4	1.14	1.20
5	1.06	1.08
6	1.03	1.05
7	1.00	1.03
8	0.98	1.00
9	0.97	0.99
10	0.95	0.98

Table 4-4. Observed and Disattenuated MC and CR Correlations for Spanish Language Operational Test Forms

Form	Item Format	Original Operational Forms	MC-CR 0.00			MC-CR 0.25		
			CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
2004	MC and CR	0.808	0.790	0.770	0.827	0.753	0.819	0.813
	Disattenuated MC and CR	0.902	0.889	0.873	0.914	0.857	0.908	0.898
2006	MC and CR	0.714	0.746	0.768	0.752	0.774	0.769	0.766
	Disattenuated MC and CR	0.800	0.833	0.844	0.836	0.854	0.852	0.853

Table 4-5. Observed and Disattenuated MC and CR Correlations for Spanish Language Pseudo-Test Forms

Form	Item Format	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
Old	MC and CR	0.760	0.781	0.779	0.786	0.730	0.776	0.724	0.746	0.778
	Disattenuated MC and CR	0.880	0.893	0.893	0.896	0.864	0.889	0.849	0.865	0.898
New	MC and CR	0.757	0.747	0.762	0.791	0.752	0.740	0.760	0.766	0.804
	Disattenuated MC and CR	0.875	0.860	0.874	0.899	0.869	0.866	0.877	0.884	0.914

Table 4-6. Eigenvalues for Spanish Language Operational Test Forms

Principal Component	Eigenvalues	
	2004	2006
1	31.38	25.71
2	8.33	7.07
3	3.29	3.22
4	2.40	2.13
5	1.62	1.42
6	1.28	1.33
7	1.23	1.23
8	1.18	1.15
9	1.10	1.11
10	1.09	1.09
11	1.04	1.03
12	1.02	1.01
13	1.02	0.99
14	0.99	0.97
15	0.98	0.97

Table 4-7. Observed and Disattenuated MC and CR Correlations for Chemistry Operational Test Forms

Form	Item Format	Original Operational Forms	MC-CR 0.00			MC-CR 0.25		
			CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
2005	MC and CR	0.866	0.859	0.858	0.862	0.862	0.884	0.878
	Disattenuated MC and CR	0.946	0.939	0.941	0.947	0.944	0.956	0.959
2007	MC and CR	0.884	0.876	0.870	0.867	0.880	0.908	0.888
	Disattenuated MC and CR	0.953	0.945	0.946	0.943	0.955	0.969	0.958

Table 4-8. Observed and Disattenuated MC and CR Correlations for Chemistry Pseudo-Test Forms

Form	Item Format	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
Old	MC and CR	0.860	0.864	0.855	0.860	0.862	0.873	0.872	0.882	0.883
	Disattenuated MC and CR	0.961	0.960	0.956	0.952	0.958	0.971	0.968	0.975	0.976
New	MC and CR	0.856	0.866	0.848	0.860	0.860	0.864	0.854	0.864	0.843
	Disattenuated MC and CR	0.954	0.959	0.950	0.959	0.956	0.955	0.953	0.954	0.943

Table 4-9. Eigenvalues for Chemistry Operational Test Forms

Principal Component	Eigenvalues	
	2005	2007
1	28.947	27.464
2	2.368	2.635
3	1.681	1.469
4	1.346	1.282
5	1.144	1.126
6	1.029	1.095
7	1.006	1.016
8	0.937	0.984
9	0.915	0.947
10	0.896	0.920

Table 4-10. Descriptive Statistics for the English Language 2004 Operational Test Form

Item Format	Descriptive Statistics	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
	N	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	34.946	36.480	36.437	35.169	36.187	37.666
	SD	9.503	9.344	9.560	9.973	9.330	9.104
	Skewness	-0.400	-0.540	-0.557	-0.449	-0.506	-0.686
	Kurtosis	2.384	2.539	2.590	2.326	2.550	2.836
	α	0.896	0.897	0.902	0.907	0.895	0.895
CR	Mean	13.962	14.651	14.588	13.732	14.224	14.704
	SD	3.908	3.880	3.794	4.034	3.714	3.910
	Skewness	-0.072	-0.199	-0.150	-0.205	-0.075	-0.263
	Kurtosis	2.892	3.071	2.969	2.969	3.005	3.269
	α	0.637	0.620	0.604	0.659	0.598	0.618
CI	Mean	13.277	13.680	13.603	13.241	13.538	13.959
	SD	3.256	3.202	3.269	3.450	3.209	3.082
	Skewness	-0.856	-0.951	-0.987	-0.923	-0.971	-1.095
	Kurtosis	3.385	3.558	3.649	3.358	3.806	4.062
CO	Mean	48.908	51.131	51.025	48.901	50.411	52.370
	SD	12.161	11.953	12.169	12.807	11.790	11.751
	Skewness	-0.343	-0.500	-0.514	-0.430	-0.414	-0.631
	Kurtosis	2.482	2.735	2.727	2.515	2.644	3.009
	α	0.883	0.880	0.887	0.894	0.880	0.877

Table 4-11. Descriptive Statistics for the English Language 2007 Operational Test Form

Item Format	Descriptive Statistics	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
	N	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	33.251	32.448	30.072	33.271	32.316	31.528
	SD	9.238	9.277	9.312	8.840	9.023	9.080
	Skewness	-0.381	-0.318	-0.144	-0.325	-0.270	-0.190
	Kurtosis	2.506	2.501	2.350	2.448	2.451	2.379
	α	0.893	0.892	0.888	0.882	0.885	0.884
Wt. MC	Mean	33.890	33.072	30.650	33.910	32.938	32.135
	SD	9.416	9.456	9.491	9.010	9.196	9.254
	Skewness	-0.381	-0.318	-0.144	-0.325	-0.270	-0.190
	Kurtosis	2.506	2.501	2.350	2.448	2.451	2.379
CR	Mean	13.565	13.206	12.299	14.159	13.819	13.348
	SD	4.124	4.089	4.003	3.970	3.980	4.045
	Skewness	-0.125	-0.197	-0.173	-0.200	-0.143	-0.176
	Kurtosis	3.268	3.102	2.995	2.949	2.968	3.025
	α	0.720	0.705	0.708	0.663	0.667	0.698
CI	Mean	13.233	13.000	12.178	13.196	12.884	12.663
	SD	3.267	3.459	3.619	3.226	3.372	3.460
	Skewness	-0.867	-0.866	-0.658	-0.793	-0.755	-0.724
	Kurtosis	3.411	3.325	2.812	3.217	3.105	3.005
CO	Mean	46.816	45.654	42.372	47.429	46.135	44.877
	SD	12.221	12.194	12.095	11.660	11.871	11.955
	Skewness	-0.340	-0.329	-0.196	-0.302	-0.252	-0.192
	Kurtosis	2.686	2.703	2.495	2.567	2.583	2.560
	α	0.886	0.884	0.881	0.874	0.877	0.878
Wt. CO	Mean	47.456	46.278	42.950	48.069	46.757	45.483
	SD	12.392	12.366	12.267	11.824	12.038	12.123
	Skewness	-0.341	-0.329	-0.196	-0.303	-0.253	-0.193
	Kurtosis	2.683	2.700	2.492	2.565	2.581	2.556

Table 4-12. Effect Sizes for English Language 2004-2007 Operational Test Forms

Score Type	MC-CR 0.00			MC-CR 0.25		
	CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
Wt. CO	-0.118	-0.399	-0.661	-0.068	-0.307	-0.577
CO	-0.172	-0.454	-0.713	-0.120	-0.361	-0.632
CI	-0.014	-0.204	-0.413	-0.013	-0.199	-0.396
Wt. MC	-0.112	-0.363	-0.608	-0.133	-0.351	-0.603
MC	-0.181	-0.433	-0.675	-0.202	-0.422	-0.675
CR	-0.099	-0.363	-0.587	0.107	-0.105	-0.341
Wt. MC-CR	-0.013	0.000	-0.021	-0.239	-0.246	-0.262
MC-CR	-0.082	-0.071	-0.088	-0.308	-0.317	-0.335

Table 4-13. Equated Moments for English Language 2004-2007 Operational Test Forms

Method	Statistic	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
FE	Mean	48.784	49.032	46.550	48.647	48.334	48.215
	SD	12.197	12.521	12.765	12.308	12.137	12.669
	Skew.	-0.351	-0.410	-0.227	-0.331	-0.306	-0.377
	Kurt.	2.508	2.586	2.324	2.441	2.497	2.576
CE	Mean	48.731	48.622	45.769	48.545	47.993	47.485
	SD	12.156	12.482	12.664	12.139	12.031	12.540
	Skew.	-0.363	-0.433	-0.232	-0.276	-0.265	-0.323
	Kurt.	2.485	2.673	2.387	2.413	2.446	2.546
TS	Mean	48.620	47.976	44.709	48.037	47.219	46.391
	SD	12.150	12.479	12.339	11.754	11.901	12.353
	Skew.	-0.341	-0.376	-0.221	-0.300	-0.260	-0.260
	Kurt.	2.570	2.588	2.395	2.494	2.502	2.492
OS	Mean	48.631	48.000	44.721	48.043	47.225	46.414
	SD	12.223	12.527	12.381	11.793	11.947	12.390
	Skew.	-0.335	-0.365	-0.212	-0.296	-0.251	-0.247
	Kurt.	2.582	2.603	2.413	2.505	2.509	2.512

Table 4-14. Summary Statistics for English Language 2004-2007 Operational Test Forms

Statistic	Equating Method	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
WARMSB	FE	--	1.489	2.324	1.004	0.462	1.702
	CE	--	1.189	1.612	1.000	0.467	1.099
	TS	--	0.679	0.658	1.242	0.724	0.581
	OS	--	0.672	0.653	1.245	0.724	0.533
WASE	FE	0.408	0.427	0.445	0.422	0.417	0.461
	CE	0.467	0.522	0.519	0.462	0.458	0.524
	TS	0.358	0.380	0.367	0.366	0.360	0.385
	OS	0.341	0.361	0.348	0.348	0.339	0.366
WARMSE	FE	--	1.549	2.366	1.089	0.622	1.763
	CE	--	1.298	1.694	1.102	0.654	1.218
	TS	--	0.778	0.753	1.295	0.808	0.697
	OS	--	0.763	0.740	1.293	0.800	0.647
Difference	FE	--	1.478	2.317	0.994	0.447	1.697
	CE	--	1.181	1.613	0.998	0.455	1.085
	TS	--	0.662	0.642	1.203	0.722	0.603
	OS	--	0.653	0.634	1.206	0.723	0.552
Standardized WARMSB	FE	--	0.125	0.191	0.078	0.039	0.145
	CE	--	0.099	0.132	0.078	0.040	0.094
	TS	--	0.057	0.054	0.097	0.061	0.049
	OS	--	0.056	0.054	0.097	0.061	0.045
Standardized WASE	FE	0.034	0.036	0.037	0.033	0.035	0.039
	CE	0.038	0.044	0.043	0.036	0.039	0.045
	TS	0.029	0.032	0.030	0.029	0.030	0.033
	OS	0.028	0.030	0.029	0.027	0.029	0.031

Table 4-15. WARMSB for English Language 2004-2007 Operational Test Forms to Illustrate Effects of CI ES

Statistic	Equating Method	MC-CR 0.00			MC-CR 0.25		
		0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)
WARMSB	FE	--	1.489	2.324	--	1.319	2.664
	CE	--	1.189	1.612	--	1.115	1.982
	TS	--	0.679	0.658	--	0.560	1.150
	OS	--	0.672	0.653	--	0.558	1.160

Table 4-16. English Language 2004-2007 Operational Test Forms Percentages of Classification Consistency

Equating Method	MC-CR 0.00			MC-CR 0.25		
	CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
FE	--	87.5	75.3	95.8	96.3	81.2
CE	**100.0**	86.8	82.0	94.2	99.3	91.0
TS	**100.0**	96.3	96.3	90.7	90.7	99.3
OS	**100.0**	93.7	96.3	89.9	90.7	99.3

Table 4-17. Descriptive Statistics for the Spanish Language 2004 Operational Test Form

Item Format	Descriptive Statistics	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
	N	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	61.453	61.445	53.943	62.079	55.470	50.285
	SD	14.494	14.130	16.123	13.633	15.875	15.807
	Skewness	-0.533	-0.572	-0.200	-0.599	-0.291	0.043
	Kurtosis	2.844	2.874	2.219	3.080	2.317	2.195
	α	0.928	0.924	0.936	0.920	0.935	0.931
CR	Mean	38.628	38.329	31.815	39.582	34.406	29.800
	SD	9.917	9.738	11.189	9.092	11.450	11.199
	Skewness	-0.889	-0.861	-0.224	-1.175	-0.460	0.037
	Kurtosis	3.380	3.317	2.283	4.282	2.269	2.265
	α	0.851	0.841	0.873	0.839	0.871	0.880
CI	Mean	16.803	16.885	15.304	16.858	15.248	14.355
	SD	4.736	4.675	5.190	4.556	4.931	5.043
	Skewness	-0.285	-0.356	-0.114	-0.273	-0.164	0.044
	Kurtosis	2.505	2.602	2.269	2.566	2.350	2.342
CO	Mean	100.082	99.774	85.757	101.661	89.876	80.085
	SD	23.144	22.503	26.141	21.337	26.096	25.749
	Skewness	-0.726	-0.729	-0.229	-0.905	-0.397	0.036
	Kurtosis	3.171	3.142	2.247	3.718	2.334	2.225
	α	0.943	0.939	0.951	0.936	0.949	0.949

Table 4-18. Descriptive Statistics for the Spanish Language 2006 Operational Test Form

Item Format	Descriptive Statistics	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
	N	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	50.976	53.959	52.506	50.925	49.103	49.539
	SD	12.192	12.537	12.052	12.680	12.895	12.501
	Skewness	-0.416	-0.665	-0.544	-0.397	-0.254	-0.352
	Kurtosis	2.548	2.805	2.736	2.462	2.279	2.476
	α	0.917	0.928	0.918	0.923	0.923	0.919
Wt. MC	Mean	61.171	64.751	63.008	61.110	58.924	59.447
	SD	14.631	15.044	14.462	15.216	15.474	15.002
	Skewness	-0.416	-0.665	-0.544	-0.397	-0.254	-0.352
	Kurtosis	2.548	2.805	2.736	2.462	2.279	2.476
CR	Mean	38.552	40.766	38.631	36.038	34.214	33.717
	SD	11.282	11.647	11.477	11.994	11.536	11.012
	Skewness	-0.762	-0.877	-0.633	-0.312	-0.248	-0.188
	Kurtosis	2.841	3.006	2.682	2.362	2.371	2.467
	α	0.874	0.892	0.882	0.889	0.882	0.878
CI	Mean	16.764	17.953	17.458	16.975	16.319	16.462
	SD	5.038	5.198	5.003	5.229	5.281	5.160
	Skewness	-0.274	-0.519	-0.450	-0.309	-0.204	-0.291
	Kurtosis	2.306	2.511	2.545	2.297	2.226	2.422
CO	Mean	89.527	94.725	91.137	86.963	83.317	83.257
	SD	21.933	22.740	22.023	23.237	22.977	22.103
	Skewness	-0.659	-0.830	-0.640	-0.403	-0.286	-0.328
	Kurtosis	2.871	3.105	2.855	2.490	2.386	2.533
	α	0.938	0.947	0.940	0.944	0.943	0.941
Wt. CO	Mean	99.723	105.517	101.638	97.148	93.138	93.165
	SD	24.238	25.122	24.301	25.647	25.433	24.485
	Skewness	-0.639	-0.818	-0.635	-0.408	-0.285	-0.336
	Kurtosis	2.850	3.091	2.856	2.495	2.379	2.532

Table 4-19. Effect Sizes for Spanish Language 2004-2006 Operational Test Forms

Score Type	MC-CR 0.00			MC-CR 0.25		
	CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
Wt. CO	-0.015	0.241	0.629	-0.191	0.127	0.521
CO	-0.468	-0.223	0.223	-0.659	-0.267	0.132
CI	-0.008	0.216	0.423	0.024	0.210	0.413
Wt. MC	-0.019	0.227	0.592	-0.067	0.220	0.595
MC	-0.783	-0.561	-0.101	-0.848	-0.440	-0.052
CR	-0.007	0.227	0.602	-0.333	-0.017	0.353
Wt. MC-CR	-0.012	-0.001	-0.009	0.266	0.237	0.242
MC-CR	-0.775	-0.788	-0.702	-0.514	-0.424	-0.405

Table 4-20. Equated Moments for Spanish Language 2004-2006 Operational Test Forms

Method	Statistic	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
FE	Mean	99.774	103.644	95.072	101.587	94.444	89.446
	SD	24.304	23.873	24.908	23.786	26.956	25.991
	Skew.	-0.699	-0.848	-0.518	-0.872	-0.473	-0.257
	Kurt.	3.009	3.271	2.646	3.403	2.428	2.311
CE	Mean	99.615	104.226	96.467	101.376	95.086	90.801
	SD	24.649	24.766	24.786	24.833	27.319	26.188
	Skew.	-0.662	-0.869	-0.577	-0.845	-0.452	-0.318
	Kurt.	2.856	3.245	2.741	3.240	2.377	2.342
TS	Mean	98.416	104.263	99.102	98.920	94.617	93.127
	SD	24.945	26.209	24.416	28.256	28.427	26.622
	Skew.	-0.437	-0.617	-0.447	-0.699	-0.539	-0.286
	Kurt.	2.402	2.491	2.572	2.667	2.346	2.260
OS	Mean	98.368	104.238	99.061	98.982	94.692	93.133
	SD	24.926	26.205	24.385	28.147	28.305	26.550
	Skew.	-0.464	-0.638	-0.458	-0.707	-0.535	-0.292
	Kurt.	2.460	2.551	2.586	2.714	2.361	2.283

Table 4-21. Summary Statistics for Spanish Language 2004-2006 Operational Test Forms

Statistic	Equating Method	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
WARMSB	FE	--	2.365	6.639	5.934	2.735	4.032
	CE	--	1.601	5.069	5.904	3.444	2.568
	TS	--	0.435	2.113	5.805	6.033	2.790
	OS	--	0.397	2.064	5.706	5.837	2.642
WASE	FE	0.870	0.847	0.836	0.899	0.865	0.827
	CE	1.022	1.011	0.969	1.058	1.009	0.985
	TS	0.912	0.921	0.750	1.329	0.839	0.754
	OS	0.892	0.907	0.737	1.296	0.831	0.744
WARMSE	FE	--	2.512	6.691	6.002	2.868	4.116
	CE	--	1.893	5.161	5.998	3.589	2.750
	TS	--	1.018	2.242	5.955	6.091	2.890
	OS	--	0.990	2.191	5.852	5.895	2.745
Difference	FE	--	2.379	6.658	5.922	2.683	4.057
	CE	--	1.574	5.046	5.925	3.430	2.534
	TS	--	0.507	1.639	6.048	6.022	3.150
	OS	--	0.463	1.562	5.911	5.825	3.028
Standardized WARMSB	FE	--	0.111	0.254	0.256	0.122	0.154
	CE	--	0.075	0.194	0.255	0.153	0.098
	TS	--	0.020	0.081	0.251	0.268	0.107
	OS	--	0.019	0.079	0.247	0.259	0.101
Standardized WASE	FE	0.035	0.040	0.032	0.039	0.038	0.032
	CE	0.041	0.047	0.037	0.046	0.045	0.038
	TS	0.037	0.043	0.029	0.057	0.037	0.029
	OS	0.036	0.043	0.028	0.056	0.037	0.028

Table 4-22. WARMSB for Spanish Language 2004-2006 Operational Test Forms to Illustrate Effects of CI ES

Statistic	Equating Method	MC-CR 0.00			MC-CR 0.25		
		0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)
WARMSB	FE	--	2.365	6.639	--	5.917	11.359
	CE	--	1.601	5.069	--	4.648	9.228
	TS	--	0.435	2.113	--	1.186	4.632
	OS	--	0.397	2.064	--	1.090	4.590

Table 4-23. Spanish Language 2004-2006 Operational Test Forms Percentages of Classification Consistency

Equating Method	MC-CR 0.00			MC-CR 0.25		
	CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
FE	--	88.7	65.6	73.0	85.9	88.3
CE	**97.1**	91.8	71.5	74.4	82.0	95.5
TS	**91.5**	97.6	91.5	73.9	72.2	84.6
OS	**91.5**	97.6	91.5	72.0	74.3	84.6

Table 4-24. Descriptive Statistics for the Chemistry 2005 Operational Test Form

Item Format	Descriptive Statistics	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
	N	1,500	1,500	1,500	1,500	1,500	1,500
MC	Mean	46.043	45.475	41.939	39.817	38.073	38.015
	SD	15.873	15.862	15.853	15.954	16.868	16.281
	Skewness	-0.370	-0.260	-0.073	0.114	0.162	0.190
	Kurtosis	2.042	1.958	1.862	1.895	1.777	1.875
	α	0.949	0.948	0.946	0.946	0.952	0.948
CR	Mean	30.520	30.380	26.998	25.377	24.015	24.413
	SD	13.913	13.628	13.670	14.035	14.913	14.197
	Skewness	-0.357	-0.270	-0.059	0.064	0.075	0.049
	Kurtosis	2.200	2.133	2.058	2.031	1.882	1.995
	α	0.881	0.878	0.876	0.880	0.898	0.885
CI	Mean	15.579	15.361	14.064	13.292	12.709	12.713
	SD	5.489	5.621	5.575	5.713	5.887	5.742
	Skewness	-0.400	-0.283	-0.096	0.016	0.129	0.133
	Kurtosis	2.131	2.075	1.927	1.951	1.833	1.924
CO	Mean	76.563	75.855	68.937	65.195	62.088	62.428
	SD	28.721	28.430	28.493	28.940	30.852	29.539
	Skewness	-0.392	-0.274	-0.072	0.090	0.117	0.122
	Kurtosis	2.148	2.040	1.944	1.963	1.817	1.923
	α	0.930	0.930	0.929	0.929	0.935	0.931

Table 4-25. Descriptive Statistics for the Chemistry 2007 Operational Test Form

Item Format	Descriptive Statistics	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
	N	1,500	1,500	1,500	1,500	1,500	1,500
MC	Mean	44.205	46.081	45.958	37.571	39.061	42.157
	SD	15.363	14.457	14.608	15.163	16.326	15.509
	Skewness	-0.335	-0.419	-0.336	0.176	0.116	-0.136
	Kurtosis	2.073	2.273	2.174	1.925	1.838	1.957
	α	0.943	0.937	0.938	0.939	0.948	0.943
CR	Mean	28.844	31.007	30.739	21.908	23.413	26.681
	SD	15.630	14.887	14.981	14.725	16.327	15.435
	Skewness	-0.222	-0.272	-0.259	0.252	0.168	-0.043
	Kurtosis	1.984	2.119	2.117	2.077	1.844	1.961
	α	0.911	0.901	0.900	0.903	0.926	0.911
Wt. CR	Mean	27.914	30.007	29.748	21.201	22.658	25.820
	SD	15.126	14.407	14.498	14.250	15.800	14.937
	Skewness	-0.222	-0.272	-0.259	0.252	0.168	-0.043
	Kurtosis	1.984	2.119	2.117	2.077	1.844	1.961
CI	Mean	15.647	16.357	16.329	13.407	13.801	14.999
	SD	5.589	5.240	5.158	5.637	5.969	5.590
	Skewness	-0.466	-0.566	-0.528	0.025	-0.052	-0.269
	Kurtosis	2.216	2.450	2.425	1.949	1.877	2.010
CO	Mean	73.049	77.088	76.697	59.479	62.474	68.838
	SD	30.018	28.371	28.589	28.975	31.891	30.067
	Skewness	-0.290	-0.354	-0.321	0.215	0.137	-0.098
	Kurtosis	2.047	2.182	2.189	1.993	1.820	1.949
	α	0.927	0.923	0.923	0.928	0.935	0.930
Wt. CO	Mean	72.119	76.088	75.706	58.772	61.719	67.977
	SD	29.530	27.906	28.122	28.515	31.377	29.583
	Skewness	-0.291	-0.356	-0.322	0.214	0.137	-0.099
	Kurtosis	2.048	2.183	2.190	1.991	1.820	1.948

Table 4-26. Effect Sizes for Chemistry 2005-2007 Operational Test Forms

Score Type	MC-CR 0.00			MC-CR 0.10		
	CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
Wt. CO	-0.153	0.008	0.239	-0.224	-0.012	0.188
CO	-0.120	0.043	0.272	-0.197	0.012	0.215
CI	0.012	0.183	0.422	0.020	0.184	0.404
MC	-0.118	0.040	0.264	-0.144	0.060	0.261
Wt. CR	-0.179	-0.027	0.195	-0.295	-0.088	0.097
CR	-0.113	0.044	0.261	-0.241	-0.039	0.153
Wt. MC-CR	0.062	0.067	0.069	0.151	0.148	0.164
MC-CR	-0.004	-0.004	0.003	0.097	0.098	0.108

Table 4-27. Equated Moments for Chemistry 2005-2007 Operational Test Forms

Method	Statistic	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
FE	Mean	76.991	80.638	79.495	65.695	67.410	73.329
	SD	28.941	26.702	26.651	28.728	31.097	28.941
	Skew.	-0.426	-0.503	-0.473	0.080	-0.065	-0.268
	Kurt.	2.173	2.351	2.379	1.959	1.837	2.052
CE	Mean	76.940	80.943	80.525	65.742	67.816	74.202
	SD	29.166	26.492	26.445	28.577	31.160	28.726
	Skew.	-0.445	-0.554	-0.493	0.093	-0.064	-0.279
	Kurt.	2.196	2.410	2.418	1.961	1.857	2.062
TS	Mean	77.061	81.407	81.233	65.477	67.919	74.855
	SD	28.857	25.388	25.799	28.838	31.503	28.338
	Skew.	-0.520	-0.491	-0.563	-0.067	-0.141	-0.381
	Kurt.	2.256	2.347	2.481	1.901	1.796	2.119
OS	Mean	77.131	81.450	81.297	65.600	68.023	74.968
	SD	28.893	25.505	25.885	28.820	31.489	28.335
	Skew.	-0.505	-0.482	-0.547	-0.043	-0.123	-0.357
	Kurt.	2.235	2.338	2.460	1.885	1.780	2.084

Table 4-28. Summary Statistics for Chemistry 2005-2007 Operational Test Forms

Statistic	Equating Method	MC-CR 0.00			MC-CR 0.25		
		CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
WARMSB	FE	--	0.790	1.313	2.230	1.366	0.889
	CE	--	1.128	1.214	2.658	1.939	1.703
	TS	--	1.635	1.269	1.529	1.630	1.957
	OS	--	1.569	1.254	1.555	1.617	1.985
WASE	FE	0.904	0.866	0.843	0.877	0.880	0.880
	CE	1.029	0.983	0.970	1.012	1.051	1.009
	TS	0.726	0.821	0.616	0.619	0.629	1.002
	OS	0.715	0.814	0.607	0.608	0.619	0.985
WARMSE	FE	--	1.172	1.560	2.396	1.625	1.251
	CE	--	1.496	1.554	2.844	2.206	1.979
	TS	--	1.830	1.410	1.649	1.747	2.199
	OS	--	1.767	1.394	1.669	1.732	2.216
Difference	FE	--	0.745	1.311	2.193	1.333	0.846
	CE	--	1.096	1.183	2.621	1.878	1.692
	TS	--	1.638	1.300	1.548	1.609	2.018
	OS	--	1.570	1.287	1.576	1.598	2.045
Standardized WARMSB	FE	--	0.028	0.046	0.077	0.044	0.030
	CE	--	0.040	0.043	0.092	0.063	0.058
	TS	--	0.058	0.045	0.053	0.053	0.066
	OS	--	0.055	0.044	0.054	0.052	0.067
Standardized WASE	FE	0.031	0.030	0.030	0.030	0.029	0.030
	CE	0.036	0.035	0.034	0.035	0.034	0.034
	TS	0.025	0.029	0.022	0.021	0.020	0.034
	OS	0.025	0.029	0.021	0.021	0.020	0.033

Table 4-29. WARMSB for Chemistry 2005-2007 Operational Test Forms to Illustrate Effects of CI ES

Statistic	Equating Method	MC-CR 0.00			MC-CR 0.25		
		0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)
WARMSB	FE	--	0.790	1.313	--	1.673	1.978
	CE	--	1.128	1.214	--	1.632	1.511
	TS	--	1.635	1.269	--	0.266	0.729
	OS	--	1.569	1.254	--	0.318	0.735

Table 4-30. Chemistry 2005-2007 Operational Test Form Percentages of Classification Consistency

Equating Method	MC-CR 0.00			MC-CR 0.25		
	CI 0.00	CI 0.20	CI 0.40	CI 0.00	CI 0.20	CI 0.40
FE	--	96.8	96.6	92.1	92.9	96.5
CE	**97.3**	98.7	96.8	90.5	95.4	94.4
TS	**97.9**	96.1	96.3	93.0	92.0	91.6
OS	**97.9**	97.3	96.9	93.0	92.0	91.6

Table 4-31. Descriptive Statistics for the Old English Language Pseudo-Test Form

Score	Statistic	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
	N	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	26.266	27.661	26.329	25.909	24.298	25.918	26.633	26.455
	SD	6.242	5.409	6.035	5.998	6.243	6.226	5.622	5.821
	Skew	-0.832	-0.922	-0.735	-0.669	-0.469	-0.691	-0.736	-0.716
	Kurt	3.184	3.609	3.012	2.859	2.584	2.921	3.183	2.937
	α	0.854	0.820	0.844	0.836	0.838	0.851	0.820	0.832
CR	Mean	9.734	10.184	9.789	9.706	8.661	9.435	9.723	9.680
	SD	2.878	2.686	2.812	2.720	2.869	2.968	2.789	2.769
	Skew	-0.199	-0.270	-0.093	-0.099	-0.179	-0.198	-0.256	-0.153
	Kurt	3.059	2.996	2.840	3.070	3.022	2.994	3.150	2.972
	α	0.592	0.493	0.555	0.544	0.587	0.598	0.545	0.529
NCR CI	Mean	12.903	13.558	12.992	12.798	12.143	12.824	13.121	13.033
	SD	3.074	2.701	3.005	3.022	3.239	3.115	2.799	2.941
	Skew	-1.053	-1.185	-0.959	-0.959	-0.778	-0.980	-1.039	-0.982
	Kurt	3.900	4.602	3.572	3.659	3.094	3.602	4.014	3.689
	CO Corr	0.857	0.825	0.847	0.839	0.855	0.855	0.829	0.832
FCR CI	Mean	10.757	11.297	10.723	10.598	9.799	10.556	10.890	10.746
	SD	2.835	2.574	2.735	2.682	2.908	2.895	2.633	2.711
	Skew	-0.503	-0.488	-0.382	-0.354	-0.357	-0.515	-0.459	-0.517
	Kurt	3.005	3.131	2.969	2.961	2.792	3.190	3.141	3.263
	CO Corr	0.862	0.817	0.846	0.846	0.863	0.863	0.831	0.837
Full CI	Mean	19.381	20.335	19.473	19.234	18.039	19.172	19.654	19.483
	SD	4.355	3.751	4.173	4.103	4.499	4.392	3.908	4.029
	Skew	-0.933	-0.924	-0.803	-0.783	-0.687	-0.855	-0.886	-0.825
	Kurt	3.754	4.132	3.599	3.622	3.152	3.594	4.036	3.662
	CO Corr	0.926	0.898	0.915	0.912	0.923	0.925	0.908	0.906
CO	Mean	36.000	37.844	36.118	35.616	32.958	35.353	36.356	36.135
	SD	8.138	6.984	7.815	7.615	8.062	8.205	7.355	7.550
	Skew	-0.755	-0.783	-0.604	-0.563	-0.432	-0.608	-0.656	-0.614
	Kurt	3.168	3.586	3.035	2.952	2.745	3.005	3.314	2.949
	α	0.834	0.787	0.821	0.813	0.821	0.831	0.798	0.809

Table 4-32. Descriptive Statistics for the New English Language Pseudo-Test Form

Score	Statistic	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
	N	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	25.893	26.096	23.346	21.672	23.659	24.124	23.657	21.926
	SD	6.901	6.264	7.320	7.257	7.183	7.306	6.837	7.353
	Skew	-0.801	-0.833	-0.454	-0.226	-0.509	-0.593	-0.541	-0.254
	Kurt	2.930	3.355	2.367	2.173	2.455	2.541	2.548	2.191
	α	0.881	0.854	0.882	0.874	0.878	0.885	0.864	0.878
CR	Mean	9.874	9.788	8.855	8.241	9.148	9.325	9.118	8.557
	SD	2.933	2.661	2.918	2.827	2.820	2.875	2.658	2.826
	Skew	-0.193	-0.028	-0.086	-0.090	-0.094	-0.066	-0.164	-0.122
	Kurt	2.965	3.029	2.916	2.940	2.811	2.956	2.955	2.908
	α	0.587	0.475	0.567	0.544	0.515	0.555	0.492	0.517
NCR CI	Mean	12.859	13.043	11.840	11.175	11.967	12.126	12.017	11.199
	SD	3.126	2.918	3.422	3.518	3.369	3.384	3.229	3.551
	Skew	-1.003	-1.037	-0.687	-0.470	-0.723	-0.832	-0.790	-0.511
	Kurt	3.721	4.040	2.839	2.454	2.911	3.145	3.121	2.569
	CO Corr	0.866	0.853	0.882	0.881	0.873	0.888	0.866	0.880
FCR CI	Mean	10.670	10.705	9.642	8.935	9.848	10.070	9.861	9.168
	SD	2.824	2.674	3.002	2.953	2.890	2.976	2.776	2.996
	Skew	-0.505	-0.447	-0.276	-0.180	-0.320	-0.359	-0.280	-0.196
	Kurt	2.994	3.056	2.651	2.531	2.792	2.740	2.675	2.562
	CO Corr	0.861	0.840	0.872	0.858	0.853	0.866	0.845	0.861
Full CI	Mean	19.291	19.451	17.708	16.646	18.007	18.259	18.015	16.857
	SD	4.377	3.983	4.715	4.768	4.569	4.694	4.382	4.838
	Skew	-0.898	-0.857	-0.610	-0.473	-0.634	-0.748	-0.680	-0.465
	Kurt	3.589	3.942	2.921	2.608	3.024	3.173	3.139	2.673
	CO Corr	0.927	0.916	0.934	0.934	0.926	0.936	0.925	0.927
CO	Mean	35.767	35.884	32.202	29.913	32.807	33.449	32.775	30.483
	SD	8.802	7.841	9.284	9.073	8.940	9.157	8.466	9.180
	Skew	-0.732	-0.677	-0.388	-0.225	-0.420	-0.510	-0.459	-0.226
	Kurt	2.960	3.325	2.466	2.313	2.533	2.603	2.601	2.330
	α	0.856	0.825	0.864	0.856	0.854	0.863	0.843	0.858

Table 4-33. Effect Sizes for English Language Pseudo-Test Forms

ES	Single Group	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
N	15,820	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
CO	-0.011	-0.027	-0.264	-0.457	-0.681	-0.018	-0.219	-0.452	-0.673
NCR CI	0.000	-0.014	-0.183	-0.358	-0.495	-0.053	-0.215	-0.365	-0.563
FCR CI	0.000	-0.031	-0.226	-0.377	-0.590	0.017	-0.165	-0.380	-0.552
Full CI	0.000	-0.021	-0.229	-0.396	-0.582	-0.007	-0.201	-0.395	-0.590
MC	-0.045	-0.057	-0.267	-0.445	-0.637	-0.095	-0.264	-0.476	-0.683
CR	0.069	0.048	-0.148	-0.326	-0.528	0.171	-0.038	-0.222	-0.401
MC-CR	-0.115	-0.105	-0.119	-0.119	-0.108	-0.266	-0.226	-0.253	-0.282

Table 4-34. Equated Moments for English Language Pseudo-Test Forms (NCR)

Method	Statistic	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
FE	Mean	35.59	35.89	36.77	33.58	32.17	32.60	33.79	33.97	32.25
	SD	7.85	8.26	7.34	8.57	8.46	8.27	8.66	8.14	8.52
	Skew	-0.63	-0.75	-0.67	-0.45	-0.25	-0.40	-0.50	-0.51	-0.24
	Kurt	3.09	3.18	3.25	2.72	2.47	2.63	2.73	2.85	2.34
CE	Mean	35.58	35.89	36.50	33.17	31.65	32.56	33.52	33.49	31.60
	SD	7.84	8.31	7.41	8.58	8.44	8.32	8.68	8.18	8.55
	Skew	-0.63	-0.73	-0.71	-0.43	-0.21	-0.39	-0.52	-0.52	-0.19
	Kurt	3.07	3.14	3.40	2.70	2.43	2.60	2.79	2.91	2.34
TS	Mean	35.59	35.79	36.34	32.80	31.00	32.45	33.35	33.06	30.88
	SD	7.80	8.35	7.42	8.82	8.60	8.32	8.89	8.15	8.87
	Skew	-0.65	-0.76	-0.71	-0.43	-0.27	-0.38	-0.48	-0.48	-0.29
	Kurt	3.20	3.21	3.57	2.63	2.43	2.69	2.71	2.77	2.49
OS	Mean	35.57	35.78	36.33	32.80	31.01	32.43	33.33	33.06	30.89
	SD	7.83	8.33	7.46	8.84	8.62	8.34	8.87	8.16	8.86
	Skew	-0.66	-0.75	-0.71	-0.42	-0.27	-0.40	-0.50	-0.47	-0.29
	Kurt	3.17	3.13	3.51	2.60	2.41	2.66	2.69	2.73	2.46

Table 4-35. Equated Moments for English Language Pseudo-Test Forms (FCR)

