Iowa Research Online
The University of Iowa's Institutional Repository

University of Iowa
Iowa Research Online

Theses and Dissertations

2010

Observed score and true score equating procedures for multidimensional item response theory

Bradley Grant Brossman
University of Iowa

Copyright 2010 Bradley Grant Brossman

This dissertation is available at Iowa Research Online: http://ir.uiowa.edu/etd/469

Recommended Citation
Brossman, Bradley Grant. "Observed score and true score equating procedures for multidimensional item response theory." PhD (Doctor of Philosophy) thesis, University of Iowa, 2010.
http://ir.uiowa.edu/etd/469.

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

Part of the Educational Psychology Commons

OBSERVED SCORE AND TRUE SCORE EQUATING PROCEDURES FOR MULTIDIMENSIONAL ITEM RESPONSE THEORY

by
Bradley Grant Brossman

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

May 2010

Thesis Supervisors: Associate Professor Timothy N. Ansley
Adjunct Assistant Professor Won-Chan Lee

ABSTRACT

The purpose of this research was to develop observed score and true score equating procedures to be used in conjunction with the Multidimensional Item Response Theory (MIRT) framework. Currently, MIRT scale linking procedures exist to place item parameter estimates and ability estimates on the same scale after separate calibrations are conducted. These procedures account for indeterminacies in (1) translation, (2) dilation, (3) rotation, and (4) correlation. However, no procedures currently exist to equate number correct scores after parameter estimates are placed on the same scale. This research sought to fill this void in the current psychometric literature.

Three equating procedures—two observed score procedures and one true score procedure—were created and described in detail. One observed score procedure was presented as a direct extension of unidimensional IRT observed score equating, and is referred to as the "Full MIRT Observed Score Equating Procedure." The true score procedure and the second observed score procedure incorporated the statistical definition of the "direction of best measurement" in an attempt to equate exams using unidimensional IRT (UIRT) equating principles. These procedures are referred to as the "Unidimensional Approximation of MIRT True Score Equating Procedure" and the "Unidimensional Approximation of MIRT Observed Score Equating Procedure," respectively.

Three exams within the Iowa Test of Educational Development (ITED) Form A and Form B batteries were used to conduct UIRT observed score and true score equating, MIRT observed score and true score equating, and equipercentile equating. The equipercentile equating procedure was conducted for the purpose of comparison since this procedure does not explicitly violate the IRT assumption of unidimensionality.

Results indicated that the MIRT equating procedures performed more similarly to the equipercentile equating procedure than the UIRT equating procedures, presumably due to the violation of the unidimensionality assumption under the UIRT equating procedures. Future

studies are expected to address how the MIRT procedures perform under varying levels of multidimensionality (weak, moderate, strong), varying frameworks of dimensionality (simple structure vs. complex structure), and number of dimensions, among other conditions.

Abstract Approved:

_____
Timothy N. Ansley, Thesis Supervisor

_____
Title and Department

_____
Date

_____
Won-Chan Lee, Thesis Supervisor

_____
Title and Department

_____
Date

OBSERVED SCORE AND TRUE SCORE EQUATING PROCEDURES FOR MULTIDIMENSIONAL ITEM RESPONSE THEORY

by

Bradley Grant Brossman

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

May 2010

Thesis Supervisors: Associate Professor Timothy N. Ansley
Adjunct Assistant Professor Won-Chan Lee

Copyright by

BRADLEY GRANT BROSSMAN

2010

All Rights Reserved

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

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

Bradley Grant Brossman

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 May 2010 graduation.

Thesis Committee: _____
Timothy N. Ansley, Thesis Supervisor

_____
Won-Chan Lee, Thesis Supervisor

_____
Robert L. Brennan

_____
Michael J. Kolen

_____
Dale Zimmerman

To Fred, Nancy, Jessica, Nick, and Archie

I have fought the good fight, I have finished the course...

2 Timothy 4:7

ACKNOWLEDGMENTS

I would like to thank my dissertation advisors, Dr. Timothy Ansley and Dr. Won-Chan Lee, for their help and guidance throughout the dissertation writing process. Dr. Timothy Ansley always provided invaluable feedback pertaining to my dissertation work. Countless hours were spent in Dr. Won-Chan Lee's office, collaborating on research ideas, determining mathematical solutions, and being delivered coffee from fellow graduate students in the department. I would also like to thank my dissertation members, Dr. Michael Kolen, Dr. Robert Brennan, and Dr. Dale Zimmerman for their service in regards to my dissertation progress. Having been given the opportunity to view my dissertation research through their eyes has definitely opened up many ideas that otherwise would never have entered my mind.