Method	Statistic	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
FE	Mean	35.59	35.80	36.51	33.45	31.59	33.09	34.13	33.95	32.33
	SD	7.85	8.17	7.40	8.67	8.55	7.99	8.52	7.94	8.42
	Skew	-0.63	-0.76	-0.74	-0.45	-0.29	-0.41	-0.48	-0.47	-0.25
	Kurt	3.09	3.24	3.42	2.65	2.47	2.70	2.69	2.76	2.30
CE	Mean	35.58	35.75	36.24	32.98	30.84	33.09	33.96	33.49	31.70
	SD	7.84	8.17	7.41	8.78	8.64	8.00	8.50	7.90	8.42
	Skew	-0.63	-0.77	-0.73	-0.41	-0.26	-0.40	-0.44	-0.41	-0.17
	Kurt	3.07	3.25	3.43	2.57	2.43	2.74	2.59	2.63	2.22
TS	Mean	35.59	35.68	36.29	32.83	30.70	32.42	33.47	32.81	30.81
	SD	7.80	8.40	7.50	8.95	8.54	8.06	8.77	7.99	8.52
	Skew	-0.65	-0.75	-0.71	-0.43	-0.25	-0.38	-0.50	-0.45	-0.30
	Kurt	3.20	3.17	3.53	2.60	2.43	2.75	2.75	2.79	2.56
OS	Mean	35.57	35.66	36.29	32.84	30.70	32.40	33.45	32.79	30.82
	SD	7.83	8.38	7.53	8.96	8.57	8.10	8.76	8.02	8.54
	Skew	-0.66	-0.74	-0.70	-0.43	-0.25	-0.40	-0.51	-0.46	-0.29
	Kurt	3.17	3.11	3.49	2.58	2.41	2.70	2.72	2.74	2.51

Table 4-36. Summary Statistics for English Language Pseudo-Test Forms (NCR)

Statistic	Equating Method	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	0.192	0.856	1.111	1.819	0.487	0.202	0.992	1.367
	CE	0.198	0.632	0.713	1.327	0.540	0.296	0.617	0.726
	TS	0.197	0.404	0.355	0.632	0.682	0.470	0.293	0.284
	OS	0.139	0.413	0.385	0.682	0.652	0.420	0.264	0.244
WASE	FE	0.279	0.275	0.287	0.311	0.279	0.283	0.306	0.290
	CE	0.311	0.319	0.323	0.343	0.311	0.323	0.352	0.353
	TS	0.252	0.243	0.252	0.270	0.249	0.246	0.252	0.272
	OS	0.234	0.219	0.234	0.246	0.227	0.227	0.230	0.250
WARMSE	FE	0.338	0.899	1.147	1.845	0.562	0.348	1.038	1.398
	CE	0.368	0.708	0.783	1.371	0.623	0.438	0.711	0.807
	TS	0.320	0.472	0.435	0.688	0.726	0.531	0.386	0.393
	OS	0.272	0.467	0.451	0.725	0.690	0.478	0.350	0.349
Difference	FE	0.181	0.852	1.095	1.816	0.489	0.181	0.984	1.370
	CE	0.183	0.618	0.699	1.330	0.527	0.269	0.623	0.741
	TS	0.179	0.416	0.313	0.616	0.676	0.497	0.300	0.312
	OS	0.117	0.426	0.336	0.665	0.645	0.445	0.269	0.263
Standardized WARMSB	FE	0.024	0.123	0.142	0.239	0.060	0.025	0.135	0.181
	CE	0.024	0.091	0.091	0.174	0.067	0.036	0.084	0.096
	TS	0.024	0.058	0.045	0.083	0.085	0.057	0.040	0.038
	OS	0.017	0.059	0.049	0.090	0.081	0.051	0.036	0.032
Standardized WASE	FE	0.034	0.039	0.037	0.041	0.035	0.035	0.042	0.038
	CE	0.038	0.046	0.041	0.045	0.039	0.039	0.048	0.047
	TS	0.031	0.035	0.032	0.036	0.031	0.030	0.034	0.036
	OS	0.029	0.031	0.030	0.032	0.028	0.028	0.031	0.033

Table 4-37. Summary Statistics for English Language Pseudo-Test Forms (FCR)

Statistic	Equating Method	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	0.257	0.625	0.977	1.249	0.376	0.481	0.932	1.469
	CE	0.328	0.394	0.537	0.543	0.406	0.378	0.493	0.917
	TS	0.282	0.403	0.431	0.355	0.766	0.336	0.304	0.176
	OS	0.239	0.407	0.457	0.392	0.738	0.290	0.271	0.121
WASE	FE	0.275	0.287	0.294	0.326	0.267	0.274	0.291	0.278
	CE	0.304	0.328	0.334	0.378	0.299	0.319	0.327	0.336
	TS	0.371	0.294	0.306	0.400	0.351	0.375	0.363	0.354
	OS	0.349	0.267	0.287	0.374	0.329	0.365	0.338	0.331
WARMSE	FE	0.377	0.688	1.020	1.291	0.461	0.554	0.977	1.495
	CE	0.447	0.513	0.633	0.662	0.504	0.494	0.592	0.976
	TS	0.466	0.498	0.529	0.535	0.843	0.504	0.474	0.395
	OS	0.423	0.487	0.540	0.542	0.808	0.466	0.434	0.353
Difference	FE	0.250	0.617	0.960	1.236	0.385	0.476	0.926	1.473
	CE	0.322	0.372	0.519	0.535	0.405	0.381	0.491	0.917
	TS	0.263	0.410	0.423	0.328	0.763	0.332	0.315	0.195
	OS	0.217	0.416	0.432	0.365	0.733	0.281	0.281	0.136
Standardized WARMSB	FE	0.032	0.090	0.125	0.164	0.047	0.059	0.127	0.195
	CE	0.040	0.056	0.069	0.071	0.050	0.046	0.067	0.121
	TS	0.035	0.058	0.055	0.047	0.095	0.041	0.041	0.023
	OS	0.029	0.058	0.059	0.052	0.092	0.035	0.037	0.016
Standardized WASE	FE	0.034	0.041	0.038	0.043	0.033	0.033	0.040	0.037
	CE	0.037	0.047	0.043	0.050	0.037	0.039	0.045	0.044
	TS	0.046	0.042	0.039	0.052	0.044	0.046	0.049	0.047
	OS	0.043	0.038	0.037	0.049	0.041	0.044	0.046	0.044

Table 4-38. WARMSB for English Language Pseudo-Test Forms to Illustrate Effects of CI ES

CI	Equating Method	MC-CR 0.00				MC-CR 0.25			
		0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.60 (CI)	0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.60 (CI)
NCR CI	FE	--	0.842	1.076	1.810	--	0.641	1.469	1.818
	CE	--	0.620	0.715	1.386	--	0.433	1.076	1.219
	TS	--	0.420	0.486	0.786	--	0.451	0.744	0.717
	OS	--	0.432	0.471	0.784	--	0.432	0.740	0.723
FCR CI	FE	--	0.725	1.005	1.259	--	0.582	1.013	1.391
	CE	--	0.567	0.700	0.663	--	0.480	0.627	0.912
	TS	--	0.505	0.652	0.679	--	0.668	0.611	0.582
	OS	--	0.519	0.654	0.653	--	0.638	0.596	0.600

Table 4-39. English Language Pseudo-Test Forms Percentages of Classification Consistency

CI	Equating Method	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
NCR CI	FE		100.0	89.3	84.1	75.0	87.1	100.0	85.1	81.2
	CE		100.0	90.7	94.6	84.1	87.1	96.5	93.4	89.3
	TS	**96.5**	94.9	92.5	92.5	91.1	90.7	90.0	96.5	96.5
	OS	**96.5**	96.5	92.5	92.5	91.1	90.7	95.1	96.5	96.5
FCR CI	FE		96.5	90.7	85.1	85.1	90.0	98.6	89.3	84.4
	CE		96.5	96.0	94.6	90.7	90.0	98.6	100.0	89.3
	TS	**96.5**	94.9	92.5	92.5	96.5	90.7	91.4	91.4	94.9
	OS	**96.5**	94.9	92.5	92.5	92.5	90.7	96.5	91.4	94.9

Table 4-40. Descriptive Statistics for the Old Spanish Language Pseudo-Test Form

Score	Statistic	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
	N	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	42.979	45.534	44.700	47.225	40.706	44.251	44.928	47.185
	SD	11.737	10.925	11.139	11.180	11.701	11.900	11.961	10.430
	Skew	-0.317	-0.383	-0.424	-0.556	-0.217	-0.364	-0.388	-0.494
	Kurt	2.431	2.654	2.611	2.863	2.328	2.387	2.511	2.816
	α	0.912	0.904	0.905	0.912	0.909	0.918	0.920	0.898
CR	Mean	24.304	25.980	24.792	26.522	22.616	25.182	25.561	27.199
	SD	7.356	6.666	6.959	6.979	7.497	7.291	7.398	6.275
	Skew	-0.502	-0.701	-0.422	-0.733	-0.373	-0.693	-0.658	-0.904
	Kurt	2.634	3.239	2.548	3.140	2.482	2.992	2.904	3.886
	α	0.833	0.803	0.822	0.823	0.841	0.830	0.837	0.795
CO	Mean	67.283	71.514	69.492	73.746	63.322	69.433	70.489	74.384
	SD	18.050	16.407	16.975	17.179	18.171	18.165	18.357	15.611
	Skew	-0.400	-0.559	-0.447	-0.660	-0.318	-0.517	-0.522	-0.691
	Kurt	2.545	3.046	2.662	3.059	2.376	2.686	2.722	3.264
	α	0.933	0.923	0.927	0.931	0.933	0.935	0.938	0.920

Table 4-41. CI Descriptive Statistics for the Old Spanish Language Pseudo-Test Form

Score	Statistic	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
	N	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
NCR MT CI	Mean	20.439	21.641	21.327	22.524	19.348	21.041	21.407	22.365
	SD	6.040	5.763	5.812	5.789	6.003	6.111	6.087	5.439
	Skew	-0.235	-0.345	-0.350	-0.457	-0.139	-0.291	-0.299	-0.385
	Kurt	2.361	2.523	2.501	2.647	2.319	2.352	2.385	2.534
	CO Corr.	0.920	0.905	0.911	0.919	0.918	0.920	0.920	0.902
NCR SM CI	Mean	20.687	21.894	21.632	22.798	19.717	21.197	21.509	22.551
	SD	6.251	5.969	6.072	5.997	6.304	6.361	6.379	5.729
	Skew	-0.299	-0.399	-0.439	-0.543	-0.228	-0.327	-0.351	-0.473
	Kurt	2.363	2.510	2.505	2.685	2.263	2.326	2.409	2.679
	CO Corr.	0.917	0.902	0.911	0.916	0.923	0.922	0.921	0.906
NCR DS CI	Mean	17.920	18.992	18.971	20.103	17.117	18.385	18.729	19.602
	SD	6.140	6.193	6.120	6.235	6.046	6.359	6.444	6.011
	Skew	0.026	-0.073	-0.151	-0.205	0.080	0.009	0.025	-0.094
	Kurt	2.317	2.320	2.300	2.361	2.339	2.190	2.274	2.279
	CO Corr.	0.888	0.876	0.887	0.892	0.895	0.897	0.893	0.883
FCR CI	Mean	21.111	22.459	21.829	23.182	19.817	21.864	22.147	23.244
	SD	5.931	5.489	5.629	5.609	5.924	5.982	6.059	5.170
	Skew	-0.401	-0.553	-0.432	-0.587	-0.290	-0.507	-0.488	-0.602
	Kurt	2.539	2.968	2.703	2.928	2.457	2.681	2.597	3.045
	CO Corr.	0.939	0.925	0.932	0.935	0.937	0.940	0.942	0.920
Full CI	Mean	42.171	44.891	43.777	46.440	39.779	43.491	44.143	46.574
	SD	11.783	10.892	11.251	11.309	11.930	11.930	12.018	10.349
	Skew	-0.340	-0.520	-0.416	-0.600	-0.286	-0.460	-0.443	-0.588
	Kurt	2.467	2.930	2.607	2.925	2.344	2.608	2.602	3.010

Table 4-42. Descriptive Statistics for the New Spanish Language Pseudo-Test Form

Score	Statistic	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
	N	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
MC	Mean	42.889	43.079	39.548	38.998	39.724	40.777	39.117	39.965
	SD	11.560	12.013	12.043	12.659	11.554	11.962	12.306	11.724
	Skew	-0.320	-0.297	-0.083	-0.047	-0.102	-0.177	-0.083	-0.133
	Kurt	2.497	2.405	2.269	2.157	2.376	2.409	2.255	2.384
	α	0.908	0.916	0.911	0.920	0.903	0.911	0.915	0.906
CR	Mean	26.126	26.415	23.321	23.128	22.059	23.039	21.874	22.666
	SD	7.713	8.095	8.269	8.813	7.508	7.898	7.953	7.764
	Skew	-0.724	-0.663	-0.180	-0.169	-0.063	-0.118	-0.019	-0.093
	Kurt	2.787	2.539	2.113	2.002	2.481	2.299	2.266	2.296
	α	0.804	0.820	0.823	0.841	0.835	0.833	0.846	0.828
CO	Mean	69.015	69.494	62.869	62.126	61.783	63.816	60.991	62.631
	SD	18.029	18.910	19.130	20.429	17.871	18.694	19.223	18.294
	Skew	-0.529	-0.475	-0.149	-0.118	-0.095	-0.164	-0.063	-0.139
	Kurt	2.661	2.488	2.229	2.103	2.447	2.403	2.277	2.383
	α	0.924	0.931	0.929	0.937	0.928	0.932	0.937	0.928

Table 4-43. CI Descriptive Statistics for the New Spanish Language Pseudo-Test Form

Score	Statistic	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
	N	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
NCR MT CI	Mean	20.572	20.826	19.293	19.009	19.637	20.091	19.279	19.649
	SD	5.733	5.946	5.933	6.131	5.872	6.004	6.258	5.924
	Skew	-0.206	-0.244	-0.063	-0.038	-0.124	-0.159	-0.089	-0.137
	Kurt	2.426	2.375	2.384	2.265	2.321	2.365	2.236	2.384
	CO Corr.	0.882	0.892	0.896	0.908	0.904	0.907	0.910	0.895
NCR SM CI	Mean	20.645	20.968	19.501	19.167	20.055	20.443	19.646	20.009
	SD	6.003	6.156	6.210	6.420	6.131	6.249	6.517	6.184
	Skew	-0.282	-0.290	-0.128	-0.094	-0.224	-0.259	-0.196	-0.209
	Kurt	2.431	2.363	2.264	2.200	2.310	2.347	2.216	2.288
	CO Corr.	0.874	0.885	0.891	0.909	0.908	0.905	0.916	0.899
NCR DS CI	Mean	17.734	18.118	16.878	16.633	17.585	17.964	17.218	17.461
	SD	6.023	6.152	5.936	6.150	5.940	6.154	6.237	6.026
	Skew	0.079	0.041	0.167	0.207	0.032	0.007	0.094	0.071
	Kurt	2.374	2.352	2.397	2.343	2.410	2.349	2.294	2.366
	CO Corr.	0.835	0.846	0.852	0.877	0.886	0.881	0.895	0.872
FCR CI	Mean	21.173	21.577	19.860	19.566	19.934	20.481	19.631	19.997
	SD	5.648	5.843	5.899	6.198	5.716	5.890	6.128	5.812
	Skew	-0.374	-0.425	-0.176	-0.146	-0.158	-0.225	-0.156	-0.209
	Kurt	2.640	2.578	2.399	2.310	2.524	2.565	2.379	2.537
	CO Corr.	0.921	0.931	0.928	0.935	0.928	0.930	0.935	0.924
Full CI	Mean	42.271	42.865	39.444	38.936	39.891	40.903	39.251	39.994
	SD	11.106	11.538	11.698	12.374	11.475	11.819	12.277	11.583
	Skew	-0.372	-0.382	-0.140	-0.119	-0.153	-0.213	-0.120	-0.170
	Kurt	2.617	2.562	2.351	2.232	2.460	2.441	2.268	2.436

Table 4-44. Effect Sizes for Spanish Language Pseudo-Test Forms

ES	Single Group	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
N	19,010	1,900	1,900	1,900	1,900	1,900	1,900	1,900	1,900
CO	0.063	0.096	-0.114	-0.366	-0.616	-0.085	-0.305	-0.505	-0.691
NCR MT	0.000	0.023	-0.139	-0.346	-0.590	0.049	-0.157	-0.345	-0.478
NCR SM	0.000	-0.007	-0.153	-0.347	-0.585	0.054	-0.120	-0.289	-0.427
NCR DS	0.000	-0.031	-0.142	-0.347	-0.561	0.078	-0.067	-0.238	-0.356
FCR	0.000	0.011	-0.156	-0.342	-0.612	0.020	-0.233	-0.413	-0.590
Full CI	0.000	0.009	-0.181	-0.378	-0.633	0.010	-0.218	-0.403	-0.599
MC	-0.047	-0.008	-0.214	-0.444	-0.689	-0.084	-0.291	-0.479	-0.651
CR	0.223	0.242	0.059	-0.193	-0.427	-0.074	-0.282	-0.480	-0.642
MC-CR	-0.270	-0.250	-0.273	-0.252	-0.262	-0.010	-0.009	0.001	-0.009

Table 4-45. Equated Moments for Spanish Language Pseudo-Test Forms (NCR MT)

Method	Statistic	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
FE	Mean	70.47	67.68	69.39	64.07	64.06	64.10	66.88	64.57	67.23
	SD	17.25	17.27	16.99	17.55	18.85	17.80	18.13	19.21	17.68
	Skew	-0.52	-0.39	-0.49	-0.22	-0.29	-0.31	-0.42	-0.31	-0.49
	Kurt	2.80	2.62	2.84	2.44	2.38	2.42	2.61	2.34	2.74
CE	Mean	70.47	67.75	69.14	63.53	63.15	64.22	66.71	64.00	66.35
	SD	17.24	17.09	17.16	17.57	18.84	17.72	18.05	19.37	18.00
	Skew	-0.52	-0.39	-0.49	-0.19	-0.22	-0.30	-0.41	-0.31	-0.49
	Kurt	2.81	2.66	2.84	2.46	2.27	2.40	2.60	2.32	2.72
TS	Mean	70.44	68.22	68.55	62.92	61.88	63.91	65.52	62.53	64.31
	SD	16.95	16.02	17.00	17.29	18.85	17.37	18.05	19.42	18.49
	Skew	-0.53	-0.46	-0.38	-0.19	-0.18	-0.23	-0.35	-0.21	-0.32
	Kurt	2.87	2.82	2.68	2.46	2.23	2.55	2.45	2.19	2.40
OS	Mean	70.41	68.20	68.52	62.93	61.89	63.95	65.56	62.55	64.34
	SD	16.94	16.02	17.00	17.28	18.86	17.34	18.05	19.43	18.47
	Skew	-0.53	-0.44	-0.38	-0.18	-0.18	-0.21	-0.35	-0.22	-0.33
	Kurt	2.84	2.77	2.65	2.43	2.23	2.53	2.46	2.21	2.43

Table 4-46. Equated Moments for Spanish Language Pseudo-Test Forms (NCR SM)

Method	Statistic	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
FE	Mean	70.47	66.84	69.50	64.39	65.02	64.65	68.51	66.63	69.36
	SD	17.247	17.82	16.59	17.08	18.30	17.47	17.84	18.56	16.70
	Skew	-0.52	-0.36	-0.48	-0.23	-0.32	-0.36	-0.52	-0.43	-0.57
	Kurt	2.80	2.53	2.90	2.49	2.46	2.50	2.77	2.49	2.95
CE	Mean	70.47	66.81	69.22	63.71	63.93	64.80	68.40	66.22	68.57
	SD	17.24	17.74	16.65	17.01	18.33	17.66	17.85	18.66	17.12
	Skew	-0.52	-0.36	-0.47	-0.18	-0.25	-0.37	-0.57	-0.43	-0.60
	Kurt	2.81	2.57	2.89	2.51	2.34	2.54	2.89	2.46	2.91
TS	Mean	70.44	68.00	68.33	63.03	61.75	63.99	65.91	63.33	64.81
	SD	16.95	16.10	16.94	17.03	18.89	17.30	17.83	18.99	18.24
	Skew	-0.53	-0.44	-0.37	-0.20	-0.17	-0.23	-0.37	-0.26	-0.35
	Kurt	2.87	2.80	2.68	2.48	2.23	2.56	2.49	2.24	2.45
OS	Mean	70.41	67.98	68.29	63.04	61.75	64.02	65.95	63.37	64.86
	SD	16.94	16.11	16.95	17.03	18.90	17.27	17.83	19.00	18.22
	Skew	-0.53	-0.43	-0.37	-0.19	-0.17	-0.22	-0.37	-0.26	-0.36
	Kurt	2.84	2.75	2.65	2.45	2.23	2.54	2.50	2.26	2.48

Table 4-47. Equated Moments for Spanish Language Pseudo-Test Forms (NCR DS)

Method	Statistic	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
FE	Mean	70.47	67.23	69.18	64.15	64.26	64.27	67.48	65.53	67.95
	SD	17.247	17.42	17.06	17.46	18.70	17.68	18.10	19.05	17.43
	Skew	-0.52	-0.39	-0.51	-0.20	-0.27	-0.31	-0.47	-0.37	-0.50
	Kurt	2.80	2.59	2.90	2.41	2.33	2.43	2.65	2.39	2.71
CE	Mean	70.47	67.21	68.92	63.53	63.28	64.34	67.32	65.10	67.26
	SD	17.24	17.30	17.16	17.48	18.74	17.62	17.97	19.05	17.56
	Skew	-0.52	-0.39	-0.49	-0.18	-0.22	-0.30	-0.48	-0.37	-0.48
	Kurt	2.81	2.64	2.88	2.44	2.26	2.45	2.70	2.35	2.60
TS	Mean	70.44	67.41	68.75	63.14	62.04	64.58	67.46	64.87	66.79
	SD	16.95	17.00	16.44	16.73	18.82	17.08	17.33	18.52	16.99
	Skew	-0.53	-0.40	-0.40	-0.21	-0.18	-0.26	-0.46	-0.33	-0.48
	Kurt	2.87	2.67	2.75	2.51	2.24	2.60	2.62	2.34	2.71
OS	Mean	70.41	67.38	68.72	63.15	62.04	64.63	67.52	64.92	66.86
	SD	16.94	16.97	16.46	16.74	18.83	17.06	17.33	18.52	16.99
	Skew	-0.53	-0.40	-0.40	-0.20	-0.18	-0.24	-0.44	-0.32	-0.46
	Kurt	2.84	2.65	2.71	2.47	2.24	2.57	2.61	2.34	2.70

Table 4-48. Equated Moments for Spanish Language Pseudo-Test Forms (FCR)

Method	Statistic	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
FE	Mean	70.47	67.47	69.08	63.95	63.35	63.62	65.53	63.33	65.35
	SD	17.247	17.29	17.35	17.84	19.17	17.64	18.06	18.74	17.84
	Skew	-0.52	-0.37	-0.45	-0.22	-0.23	-0.23	-0.28	-0.21	-0.35
	Kurt	2.80	2.60	2.70	2.38	2.33	2.43	2.51	2.36	2.61
CE	Mean	70.47	67.47	68.87	63.53	62.61	63.68	65.24	62.91	64.53
	SD	17.24	17.17	17.48	17.85	19.21	17.53	17.95	18.62	17.88
	Skew	-0.52	-0.37	-0.44	-0.19	-0.21	-0.20	-0.24	-0.20	-0.31
	Kurt	2.81	2.64	2.65	2.35	2.32	2.45	2.55	2.39	2.60
TS	Mean	70.44	67.50	68.58	63.06	61.96	64.13	65.67	63.34	64.81
	SD	16.95	16.71	17.25	17.37	18.64	17.54	17.85	18.69	17.86
	Skew	-0.53	-0.41	-0.38	-0.19	-0.18	-0.23	-0.36	-0.26	-0.36
	Kurt	2.87	2.70	2.65	2.45	2.25	2.54	2.48	2.27	2.49
OS	Mean	70.41	67.47	68.55	63.07	61.96	64.17	65.70	63.37	64.84
	SD	16.94	16.69	17.24	17.35	18.65	17.50	17.86	18.70	17.85
	Skew	-0.53	-0.41	-0.38	-0.19	-0.18	-0.22	-0.36	-0.26	-0.36
	Kurt	2.84	2.68	2.63	2.42	2.25	2.52	2.48	2.28	2.51

Table 4-49. Summary Statistics for Spanish Language Pseudo-Test Forms (NCR MT)

Statistic	Equating Method	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	1.033	1.146	1.738	2.542	3.366	3.986	4.417	5.420
	CE	0.904	0.929	1.219	1.576	3.398	3.837	4.027	4.681
	TS	0.493	0.519	0.623	0.416	2.758	2.681	2.847	3.035
	OS	0.485	0.463	0.651	0.431	2.805	2.719	2.866	3.051
WASE	FE	0.521	0.555	0.554	0.623	0.542	0.560	0.597	0.612
	CE	0.615	0.654	0.646	0.708	0.603	0.651	0.689	0.728
	TS	0.467	0.531	0.521	0.437	0.423	0.409	0.453	0.609
	OS	0.450	0.514	0.505	0.423	0.410	0.397	0.443	0.590
WARMSE	FE	1.157	1.274	1.824	2.617	3.409	4.025	4.457	5.454
	CE	1.093	1.136	1.379	1.728	3.451	3.892	4.085	4.738
	TS	0.679	0.743	0.812	0.604	2.790	2.712	2.883	3.096
	OS	0.662	0.691	0.824	0.604	2.835	2.748	2.900	3.107
Difference	FE	1.017	1.143	1.745	2.518	3.343	3.967	4.414	5.391
	CE	0.879	0.918	1.198	1.562	3.400	3.811	4.010	4.680
	TS	0.547	0.481	0.429	0.395	2.712	2.683	2.885	3.022
	OS	0.536	0.430	0.455	0.408	2.760	2.718	2.902	3.035
Standardized WARMSB	FE	0.057	0.070	0.095	0.155	0.185	0.219	0.311	0.319
	CE	0.050	0.057	0.066	0.096	0.187	0.211	0.284	0.276
	TS	0.027	0.032	0.034	0.025	0.152	0.148	0.201	0.179
	OS	0.027	0.028	0.035	0.026	0.154	0.150	0.202	0.180
Standardized WASE	FE	0.029	0.034	0.030	0.038	0.030	0.031	0.042	0.036
	CE	0.034	0.040	0.035	0.043	0.033	0.036	0.049	0.043
	TS	0.026	0.032	0.028	0.027	0.023	0.023	0.032	0.036
	OS	0.025	0.031	0.027	0.026	0.023	0.022	0.031	0.035

Table 4-50. Summary Statistics for Spanish Language Pseudo-Test Forms (NCR SM)

Statistic	Equating Method	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	1.342	1.053	1.850	2.726	3.434	4.602	5.329	6.090
	CE	1.312	0.841	1.312	1.748	3.420	4.450	4.972	5.442
	TS	0.474	0.643	0.824	0.369	2.787	2.959	3.190	3.267
	OS	0.467	0.592	0.838	0.377	2.837	3.006	3.233	3.302
WASE	FE	0.527	0.564	0.564	0.608	0.530	0.554	0.582	0.566
	CE	0.631	0.667	0.639	0.692	0.605	0.642	0.668	0.709
	TS	0.383	0.456	0.454	0.450	0.423	0.408	0.441	0.648
	OS	0.365	0.438	0.439	0.436	0.409	0.395	0.429	0.627
WARMSE	FE	1.442	1.194	1.934	2.793	3.475	4.635	5.360	6.117
	CE	1.456	1.073	1.460	1.880	3.474	4.496	5.016	5.488
	TS	0.610	0.789	0.941	0.582	2.819	2.987	3.220	3.330
	OS	0.593	0.736	0.946	0.576	2.867	3.032	3.262	3.361
Difference	FE	1.325	1.020	1.851	2.703	3.414	4.586	5.330	6.065
	CE	1.290	0.836	1.304	1.737	3.421	4.422	4.965	5.440
	TS	0.489	0.582	0.645	0.413	2.750	2.937	3.201	3.247
	OS	0.479	0.533	0.656	0.416	2.801	2.982	3.243	3.281
Standardized WARMSB	FE	0.074	0.064	0.101	0.166	0.189	0.253	0.376	0.359
	CE	0.073	0.051	0.071	0.107	0.188	0.245	0.351	0.321
	TS	0.026	0.039	0.045	0.023	0.153	0.163	0.225	0.192
	OS	0.026	0.036	0.046	0.023	0.156	0.166	0.228	0.195
Standardized WASE	FE	0.029	0.034	0.031	0.037	0.029	0.031	0.041	0.033
	CE	0.035	0.041	0.035	0.042	0.033	0.035	0.047	0.042
	TS	0.021	0.028	0.025	0.027	0.023	0.022	0.031	0.038
	OS	0.020	0.027	0.024	0.027	0.023	0.022	0.030	0.037

Table 4-51. Summary Statistics for Spanish Language Pseudo-Test Forms (NCR DS)

Statistic	Equating Method	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	1.867	1.521	2.174	3.595	3.833	5.556	6.312	7.417
	CE	1.875	1.363	1.649	2.496	3.912	5.533	5.985	6.761
	TS	1.139	0.976	1.180	0.609	3.237	4.418	4.449	4.428
	OS	1.119	0.916	1.177	0.630	3.297	4.484	4.514	4.494
WASE	FE	0.581	0.602	0.589	0.656	0.569	0.595	0.619	0.615
	CE	0.697	0.714	0.692	0.784	0.648	0.697	0.724	0.759
	TS	0.473	0.572	0.569	0.595	0.460	0.470	0.530	0.478
	OS	0.457	0.553	0.554	0.582	0.445	0.456	0.519	0.464
WARMSE	FE	1.955	1.636	2.252	3.655	3.875	5.588	6.342	7.442
	CE	2.000	1.539	1.788	2.616	3.965	5.577	6.028	6.804
	TS	1.233	1.131	1.310	0.851	3.269	4.443	4.481	4.453
	OS	1.208	1.070	1.301	0.858	3.327	4.507	4.544	4.518
Difference	FE	1.858	1.506	2.167	3.546	3.801	5.554	6.323	7.416
	CE	1.857	1.358	1.635	2.479	3.907	5.514	5.970	6.760
	TS	1.067	0.947	0.939	0.440	3.231	4.338	4.431	4.725
	OS	1.046	0.891	0.932	0.461	3.292	4.402	4.495	4.805
Standardized WARMSB	FE	0.103	0.093	0.118	0.219	0.211	0.306	0.445	0.437
	CE	0.104	0.083	0.090	0.152	0.215	0.305	0.422	0.398
	TS	0.063	0.059	0.064	0.037	0.178	0.243	0.314	0.261
	OS	0.062	0.056	0.064	0.038	0.181	0.247	0.318	0.265
Standardized WASE	FE	0.032	0.037	0.032	0.040	0.031	0.033	0.044	0.036
	CE	0.039	0.044	0.038	0.048	0.036	0.038	0.051	0.045
	TS	0.026	0.035	0.031	0.036	0.025	0.026	0.037	0.028
	OS	0.025	0.034	0.030	0.036	0.024	0.025	0.037	0.027

Table 4-52. Summary Statistics for Spanish Language Pseudo-Test Forms (FCR)

Statistic	Equating Method	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	1.128	0.657	1.681	1.810	2.712	2.524	2.964	3.501
	CE	1.029	0.466	1.277	1.123	2.655	2.177	2.520	2.714
	TS	0.894	0.454	0.705	0.432	3.023	2.747	3.079	3.075
	OS	0.885	0.399	0.739	0.440	3.067	2.792	3.130	3.125
WASE	FE	0.482	0.514	0.511	0.560	0.503	0.502	0.534	0.552
	CE	0.560	0.596	0.588	0.638	0.553	0.589	0.608	0.663
	TS	0.374	0.418	0.403	0.423	0.411	0.400	0.413	0.527
	OS	0.359	0.398	0.390	0.409	0.398	0.389	0.402	0.507
WARMSE	FE	1.227	0.835	1.757	1.895	2.758	2.573	3.012	3.544
	CE	1.171	0.757	1.405	1.291	2.712	2.255	2.593	2.794
	TS	0.969	0.617	0.812	0.605	3.051	2.776	3.107	3.120
	OS	0.955	0.563	0.836	0.600	3.092	2.819	3.155	3.166
Difference	FE	1.116	0.634	1.675	1.767	2.702	2.513	2.966	3.491
	CE	1.006	0.431	1.273	1.104	2.651	2.158	2.512	2.710
	TS	0.831	0.399	0.554	0.345	2.988	2.723	3.058	3.043
	OS	0.821	0.351	0.587	0.353	3.032	2.765	3.107	3.092
Standardized WARMSB	FE	0.062	0.040	0.092	0.111	0.149	0.139	0.209	0.206
	CE	0.057	0.028	0.070	0.069	0.146	0.120	0.178	0.160
	TS	0.050	0.028	0.038	0.026	0.166	0.151	0.217	0.181
	OS	0.049	0.024	0.040	0.027	0.169	0.154	0.221	0.184
Standardized WASE	FE	0.027	0.031	0.028	0.034	0.028	0.028	0.038	0.033
	CE	0.031	0.036	0.032	0.039	0.030	0.032	0.043	0.039
	TS	0.021	0.025	0.022	0.026	0.023	0.022	0.029	0.031
	OS	0.020	0.024	0.021	0.025	0.022	0.021	0.028	0.030

Table 4-53. WARMSB for Spanish Language Pseudo-Test Forms to Illustrate Effects of CI ES

CI	Equating Method	MC-CR 0.00				MC-CR 0.25			
		0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.60 (CI)	0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.60 (CI)
NCR MT CI	FE	--	2.263	2.806	3.725	--	1.176	1.725	3.031
	CE	--	1.812	2.020	2.494	--	1.131	1.175	2.107
	TS	--	0.445	0.637	1.066	--	0.620	1.202	1.179
	OS	--	0.429	0.693	1.084	--	0.571	1.171	1.124
NCR SM CI	FE	--	2.615	3.488	4.502	--	1.641	2.650	3.574
	CE	--	2.259	2.834	3.363	--	1.560	2.143	2.694
	TS	--	0.472	0.866	0.993	--	0.633	0.866	1.007
	OS	--	0.453	0.927	1.015	--	0.587	0.827	0.962
NCR DS CI	FE	--	3.761	4.632	6.229	--	2.397	3.537	4.882
	CE	--	3.538	3.923	5.031	--	2.233	2.909	3.795
	TS	--	2.022	2.529	1.841	--	1.603	1.602	1.623
	OS	--	1.950	2.521	1.848	--	1.590	1.586	1.612
FCR CI	FE	--	1.844	2.898	3.097	--	0.528	0.952	1.415
	CE	--	1.455	2.321	2.193	--	0.820	0.811	0.813
	TS	--	0.794	1.692	1.362	--	0.758	0.725	0.754
	OS	--	0.774	1.731	1.379	--	0.707	0.658	0.716

Table 4-54. Spanish Language Pseudo-Test Forms Percentages of Classification Consistency

CI	Equating Method	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
NCR MT CI	FE		92.2	94.3	88.5	82.8	75.7	69.2	64.3	61.3
	CE		93.3	95.9	92.6	90.2	75.7	70.5	69.2	66.2
	TS	**97.5**	100.0	96.6	100.0	99.1	78.5	81.0	77.2	77.2
	OS	**98.4**	99.1	95.7	96.8	97.7	75.4	79.2	76.3	76.3
NCR SM CI	FE		91.9	96.6	88.5	81.2	75.7	66.2	58.7	57.9
	CE		91.9	95.1	91.7	90.9	73.5	70.3	62.9	62.7
	TS	**97.5**	98.2	96.1	100.0	99.1	78.5	80.1	75.0	75.0
	OS	**98.4**	97.3	97.3	99.1	100.0	75.4	77.1	74.1	74.1
NCR DS CI	FE		89.1	95.7	89.3	80.4	73.5	62.7	56.7	53.8
	CE		88.0	94.1	96.6	84.3	71.2	62.7	56.7	56.7
	TS	**97.5**	94.8	97.0	99.1	96.1	75.4	71.4	68.3	67.5
	OS	**98.4**	95.7	98.2	98.2	95.2	74.5	68.8	67.4	66.6
FCR	FE		90.1	96.8	89.4	87.6	80.0	81.6	76.2	73.3
	CE		92.2	96.8	92.9	92.0	77.2	83.8	79.1	79.8
	TS	**97.5**	94.8	96.6	94.6	100.0	74.4	80.1	76.7	76.3
	OS	**98.4**	95.7	95.7	93.7	99.1	73.5	79.2	75.8	75.4

Table 4-55. Descriptive Statistics for the Old Chemistry Pseudo-Test Form

Score	Statistic	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
	N	1,500	1,500	1,500	1,500	1,500	1,500	1,500	1,500
MC	Mean	32.104	29.890	28.211	26.727	26.829	28.271	27.631	25.781
	SD	11.549	11.205	11.667	11.586	11.611	11.540	11.871	11.779
	Skew	-0.330	-0.144	0.025	0.211	0.161	0.002	0.080	0.252
	Kurt	2.050	1.967	1.873	1.950	1.894	1.891	1.848	1.917
	α	0.930	0.922	0.927	0.925	0.926	0.926	0.930	0.928
CR	Mean	20.739	18.579	16.960	15.712	16.031	17.498	16.753	15.103
	SD	11.223	10.930	11.256	11.041	11.173	11.296	11.548	11.289
	Skew	-0.249	-0.100	0.072	0.181	0.166	0.030	0.084	0.246
	Kurt	2.010	1.872	1.835	1.891	1.833	1.820	1.747	1.823
	α	0.870	0.867	0.879	0.876	0.873	0.877	0.880	0.881
NCR CI	Mean	17.917	16.763	15.767	14.911	15.004	15.795	15.439	14.388
	SD	6.399	6.302	6.627	6.584	6.604	6.543	6.665	6.701
	Skew	-0.522	-0.331	-0.113	0.034	-0.018	-0.173	-0.088	0.088
	Kurt	2.190	2.020	1.842	1.843	1.809	1.867	1.788	1.787
	CO Corr.	0.938	0.935	0.940	0.939	0.945	0.943	0.947	0.949
FCR CI	Mean	17.041	15.922	14.845	13.981	14.155	15.010	14.561	13.559
	SD	6.592	6.504	6.790	6.713	6.763	6.750	6.844	6.890
	Skew	-0.476	-0.304	-0.086	0.049	-0.005	-0.148	-0.079	0.108
	Kurt	2.259	2.080	1.902	1.904	1.870	1.908	1.833	1.839
	CO Corr.	0.946	0.943	0.951	0.948	0.948	0.950	0.950	0.953
Full CI	Mean	32.857	30.355	28.291	26.603	26.915	28.661	27.734	25.665
	SD	13.510	13.259	13.773	13.571	13.839	13.818	14.172	14.058
	Skew	-0.417	-0.244	-0.062	0.091	0.059	-0.102	-0.020	0.155
	Kurt	2.076	1.910	1.780	1.815	1.774	1.799	1.721	1.749
	CO Corr.	0.985	0.985	0.987	0.986	0.986	0.986	0.987	0.987
CO	Mean	52.843	48.469	45.171	42.439	42.861	45.769	44.385	40.884
	SD	21.985	21.318	22.104	21.834	22.051	22.093	22.718	22.381
	Skew	-0.308	-0.141	0.038	0.189	0.150	-0.003	0.062	0.239
	Kurt	2.034	1.913	1.820	1.902	1.843	1.838	1.778	1.838
	α	0.901	0.897	0.902	0.902	0.902	0.901	0.904	0.905

Table 4-56. Descriptive Statistics for the New Chemistry Pseudo-Test Form

Score	Statistic	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
	N	1,500	1,500`	1,500	1,500	1,500`	1,500	1,500	1,500
MC	Mean	32.791	33.188	33.580	34.175	28.235	31.489	33.315	33.847
	SD	11.572	11.148	11.281	11.303	11.883	11.392	11.476	10.895
	Skew	-0.342	-0.340	-0.414	-0.487	0.112	-0.196	-0.422	-0.437
	Kurt	2.070	2.141	2.124	2.227	1.887	1.965	2.124	2.222
	α	0.931	0.925	0.928	0.929	0.930	0.926	0.930	0.921
CR	Mean	20.914	21.364	21.768	22.381	16.223	19.437	21.092	21.643
	SD	11.620	11.141	11.285	11.238	11.490	11.312	11.629	11.005
	Skew	-0.248	-0.247	-0.301	-0.348	0.181	-0.050	-0.264	-0.279
	Kurt	1.952	2.030	2.049	2.048	1.890	1.905	1.957	2.049
	α	0.876	0.863	0.867	0.870	0.881	0.867	0.881	0.868
NCR CI	Mean	17.830	18.157	18.325	18.647	15.323	17.177	18.140	18.494
	SD	6.398	6.159	6.209	6.183	6.746	6.405	6.364	6.042
	Skew	-0.498	-0.517	-0.591	-0.670	-0.049	-0.375	-0.578	-0.604
	Kurt	2.194	2.273	2.312	2.450	1.847	2.048	2.301	2.399
	CO Corr.	0.943	0.933	0.940	0.937	0.946	0.936	0.938	0.930
FCR CI	Mean	16.983	17.345	17.547	17.915	14.435	16.291	17.217	17.575
	SD	6.645	6.321	6.388	6.365	6.923	6.541	6.617	6.232
	Skew	-0.459	-0.455	-0.523	-0.609	-0.019	-0.327	-0.512	-0.508
	Kurt	2.239	2.317	2.316	2.465	1.884	2.091	2.296	2.400
	CO Corr.	0.950	0.941	0.943	0.943	0.951	0.943	0.948	0.942
Full CI	Mean	32.783	33.349	33.885	34.611	27.282	31.223	33.312	34.023
	SD	13.553	12.955	13.153	13.163	13.984	13.332	13.536	12.709
	Skew	-0.418	-0.430	-0.494	-0.559	0.045	-0.257	-0.483	-0.496
	Kurt	2.091	2.173	2.184	2.257	1.815	1.939	2.144	2.242
	CO Corr.	0.985	0.982	0.983	0.983	0.985	0.984	0.984	0.982
CO	Mean	53.705	54.552	55.348	56.556	44.459	50.926	54.407	55.491
	SD	22.403	21.428	21.761	21.736	22.565	21.857	22.305	21.024
	Skew	-0.316	-0.318	-0.387	-0.448	0.135	-0.142	-0.373	-0.378
	Kurt	2.028	2.102	2.109	2.154	1.862	1.923	2.053	2.158
	α	0.900	0.895	0.897	0.899	0.903	0.897	0.900	0.894

Table 4-57. Effect Sizes for Chemistry Pseudo-Test Forms

ES	Single Group	MC-CR 0.00				MC-CR 0.10			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
N	12,328	1,500	1,500	1,500	1,500	1,500	1,500	1,500	1,500
CO	0.040	0.039	0.285	0.464	0.648	0.072	0.235	0.445	0.673
NCR CI	0.000	-0.014	0.224	0.398	0.585	0.048	0.213	0.415	0.644
FCR CI	0.000	-0.009	0.222	0.410	0.602	0.041	0.193	0.395	0.612
Full CI	0.000	-0.005	0.228	0.415	0.599	0.026	0.189	0.403	0.624
MC	0.065	0.059	0.295	0.468	0.651	0.120	0.281	0.487	0.711
CR	0.012	0.015	0.252	0.427	0.599	0.017	0.172	0.375	0.587
MC-CR	0.053	0.044	0.043	0.041	0.052	0.103	0.109	0.112	0.124

Table 4-58. Equated Moments for Chemistry Pseudo-Test Forms (NCR)

Method	Statistic	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
FE	Mean	52.93	52.56	52.91	53.24	54.31	43.92	50.14	53.19	53.97
	SD	21.50	21.97	21.06	21.16	21.11	22.49	21.91	22.05	21.07
	Skew	-0.31	-0.29	-0.32	-0.40	-0.43	0.12	-0.20	-0.39	-0.42
	Kurt	2.06	2.04	2.05	2.12	2.25	1.85	1.93	2.07	2.19
CE	Mean	52.92	52.55	53.30	53.78	55.05	44.01	50.51	53.74	54.79
	SD	21.49	21.96	21.18	21.14	21.15	22.61	21.94	22.12	20.95
	Skew	-0.32	-0.28	-0.32	-0.42	-0.48	0.13	-0.20	-0.40	-0.43
	Kurt	2.06	2.04	2.05	2.14	2.28	1.87	1.92	2.09	2.21
TS	Mean	52.94	52.61	53.61	54.34	55.80	44.37	50.80	54.26	55.55
	SD	21.52	22.07	21.16	21.44	21.46	22.81	22.29	22.40	21.14
	Skew	-0.30	-0.26	-0.26	-0.30	-0.34	0.18	-0.15	-0.35	-0.38
	Kurt	2.07	2.03	2.10	2.09	2.10	1.85	1.90	2.04	2.15
OS	Mean	52.92	52.58	53.58	54.31	55.76	44.35	50.79	54.25	55.54
	SD	21.49	22.04	21.12	21.41	21.43	22.81	22.25	22.38	21.12
	Skew	-0.30	-0.27	-0.27	-0.31	-0.35	0.17	-0.16	-0.36	-0.39
	Kurt	2.06	2.03	2.11	2.09	2.12	1.85	1.91	2.05	2.16

Table 4-59. Equated Moments for Chemistry Pseudo-Test Forms (FCR)

Method	Statistic	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
FE	Mean	52.93	52.67	52.84	53.55	54.65	43.77	49.72	52.87	53.41
	SD	21.50	22.08	20.89	21.09	21.13	22.53	21.66	22.17	20.86
	Skew	-0.31	-0.29	-0.32	-0.40	-0.44	0.12	-0.18	-0.36	-0.38
	Kurt	2.06	2.01	2.08	2.11	2.25	1.83	1.94	2.06	2.19
CE	Mean	52.92	52.69	53.16	54.04	55.42	43.81	49.94	53.32	54.01
	SD	21.49	22.14	21.01	21.16	21.22	22.61	21.65	22.32	20.85
	Skew	-0.32	-0.29	-0.30	-0.42	-0.48	0.14	-0.18	-0.36	-0.38
	Kurt	2.06	2.00	2.08	2.10	2.26	1.86	1.94	2.07	2.20
TS	Mean	52.94	52.46	53.51	54.18	55.71	44.68	51.24	54.48	55.61
	SD	21.52	22.25	21.03	21.34	21.51	22.99	22.17	22.51	21.10
	Skew	-0.30	-0.25	-0.26	-0.30	-0.33	0.17	-0.18	-0.36	-0.38
	Kurt	2.07	2.01	2.11	2.09	2.10	1.84	1.92	2.04	2.16
OS	Mean	52.92	52.43	53.48	54.14	55.67	44.66	51.24	54.46	55.60
	SD	21.49	22.23	20.99	21.31	21.48	22.99	22.13	22.49	21.08
	Skew	-0.30	-0.26	-0.27	-0.31	-0.34	0.16	-0.18	-0.37	-0.39
	Kurt	2.06	2.01	2.11	2.10	2.11	1.84	1.92	2.05	2.17

Table 4-60. Summary Statistics for Chemistry Pseudo-Test Forms (NCR)

Statistic	Equating Method	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	0.486	0.928	1.304	1.598	0.499	0.805	0.595	0.979
	CE	0.646	0.697	0.819	1.032	0.624	0.955	0.631	0.649
	TS	0.365	0.219	0.436	0.627	0.970	1.078	0.821	1.026
	OS	0.363	0.204	0.419	0.593	0.928	1.063	0.829	1.035
WASE	FE	0.610	0.596	0.598	0.636	0.600	0.604	0.599	0.623
	CE	0.683	0.677	0.698	0.716	0.684	0.703	0.698	0.704
	TS	0.422	0.441	0.477	0.525	0.445	0.439	0.461	0.477
	OS	0.415	0.435	0.472	0.520	0.439	0.433	0.456	0.471
WARMSE	FE	0.780	1.103	1.435	1.720	0.781	1.006	0.845	1.160
	CE	0.940	0.972	1.076	1.256	0.926	1.186	0.941	0.957
	TS	0.558	0.493	0.646	0.818	1.067	1.164	0.942	1.131
	OS	0.551	0.481	0.631	0.789	1.026	1.148	0.946	1.137
Difference	FE	0.465	0.905	1.283	1.564	0.472	0.796	0.582	0.961
	CE	0.637	0.670	0.831	1.045	0.597	0.933	0.596	0.619
	TS	0.357	0.231	0.455	0.645	0.980	1.079	0.824	1.050
	OS	0.353	0.214	0.436	0.607	0.937	1.064	0.829	1.057
Standardized WARMSB	FE	0.022	0.044	0.059	0.073	0.023	0.036	0.026	0.044
	CE	0.029	0.033	0.037	0.047	0.028	0.043	0.028	0.029
	TS	0.017	0.010	0.020	0.029	0.044	0.049	0.036	0.046
	OS	0.017	0.010	0.019	0.027	0.042	0.048	0.036	0.046
Standardized WASE	FE	0.028	0.028	0.027	0.029	0.027	0.027	0.026	0.028
	CE	0.031	0.032	0.032	0.033	0.031	0.032	0.031	0.031
	TS	0.019	0.021	0.022	0.024	0.020	0.020	0.020	0.021
	OS	0.019	0.020	0.021	0.024	0.020	0.020	0.020	0.021

Table 4-61. Summary Statistics for Chemistry Pseudo-Test Forms (FCR)

Statistic	Equating Method	MC-CR 0.00				MC-CR 0.25			
		CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
WARMSB	FE	0.434	0.983	1.055	1.223	0.554	0.710	0.851	1.321
	CE	0.664	0.762	0.706	0.657	0.731	0.700	0.634	0.807
	TS	0.532	0.326	0.566	0.650	1.298	1.369	1.066	1.061
	OS	0.526	0.325	0.558	0.613	1.270	1.373	1.076	1.072
WASE	FE	0.586	0.572	0.578	0.622	0.588	0.584	0.587	0.603
	CE	0.661	0.670	0.669	0.698	0.667	0.682	0.693	0.692
	TS	0.443	0.462	0.473	0.545	0.469	0.463	0.504	0.489
	OS	0.437	0.457	0.468	0.540	0.463	0.457	0.499	0.484
WARMSE	FE	0.729	1.137	1.203	1.372	0.808	0.919	1.033	1.452
	CE	0.937	1.015	0.973	0.959	0.990	0.977	0.939	1.063
	TS	0.692	0.565	0.738	0.849	1.381	1.445	1.179	1.168
	OS	0.684	0.561	0.728	0.817	1.352	1.447	1.186	1.176
Difference	FE	0.410	0.963	1.014	1.199	0.540	0.693	0.831	1.298
	CE	0.622	0.756	0.708	0.648	0.700	0.675	0.606	0.780
	TS	0.542	0.327	0.581	0.673	1.302	1.372	1.068	1.089
	OS	0.534	0.324	0.571	0.632	1.271	1.374	1.077	1.099
Standardized WARMSB	FE	0.020	0.046	0.048	0.056	0.025	0.032	0.037	0.059
	CE	0.030	0.036	0.032	0.030	0.033	0.032	0.028	0.036
	TS	0.024	0.015	0.026	0.030	0.059	0.062	0.047	0.047
	OS	0.024	0.015	0.025	0.028	0.058	0.062	0.047	0.048
Standardized WASE	FE	0.027	0.027	0.026	0.028	0.027	0.026	0.026	0.027
	CE	0.030	0.031	0.030	0.032	0.030	0.031	0.030	0.031
	TS	0.020	0.022	0.021	0.025	0.021	0.021	0.022	0.022
	OS	0.020	0.021	0.021	0.025	0.021	0.021	0.022	0.022

Table 4-62. WARMSB for Chemistry Pseudo-Test Forms to Illustrate Effects of CI ES

CI	Equating Method	MC-CR 0.00				MC-CR 0.25			
		0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.60 (CI)	0.00-0.00 (CI)	0.00-0.20 (CI)	0.00-0.40 (CI)	0.00-0.60 (CI)
NCR CI	FE	--	0.838	1.079	1.425	--	0.721	0.837	1.171
	CE	--	0.920	0.880	1.207	--	0.786	0.768	0.737
	TS	--	0.207	0.223	0.553	--	0.031	0.379	0.437
	OS	--	0.205	0.208	0.530	--	0.031	0.381	0.393
FCR CI	FE	--	0.902	0.942	1.245	--	0.793	0.989	1.510
	CE	--	0.904	0.775	1.036	--	0.878	0.760	1.145
	TS	--	0.424	0.404	0.588	--	0.523	0.539	0.647
	OS	--	0.424	0.397	0.573	--	0.514	0.500	0.624

Table 4-63. Chemistry Pseudo-Test Forms Percentages of Classification Consistency

CI	Equating Method	Single Group	MC-CR 0.00				MC-CR 0.25			
			CI 0.00	CI 0.20	CI 0.40	CI 0.60	CI 0.00	CI 0.20	CI 0.40	CI 0.60
NCR CI	FE		96.6	94.8	92.6	91.8	98.4	97.4	95.8	95.8
	CE		96.6	96.5	95.8	92.4	94.0	94.2	97.4	96.7
	TS	**99.0**	98.2	100.0	96.6	96.6	95.0	95.0	94.0	94.0
	OS	**99.0**	98.2	100.0	96.6	96.6	95.0	94.0	94.0	94.0
FCR CI	FE		97.3	94.2	92.6	91.8	97.4	93.2	95.8	91.3
	CE		98.4	94.8	97.5	95.8	96.8	96.0	97.2	94.8
	TS	**99.0**	96.6	98.2	96.6	96.6	90.9	91.0	94.0	94.0
	OS	**99.0**	96.6	96.8	96.6	96.6	90.9	91.0	94.0	94.0

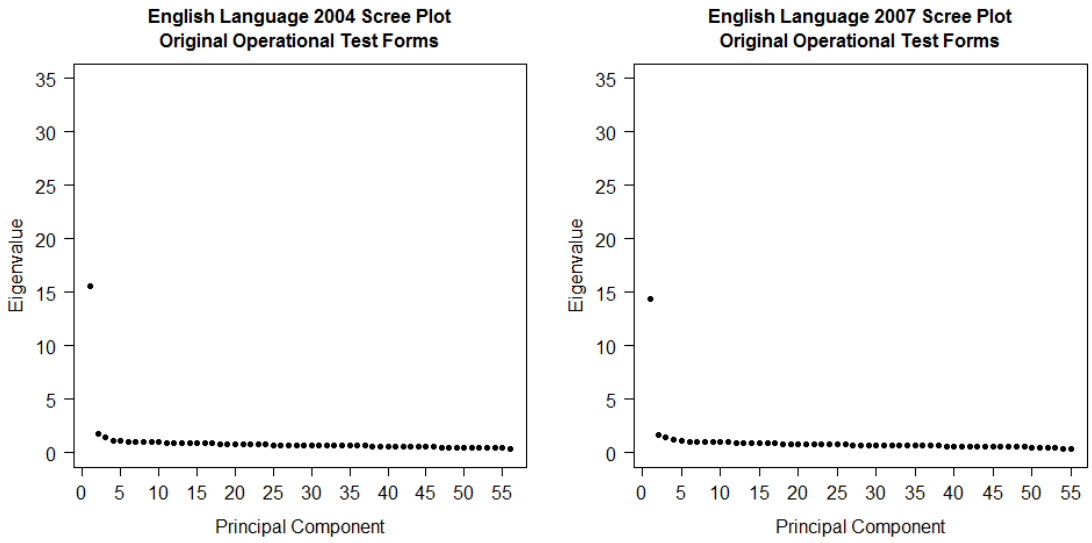

Figure 4-1. English Language operational test form scree plots.

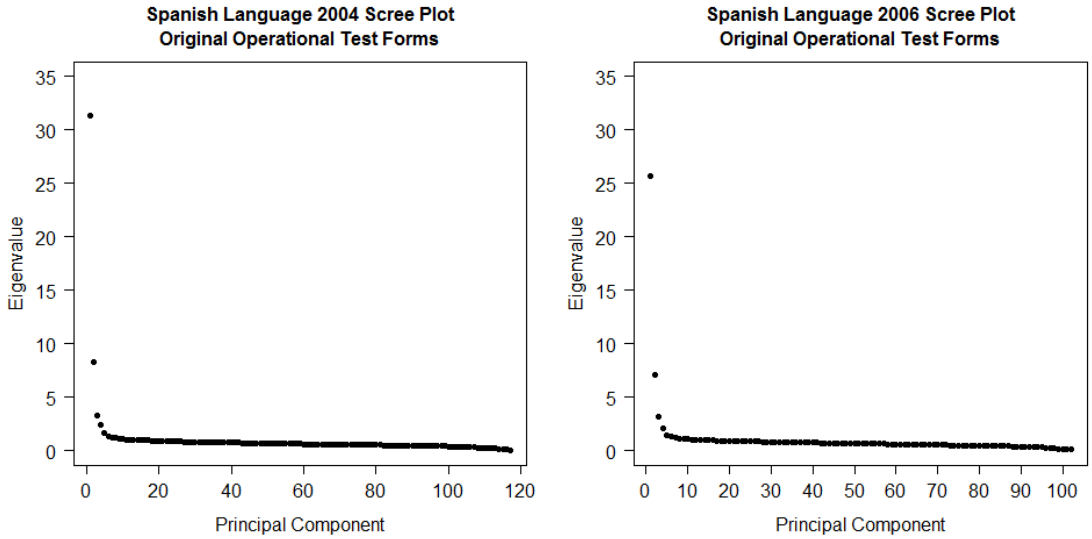

Figure 4-2. Spanish Language operational test form scree plots.

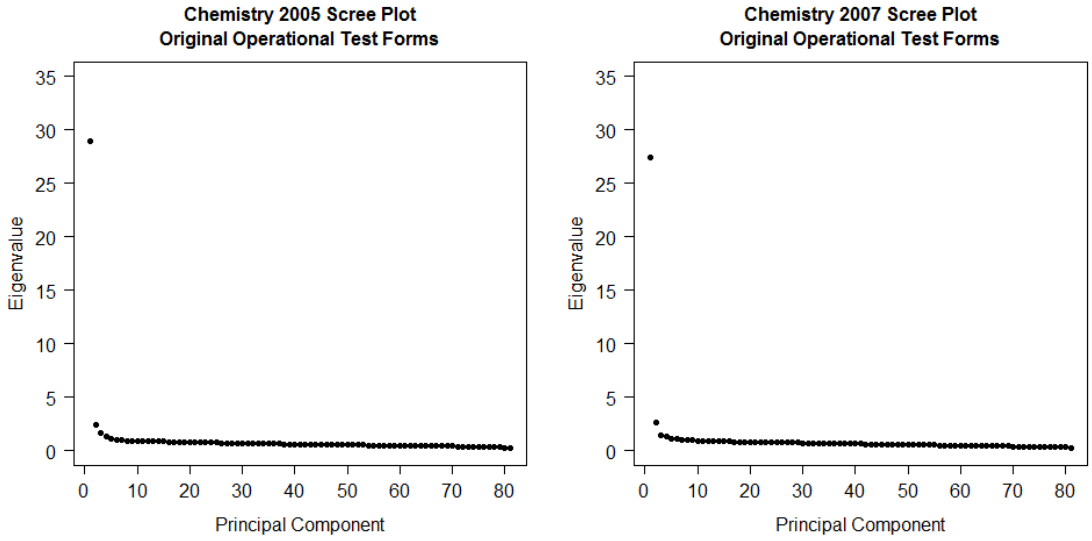

Figure 4-3. Chemistry operational test form scree plots.

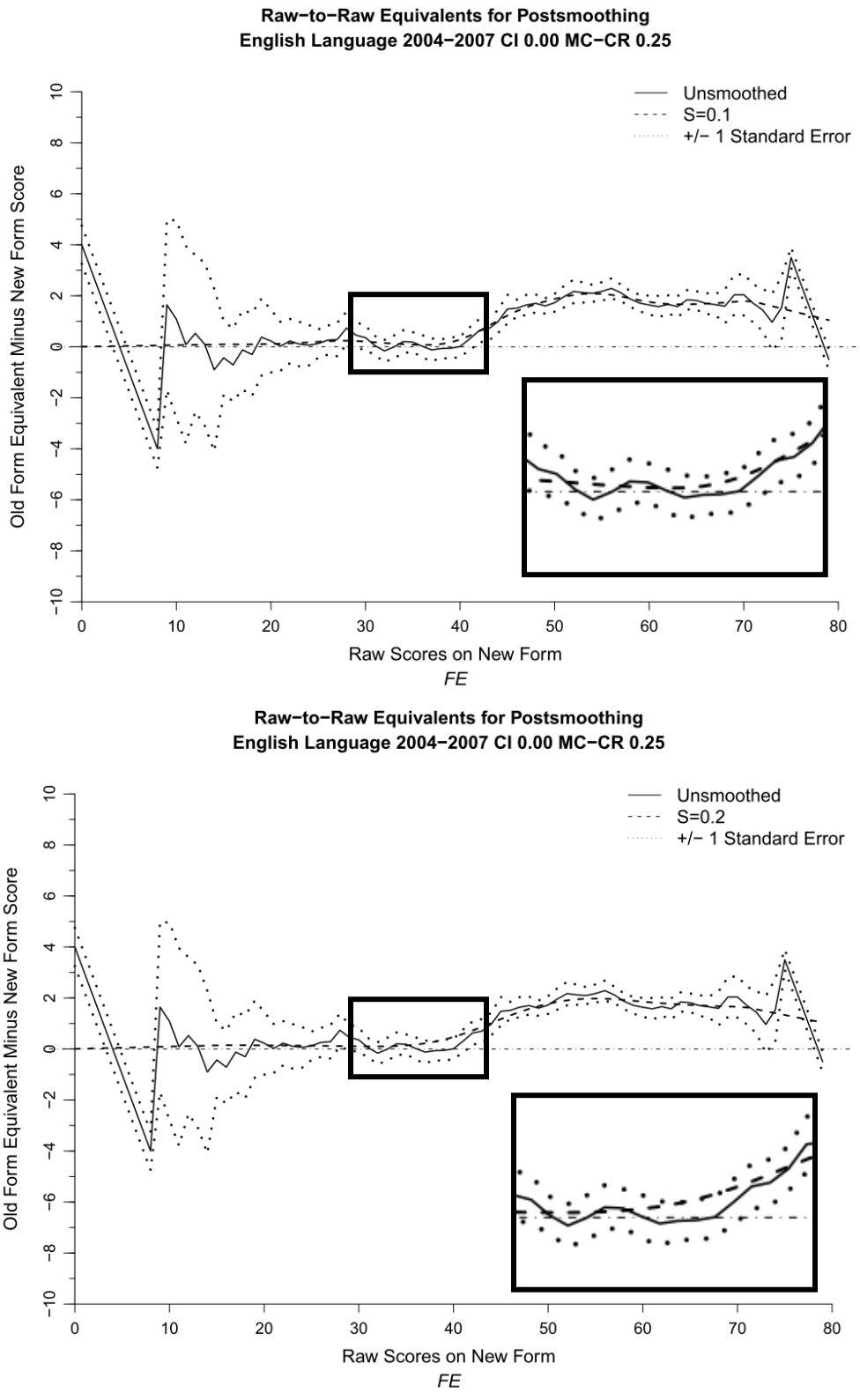

Figure 4-4. English Language operational test form smoothing for FE (CI 0.00 MC-CR 0.25).

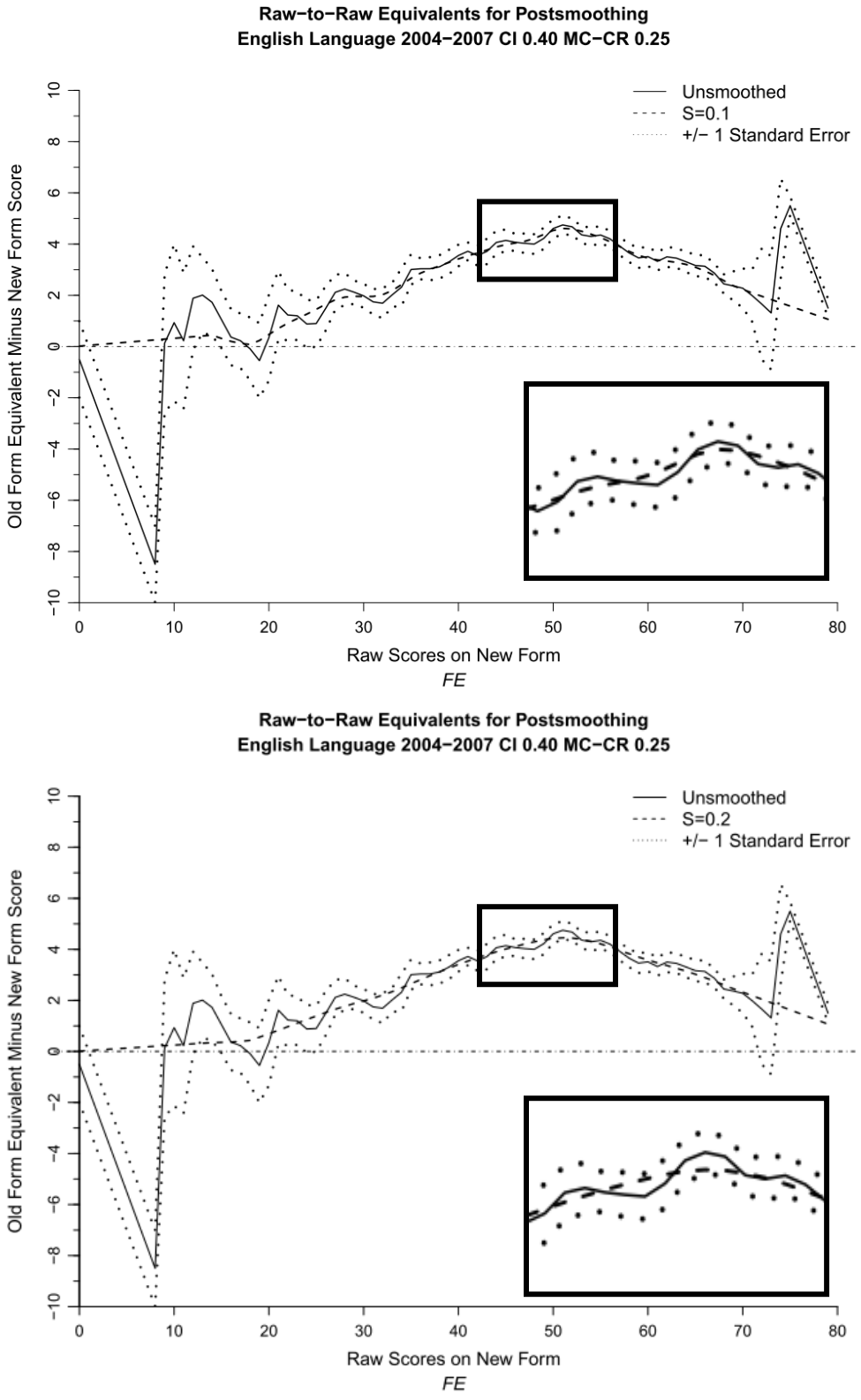

Figure 4-5. English Language operational test form smoothing for FE (CI 0.40 MC-CR 0.25).

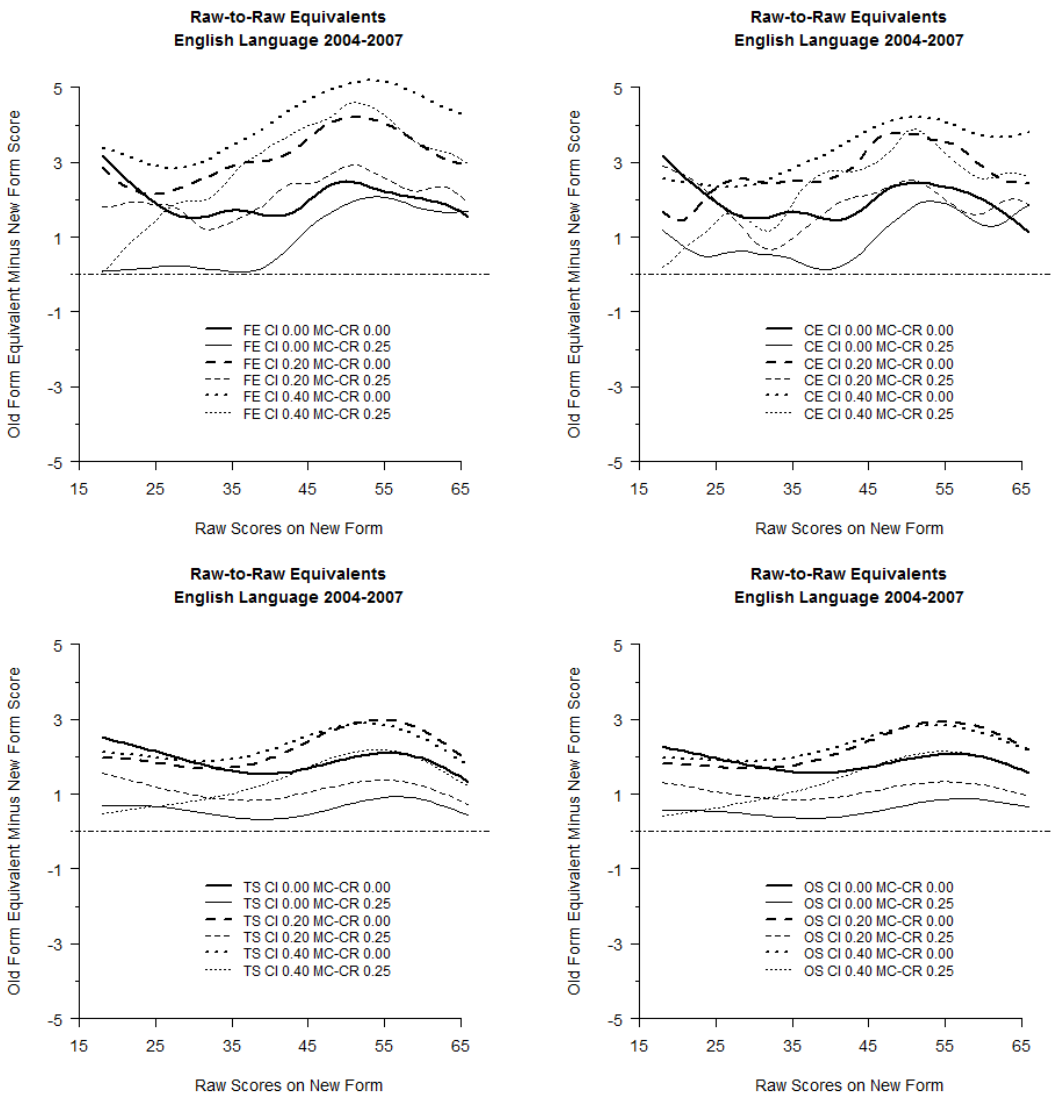

Figure 4-6. Equating relationships for English Language 2004-2007 by equating method.

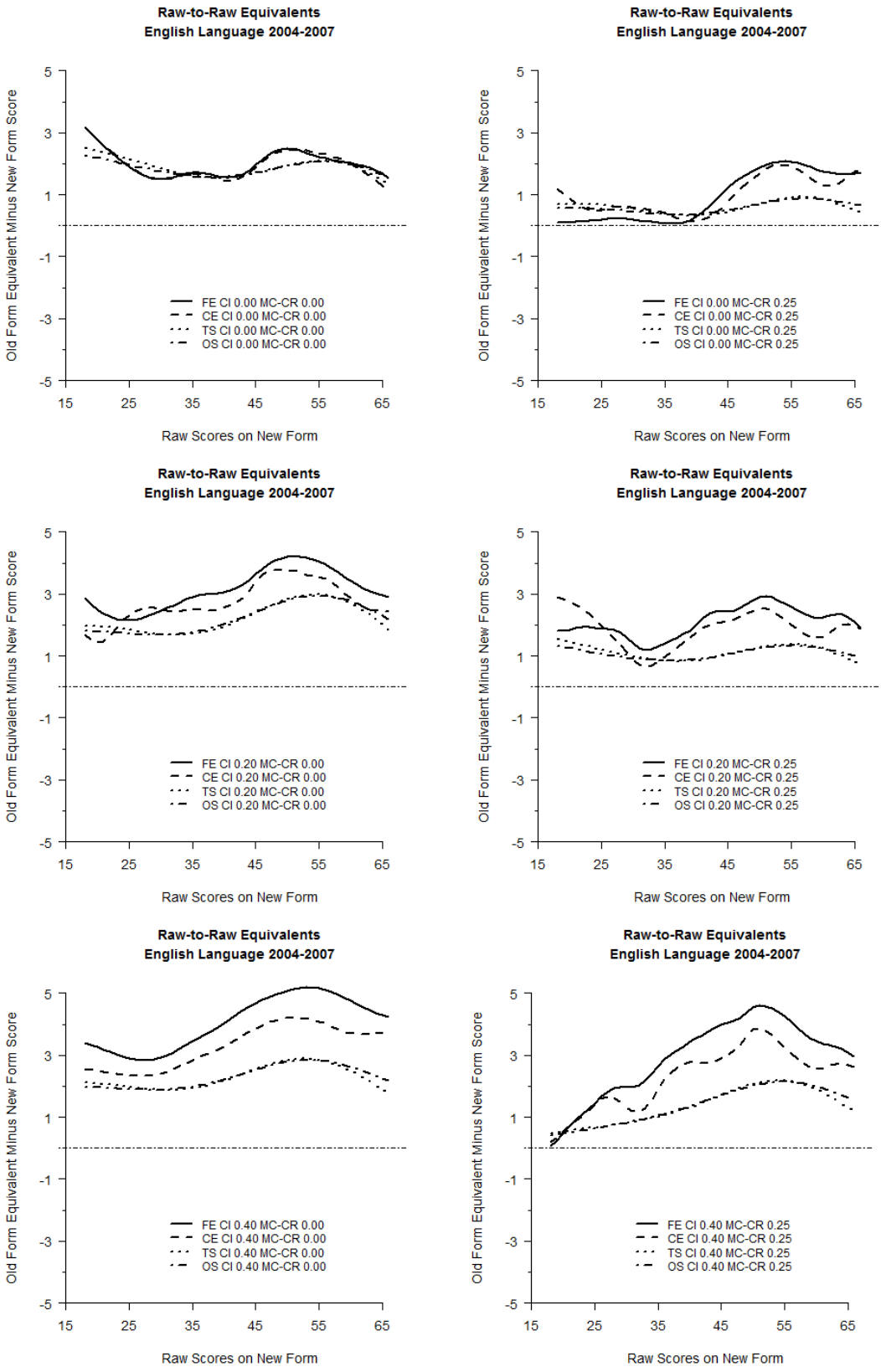

Figure 4-7. Equating relationships for English Language 2004-2007 by sample.