I would like to thank my parents, Fred and Nancy Brossman, for all of the support that they provided along the way. I certainly would not have completed my doctoral studies if I had not inherited the work ethic, dedication, and principles of living modeled by these exemplars. I would also like to thank my sister and my brother, Jessica and Nicholas. Jessica's location in Houston always provided a warm harbor for me to escape the harsh Iowa weather over Spring Break, and Nick's Playstation skills always provided an adequate challenge over Summer and Winter Breaks.

Lastly, I would like to thank the many friends and colleagues in my department that I met along the way. I certainly would not have retained my sanity if it were not for the happy hours, tailgates, coffee runs, and—of course—the many eventful moving excursions. Come tornado and flood, we made it through. So many individuals have entered my life and contributed along the way that I do not feel comfortable naming just a few. Thank you, to all.

ABSTRACT

The purpose of this research was to develop observed score and true score equating procedures to be used in conjunction with the Multidimensional Item Response Theory (MIRT) framework. Currently, MIRT scale linking procedures exist to place item parameter estimates and ability estimates on the same scale after separate calibrations are conducted. These procedures account for indeterminacies in (1) translation, (2) dilation, (3) rotation, and (4) correlation. However, no procedures currently exist to equate number correct scores after parameter estimates are placed on the same scale. This research sought to fill this void in the current psychometric literature.

Three equating procedures—two observed score procedures and one true score procedure—were created and described in detail. One observed score procedure was presented as a direct extension of unidimensional IRT observed score equating, and is referred to as the "Full MIRT Observed Score Equating Procedure." The true score procedure and the second observed score procedure incorporated the statistical definition of the "direction of best measurement" in an attempt to equate exams using unidimensional IRT (UIRT) equating principles. These procedures are referred to as the "Unidimensional Approximation of MIRT True Score Equating Procedure" and the "Unidimensional Approximation of MIRT Observed Score Equating Procedure," respectively.

Three exams within the Iowa Test of Educational Development (ITED) Form A and Form B batteries were used to conduct UIRT observed score and true score equating, MIRT observed score and true score equating, and equipercentile equating. The equipercentile equating procedure was conducted for the purpose of comparison since this procedure does not explicitly violate the IRT assumption of unidimensionality.

Results indicated that the MIRT equating procedures performed more similarly to the equipercentile equating procedure than the UIRT equating procedures, presumably due to the violation of the unidimensionality assumption under the UIRT equating procedures. Future studies are expected to address how the MIRT procedures perform under varying levels of multidimensionality (weak, moderate, strong), varying frameworks of dimensionality (simple structure vs. complex structure), and number of dimensions, among other conditions.

TABLE OF CONTENTS

CHAPTER I INTRODUCTION 1

Observed Score and True Score Equating 1
Item Response Theory 3
IRT Equating 4
Multidimensional Item Response Theory 6
Multidimensional IRT Equating 8
Research Statements 9

CHAPTER 2 LITERATURE REVIEW 11

Item Response Theory 11
Multidimensional Item Response Theory 15
Dimensionality Assessment 23
Unidimensional IRT Scale Linking 24
Moment Transformation Methods 26
Characteristic Curve Transformation Methods 29
MIRT Scale Linking 30
MIRT Scale Linking for Nonequivalent Groups 33
MIRT Scale Linking for Randomly Equivalent Groups 35
Unidimensional IRT Equating 40
IRT Observed Score Equating 41
IRT True Score Equating 42
Foundations for MIRT Equating 44
Unidimensional Approximation 44

CHAPTER 3 METHODOLOGY 50

Data and Procedures 50
Unidimensional Equating Procedures 51
Multidimensional Equating Methodology 53
Full MIRT Observed Score Equating 53
Unidimensional Approximation 55
Unidimensional Approximation of MIRT True Score Equating 58
Unidimensional Approximation of MIRT Observed Score Equating 59
Multidimensional Procedures 59
Scale Linking and Equating Assumptions 65
Other Procedures for Conducting MIRT Equating 69
Equipercentile Equating 72
Evaluation Procedures 73
Evaluation of Linking Procedures 74
Standard Error of Equating 75
Differences That Matter 77