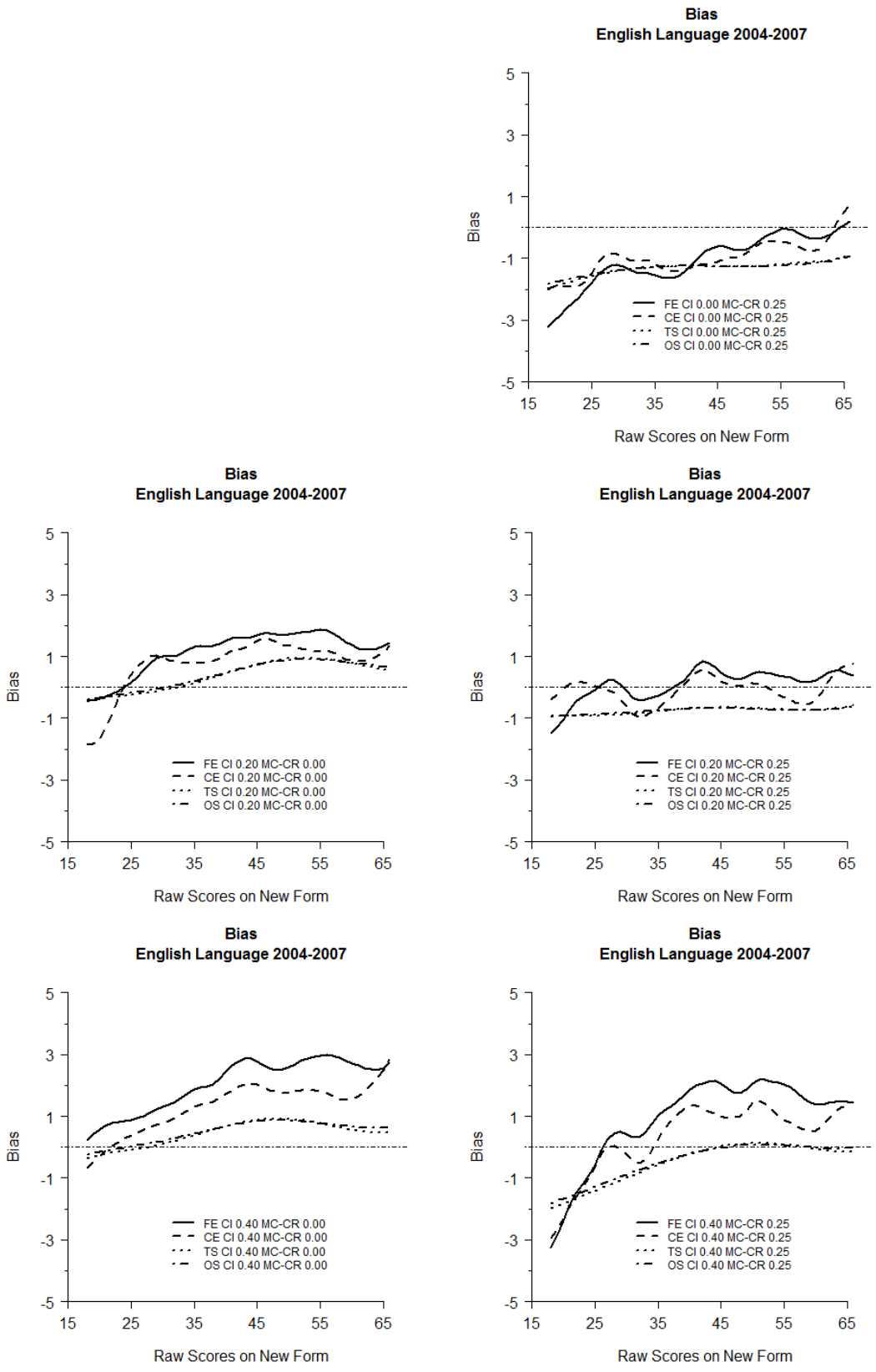

Figure 4-8. Conditional bias for English Language 2004-2007.

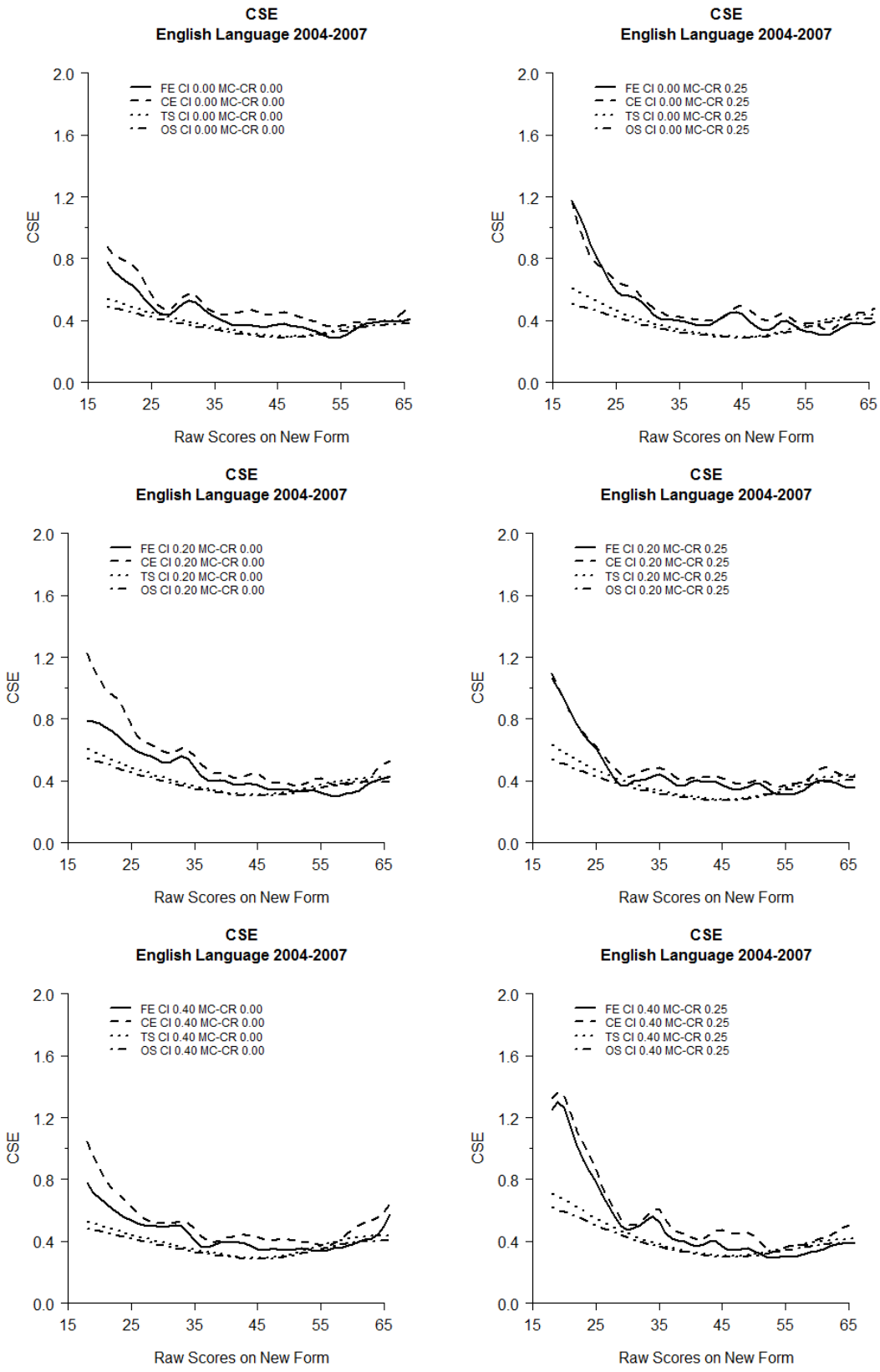

Figure 4-9. CSE for English Language 2004-2007.

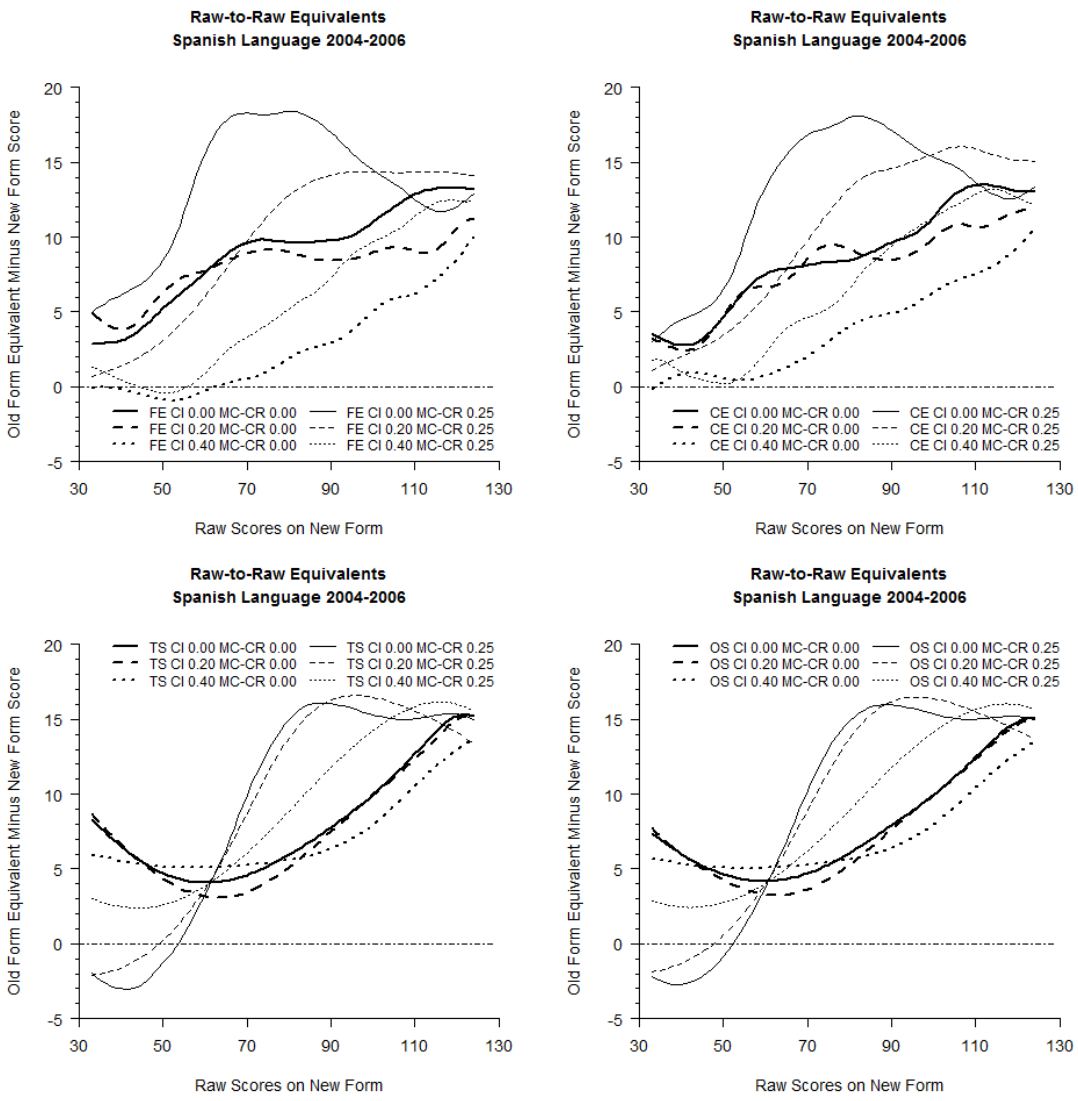

Figure 4-10. Equating relationships for Spanish Language 2004-2006 by equating method.

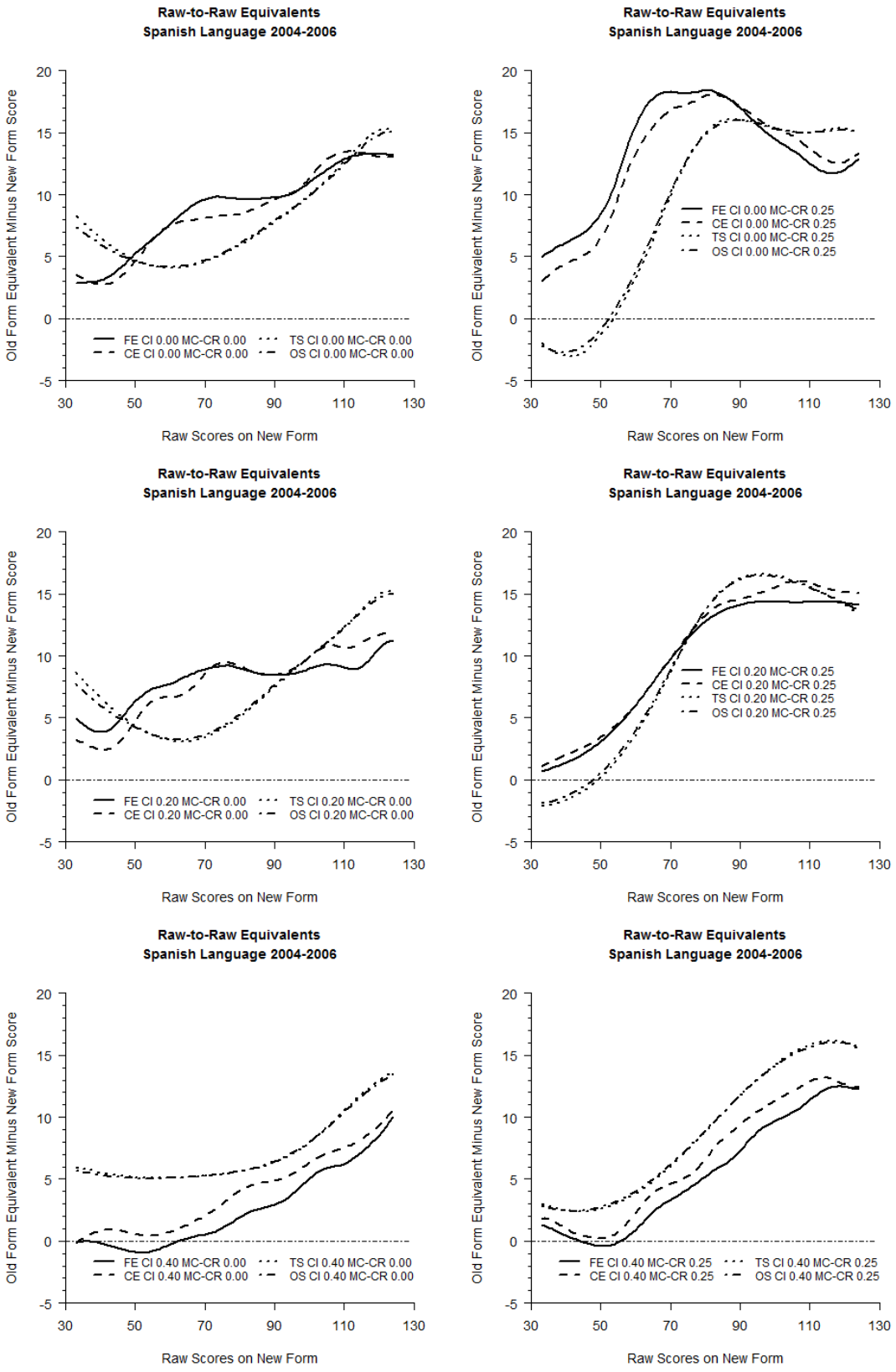

Figure 4-11. Equating relationships for Spanish Language 2004-2006 by sample.

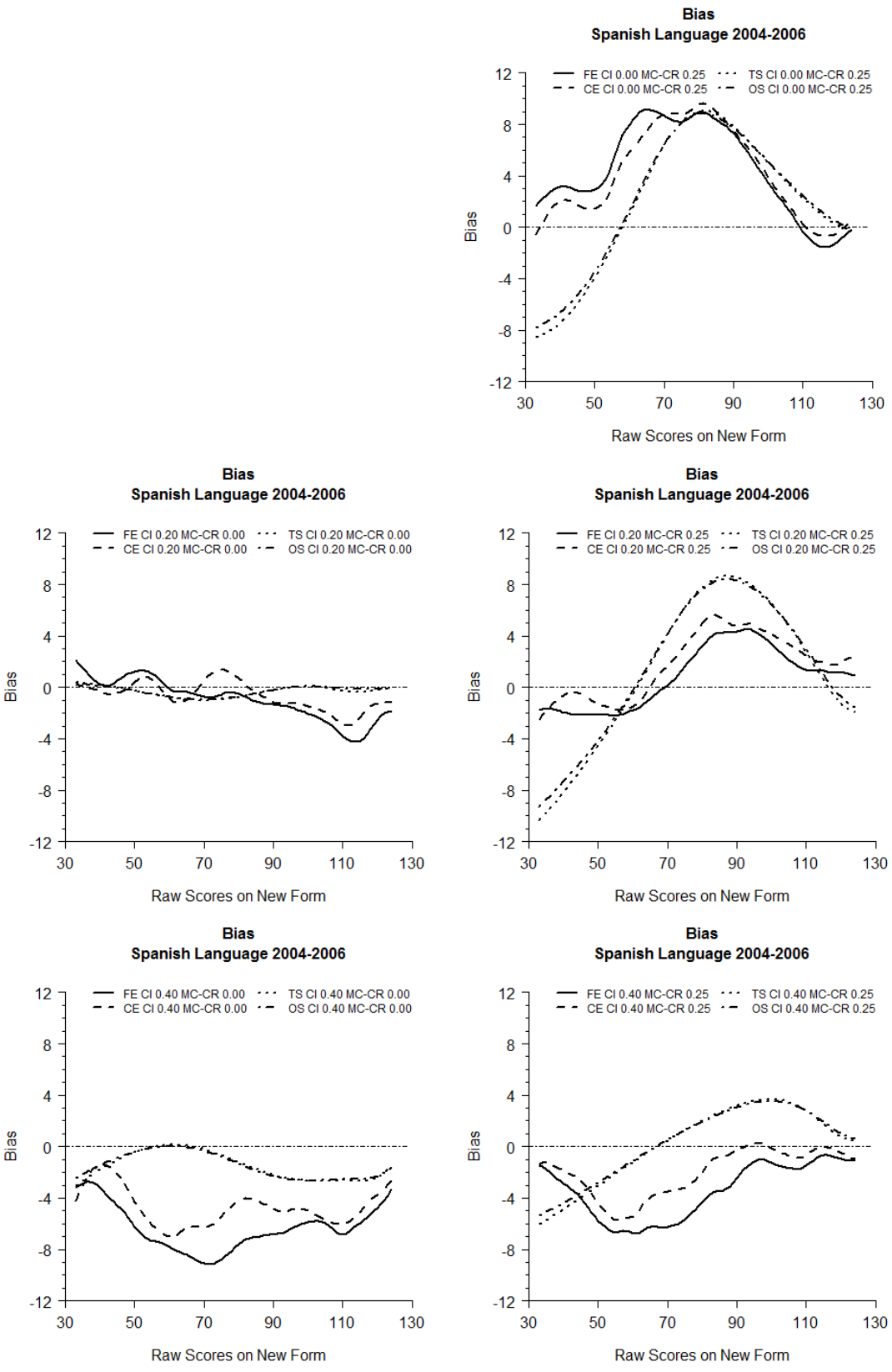

Figure 4-12. Conditional bias for Spanish Language 2004-2006.

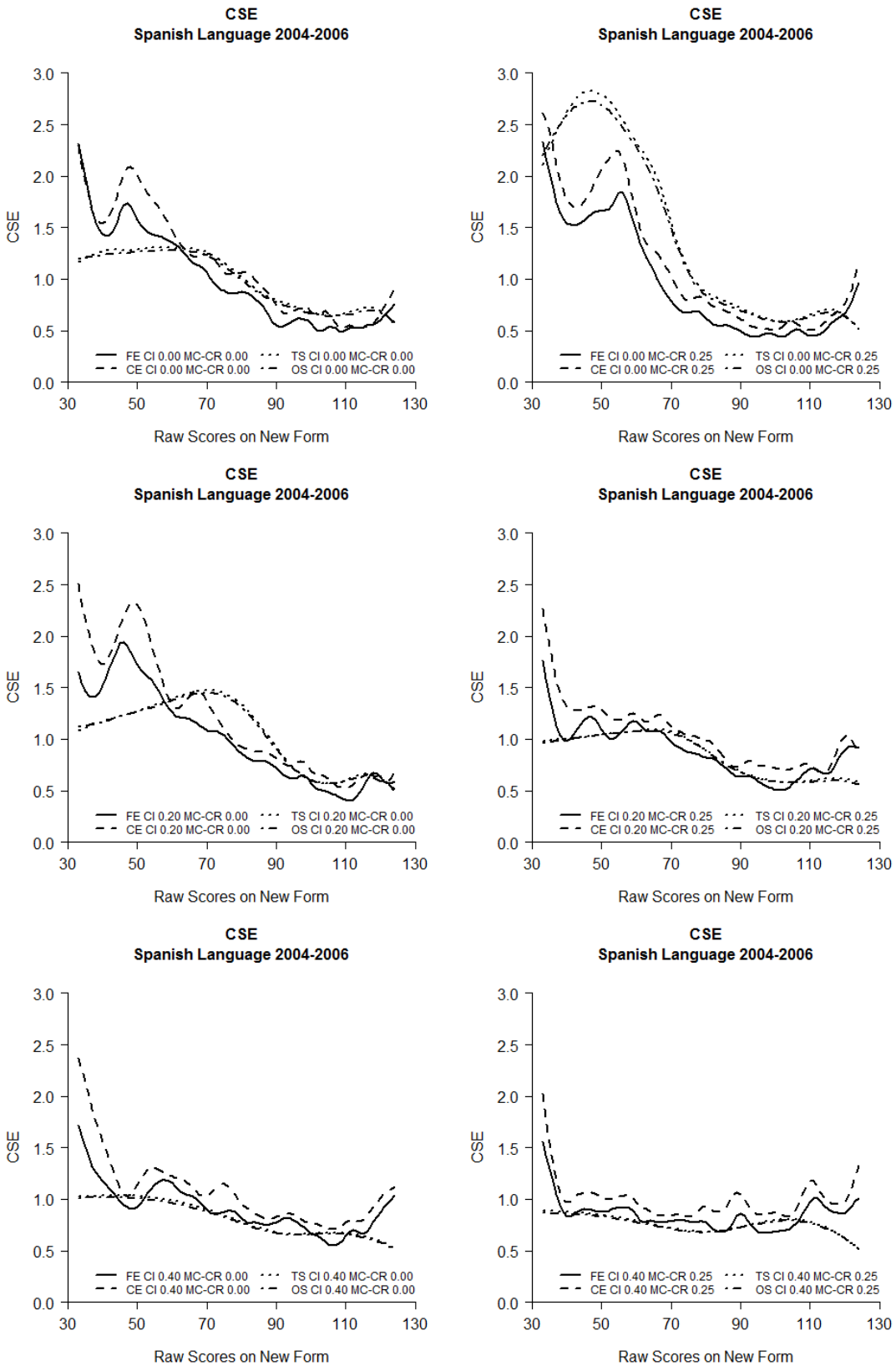

Figure 4-13. CSE for Spanish Language 2004-2006.

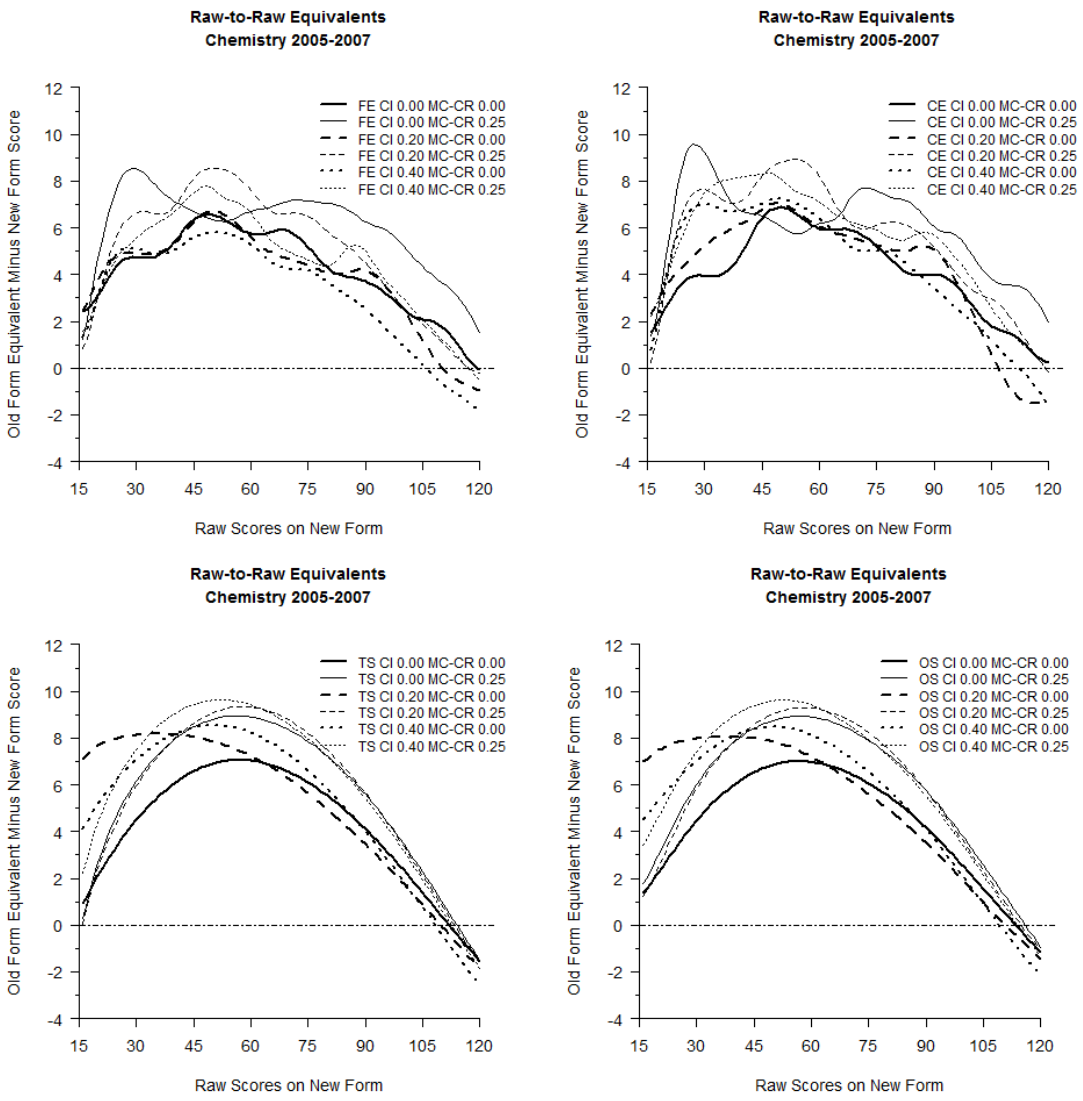

Figure 4-14. Equating relationships for Chemistry 2005-2007 by equating method.

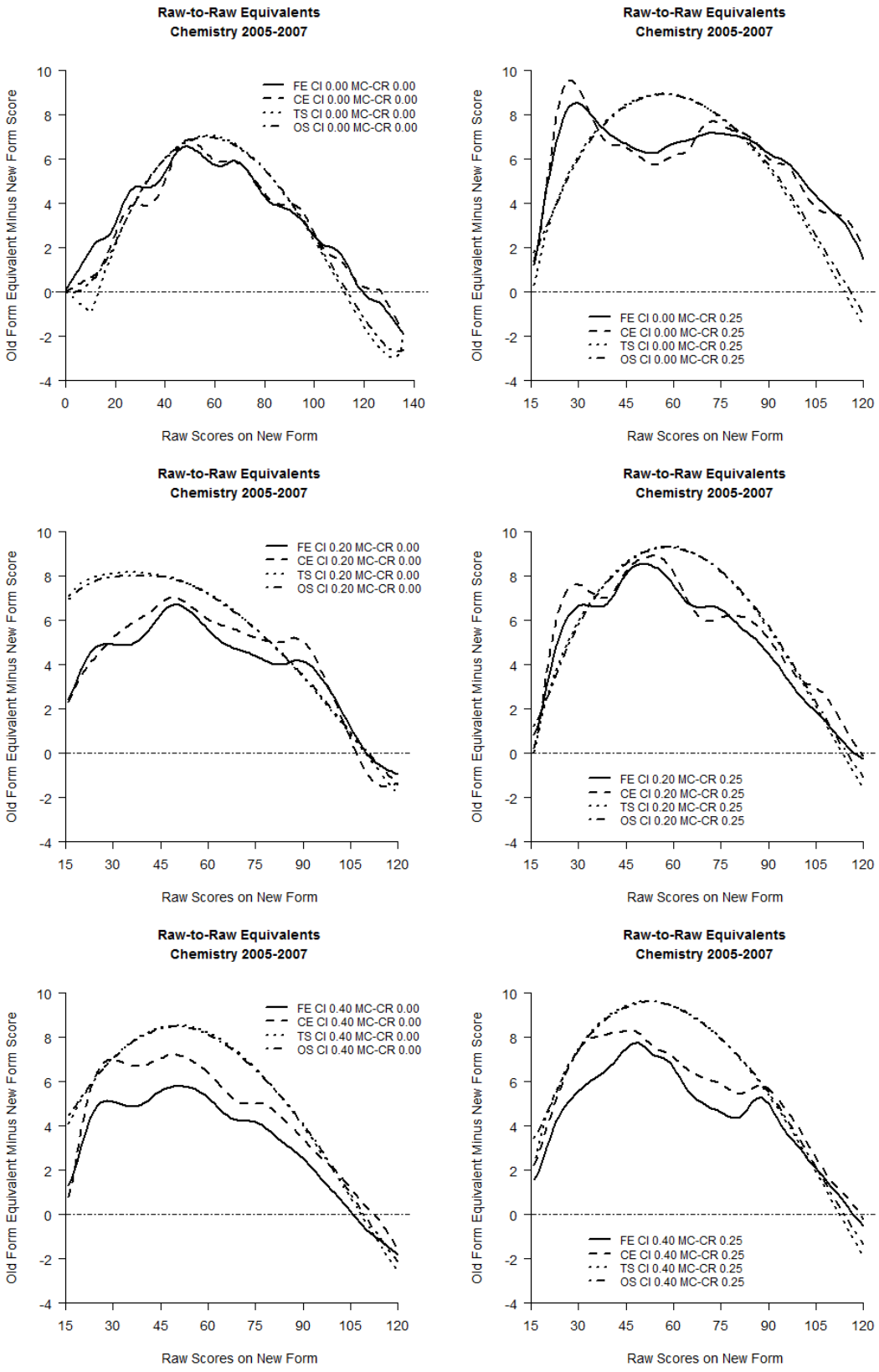

Figure 4-15. Equating relationships for Chemistry 2005-2007 by sample.

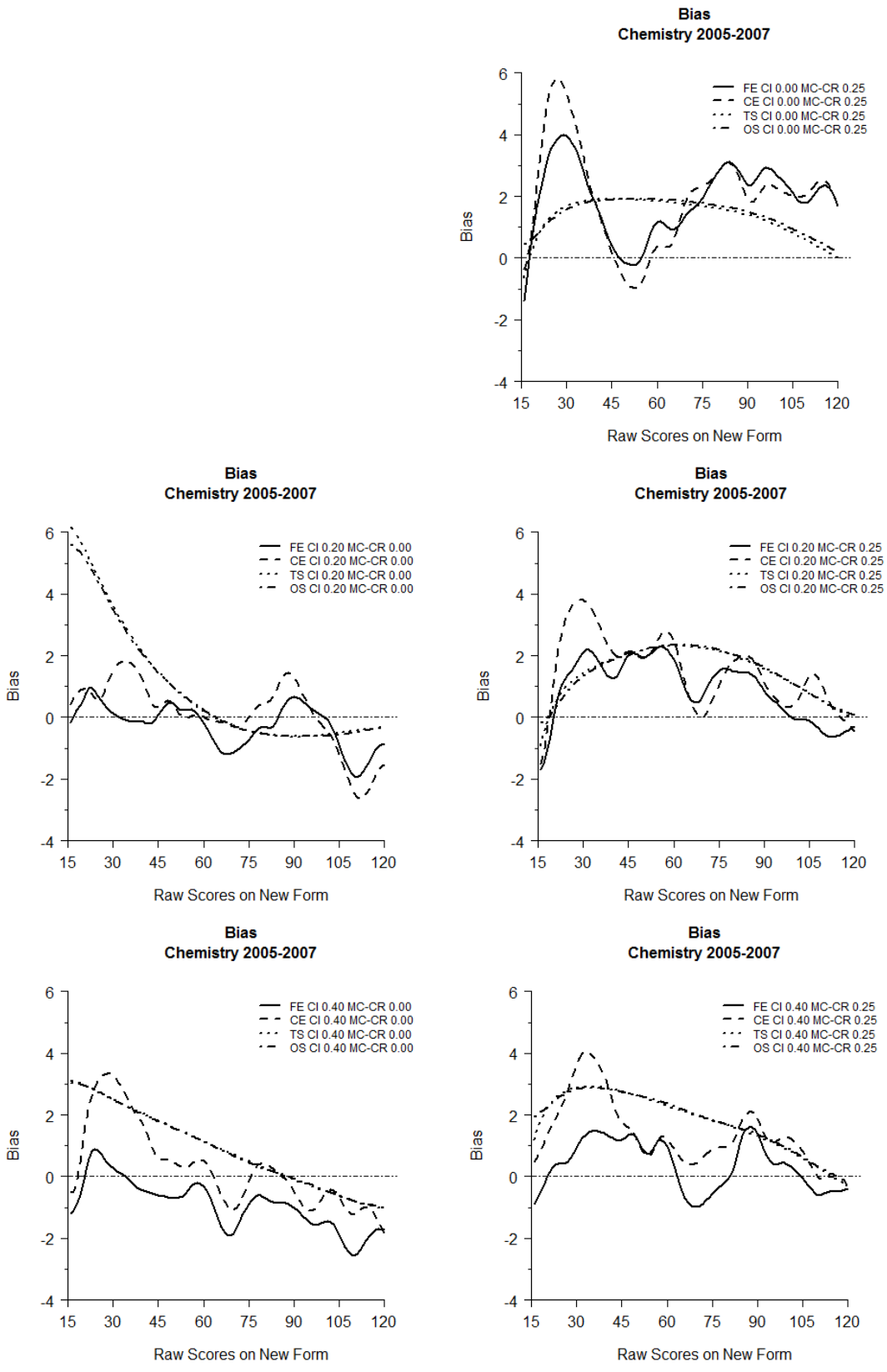

Figure 4-16. Conditional bias for Chemistry 2005-2007.

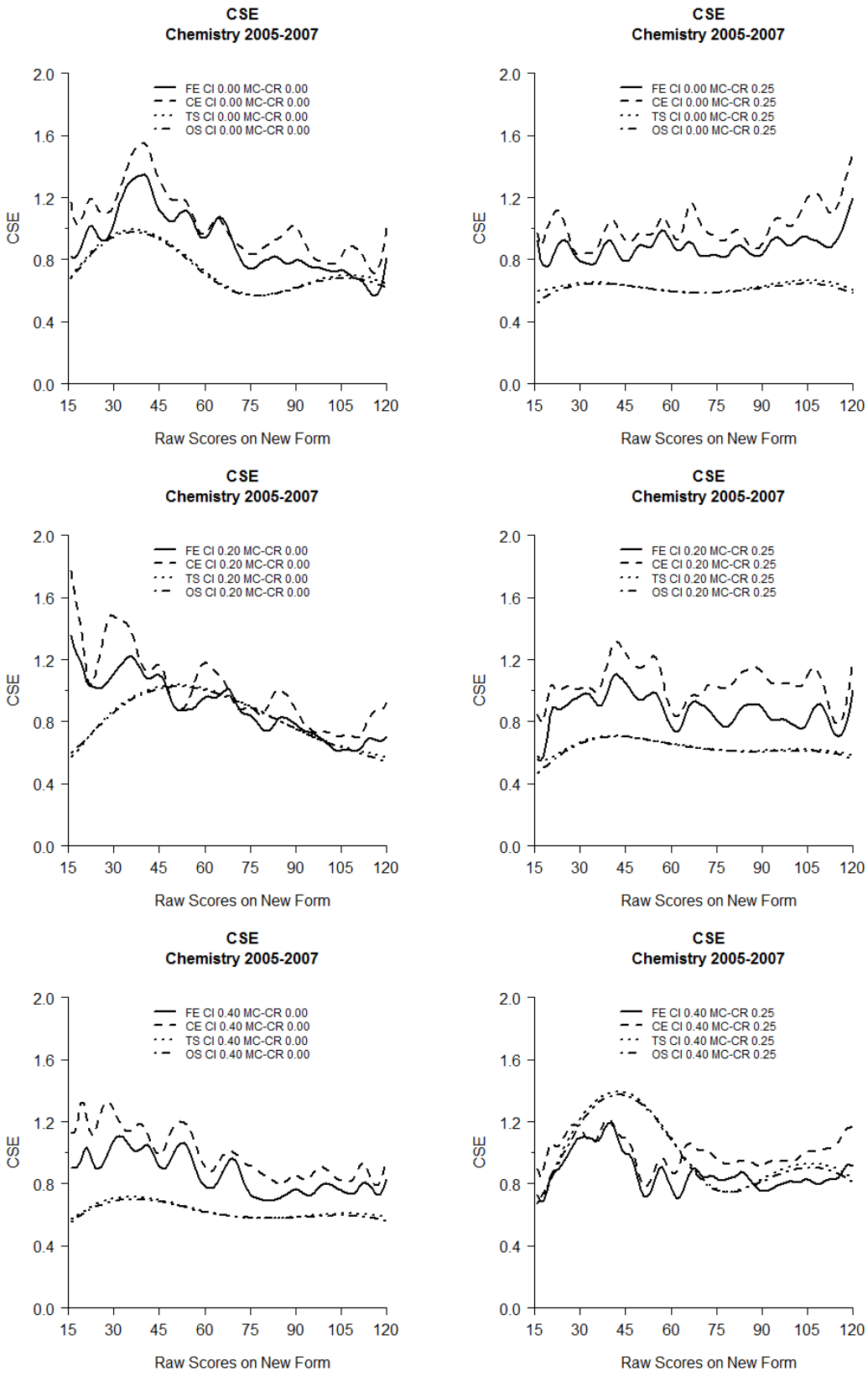

Figure 4-17. CSE for Chemistry 2005-2007.

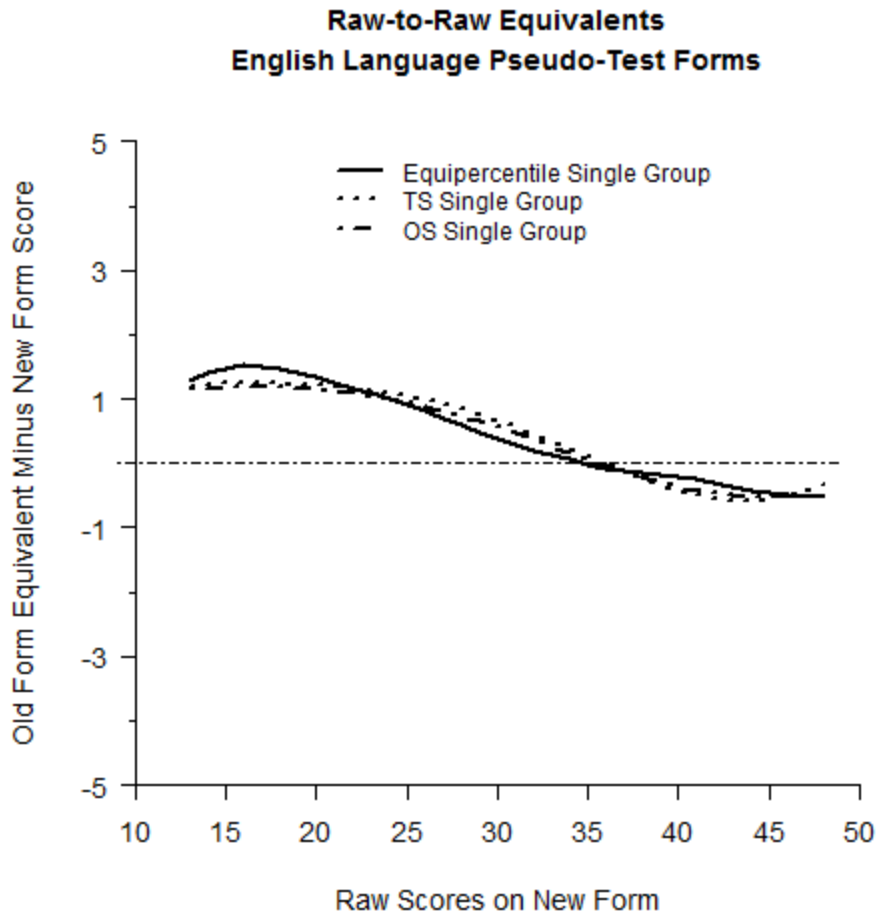

Figure 4-18. Criterion equating relationships for English Language pseudo-test forms (single group).

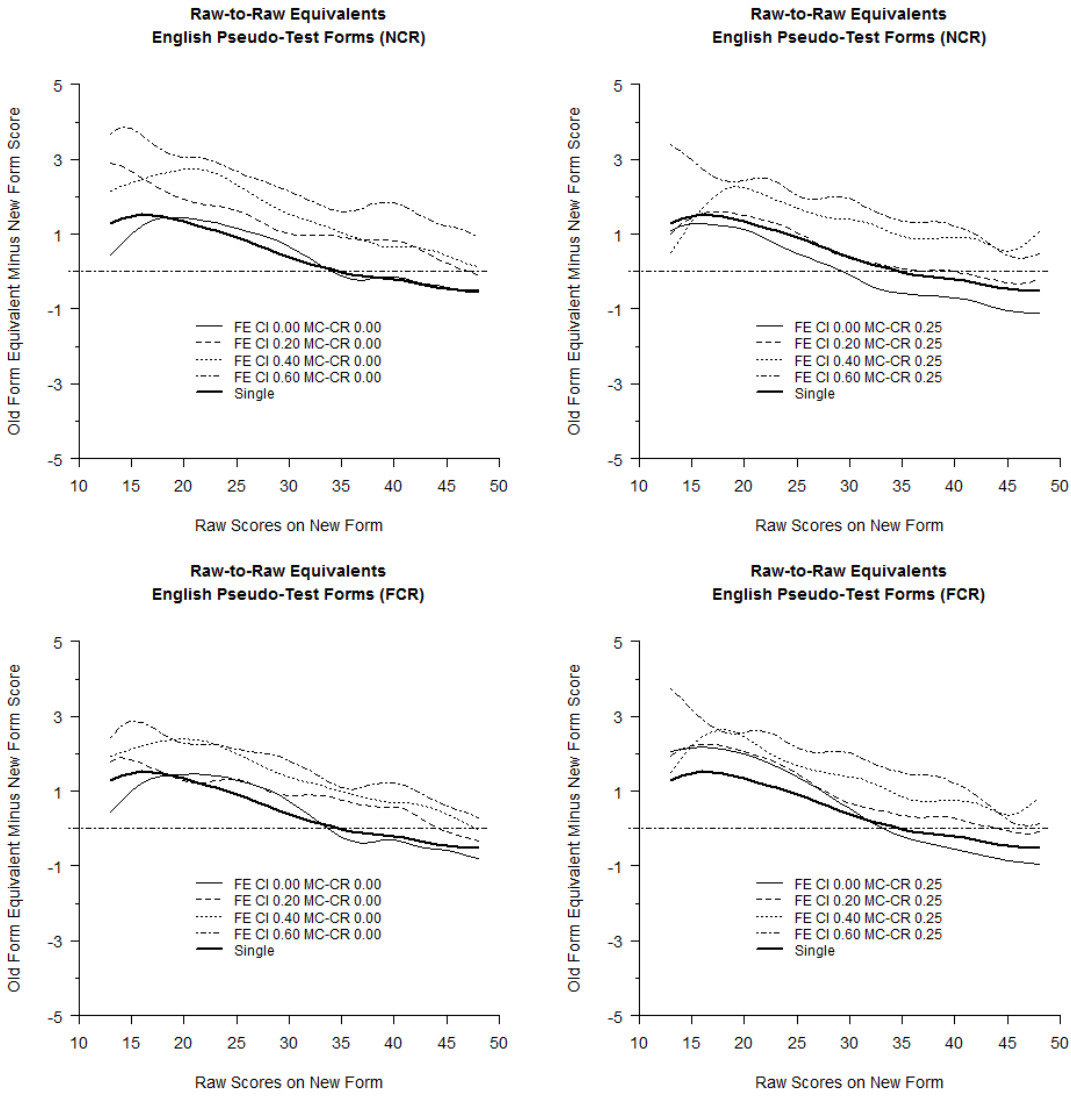

Figure 4-19. English Language pseudo-test form comparison of NCR and FCR for FE.

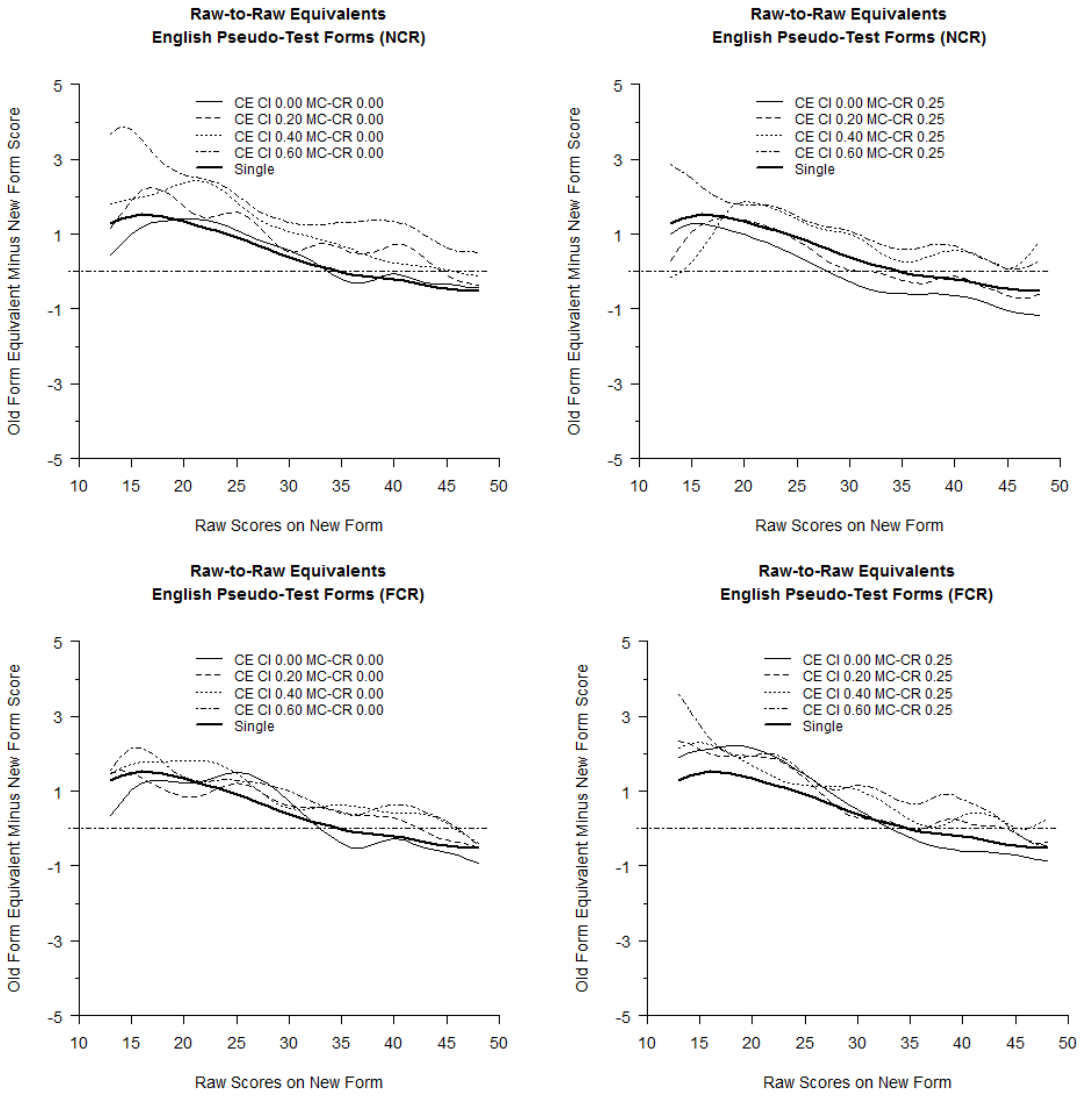

Figure 4-20. English Language pseudo-test form comparison of NCR and FCR for CE.

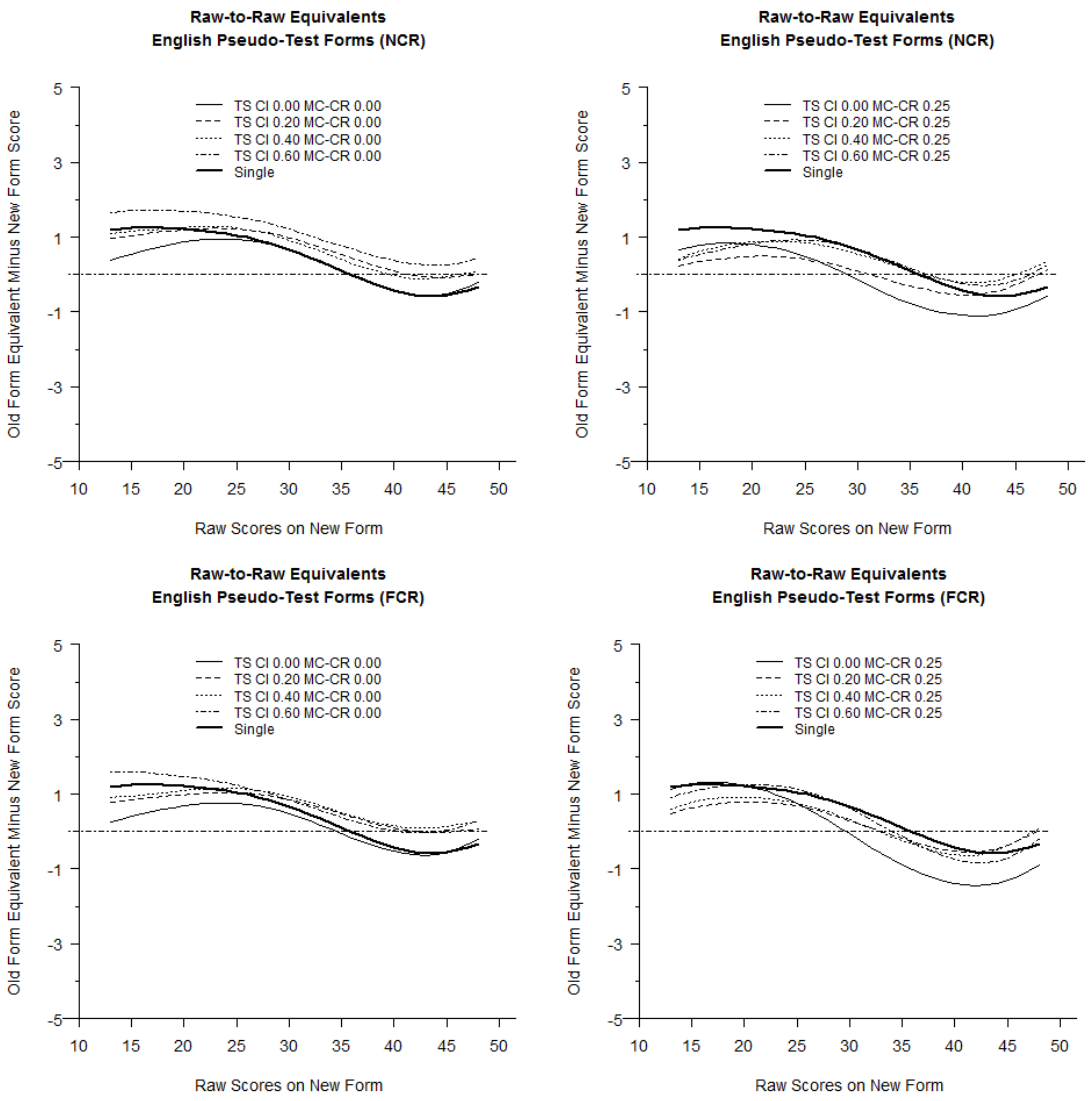

Figure 4-21. English Language pseudo-test form comparison of NCR and FCR for TS.

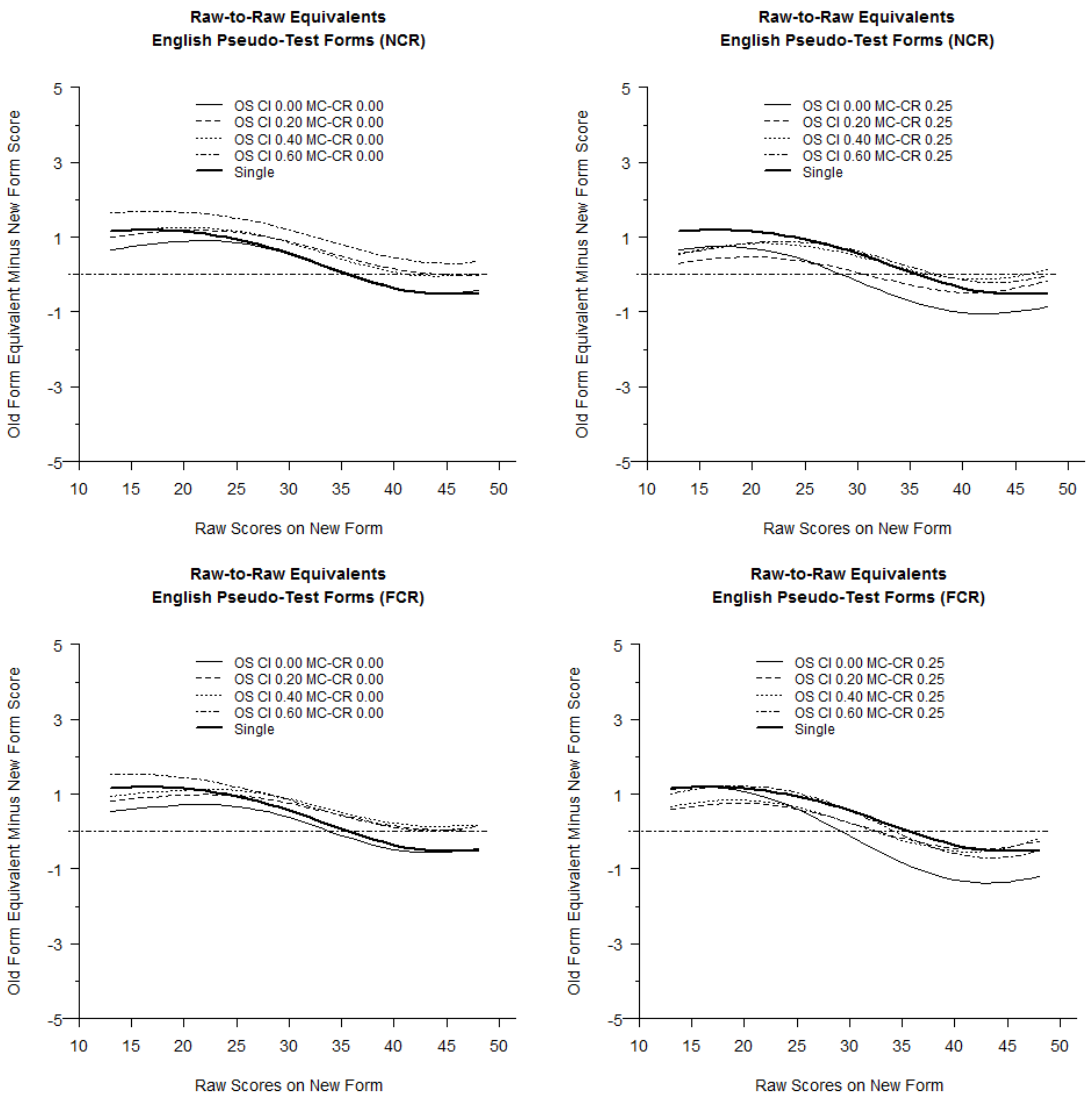

Figure 4-22. English Language pseudo-test form comparison of NCR and FCR for OS.

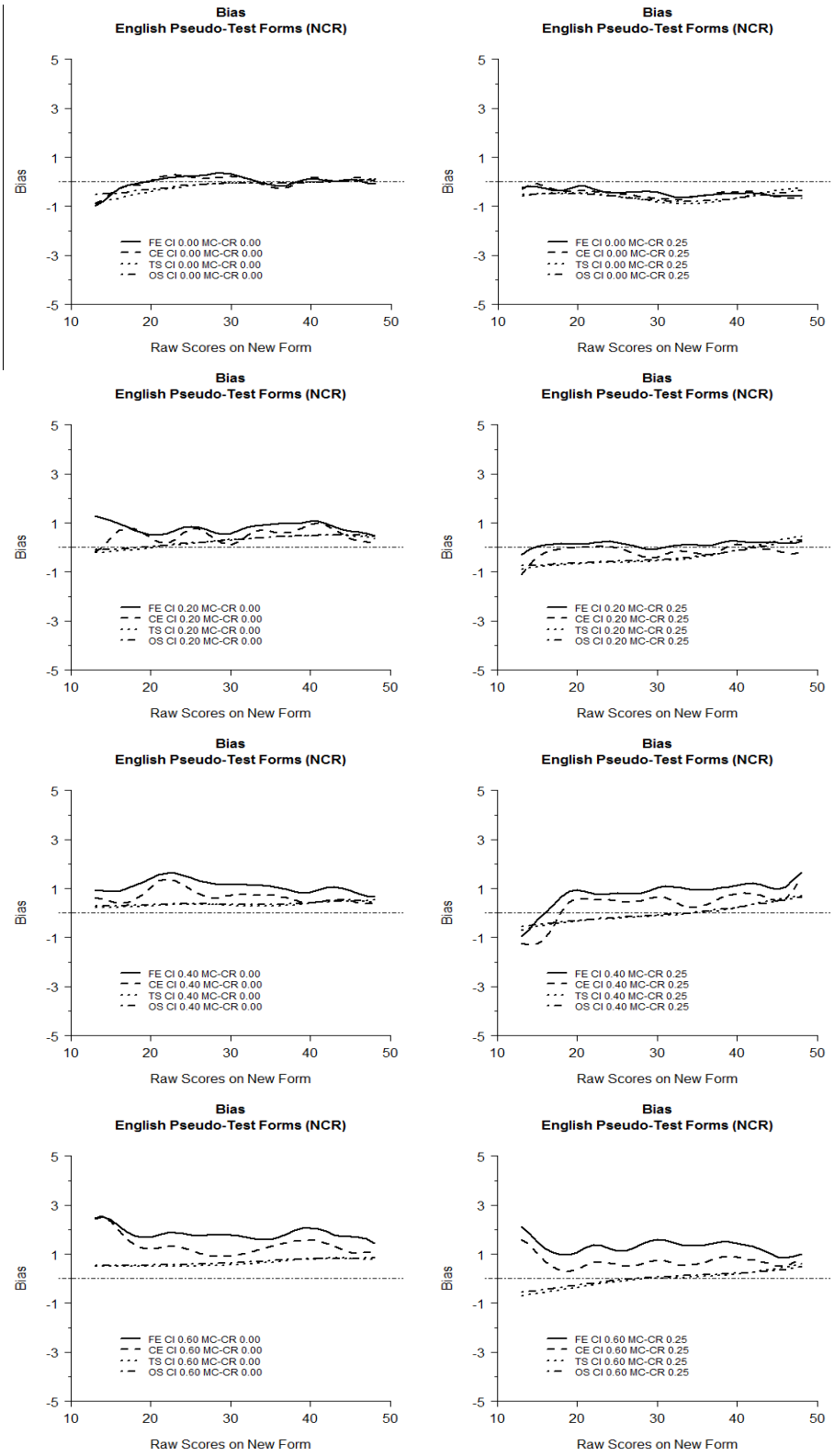

Figure 4-23. English Language pseudo-test form conditional bias for NCR.

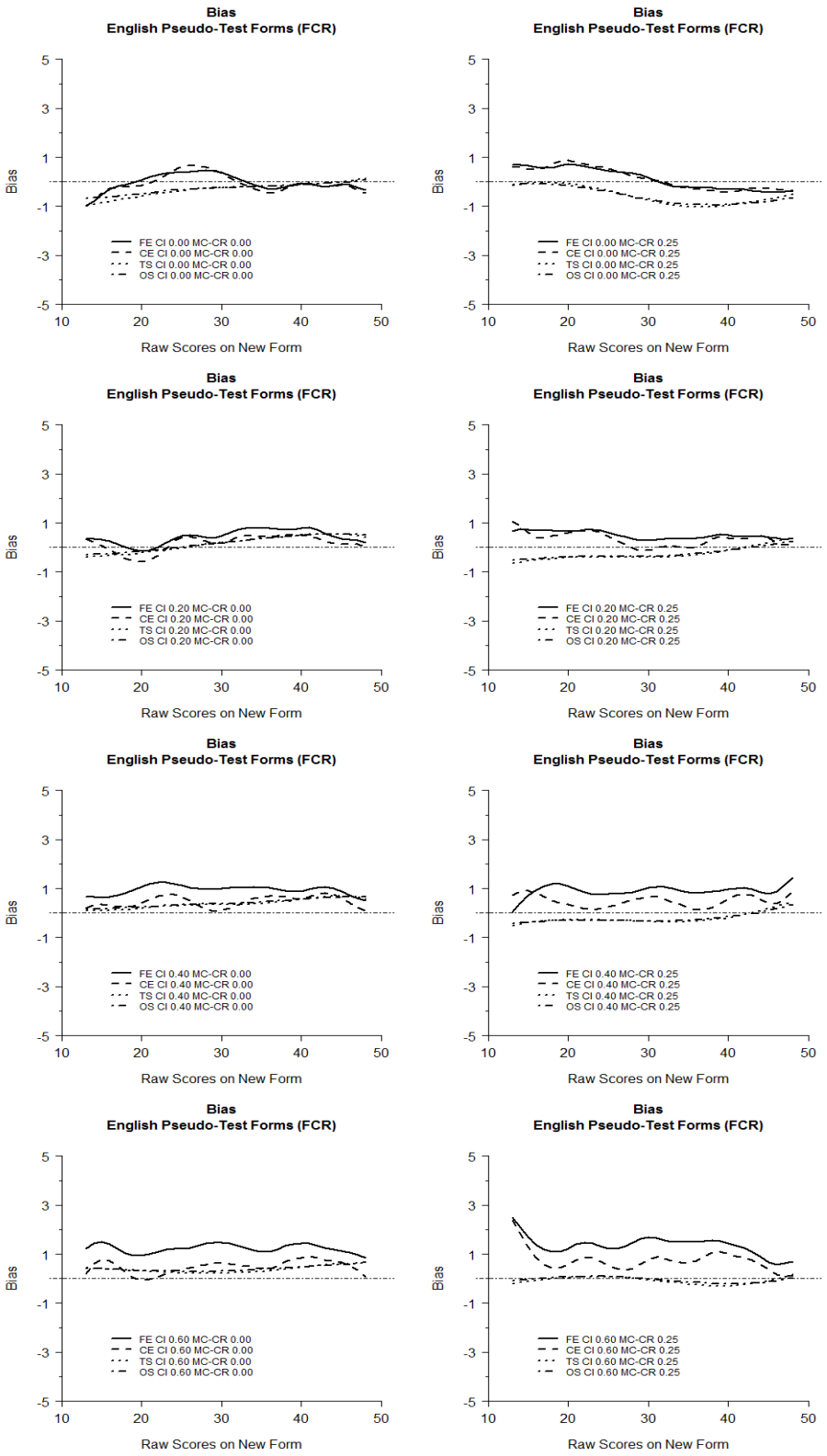

Figure 4-24. English Language pseudo-test form conditional bias for FCR.

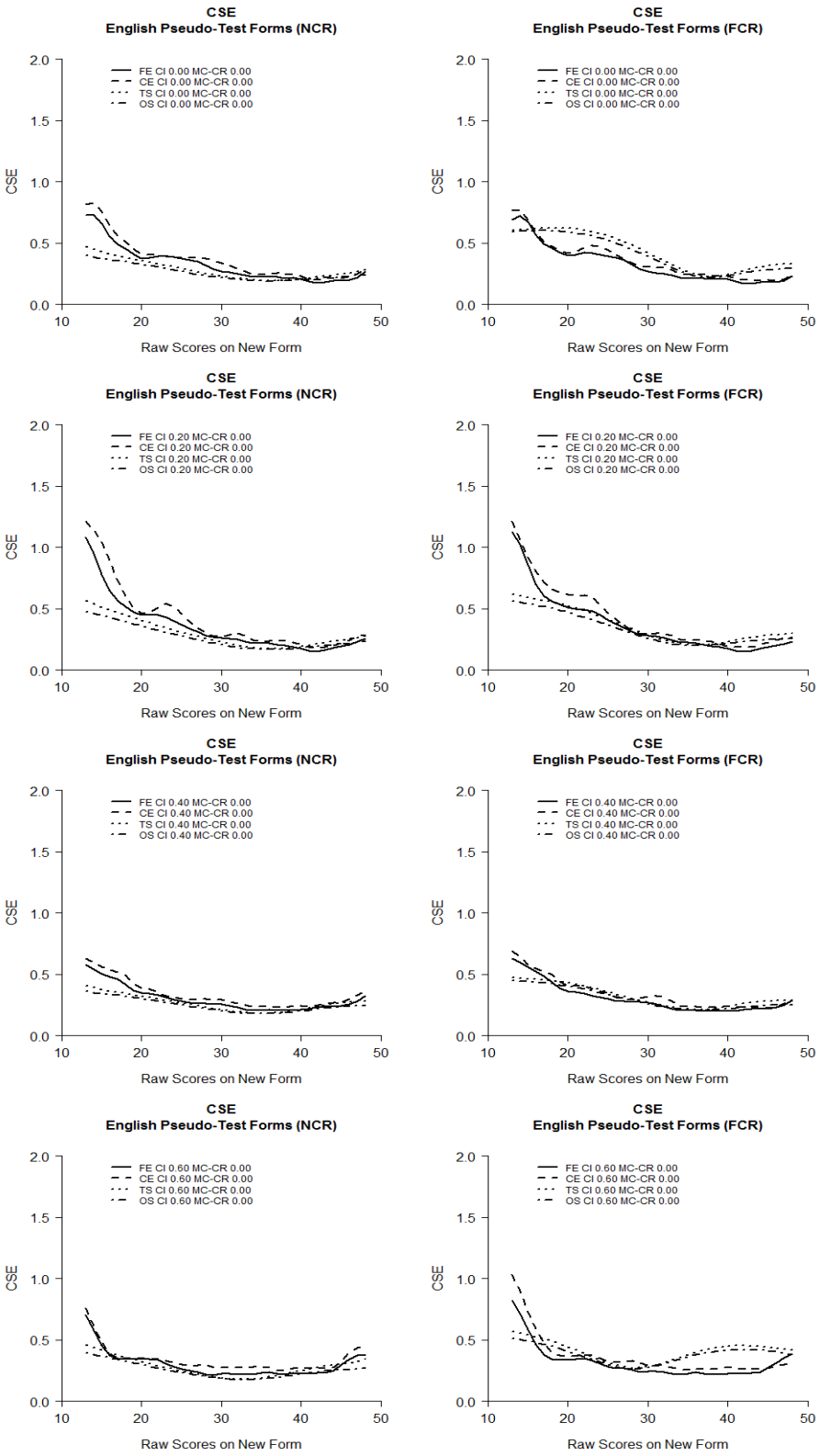

Figure 4-25. English Language pseudo-test forms CSE (NCR and FCR, MC-CR 0.00).

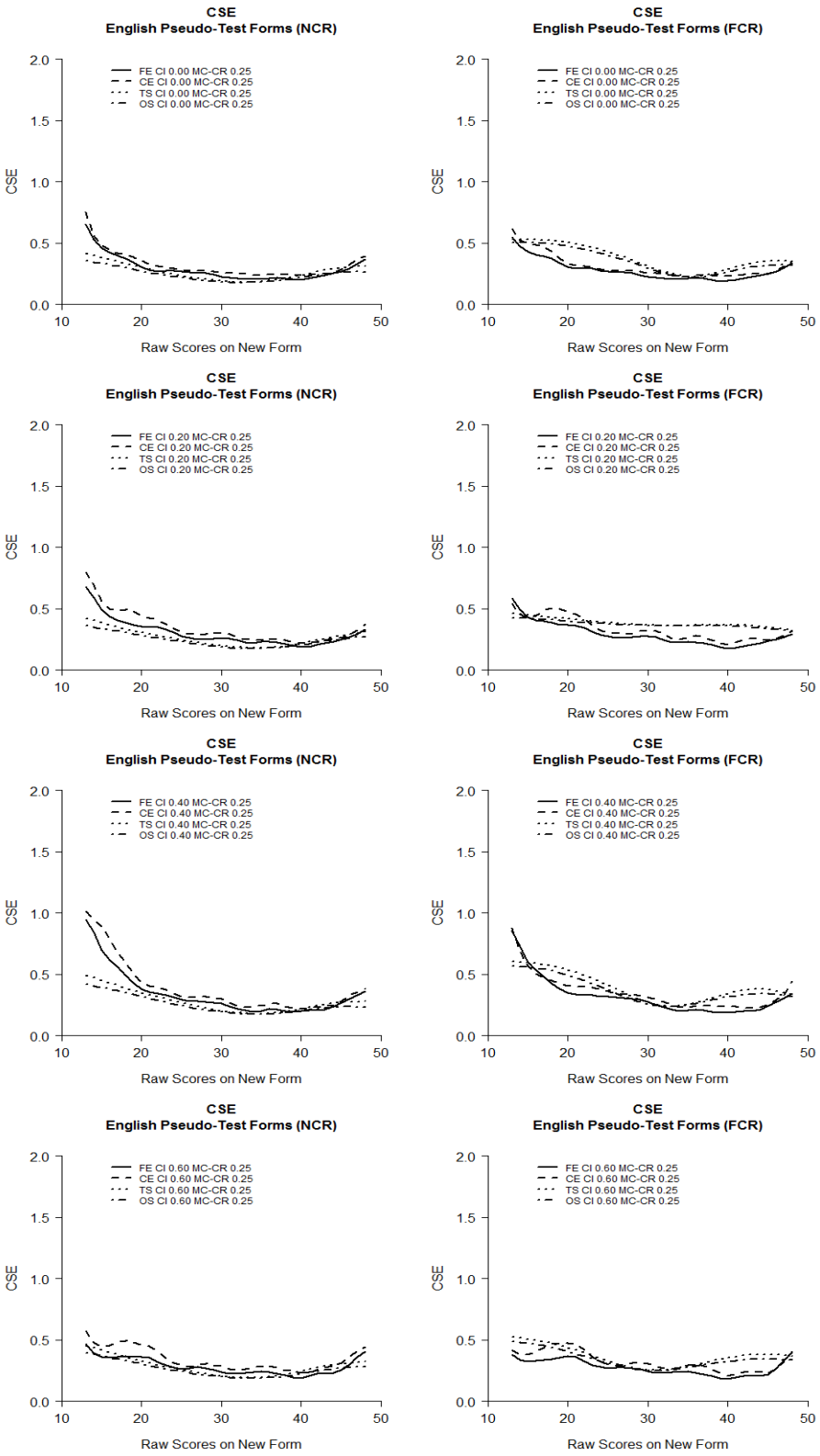

Figure 4-26. English Language pseudo-test forms CSE (NCR and FCR, MC-CR 0.25).

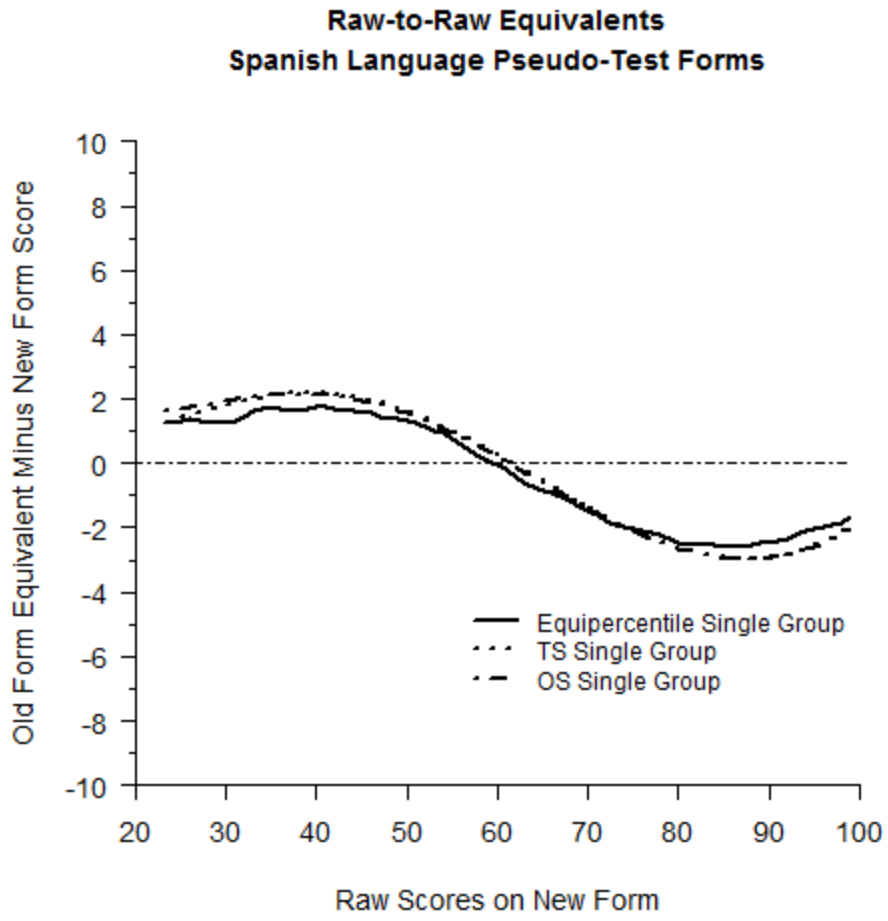

Figure 4-27. Criterion equating relationships for Spanish Language pseudo-test forms (single group).

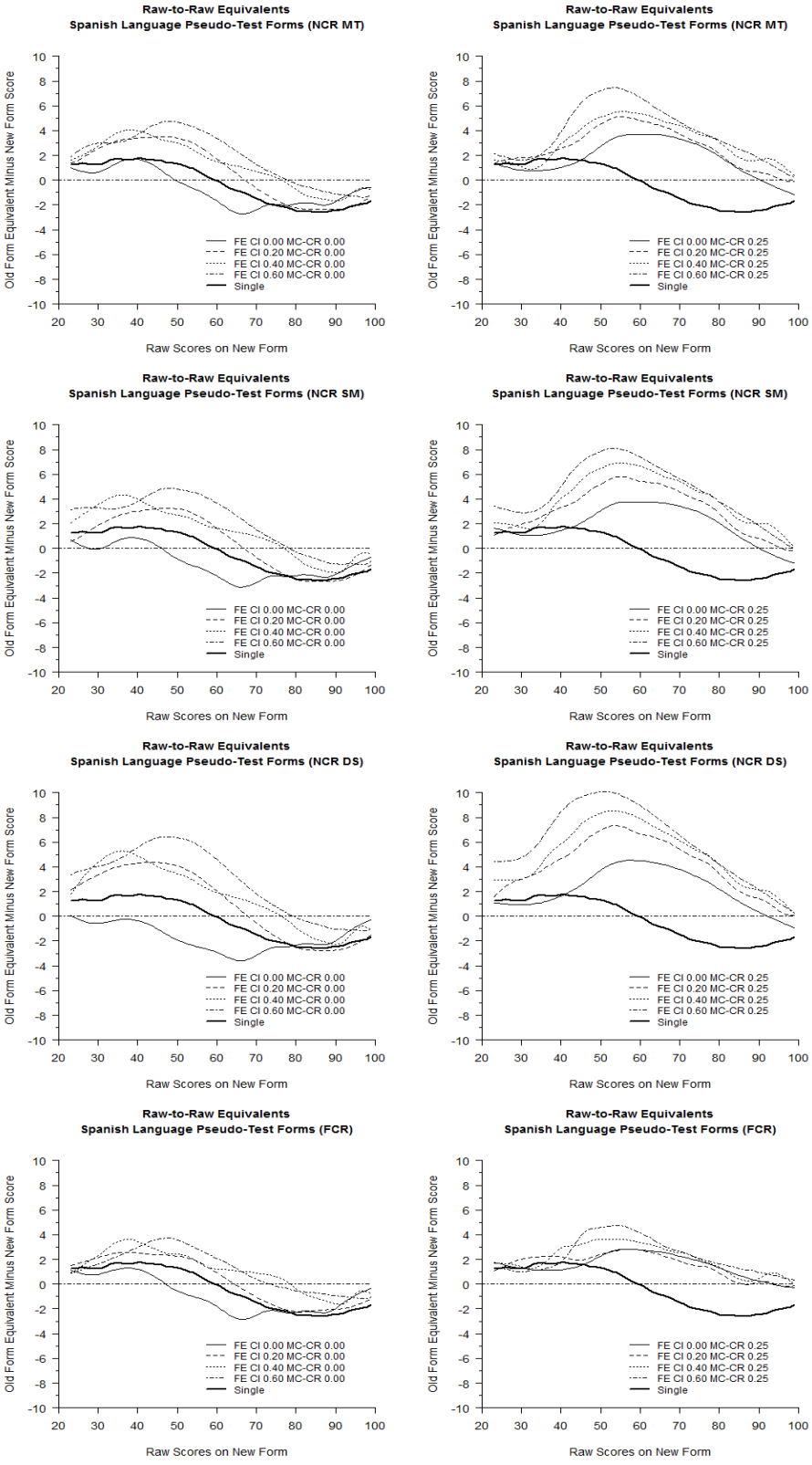

Figure 4-28. Spanish Language pseudo-test form comparison of CI ES for FE.