CHAPTER 4 RESULTS 79

Descriptive Statistics for Each Form 79
Dimensionality Assessment 80
Linking Results 82

Standard Error of Equating 86
Equating Results 87
Overall Trends 92
Equating of the Math Exams 93
Equating of the Science Exams 94
Equating of the Social Studies Exams 96
Summary of Equating Results 96

CHAPTER 5 DISCUSSION AND CONCLUSION 99

Dimensionality Assessment 99
Linking Results 103
Standard Error of Equating 107
Equating Results 109
Differences Between Unidimensional Procedures and Multidimensional Procedures 109
Differences with the Equipercentile Equating Procedure 114
Limitations and Future Studies 115
Limitations Associated with using the Equipercentile Equating Procedure as the Benchmark for Comparison 115
Limitations of using “Real” Data 118
Summary and Conclusion 120

REFERENCES 123

APPENDIX A. TABLES AND FIGURES 127

APPENDIX B. SAMPLE COMPUTER CODE 177

APPENDIX C. MULTIDIMENSIONAL EQUATING EXAMPLE 210

LIST OF TABLES

Table A-1. Descriptive Statistics for Each Form 127

Table A-2. Confirmatory and Exploratory DETECT Statistics for Math Forms 128

Table A-3. Confirmatory and Exploratory DETECT Statistics for Science Forms 128

Table A-4. Confirmatory and Exploratory DETECT Statistics for Social Studies Forms 128

Table A-5. Pre-rotation and Post-rotation Angles Between Corresponding Reference Composites 129

Table A-6. Comparison of Math ITED Classification Tables and DETECT Procedure 130

Table A-7. Comparison of Science ITED Classification Tables and DETECT Procedure 131

Table A-8. Comparison of Social Studies ITED Classification Tables and DETECT Procedure 132

Table A-9. Standard Error of Equating for Unsmoothed Equipercentile Procedure 133

Table A-10. Standard Error of Equating for Smoothed Equipercentile Procedure 134

Table A-11. Equating Results for Math Exams 135

Table A-12. Equating Results for Science Exams 137

Table A-13. Equating Results for Social Studies Exams 139

Table A-14. Differences Between Equating Results and Unsmoothed Equipercentile Results for Math Exams 141

Table A-15. Differences Between Equating Results and Unsmoothed Equipercentile Results for Science Exams 143

Table A-16. Differences Between Equating Results and Unsmoothed Equipercentile Results for Social Studies Exams 145

Table A-17. Differences Between Equating Results and Smoothed Equipercentile Results for Math Exams 147

Table A-18. Differences Between Equating Results and Smoothed Equipercentile Results for Science Exams 149

Table A-19. Differences Between Equating Results and Smoothed Equipercentile Results for Social Studies Exams 151

Table A-20. Statistics for Item Parameter Estimates 153

Table A-21. Correlations Between Unidimensional Item Parameter Estimates and Unidimensional Approximation Item Parameter Estimates 154

LIST OF FIGURES

Figure 2-1. Example Item Characteristic Curve (ICC) 12

Figure 2-2. Example Test Characteristic Curve (TCC) 14

Figure 2-3. Example Item Characteristic Surface (ICS) plot 16

Figure 2-4. Angle between direction of best measurement and coordinate axes 19

Figure 2-5. Example Test Characteristic Surface (TCS) 20

Figure 2-6. Example Reference Composite 22

Figure 2-7. A comparison of UIRT and MIRT Linking Methods 32

Figure 2-8. Graphical Representation of True Score Equating 44

Figure 5-1. Comparison of Pre-Rotation and Post-Rotation Items 106

Figure A-1. Observed Score Distributions for Math Forms 155

Figure A-2. Observed Score Distributions for Science Forms 156

Figure A-3. Observed Score Distributions for Social Studies Forms 157

Figure A-4. Differences Between Unsmoothed Equating Results and Identity Equating for Math Exams 158

Figure A-5. Differences Between Unsmoothed Equating Results and Identity Equating for Science Exams 159

Figure A-6. Differences Between Unsmoothed Equating Results and Identity Equating for Social Studies Exams 160

Figure A-7. Differences Between Smoothed Equating Results and Identity Equating for Math Exams 161

Figure A-8. Differences Between Smoothed Equating Results and Identity Equating for Science Exams 162

Figure A-9. Differences Between Smoothed Equating Results and Identity Equating for Social Studies Exams 163

Figure A-10. Differences Between Equating Results and Unsmoothed Equipercentile Results for Math Exams 164

Figure A-11. Differences Between Equating Results and Unsmoothed Equipercentile Results for Science Exams 165

Figure A-12. Differences Between Equating Results and Unsmoothed Equipercentile Results for Social Studies Exams 166

Figure A-13. Differences Between Equating Results and Smoothed Equipercentile Results for Math Exams 167

Figure A-14. Differences Between Equating Results and Smoothed Equipercentile Results for Science Exams 168

Figure A-15. Differences Between Equating Results and Smoothed Equipercentile Equating Results for Social Studies Exams 169

Figure A-16. Absolute Differences Between Equating Results and Unsmoothed Equipercentile Results for Math Exams 170

Figure A-17. Absolute Differences Between Equating Results and Unsmoothed Equipercentile Results for Science Exams 171

Figure A-18. Absolute Differences Between Equating Results and Unsmoothed Equipercentile Results for Social Studies Exams.....172

Figure A-19. Absolute Differences Between Equating Results and Smoothed Equipercentile Results for Math Exams 173

Figure A-20. Absolute Differences Between Equating Results and Smoothed Equipercentile Results for Science Exams 174

Figure A-21. Abolute Differences Between Equating Results and Smoothed Equipercentile Results for Social Studies Exams.....175

Figure A-22. Ability Distributions as the Number of Quadrature Points Increases 176

Figure B-1. TESTFACT Code for Form A Math Exam 177

Figure B-2. R Code 178

CHAPTER I

INTRODUCTION

Large-scale testing programs often create and administer parallel test forms because of item exposure and test security issues (Kolen & Brennan, 2004). Although parallel forms are constructed to be as similar as possible in terms of content and statistical specifications, often the exams differ in difficulty. As a result, an adjustment must be made to correct for differences in difficulty in order for examinee scores to be comparable across test forms. This adjustment process is known as "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).

The equating process should be designed to yield the most accurate equating relationship possible. If accurate relationships are derived from the equating process, examinees can be compared on parallel forms since differences in form difficulty have been taken into account. If the equating procedure yields inaccurate results, scores on different forms cannot be compared: there is no common metric on which scores can be evaluated (Kolen & Brennan, 2004).

Observed Score and True Score Equating

Various equating procedures have been developed and are used in conjunction with different psychometric models. Equating procedures typically fall within one of two categories: procedures that focus on observed scores (and observed score distributions), and procedures that focus on true scores. The goal of observed score equating is to make

an adjustment such that the properties of score distributions across parallel forms are as similar as possible. The result is an established equating relationship between observed scores on two parallel forms. The goal of true score equating procedures is to map true scores on one form to true scores on another form (via the definition of true score in either classical test theory or item response theory), as opposed to mapping observed scores (Kolen & Brennan, 2004).

Both observed score and true score equating procedures can be conducted in conjunction with one of several data collection methods, or "equating designs" (Kolen & Brennan, 2004). In a single-group design, for example, examinees complete both a linking form (denoted "Form A" in this study), and a base form ("Form B"). In practice, counterbalancing is typically used (half of the examinees take Form A first, and then Form B; the rest of the examinees take Form B first, and then Form A) to help account for order effect. After the test administration, scores on Form A are mapped to scores on Form B. In a random groups design, examinees take either Form A *or* Form B, not both. The assumption that both groups are randomly equivalent is exploited in order to determine an equating relationship between the forms. In the common-item nonequivalent groups design, items which are common to both Form A and Form B serve as the basis on which to compute the appropriate statistical adjustment. Although the groups are not assumed to be equal in the measured trait, performance on the common items between the two groups can be compared to determine an overall equating relationship between the forms.

Equating procedures have been developed for many combinations of data collection design (single-group, random group, and common-item nonequivalent group)

and procedure type (observed score and true score). Mean, linear, and equipercentile procedures—which focus on properties of the observed score distributions—have been developed for use with the random groups design. Extensions of the linear equating procedure—namely the Tucker and Levine observed score methods—have been developed for use with the common-item nonequivalent groups (CINEG) design. A technique known as "frequency estimation"—which can be viewed as an extension of the equipercentile procedure—has been developed for the nonequivalent groups design. The chained linear and chained equipercentile procedures—which first determine an equating relationship between Form A and common items, and then transfer the relationship to Form B—have also been created for use with the CINEG design. The Levine true score method—a procedure to be used in conjunction with the classical congeneric psychometric model—has been developed for use with the common-item nonequivalent groups design (Kolen & Brennan, 2004).