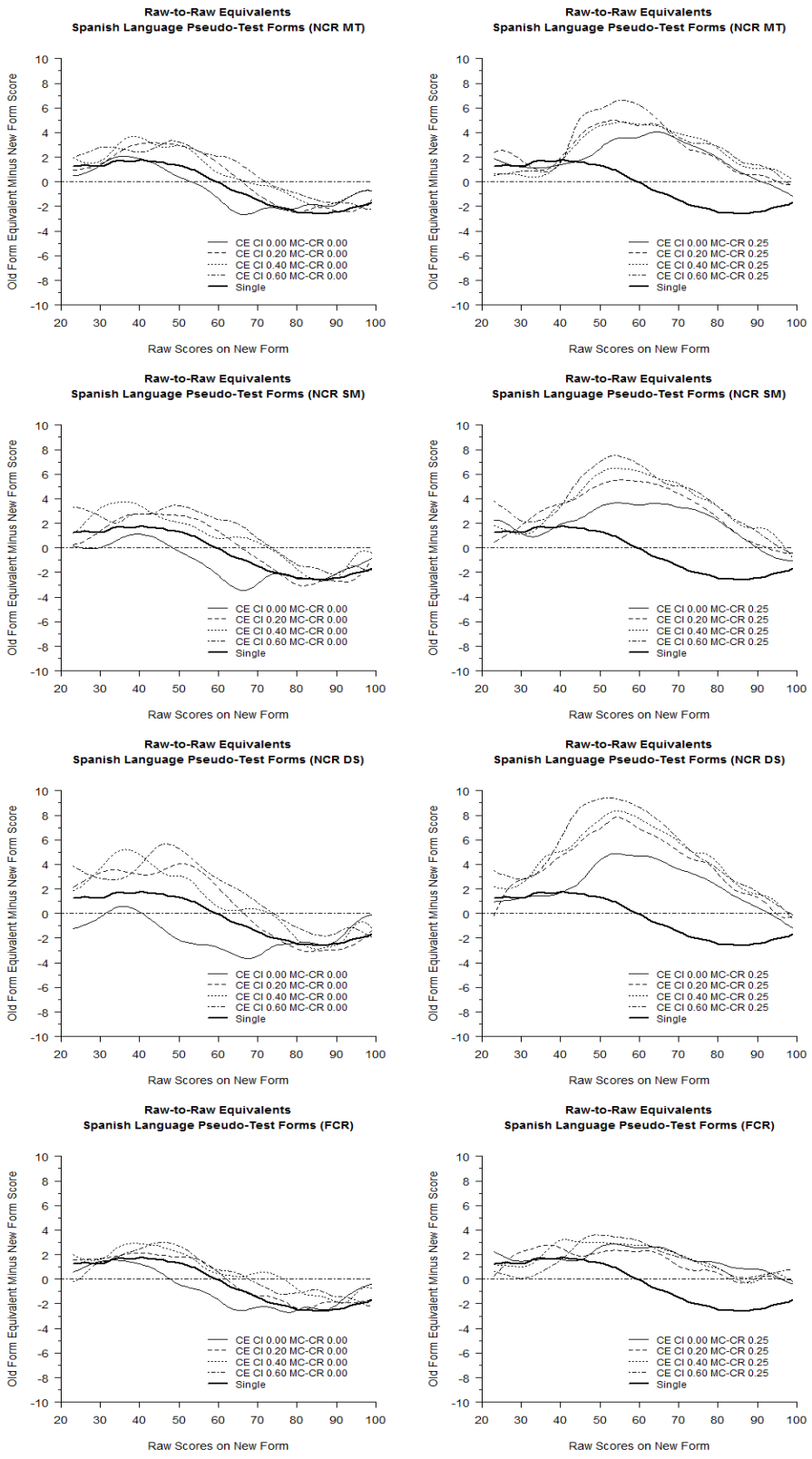

Figure 4-29. Spanish Language pseudo-test form comparison of CI ES for CE.

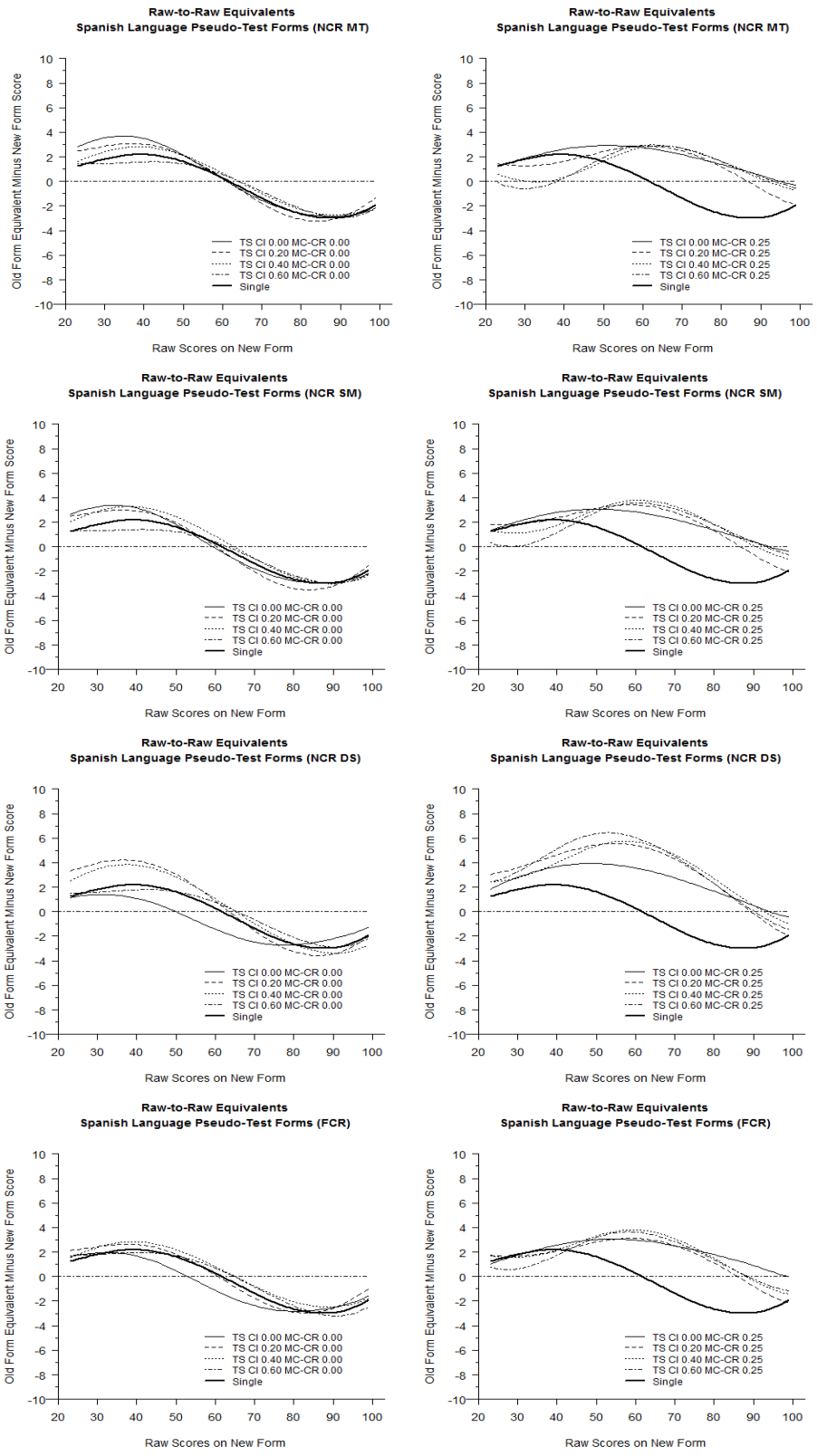

Figure 4-30. Spanish Language pseudo-test form comparison of CI ES for TS.

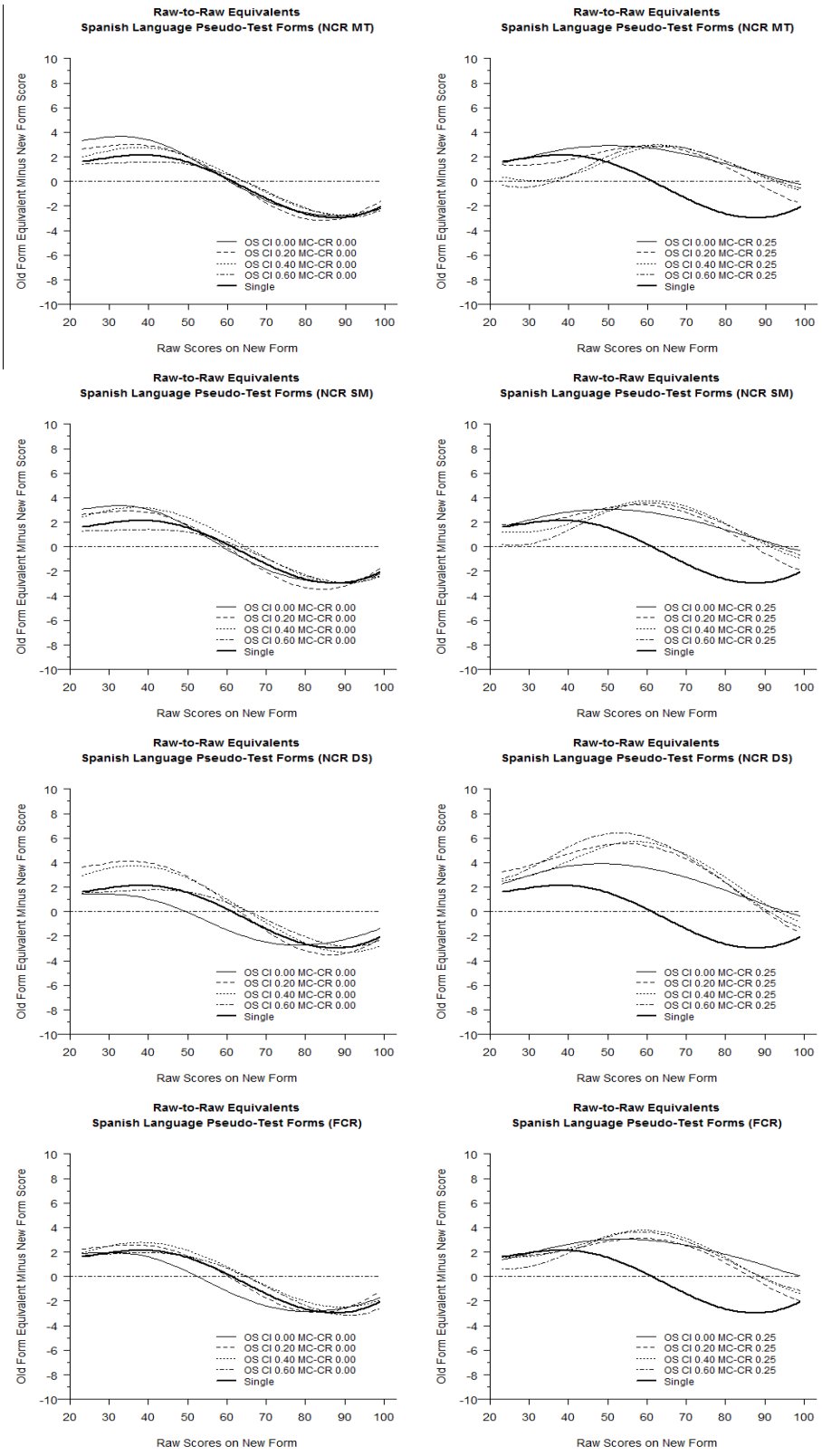

Figure 4-31. Spanish Language pseudo-test form comparison of CI ES for OS.

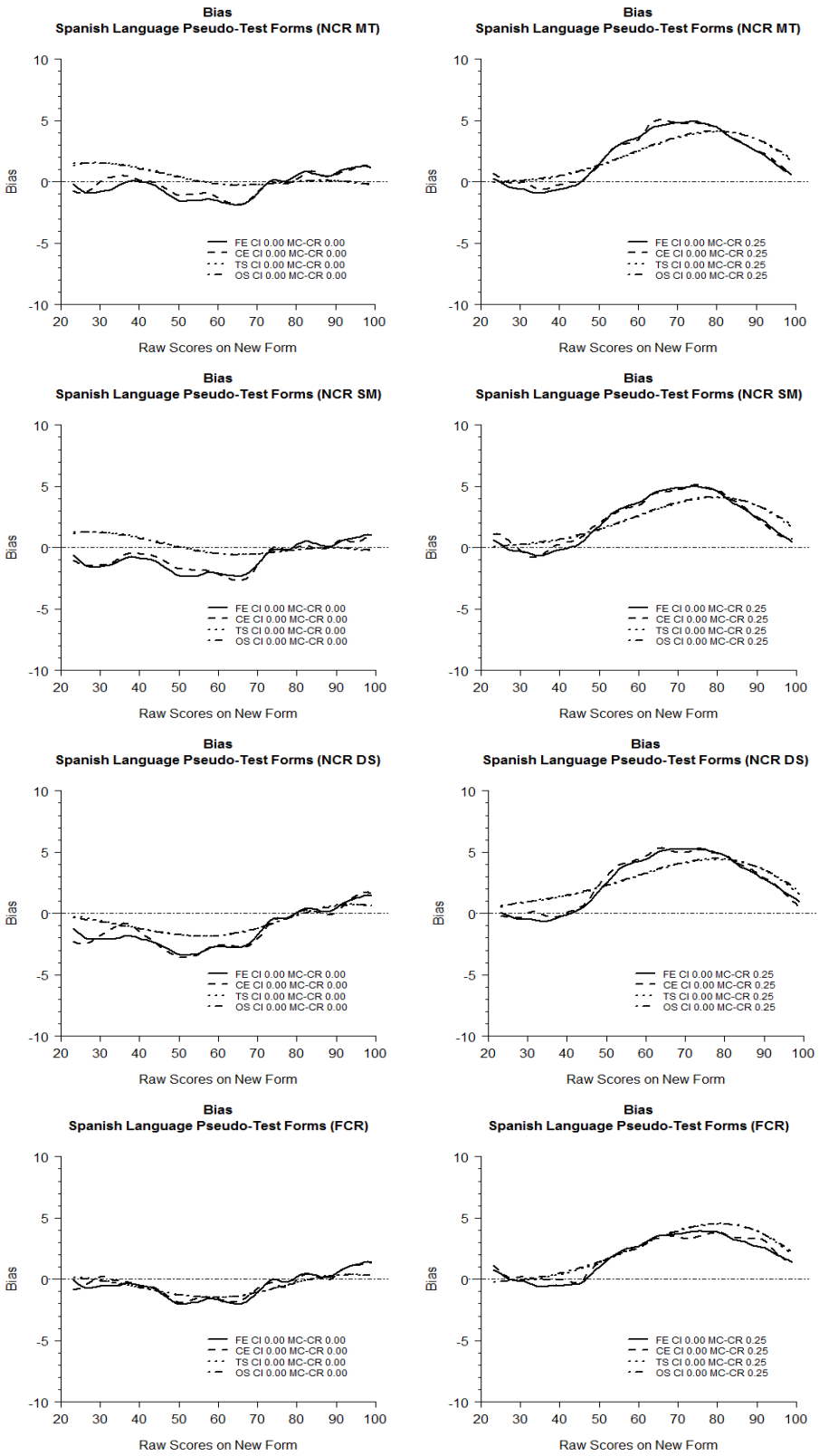

Figure 4-32. Spanish Language pseudo-test form conditional bias for CI 0.00.

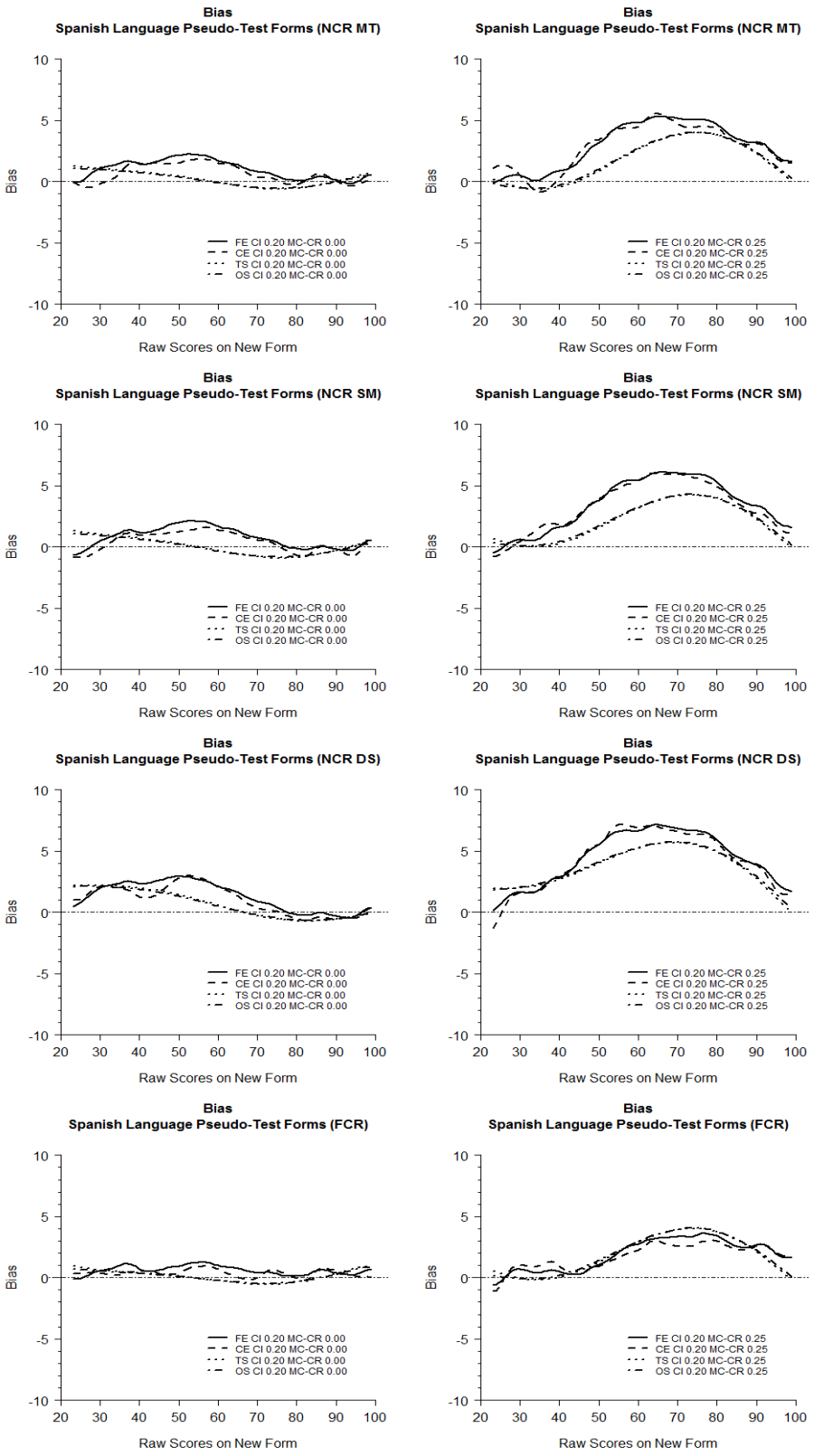

Figure 4-33. Spanish Language pseudo-test form conditional bias for CI 0.20.

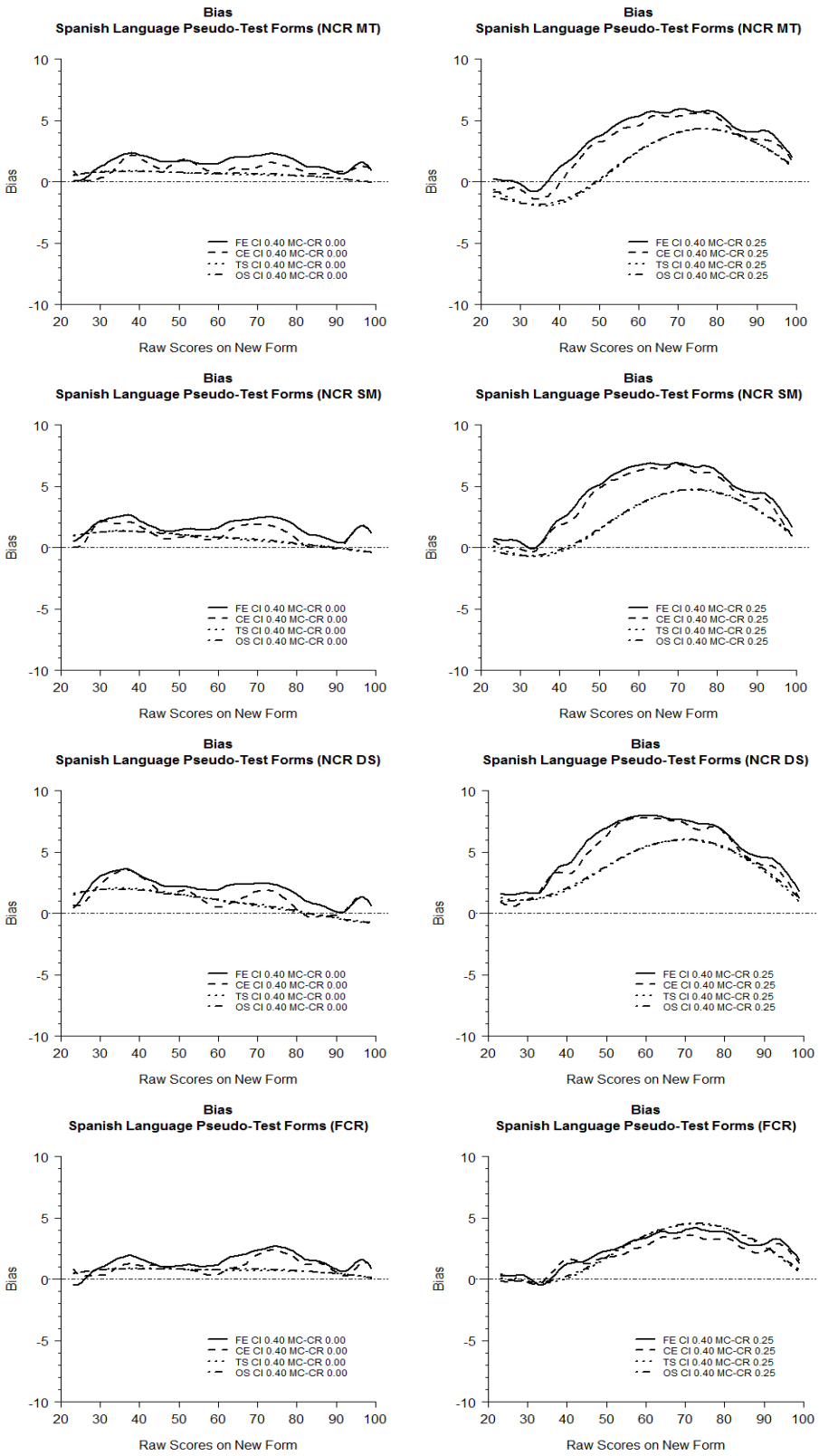

Figure 4-34. Spanish Language pseudo-test form conditional bias for CI 0.40.

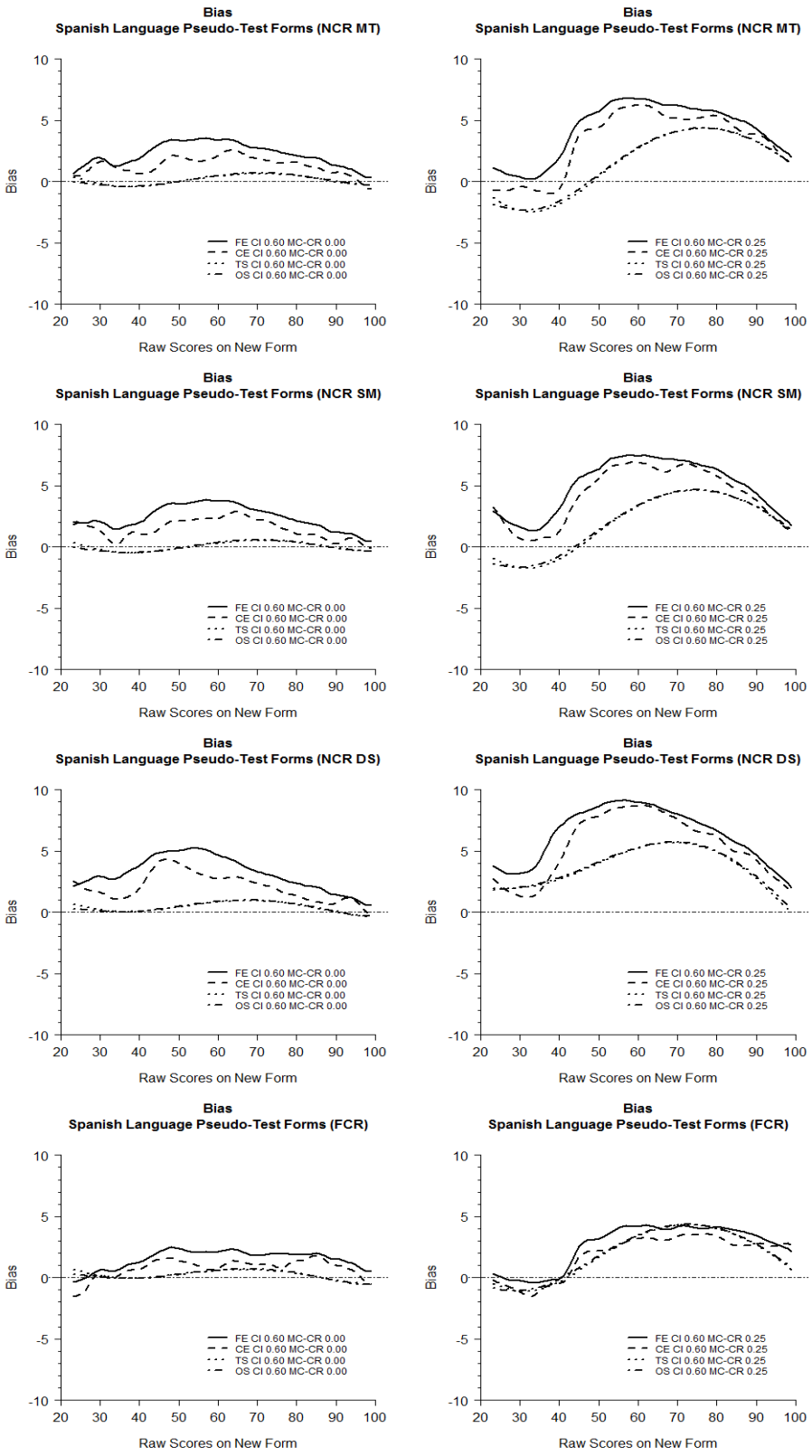

Figure 4-35. Spanish Language pseudo-test form conditional bias for CI 0.60.

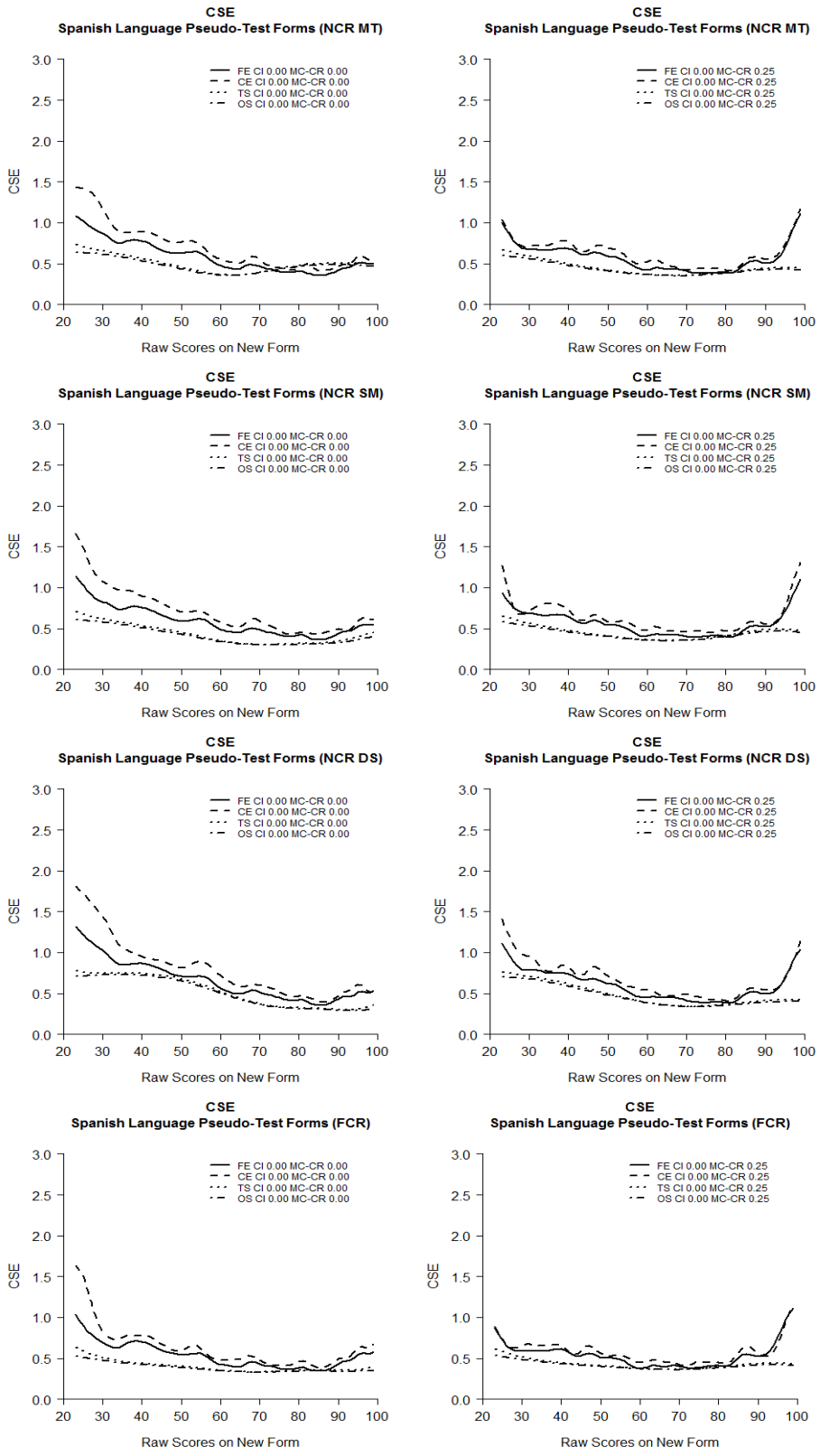

Figure 4-36. Spanish Language pseudo-test forms CSE for CI 0.00.

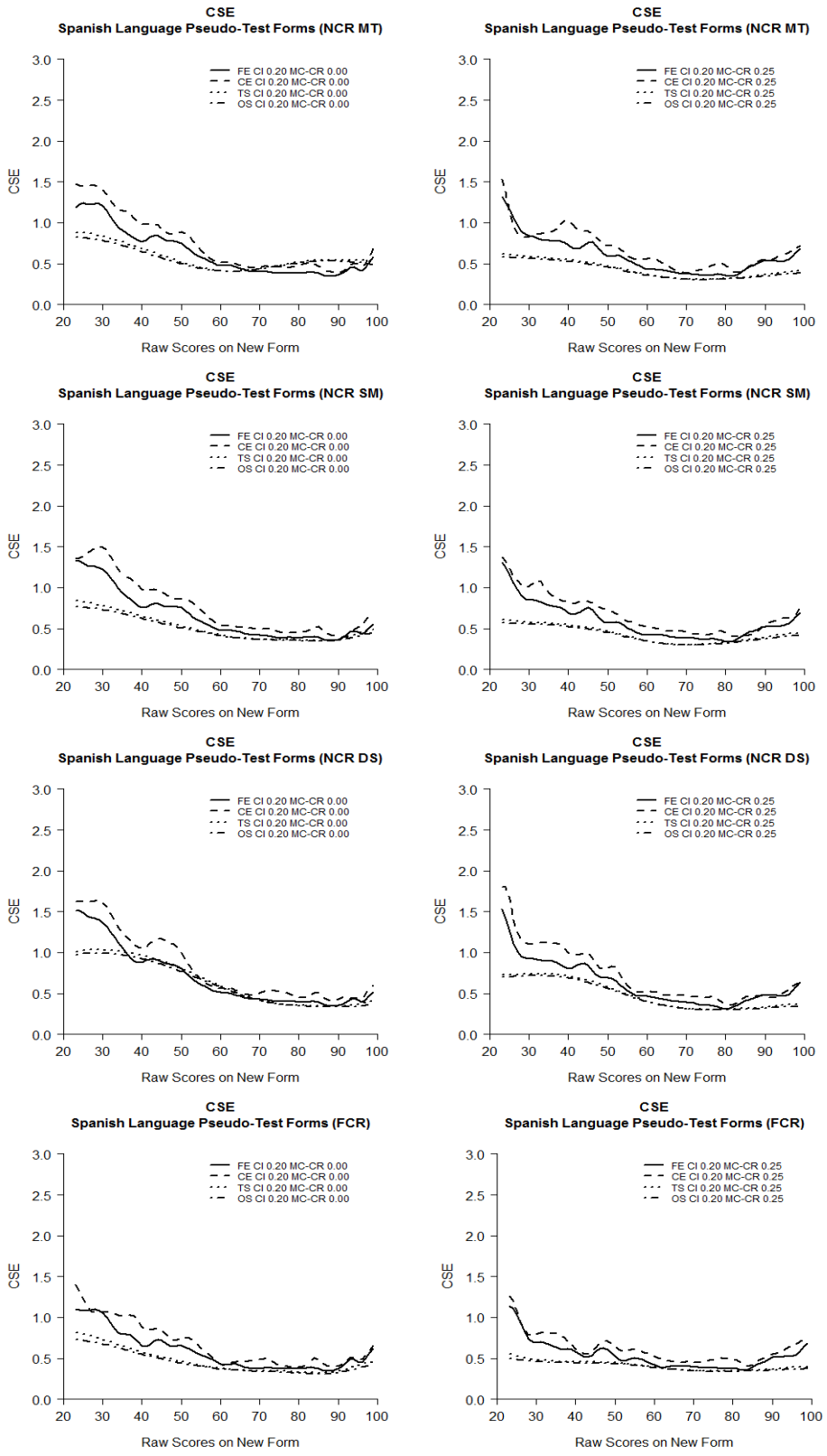

Figure 4-37. Spanish Language pseudo-test forms CSE for CI 0.20.

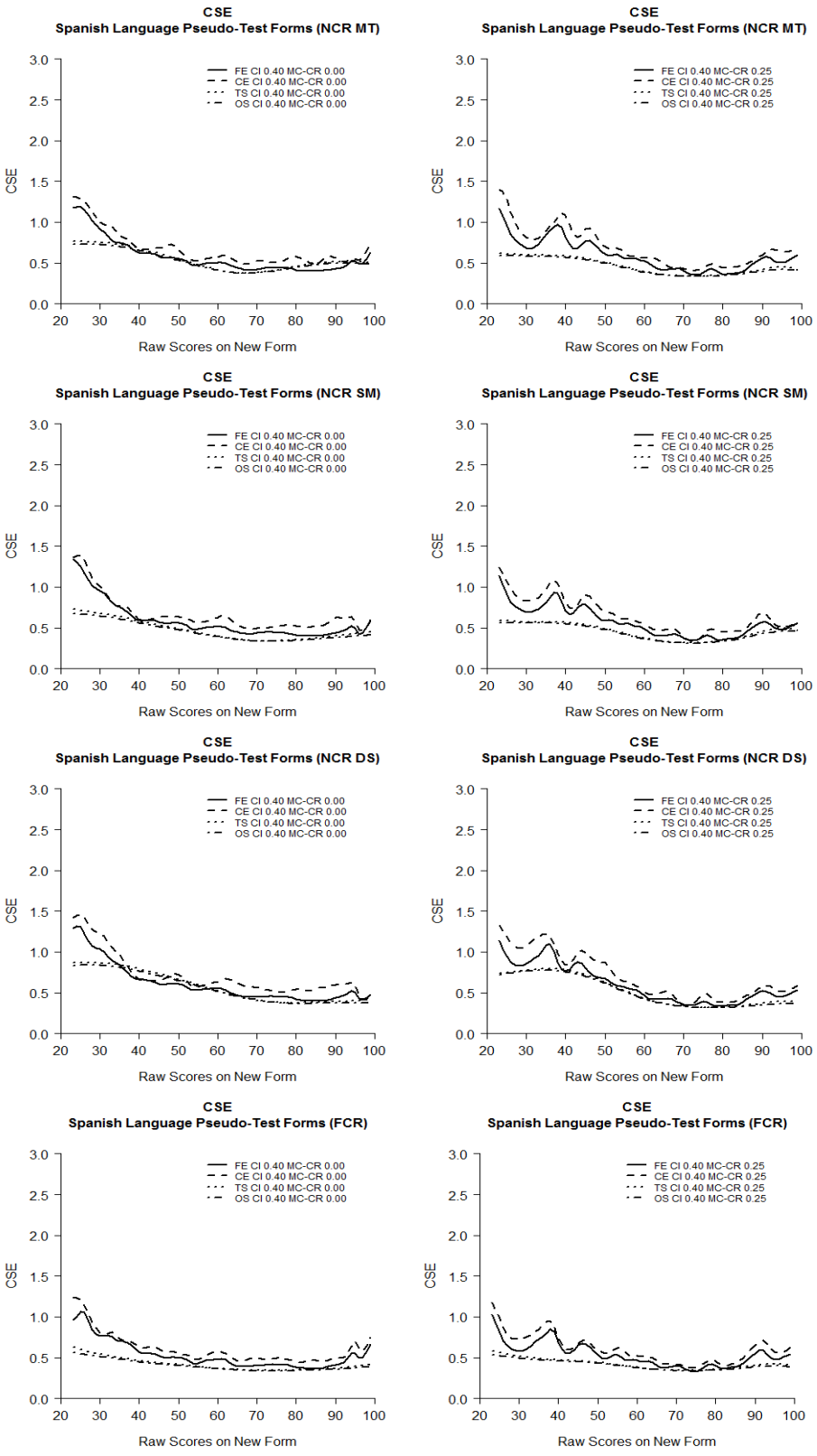

Figure 4-38. Spanish Language pseudo-test forms CSE for CI 0.40.

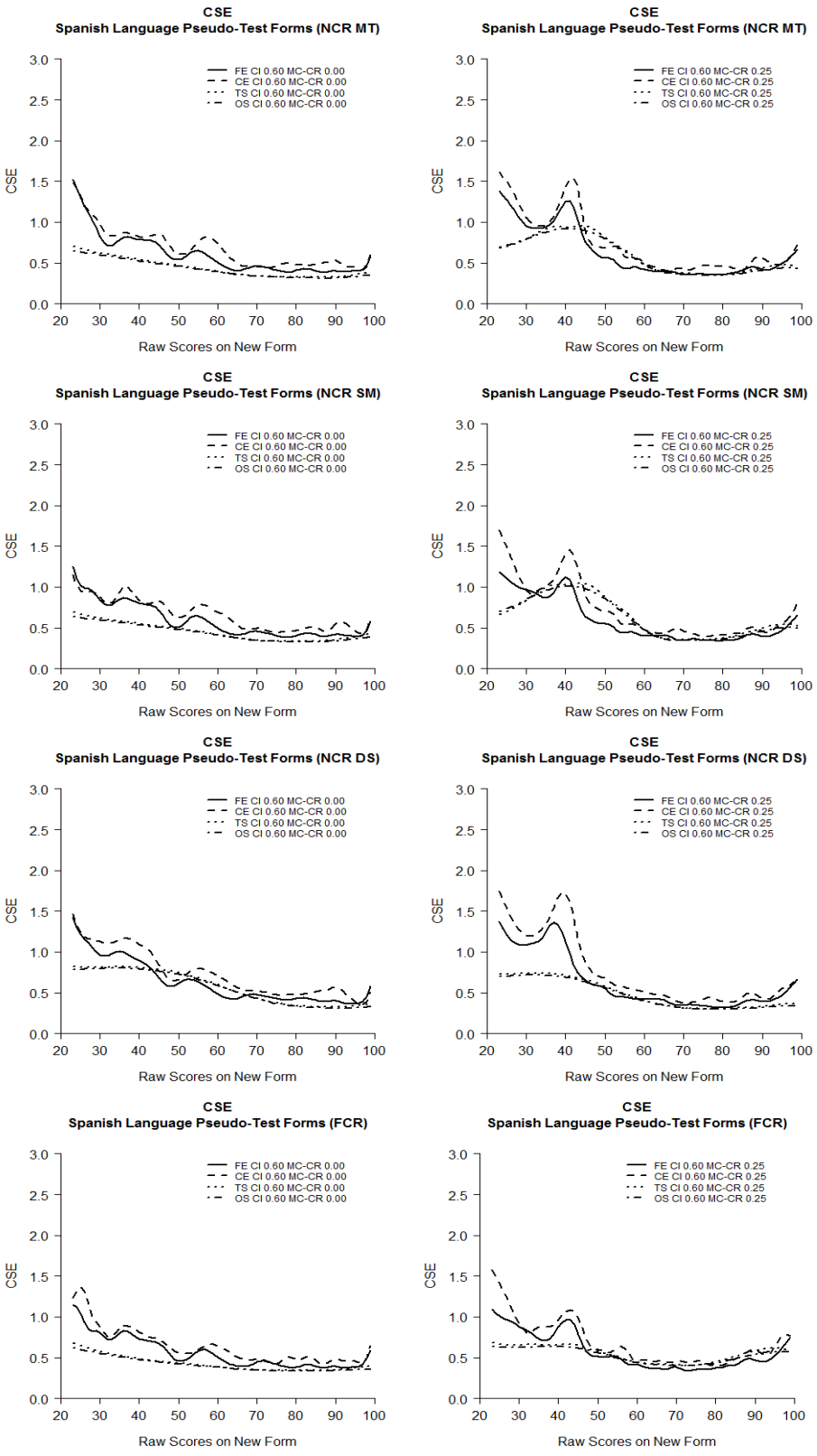

Figure 4-39. Spanish Language pseudo-test forms CSE for CI 0.60.

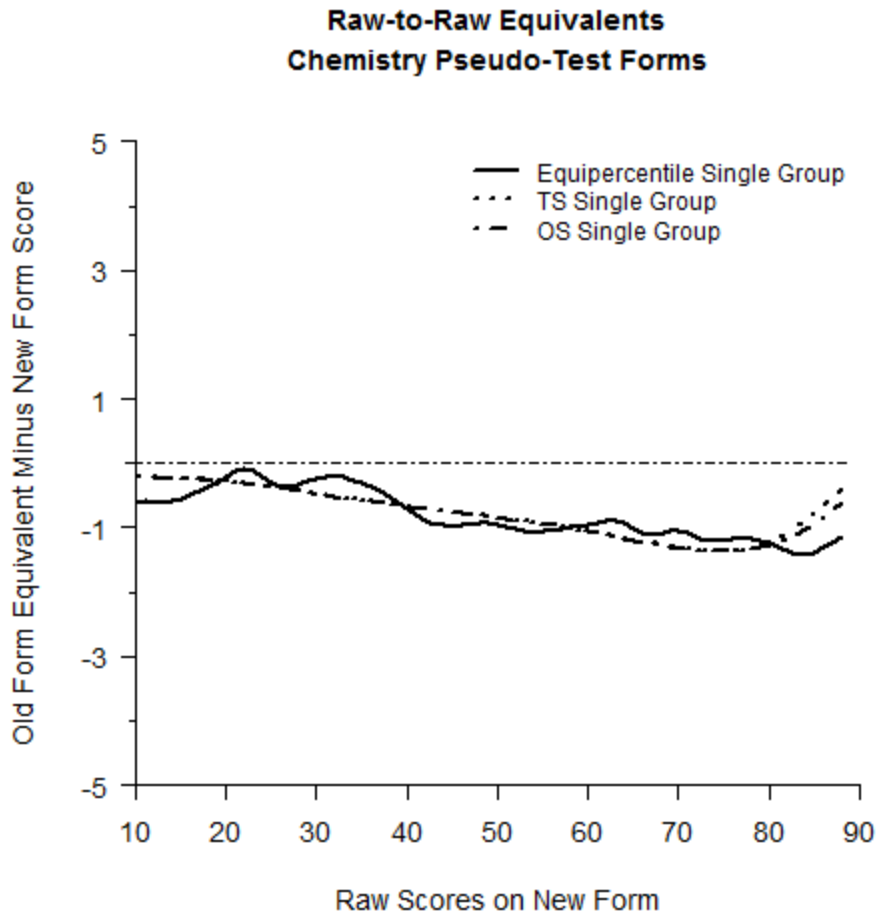

Figure 4-40. Criterion equating relationships for Chemistry pseudo-test forms (single group).

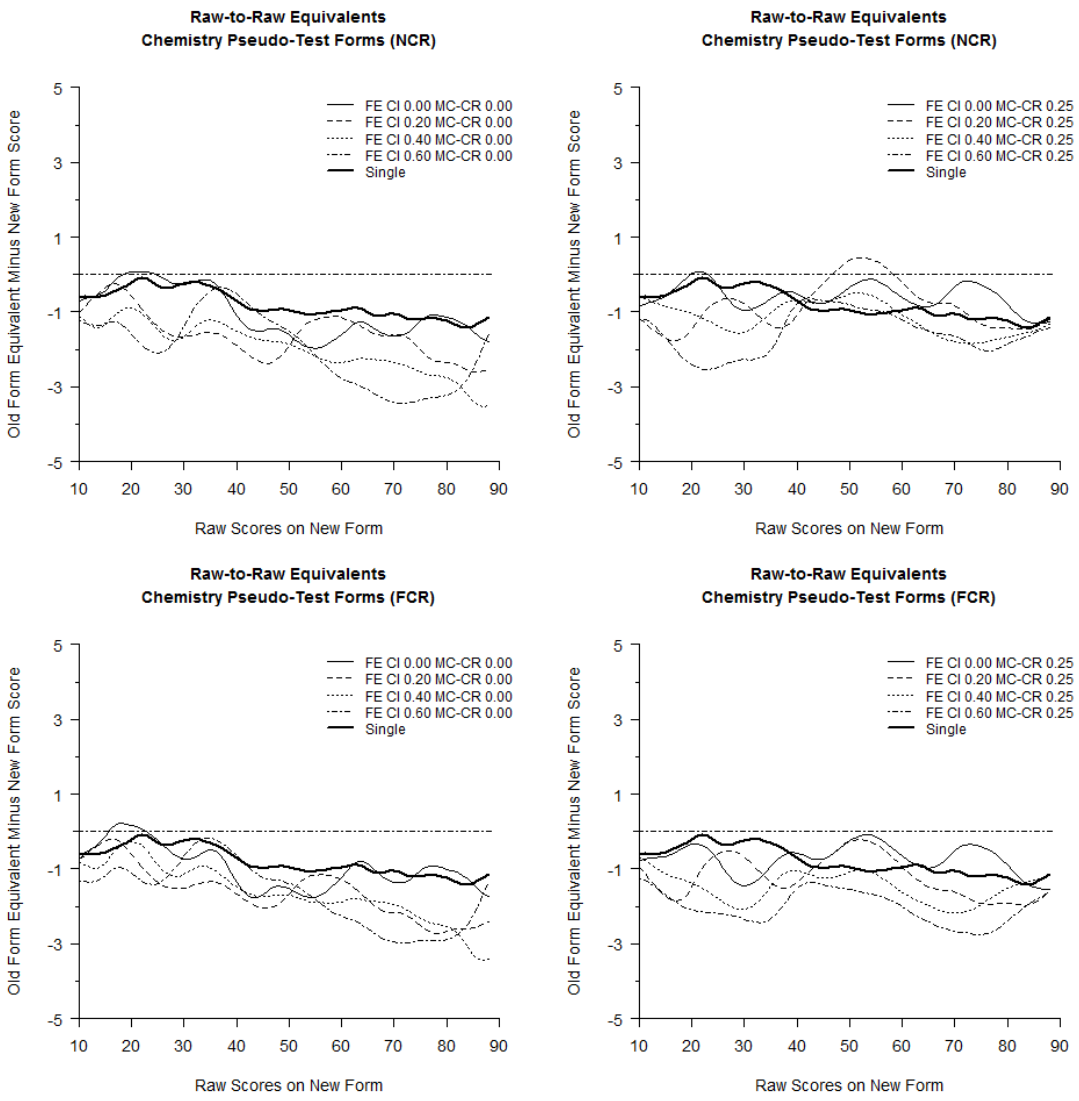

Figure 4-41. Chemistry pseudo-test form comparison of NCR and FCR for FE.

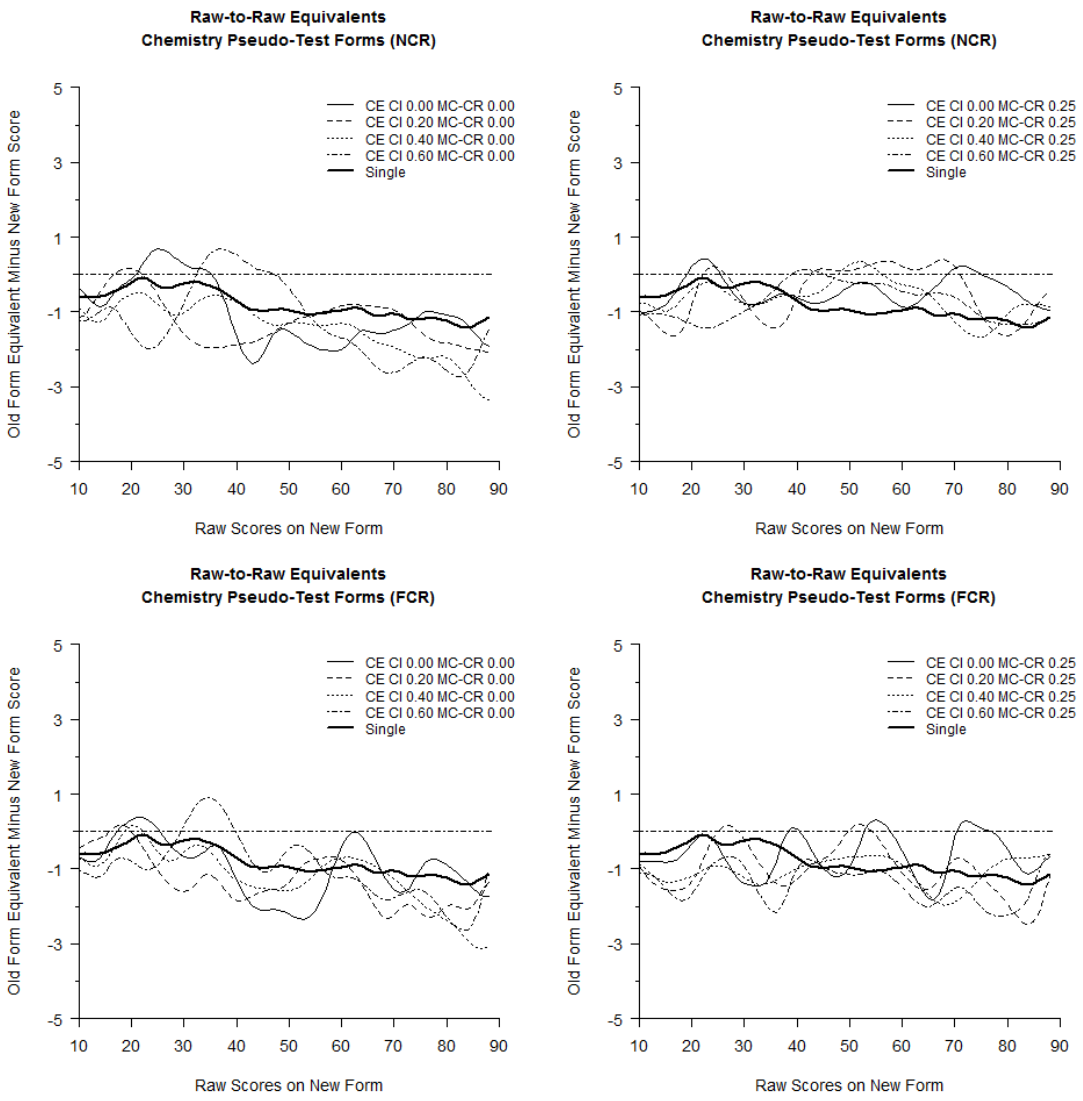

Figure 4-42. Chemistry pseudo-test form comparison of NCR and FCR for CE.

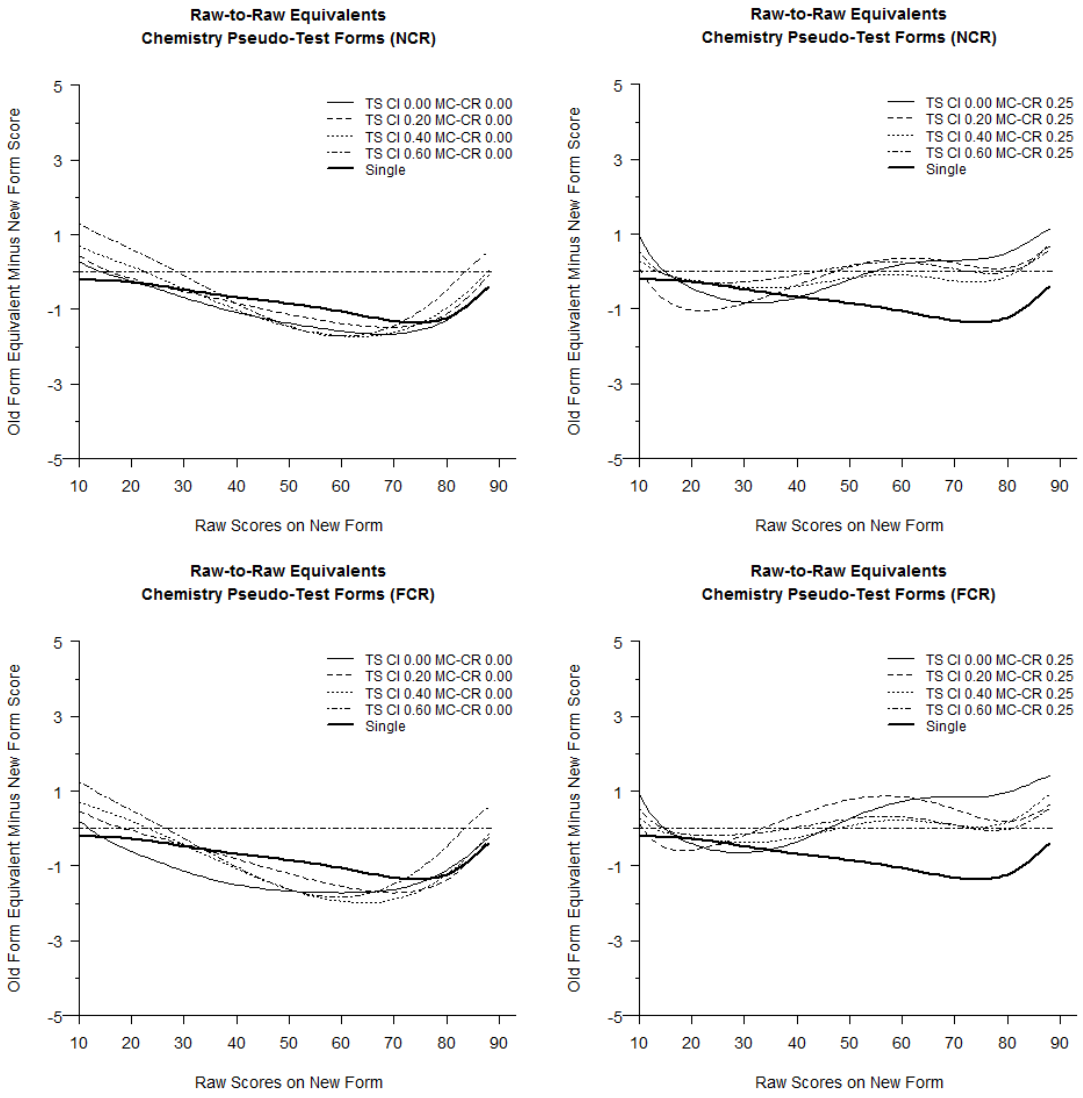

Figure 4-43. Chemistry pseudo-test form comparison of NCR and FCR for TS.

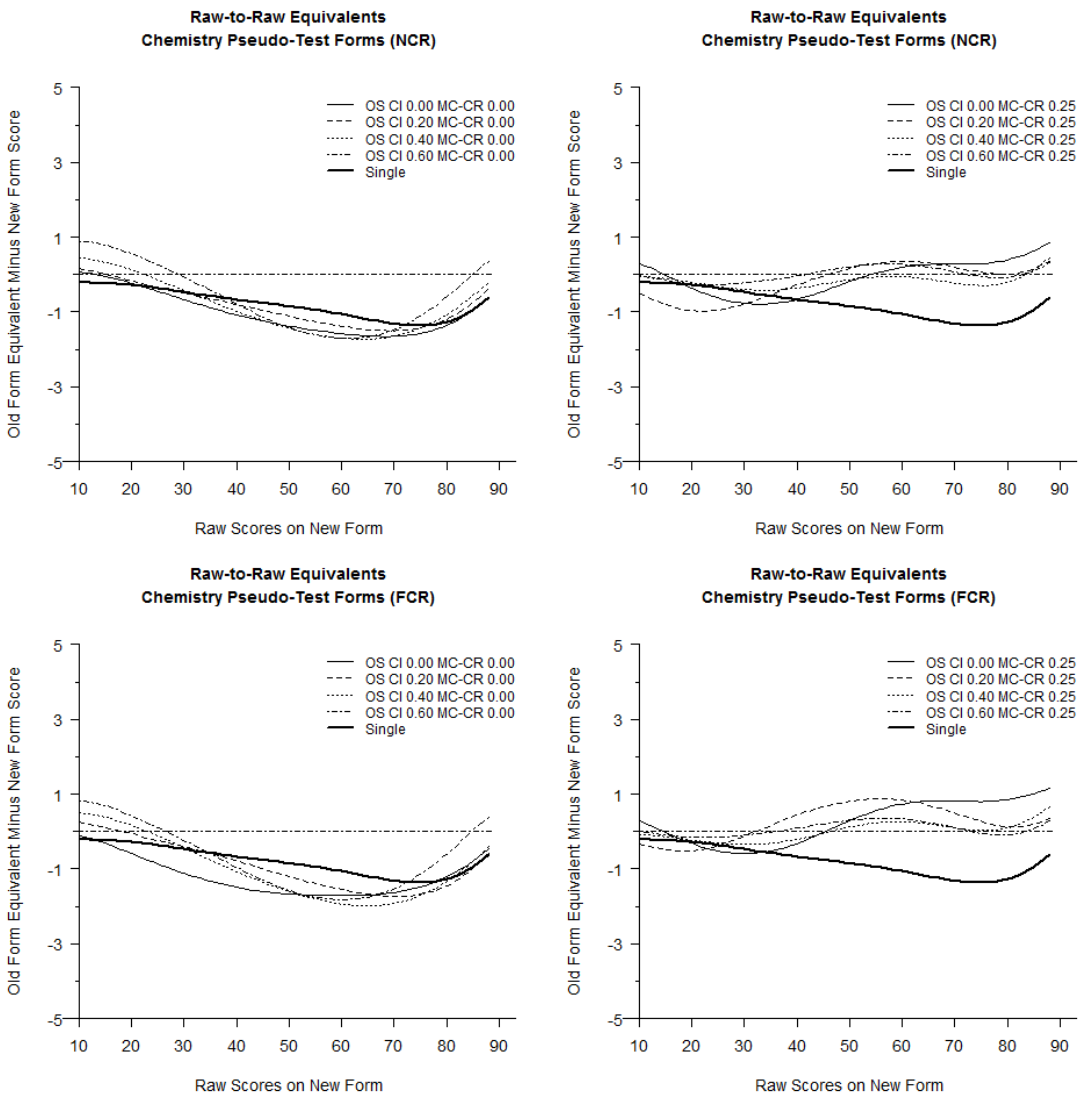

Figure 4-44. Chemistry pseudo-test form comparison of NCR and FCR for OS.

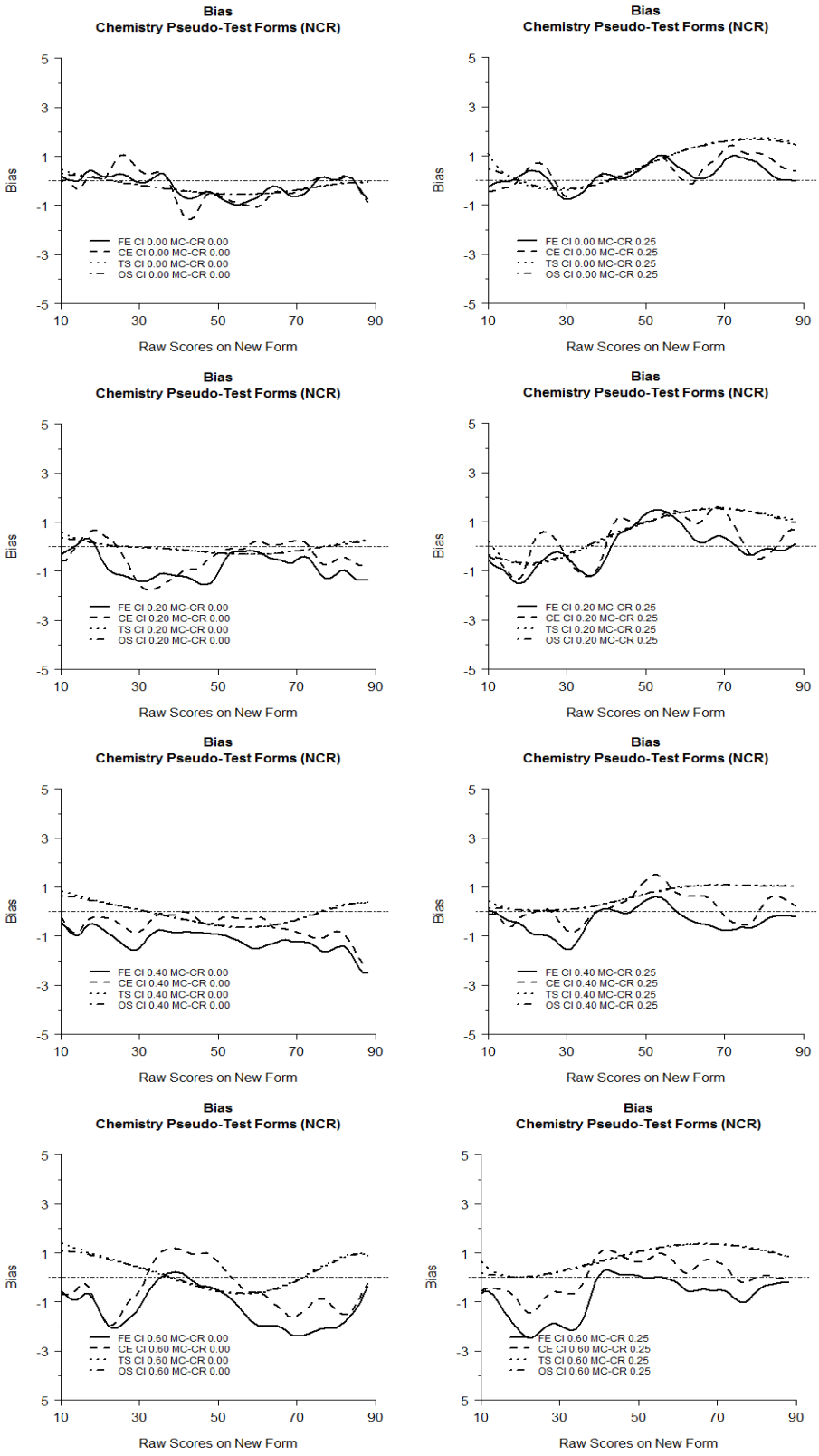

Figure 4-45. Chemistry pseudo-test form conditional bias for NCR.

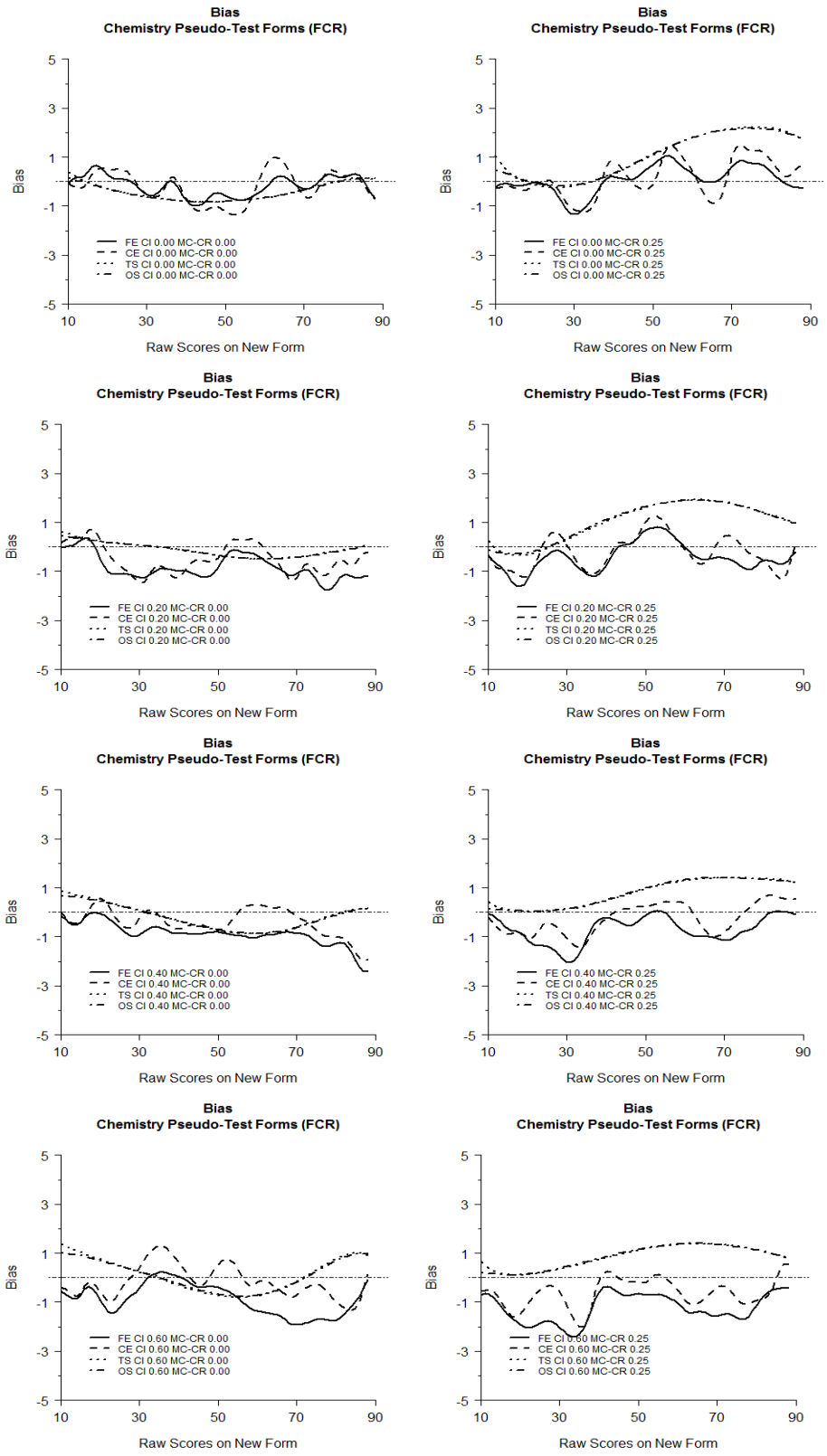

Figure 4-46. Chemistry pseudo-test form conditional bias for FCR.

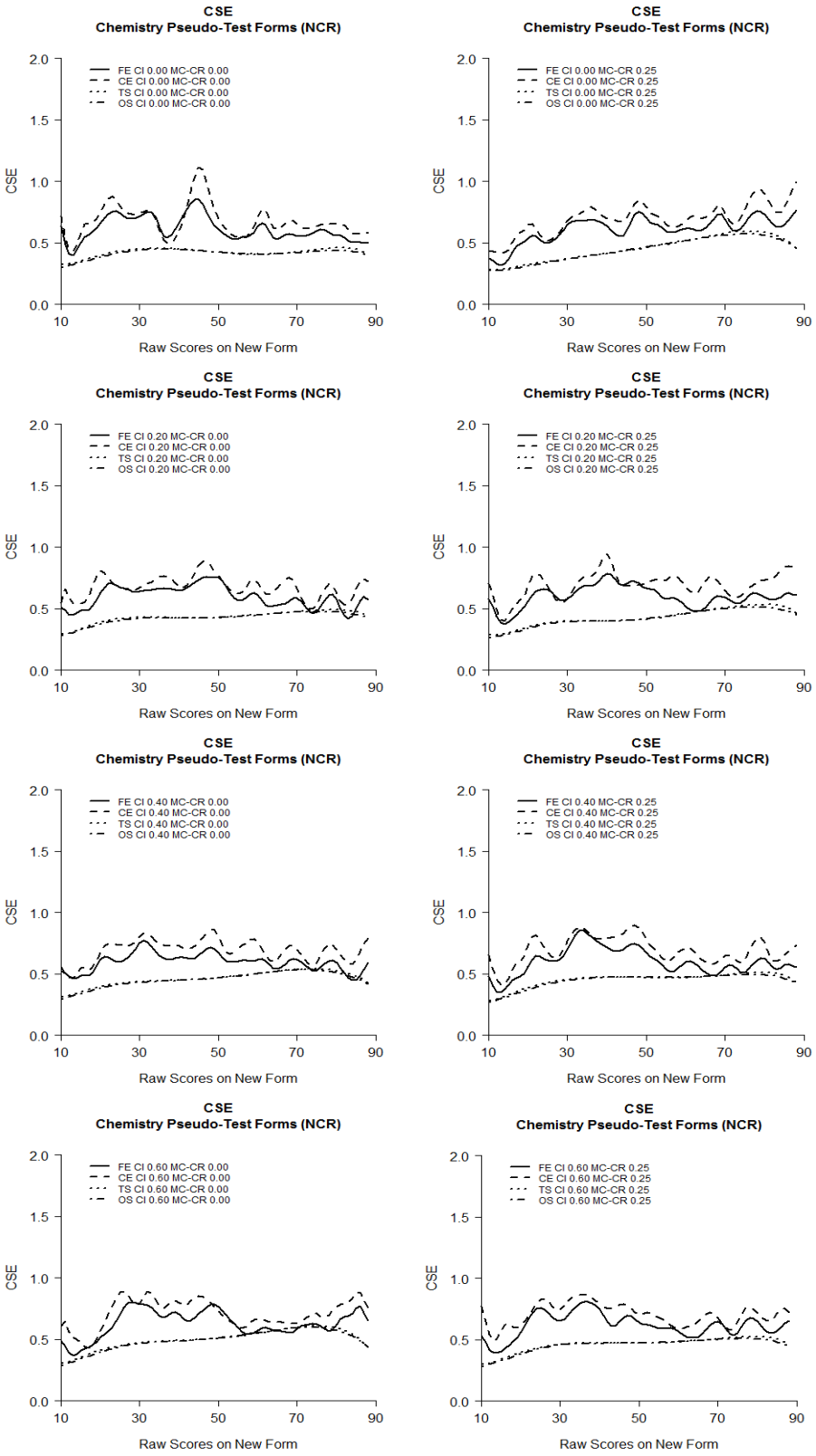

Figure 4-47. Chemistry pseudo-test forms CSE (NCR and FCR, MC-CR 0.00).

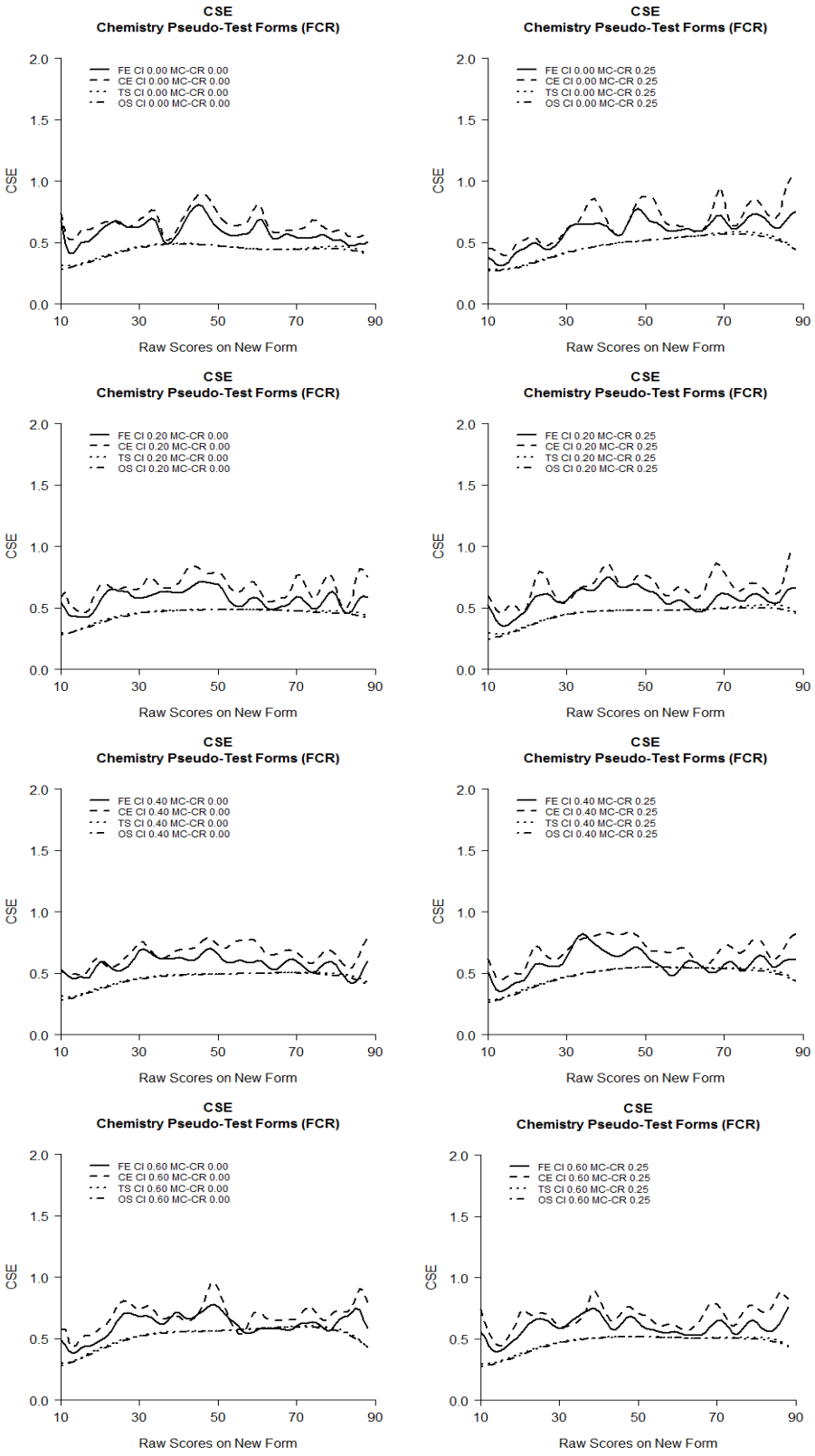

Figure 4-48. Chemistry pseudo-test forms CSE (NCR and FCR, MC-CR 0.25).

CHAPTER FIVE:
DISCUSSION

The purpose of this dissertation was to examine how characteristics of mixed-format tests might adversely impact equating and explore test characteristics that might lead to satisfactory equating with mixed-format tests. The specific factors of investigation in this dissertation included examinee group differences, equating methods, statistical and format representativeness of the common-item set, and relative difficulty of MC and CR items. Additionally, analyses were conducted for three tests: English Language, Spanish Language, and Chemistry. Further, for each of the three tests, equating analyses were conducted for both operational test forms and pseudo-test forms.

A large volume of results was presented in Chapter Four, and at times, the results may have appeared contradictory. One of the primary purposes of Chapter Five is to summarize the important findings across the analyses for the six tests. Chapter Five is divided into four sections: summary of findings, practical implications, limitations, and future research. Within the summary of findings, there are five subsections corresponding to the five research questions addressed in this dissertation. At the beginning of each subsection, the research question is restated to remind the reader. Within each research question, conclusions are separated by up to three sections: conclusions based on MC-CR 0.00, conclusions based on MC-CR 0.25, or conclusions based on MC-CR 0.00 and MC-CR 0.25. However, because the purpose of Research Question Two was to compare results based on MC-CR 0.00 and MC-CR 0.25, conclusions are not separated according to sampling conditions. For each section, there is a discussion of the conclusions reached based on the results from this dissertation, and key findings from previous studies are also incorporated into the discussion. The practical implications section places the results into a practical context by presenting some possible implications of the results. The

limitation section contains a discussion of limitations of this dissertation, and the future research section suggests ideas for future research.

Summary of Findings

The findings in this section are divided into five subsections according to the five research questions investigated in this dissertation.

Research Question One

What is the impact on equated scores when examinees on one mixed-format test form are higher in proficiency, as measured by items in common between test forms, than examinees on the other mixed-format test form?

MC-CR 0.00

As the difference in proficiency between old and new form examinee groups increased, equating relationships tended to become more biased relative to the criterion equating relationship. (The criterion equating relationship represented no difference in proficiency between groups of examinees on the old and new test forms.) However, the increase in bias was not consistent across equating methods or tests. Moreover, for the traditional equating methods, the increase in bias appeared to be impacted by the difference between the composite score and common-item score effect size rather than the magnitude of the common-item score effect size alone. Therefore, it often appeared that as the difference in proficiency between examinee groups increased, the common-item set no longer represented the total test in the same manner as for the criterion sample of examinees.

For the operational test forms, the criterion equating relationships were the equating relationships for CI 0.00 MC-CR 0.00. CI 0.00 MC-CR 0.00 was chosen as the criterion sample, because it essentially represented no difference in proficiency between examinee groups on the old and new test forms. This criterion sample also indicated that

effect sizes across old and new test forms were similar for MC and CR scores. The study condition equating relationships became increasingly different from the criterion equating relationships as the difference in proficiency between old and new form groups of examinees increased. However, as is discussed in greater detail for Research Question Three, when the common-item effect size was large, bias was typically larger for FE than for CE. Bias was typically smallest for TS and OS. Further, bias did not always increase for TS and OS as the difference in proficiency between old and new form groups of examinees increased. It is important to note, however, that a number of factors may have interacted to contribute to these findings. First, the criterion equating relationship was different for each equating method; consequently, comparisons across methods may not be reasonable. Second, only one smoothing value was selected for all bootstrap replications for the traditional equating methods. This smoothing value may not have been optimal for all replications, introducing additional bias or random error. For the pseudo-test forms, the single-group equating relationships were the criteria. Results were similar to the operational test forms. Additionally, when a common-item set composition other than an MC-only minitest was used, bias did not always consistently increase for FE and CE.

As discussed in Chapter Two, a number of previous research studies have examined the impact of group differences on the accuracy of equating. For the most part, the findings in this dissertation regarding group differences confirm findings from previous research. Many studies have found that equating tends to be more accurate when there are only small differences in proficiency between groups of examinees taking the old and new test forms (Cao 2008; Kim & Lee, 2006; Kirkpatrick, 2005; Lee et al., 2010; Wang, Lee, Brennan, & Kolen, 2008; Wu, Huang, Hu, & Harris 2009). Further, this dissertation confirms findings from previous research that CE may be less sensitive than FE to differences in group proficiency (Lee et al., 2010; Wang, Lee, Brennan, & Kolen, 2008). Little research has been conducted comparing traditional and IRT equating

methods. Although, von Davier and Wilson (2008) found that TS and CE performed similarly, and that both equating methods were relatively invariant across groups. However, both Cao (2008) and Kirkpatrick (2005) found that differences in group proficiency impacted equating results in the IRT framework.

One additional interesting finding was that bias in equating relationships, for the traditional equating methods, appeared to also be impacted by the difference between the composite score effect size and the common-item score effect size rather than the magnitude of the common-item score effect size alone. Typically, it was found for the traditional equating methods that equating relationships were more similar to the criterion when the difference between the composite score effect size and common-item score effect size was also similar to the criterion. Studies investigating the impact of group differences on equating results have not explicitly addressed this finding. This trend did not appear to influence results for the IRT equating methods; however.

As an example, consider English Language and Chemistry operational test forms. The same common-item effect sizes were studied for both English Language and Chemistry: CI 0.00, CI 0.20, and CI 0.40. Yet, values of standardized WARMSB were larger by as much as 0.15 standard deviation units for English Language as compared to Chemistry. For English Language, the differences between the composite score effect size and common-item score effect size were 0.158, 0.250, and 0.300 for CI 0.00, CI 0.20, and CI 0.40, respectively. For Chemistry, the differences were 0.132, 0.140, and 0.150 for CI 0.00, CI 0.20, and CI 0.40, respectively. For English Language, the differences varied by as much as approximately 0.15 standard deviation units across the three CI sampling conditions. For Chemistry, the differences varied by only 0.02 standard deviation units.

One unique characteristic of the Chemistry tests was that the disattenuated MC and CR correlations were near 1. One plausible hypothesis is that when the MC and CR correlation is higher for a test, the common items might be more representative of the

total test, regardless of group differences in proficiency. Further, for the operational test forms, the common-item and composite score correlation was approximately 0.84 for English and Spanish Language and approximately 0.92 for Chemistry. For the pseudo-test forms, the correlations were approximately 0.85 for English Language, 0.90 for Spanish Language, and 0.94 for Chemistry. Therefore, another plausible hypothesis is that when the common-item and composite score correlation is high, group differences in proficiency may not impact equating results to a large extent.

MC-CR 0.25

When effect sizes across old and new test forms were not similar for MC and CR scores (i.e., MC-CR 0.25), the impact of examinee group differences did not always appear to follow a consistently increasing pattern. Patterns of results for the MC-CR 0.25 sampling conditions appeared to be impacted by an interaction between the common-item effect size and the difference in relative difficulty of MC and CR items. Further, the specific criterion equating relationship used to evaluate the equating results also impacted the conclusions.

For the operational test forms, the criterion was the same as for the MC-CR 0.00 sampling conditions (i.e., CI 0.00 MC-CR 0.00). By examining values of WARMSB based on this criterion, results seemed somewhat random and inconsistent. Typically, as the difference in proficiency between old and new form groups of examinees increased, bias either decreased across all of the common-item effect sizes or decreased for CI 0.20 and then increased for CI 0.40. However, when plots of equating relationships were examined, it was evident that the difference between equating relationships increased as the difference in proficiency between groups increased. Further, when CI 0.00 MC-CR 0.25 was used as the criterion equating relationship, as the difference in proficiency between old and new form groups of examinees increased, WARMSB also increased. For the pseudo-test forms, the single-group equating relationship was the criterion. Similar to

the operational test forms, values of WARMSB did not consistently increase as the difference in proficiency between old and new form groups of examinees increased. When CI 0.00 MC-CR 0.25 was used as the criterion, WARMSB typically increased as the difference in proficiency between old and new form groups of examinees increased, although results did vary according to test, equating method, and common-item set composition.

Research Question Two

When one type of item format (i.e., MC or CR) is relatively more difficult for examinees taking one form as compared to examinees taking another form, how are the resulting equated scores impacted?

The results for this research question are inconclusive, because it was difficult to disentangle the interaction of the common-item effect size, the difference in relative difficulty of MC and CR items, and the difference between the common-item and composite score effect size. Across the test forms, results were mixed as to how relative difficulty of MC and CR items, as operationalized in this dissertation, impacted equating results. However, some results from this dissertation suggest that equating mixed-format tests with only multiple-choice common items may result in larger bias when examinees find certain item formats more difficult relative to other item formats.

The MC-CR 0.00 sampling conditions represented conditions where examinee performance was similar on the MC and CR items. The MC-CR 0.25 sampling conditions represented conditions where examinee performance was different across the MC and CR items. Therefore, it was expected that bias would be larger for the MC-CR 0.25 sampling conditions as compared to the MC-CR 0.00 sampling conditions. For all of the pseudo-test forms, values of WARMSB were larger for CI 0.00 MC-CR 0.25 as compared to CI 0.00 MC-CR 0.00. However, for the remaining common-item effect sizes, there appeared to be a complex interaction between the common-item effect size and the MC-CR effect

size. Therefore, results appeared to be dependent on the particular test. For English Language operational and pseudo-test forms and Chemistry pseudo-test forms, MC-CR 0.25 sampling conditions tended to result in smaller values of WARMSB as compared to MC-CR 0.00 sampling conditions. This may have occurred because scores on the new form were lower than scores on the old form. For Spanish Language and Chemistry operational test forms and Spanish Language pseudo-test forms, the MC-CR 0.25 conditions often resulted in larger bias than the MC-CR 0.00 conditions. The different results found across tests may be explained by the difference between the composite score and common-item score effect sizes. Often, the equating relationships that were most similar to the criterion equating relationship were also the sampling conditions for which the difference between the composite score and common-item effect size was most similar to the difference for the criterion.

Although the evidence is contradictory, results from the Spanish Language pseudo-test analyses provide rather compelling evidence that the difference in effect sizes between MC and CR items does impact the accuracy of the equating relationships. For the Spanish Language pseudo-test form criterion, the difference between effect sizes for MC and CR scores was approximately -0.27. MC items were similar in difficulty across test forms, but CR items were much easier on the new form relative to the old form. Four sampling conditions were created with an MC and CR effect size difference similar to the criterion, and four sampling conditions were created with an MC and CR effect size difference that was different from the criterion. For the sampling conditions with an MC and CR effect size different from the criterion, bias was approximately three points (or 0.15 standard deviation units) larger than the sampling conditions with an MC and CR effect size similar to the criterion. However, because other data characteristics of the test also varied across sampling conditions, and because Spanish Language was the only pseudo-test where a large difference in effect sizes between MC and CR could be created,

it is difficult to isolate whether the bias was primarily influenced by the disparity in MC and CR effect sizes.

Research Question Three

How much do equated scores vary across equating methods?

MC-CR 0.00

TS, OS, and to a lesser extent, CE, appeared to be less sensitive to group differences in proficiency than FE. Equating methods generally yielded similar results when examinee groups on the old and new forms were similar in proficiency. When examinee groups differed in proficiency across old and new test forms, FE generally resulted in larger bias than CE, TS, or OS. TS and OS also typically resulted in smaller bias than CE. The trends were fairly consistent across tests, although for the Chemistry operational test forms, WARMSB was larger for TS and OS as compared to FE and CE. As discussed previously for Research Question One, it is important not to overemphasize the difference in bias between the traditional and IRT equating methods for large common-item effect sizes. To reiterate what was discussed in Research Question One, the findings in this dissertation may have resulted from factors such as different criterion equating relationships for each equating method, choice of smoothing value, or interactions among the common-item and MC-CR effect sizes.

The results of this dissertation confirm results from a number of studies conducted for both MC and mixed-format tests that have found similar results in comparisons between FE and CE: when group differences are small, CE and FE tend to have similar results. However, CE tends to be more accurate than FE when group differences are larger (Harris & Kolen, 1990; Holland, Sinharay, von Davier, & Han, 2008; Lee et al., 2010; Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008). Research comparing IRT and traditional equating methods appears to be less prevalent; therefore, the results from this dissertation comparing traditional and IRT equating methods are

somewhat novel. However, some previous research has found that TS and CE may both be relatively group invariant (von Davier & Wilson, 2008).

MC-CR 0.25

As noted previously, for the MC-CR 0.25 sampling conditions, values of WARMSB appeared inconsistent, and interpretation of the findings was difficult. There did not appear to be consistent conclusions across tests regarding which equating methods performed better and under what circumstances. For example, for the English Language operational test forms, TS and OS appeared to perform better than FE and CE for CI 0.40 only. For Spanish Language and Chemistry, TS and OS performed better than FE and CE for only CI 0.00. Inconsistencies were also seen across the pseudo-test forms. As previously discussed, results based on the CI 0.00 MC-CR 0.00 or single group criterion sometimes obscured the trends that were evident in the figures. When CI 0.00 MC-CR 0.25 was used as the criterion, it was evident that the equating relationships for TS and OS were typically more similar to each other than equating relationships for FE and CE across common-item effect sizes. That is, when CI 0.00 MC-CR 0.25 was used as the criterion, conclusions were similar to those for MC-CR 0.00. As the common-item effect size increased, bias tended to be smaller for TS and OS as compared to FE and CE.

MC-CR 0.00 and MC-CR 0.25

Conclusions regarding the standard errors of equating were similar across the MC-CR 0.00 and MC-CR 0.25 sampling conditions, and results did not appear to be influenced by these effect sizes. OS most often resulted in the smallest standard errors and CE most often resulted in the largest standard errors. Although no previous research was found comparing standard errors of equating for traditional and IRT equating methods, Tsai, Hanson, Kolen, and Forsyth (2001) found that OS generally resulted in smaller standard errors than TS. Further, a number of studies have found that CE tends to result in larger standard errors of equating than FE (Lee et al., 2010; Sinharay & Holland,

2007; Wang, Lee, Brennan, & Kolen, 2008). The results of this dissertation confirm findings from these studies. In addition, results from this dissertation provide some novel results that suggest standard errors of equating may be smallest for IRT equating methods. However, these results may have been influenced by the specific software programs used to conduct the analyses or choice of smoothing values for the traditional equating methods.

Occasionally, TS and OS resulted in larger standard errors than the traditional equating methods. Additional investigation is needed to determine what caused this result. As described in Chapter Three, the process for calculating IRT standard errors of equating was complex and involved a series of computer software packages. It was impossible to check all of the output for accuracy. Therefore, a number of steps were taken in an attempt to ensure that the IRT analyses were accurate. When PARSCALE could not estimate a pseudo-chance item parameter, that item parameter was fixed to 0 in the command file. However, for a given test form, the same control card was used across all bootstrap replications. Additional checks were also built into the IRT standard error of equating computer code. These checks included removing replications when PARSCALE did not converge and removing replications for which the equating results seemed unreasonable. Although several checks were written into the computer programming for calculating standard errors of equating for IRT results, it is still plausible that some of the equating results were unreasonable, leading to inaccurate standard errors of equating.

Research Question Four

How do the content and statistical specifications of a test (e.g., subject area, correlation between MC and CR scores, composition of CI) impact equated scores? This research question encompassed a number of components; therefore, the research question is separated into three smaller questions that are summarized individually.

How do the subject area and/or correlation between MC and CR scores impact equated scores?

MC-CR 0.00

There is some evidence in this dissertation that higher MC and CR correlations or certain subject areas lead to less bias in equating results when examinees perform similarly on the MC and CR items. Disattenuated MC and CR correlations were in the range of approximately 0.75 to 0.80 for the English Language operational test forms, approximately 0.80 to 0.90 for Spanish Language operational test forms, and approximately 0.95 for Chemistry. Disattenuated correlations for the pseudo-test forms were similar in magnitude. Of the three operational tests, values of standardized WARMSB were generally smallest for Chemistry. Although disattenuated MC and CR correlations were lowest for English Language, English Language did not consistently result in the highest values of standardized WARMSB. However, given the unique equating situation presented by the Spanish Language operational test forms, it is not surprising that Spanish Language resulted in larger values of WARMSB. For the pseudo-test forms, Chemistry also generally resulted in the smallest values of standardized WARMSB, and English Language generally resulted in the largest values of standardized WARMSB. These results provide some evidence that either the subject area or the disattenuated MC and CR correlation impacted equating results. Given the tests investigated in this study, it is impossible to disentangle the influence of subject area from MC and CR correlation. However, Lee et al. (2010) found that equating was more accurate for a common-item set containing only MC items when the correlation between MC and CR scores was higher.

MC-CR 0.25

Given that the MC-CR 0.25 effect size was not always 0.25 in magnitude, making comparisons across the tests based on the MC-CR 0.25 conditions may not be reasonable.

For the operational test forms, Spanish Language tended to result in the largest values of standardized WARMSB; however, the unweighted MC-CR 0.25 effect size differed from the unweighted MC-CR 0.00 effect size by as much as 0.364. The unweighted MC-CR 0.25 effect sizes differed by approximately 0.25 and 0.10 for English Language and Chemistry, respectively. The larger difference in unweighted MC-CR effect sizes for Spanish Language may have accounted for the larger values of standardized WARMSB. For about half of the common-item effect size conditions, values of standardized WARMSB were smallest for English Language. For the other approximately half of the conditions, values of standardized WARMSB were smallest for Chemistry. It is not clear why English Language may have resulted in smaller values of standardized WARMSB than for Chemistry. For the pseudo-test forms, the MC-CR 0.25 effect size differed from the MC-CR 0.00 effect size by 0.15 for English Language, 0.25 for Spanish Language, and 0.05 for Chemistry. For the MC-only minitests (NCR), Chemistry tended to result in the smallest values of standardized WARMSB, and Spanish Language tended to result in the largest values of standardized WARMSB. Consequently, it is plausible that the magnitude of the MC-CR effect size impacted the values of standardized WARMSB to a greater extent than the MC and CR correlation.

How do the statistical specifications of the CI impact equated scores?

MC-CR 0.00

Analyses comparing the statistical specifications of the common items were conducted only for the Spanish Language pseudo-tests. Consequently, evidence is limited and should not be broadly generalized beyond these analyses. That said, for this dissertation, in terms of bias, the minitest (NCR MT) typically performed slightly better than either the semi-miditest (NCR SM) or the difficulty shift (NCR DS). In this dissertation, the minitest resulted in less bias than the semi-miditest for approximately half of the sampling conditions. The difficulty shift common-item set always resulted in

larger bias than the minitest or semi-miditest. Differences in WARMSB between the minitest and semi-miditest were generally less than approximately 0.40 score points, and classification consistency differed by less than 3%. Sinharay and Holland (2007) also found that the minitest and semi-miditest did not result in practically significant differences in equated scores. Differences in WARMSB between the minitest and difficulty shift common items were as large as 1 point, and classification consistency differed by as much as 6%. Cao (2008) also found that bias was lower and classification consistency rates were higher for common-item sets that did not differ in average difficulty from the total test.

It is important to note that, although every effort was made to create common-item sets that reflected the target means and standard deviations of difficulty and discrimination, it was not possible to create the common-item sets as intended. The difficulty shift common-item set represented a shift in difficulty; however, it was slightly smaller than the intended target shift in difficulty. The nuances between the minitest and the semi-miditest were more difficult to create. The standard deviation of difficulty was only slightly smaller for the semi-miditest than for the minitest. Further, the standard deviation of discrimination for the semi-miditest was slightly larger than for the minitest. Consequently, the semi-miditest created for this dissertation may not actually represent a semi-miditest as intended by Sinharay and Holland (2006, 2007).

Another finding by Sinharay and Holland (2006) was that the correlations between the common items and total test score were higher for the semi-miditest. In this dissertation, correlations were typically very similar for the minitest and semi-miditest, though, and differences in equating relationships for the minitests and semi-miditests did not appear to be influenced by the magnitude of the correlations. The difficulty shift common-item scores always resulted in the lowest correlations with the total test. Both the minitest and semi-miditest performed better than the difficulty shift common-item set. Even though the semi-miditest may not have accurately represented a semi-miditest in

this dissertation, an important conclusion from the results of this dissertation is that the specific items selected as common items *do* impact equating results. Small differences in statistical characteristics of the common items can substantially impact equating results.

MC-CR 0.25

Results were similar for the MC-CR 0.25 sampling conditions as compared to the MC-CR 0.00 sampling conditions. The minitest common-item set resulted in the smallest values of WARMSB, and the difficulty shift common-item set resulted in the largest values of WARMSB. However, values of WARMSB were as much as 1 point larger for the semi-miditest as compared to the minitest, and classification consistency differed by as much as 7%. For the difficulty shift common-item set, values of WARMSB were as much as 2 points larger as compared to the minitest, and classification consistency differed by as much as 13%. Therefore, it appeared that for the MC-CR 0.25 sampling conditions, the statistical composition of the common-item set may have had a greater impact on the accuracy of the equating results. Similar to the MC-CR 0.00 sampling conditions, common-item and total test score correlations were similar for the minitest and semi-miditest. Correlations were always lowest for the difficulty shift common-item set.

How does the format of the CI impact equated scores?

MC-CR 0.00

One primary focus of this dissertation was investigation of whether the inclusion of CR items in addition to MC items in a common-item set results in more accurate equating relationships than MC items alone. For each of the pseudo-test forms, equating was conducted using an MC-only minitest and a common-item set that contained both MC and CR items. In general, across all three pseudo-test forms, the common-item set containing both MC and CR items tended to result in smaller values of WARMSB as compared to the MC-only common-item set. However, when the common-item effect

size was small (i.e., CI 0.00 or CI 0.20), bias was not always smaller for the common-item set containing both item formats. Further, for the TS and OS equating methods, the common-item set containing both item formats did not always reduce bias over the MC-only common-item set.

MC-CR 0.25

For the English Language and Chemistry pseudo-test forms, conclusions were split. For approximately half of the common-item effect size levels for English Language, the common-item set containing both item formats resulted in less bias than the MC-only common-item set. This result occurred for all four equating methods. For Chemistry, bias was always smaller for the MC-only common-item set as compared to the common-item set containing both item formats. For Spanish Language, however, the common-item set containing both MC and CR items resulted in substantially smaller values of WARMSB as compared to the MC-only minitest for FE and CE. However, for TS and OS, values of WARMSB were larger for the common-item set containing both item formats.

One plausible explanation for the English Language and Chemistry results is that the difference between the MC-CR 0.00 and MC-CR 0.25 sampling conditions was only approximately 0.15 for English Language and 0.05 for Chemistry. Another plausible explanation is that for both the Chemistry and English Language tests, only one CR item was selected for the common-item set because there were a limited number of CR items to choose from, and the point values of the CR items were all large. For Spanish Language, two different item formats were selected for the common-item set. Additionally, for Spanish Language, it was possible to have the number of CR points in the common-item set be nearly proportional to the number of CR points on the total test. For Chemistry, the proportion of CR points in the common-item set was smaller than the proportion of CR points on the total test. For English Language, the proportion of CR points in the common-item set was larger than the proportion of CR points on the total

test. Last, for both English Language and Chemistry, the average difficulty of the common-item sets was different from the average difficulty for the total test. Additionally, for English Language, the average difficulty of the common-item set containing both MC and CR items was more similar to the total test than the average difficulty for the MC-only common-item set.

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

Research Question Five

To what extent do analyses with two different classes of data, operational test forms and pseudo-test forms, result in the same findings?

The two classes of data resulted in similar conclusions; however, across the three tests, there were some differences in the results and interpretations based on operational test forms as compared to pseudo-test forms. All comparisons for the pseudo-test forms are based on the conditions for which the common-item set contained only MC items

(NCR). Additionally, because the magnitude of the difference between the MC-CR 0.00 and MC-CR 0.25 effect sizes differed between the operational and pseudo-test forms, comparisons are based only on the MC-CR 0.00 sampling conditions.

English Language

For both the English Language operational and pseudo-test form analyses, the conclusions based on the operational and pseudo-test forms were generally the same. WARMSB was largest for FE and smallest for TS and OS. For the MC-CR 0.00 sampling conditions, standardized values of WARMSB differed by less than 0.05 between the operational test and pseudo-test forms. Additionally, standard errors of equating were similar, although they were smaller in magnitude for the pseudo-test forms. However, for the operational test forms, standard errors of equating were always smallest for OS. For the pseudo-test forms, standard errors of equating were smallest for OS for only half of the MC-CR 0.00 sampling conditions.

Spanish Language

Not surprisingly, results for Spanish Language operational and pseudo-test forms were less similar than those for English Language, given that the operational Spanish Language 2006 test form contained 15 fewer MC items than the 2004 Spanish Language test form. The Spanish Language operational test forms represented a unique situation where equating, in the strictest sense, could not be done. Standardized values of WARMSB were smaller by as much as 0.10 standard deviation units on the pseudo-test forms as compared to the operational test forms. Standard errors of equating generally differed by less than 0.02 standard deviation units for operational test forms as compared to pseudo-test forms. For the MC-CR 0.00 sampling conditions, values of WASE for the operational test forms were only smallest for OS for one of the three MC-CR 0.00 sampling conditions. For the pseudo-test forms, values of WASE were smallest for OS for all of the MC-CR 0.00 sampling conditions.

Chemistry

In general, for Chemistry, similar results were seen across operational and pseudo-test forms. For FE and CE, similar patterns of WARMSB were seen across the operational test and pseudo-test forms. WARMSB steadily increased as the common-item effect size increased for the MC-CR 0.00 sampling conditions. Additionally, standardized values of WARMSB were within approximately 0.01 standard deviations units between operational and pseudo-test forms. However, patterns of WARMSB were somewhat different for TS and OS across operational test and pseudo-test forms. Standardized values of WASE were also within approximately 0.01 standard deviations units across operational test and pseudo-test forms. Further, for all MC-CR 0.00 sampling conditions, both operational test and pseudo-test forms led to the conclusions that WASE was smallest for OS and largest for CE. Also, across both classes of data, classification consistency was 90% or higher for all equating methods and sampling conditions.

As described previously in Chapter Two, most research has not compared operational and pseudo-test forms. Sinharay and Holland (2006, 2007) used operational test, pseudo-test, and simulated test forms for MC-only tests. Similar to research in this dissertation, they found that the different classes of data led to similar conclusions.

Practical Implications

This section is divided into two subsections corresponding to two practical implications from this dissertation: choice of equating method and composition of the common-item set.

Choice of Equating Method

When samples of examinees were relatively similar in proficiency across old and new test forms, all of the equating methods resulted in similar equated scores. Therefore, in practice, choice of equating method might be of little consequence when examinee groups are similar in proficiency. However, standard errors of equating were typically

larger for CE as compared to FE, TS, or OS. Consequently, when groups are similar in proficiency, FE, TS or OS might be preferred over CE. As the difference in proficiency between old and new form groups of examinees increased, TS, OS, and even CE were less sensitive to differences than FE. Therefore, for operational equating, it *may* be advisable to consider using TS, OS, or CE when examinee groups differ substantially in proficiency. However, because different criterion equating relationships were used for each of the equating methods, comparisons of bias across equating methods may not be reasonable. Further, because only a limited number of effect sizes were considered for this dissertation, it is not possible to determine how large an effect size must be before equating results are no longer reasonable.

Composition of the Common-Item Set

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

In terms of the statistical characteristics of the common-item set, results from this dissertation suggest that for mixed-format tests, MC-only common-item sets that are

constructed to be mini versions of the MC portion of the total test result in more accurate equating than other MC-only common-item set compositions. However, results for MC-only common-item set compositions were available for only one test; consequently, it is also not advisable to draw conclusions based on these limited results.

Limitations

This section is divided into a number of smaller subsections discussing the primary limitations of this dissertation.

Resampling Operational Data

After examining tables of effect sizes, calculating differences among effect sizes, studying descriptive statistics and equated moments, and analyzing plots and overall summary statistics, the limitations of resampling operational data were prominent. Disentangling the many data characteristics that may have led to the results found in this dissertation was difficult at best and impossible at worst. Although sampling was conducted to create various levels of effect sizes, the sampling process also resulted in differences in standard deviations, skewness, and kurtosis. Furthermore, because of the constraint of simultaneously creating different levels of common-item effect sizes and differences in the effect sizes of MC and CR items, it was not possible to hold old form means constant across sampling conditions. The lack of control undoubtedly limited the extent to which conclusions could be drawn about the impact of some of the conditions on the equating results. Consequently, the situations to which the results can be generalized were also limited.

A second limitation of the resampling methodology was that a number of the effect size combinations were difficult to create for certain tests. For other tests, such as Chemistry, some of the conditions simply could not be created. Given that it was difficult to obtain a sample of examinees for some of the effect size combinations, obtaining multiple samples containing different examinees was not possible for those conditions.

Therefore, bootstrap replications were used to create different samples of examinees rather than drawing new samples of examinees from the original full sample of examinees. Consequently, bias and RMSE were not calculated in the manner in which these statistics are typically calculated. Further, examinees were sampled based on two demographic variables in an effort to eliminate correlated error that could result from sampling examinees based on their common-item scores. However, because examinees were sampled from demographic groups until the target common-item effect size was obtained, it is possible that correlated errors were still introduced into the sampling process.

MC and CR Correlation Confounded with Subject Area

One of the limitations of the operational test forms selected for this dissertation was that the MC and CR correlation levels and subject areas for the tests were completely confounded. Consequently, it was impossible to disentangle the influence of the subject area test characteristics from the influence of the MC and CR correlations.

Score Weighting

Another limitation exclusive to the operational test forms was that the number of score points differed across new and old test forms. Operationally, each item type received a different weight, which ensured that the number of score points was the same for old and new test forms. However, the operational weighting scheme resulted in non-integer scores, which are not easily handled with currently available psychometric software. Consequently, the decision was made for this dissertation to use summed score weighting, which resulted in the old and new test forms containing different numbers of score points.

Criterion Equating Relationships

In this dissertation, the criterion equating relationship differed for each equating method. Because each equating method had a different criterion equating relationship, comparing bias across equating methods may not be advisable. Although the criterion equating relationships were similar across equating methods, differences among equating relationships existed. Consequently, it is not known which of the four criterion equating relationships (or, three criterion equating relationships for the pseudo-test forms) was accurate. Therefore, although the IRT equating methods may have appeared to result in less bias relative to the traditional equating methods, this conclusion may have been an artifact of the different criterion equating relationships.

Choice of Smoothing Value

Only one cubic spline postsmoothing value was selected in this dissertation. Although the choice of smoothing value represented a reasonable choice, it may not have been optimal across all equating methods, tests, or bootstrap replications. Selection of different smoothing values may have resulted in increases or decreases in bias and standard errors of equating for the traditional equating methods. Consequently, results for the traditional and IRT equating methods may have been more similar in terms of either bias or standard errors of equating.

Common-Item Effect Sizes

This dissertation examined only a limited number of common-item effect sizes. The effect sizes selected for this dissertation were chosen to represent small, moderate, and large effect sizes that might be seen in practice. However, because the increments between effect sizes were large, it was not possible to determine the precise effect size for which equating may no longer be reasonable.

Pseudo-Test Form and Common-Item Set Construction

Another limitation of this study was the manner in which pseudo-test forms were created. Limited content information was available for constructing the pseudo-test forms; consequently, the extent to which the pseudo-test forms represented the content of the operational tests is not well known. Additionally, pseudo-test forms were only constructed with an MC to CR point ratio similar to the operational test forms. Further, for a given common-item set composition (e.g., NCR, FCR, NCR SM), only one version was created. For example, for English Language, only one set of multiple-choice items was selected for NCR. Additionally, it was not always possible to create common-item sets containing the correct statistical characteristics or proportion of CR points. For example, the semi-miditest (NCR SM) for Spanish Language was very similar statistically to the minitest (NCR MT) for Spanish Language. For Chemistry, the proportion of CR points in the common-item set containing both MC and CR items (FCR) was not the same as the proportion of CR points in the total test.

Future Research

A variety of issues concerning equating mixed-format tests were considered in this dissertation. Although a number of limitations with the current research were discussed in the previous section, the diverse array of topics provides a rich starting point for future research. The current section describes some of the implications for future research from this dissertation.

Simulation Study

A number of the limitations in this dissertation can likely be addressed only through a simulation study. One of the main limitations was choice of criterion equating relationship. Employing a simulation study in future research will allow for the creation of a true population equating relationship.

A second primary limitation of this dissertation was use of the resampling methodology. The complex interactions among effect sizes for the various item formats illustrates the need to hold certain scores and effect sizes constant while manipulating others. A future simulation study should investigate alternative ways of calculating relative difficulty of MC and CR items. Specifically, examinees could be simulated with different proficiency levels across the two latent dimensions represented by MC and CR items. A simulation study could also incorporate various levels of MC and CR correlations for a given subject area. Differences in examinee proficiency on MC and CR items could also be simulated for various levels of MC and CR correlations. Also, different levels of MC and CR correlations in combination with levels of composite score and common-item score correlations may be informative. Another consideration for a simulation study would be to hold constant the difference between the composite score effect size and common-item score effect size as the common-item score effect size increased.

In particular, a future simulation study based on the Spanish Language test forms may be a logical starting point. The Spanish Language tests contain a variety of CR item formats, allowing for creation of various compositions of common-item sets. Additionally, results for the Spanish Language pseudo-test forms yielded rather compelling evidence in this dissertation, especially for the different common-item set compositions. A simulation study based on the Spanish Language test forms would allow for the opportunity to verify through replication some of the findings in this dissertation for Spanish Language. Additionally, a simulation study paired with the operational and pseudo-test analyses may provide both informative and more comprehensive results.

Resampling Considerations

Although a simulation study may be the best approach to further investigate some of the findings in this dissertation, future research could also be conducted with changes

to the resampling study. In this dissertation, common-item effect sizes were created by sampling examinees based on all items in common between test forms. However, different methods could also be considered, such as sampling based on MC common items, CR common items, or both MC and CR common items. Future research could also consider common-item effect sizes as compared to effect sizes for non-common items.

MC and CR Correlation Confounded with Subject Area

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

Score Weighting

Summed score weighting was considered for this dissertation; however, that was just one of several weighting schemes that could have been considered. Future research could incorporate research investigating the impact of rounding when non-integer weighting is used. Powers, Liu, Hagge, He, and Kolen (2010) investigated three different methods of rounding and found little difference in equating results among the rounding schemes. However, other rounding options could be investigated. Additionally, different combinations of integer or non-integer weights should be incorporated into future research to investigate how the relative weight of MC or CR items impacts equating accuracy.

Common-Item Effect Sizes

As noted previously, this dissertation investigated only a limited number of common-item effect sizes. Future research should focus on incrementally increasing the common-item effect sizes by 0.05 in order to investigate at what point the common-item effect size is too large for equating methods to produce reasonable results.

Choice of Smoothing Value

Future research should also consider analyses with different smoothing values for FE and CE. Selecting larger smoothing values may result in standard errors of equating for the traditional equating methods that are more similar to those for the IRT equating methods. Further, selecting different smoothing values would also likely impact the amount of bias in the equating relationships.

Pseudo-Test Form and Common-Item Set Construction

Future research with pseudo-tests should incorporate various compositions of pseudo-test forms by varying the proportion of MC points to CR points or changing the length of the total test. For each of the common-item compositions (e.g., NCR, FCR), multiple versions should be created in order to determine whether results are similar when different items are chosen as common items. Additional factors, such as length of the common-item set could also be manipulated.

Other Considerations for Future Research

A number of decisions were made in this dissertation that were not factors of investigation, such as choice of old and new form weights for FE and OS, software programs, or prior distributions for IRT item parameter estimation. Choice of new form weights for FE and OS could be compared in future research studies, especially when the difference in proficiency between old and new form examinees is large. Future research could also implement different software programs, such as MULTILOG for IRT item

parameter estimation or *Equating Recipes* for IRT equating. Different prior distributions could also be considered in either MULTILOG or PARSCALE in order to improve item parameter estimation. Further, in this dissertation, it was found that bias calculated based on bootstrap replications and difference based on one sample yielded similar results. Consequently, replications in future resampling studies may be unnecessary unless comparison of standard errors of equating is desired.

Conclusion

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

REFERENCES

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

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

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

Brennan, R. L., Wang, T., Kim, S.-H., & Seol, J. (2009). *Equating recipes*. [Computer program]. Iowa City: University of Iowa.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*, 13-20.

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

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement, 54*, 8-20.

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

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225-244.

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

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

Ebel, R. L. (1980). *Practical problems in educational measurement.* Lexington, MA: D. C. Heath.

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

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

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*, 195-208.

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

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

Haladyna, T. M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5*(1), 73-88.

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.

Harris, D. J. (1991). *Equating with non-representative common item sets and non-equivalent groups.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

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

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

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 129-164.

Holland, P. W., Sinharay, S., von Davier, A.A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement, 45*(1), 17-43.

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

Kamata, A., & Tate, R. (2005). The performance of a method for the long-term equating of mixed-format assessment. *Journal of Educational Measurement, 42*, 193-213.

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

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

Kim, S.-H., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*(1), 53-76.

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

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

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common item equating with nonrandom groups. *Journal of Educational Measurement, 22*, 197-206.

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

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

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531-578). Westport, CT: American Council on Education/Praeger.

Lee, W.-C., Hagge, S., He, Y., Kolen, M., & Wang, W. (May, 2010). *Equating mixed-format tests using dichotomous anchor items.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 119-158). Washington, DC: American Council on Education.

Linn, R. L., Baker, E. L., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), p. 15-21.

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

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

Lukhele, R., Thissen, D., & Wainer, H. (1993). *On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests.* Technical Report (RR-93-6). Princeton, N.J.: Educational Testing Service. (ED385544)

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139-160.

Mislevy, R. J. (1993). A framework for studying differences between multiple-choice and constructed-response test items. In R. E. Bennett and W. C. Ward (Eds.). *Construction versus choice in cognitive measurement* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Powers, S., Liu, C., Hagge, S., He, Y., & Kolen, M. (2010). *A comparison of non-linear equating methods for mixed-format tests.* Poster presented at the annual conference of the American Educational Research Association, Denver, CO.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1982). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Petersen, 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* 9pp. 71-135). New York: Academic Press, Inc.

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

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

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

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

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing scores in test and questionnaire data. *Multivariate Behavioral Research, 38*(4), 505-528.

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

Sinharay, S., & Holland, P. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

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

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

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

Tate, R. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36,* 336-346/

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

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

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

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and constructed-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement,* 31(2), 113-123.

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.

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

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*(1), 11-26.

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

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

Wang, T., Lee, W.-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.

Whitney, D. R. (1989). Educational admissions and placement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 515-525). New York, NY: Macmillan.

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