Item Response Theory

Whereas the methods previously described either do not require the explicit use of a psychometric model or are closely aligned with the classical test theory (CTT) framework, another class of equating procedures has been developed to be used in conjunction with the item response theory (IRT) framework. Item response theory consists of a family of probabilistic models which relates an examinee's proficiency level (θ) to the probability of answering an item within a particular category (Lord, 1980). Items are typically classified as either dichotomous, meaning that there are only two response categories (usually denoted "correct" and "incorrect"); or polytomous, meaning that responses can be scored in one of several categories.

Two common mathematical forms that are used to model the probability of a correct response for dichotomous items are logistic models and normal ogive models (Lord, 1980). Logistic models are of the form:

$$p_{ij}(\theta_i) = c_j + (1 - c_j)\frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_i - b_j)}} \tag{1.1}$$

Normal ogive models are of the form:

$$p_{ij}(\theta_i) = c_j + (1 - c_j)\int_{-\infty}^{a_j(\theta_i - b_j)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} = c_j + (1 - c_j)\Phi\left(a_j(\theta_i - b_j)\right) \tag{1.2}$$

For both models, a_j represents the item discrimination parameter, b_j represents the item difficulty parameter, c_j represents the lower asymptote parameter, θ_i represents a specific proficiency level, and, for the normal ogive model, Φ represents the standard normal cumulative distribution function (CDF) (Lord, 1980).

IRT Equating

Before equating can be conducted in the IRT framework, item parameters and examinee abilities must first be estimated. Due to the scale indeterminacy property of item response theory, any linear transformation of the ability scale will yield the same probabilistic relationships, assuming that both the ability scale and item parameters are transformed. To resolve the indeterminacy issue, IRT calibrations are typically specified to yield an ability distribution with mean of 0 and standard deviation of 1 (Kolen & Brennan, 2004). Therefore, when item parameters and abilities are calibrated separately for groups of examinees who are different with respect to the measured ability (i.e., nonequivalent groups) the estimates will be placed on different—yet linearly related—ability scales. "In item response theory the item parameters are invariant from group to group as long as the ability scale is not changed…Similarly, ability is invariant across

tests of the same psychological dimension as long as the ability scale is not changed" (Lord, p. 38). Therefore, a scale linking procedure must first be conducted to place item parameter estimates and ability estimates from the Form A scale to that of the Form B scale before equating can be conducted (Kolen & Brennan, 2004).

When parameter estimates are placed on different scales due to separate calibrations under the nonequivalent groups design, the scale of Form A and the scale of Form B may differ in two aspects: the scales may differ in (1) origin (i.e., mean) and (2) unit of measurement (i.e., standard deviation). Scale linking is not required when item parameters and abilities are estimated in the same calibration for both groups (concurrent calibration), or when separate calibrations are conducted under the random groups design. If separate calibrations are conducted under the random groups design, examinees that were administered different forms are assumed to be equivalent on the measured trait. That is, examinees that were administered different tests are assumed to have equal means and standard deviations on the ability scale. If both calibrations are specified to yield a standard normal ability distribution, no scale linking method must be employed since means and standard deviations are specified to be equal across groups (Kolen & Brennan, 2004).

Both observed score equating and true score equating can be conducted within the IRT framework. To conduct IRT observed score equating, conditional observed score distributions are first estimated at each specified ability level. The conditional distributions are then multiplied by the ability density and integrated (or summed) across all ability levels to produce an estimated marginal observed score distribution. Once this process has been completed for both Form A and Form B, equipercentile equating is

conducted to determine an equating relationship between the two forms (Kolen & Brennan, 2004).

IRT true score equating establishes a relationship between true scores on both forms. First, a true score on Form A is selected. Using an iterative procedure such as the Newton-Raphson method, the ability level (θ) associated with this true score is estimated. The true score on Form B that corresponds to this ability level is then estimated using the IRT definition of true score—i.e. $\tau_B(\theta_i) = \sum_{j:B} p_{ij}(\theta_i; a_j, b_j, c_j)$. This procedure is typically conducted for each integer raw score on Form A in order to establish a relationship between true scores on both forms (Kolen & Brennan, 2004).

Multidimensional Item Response Theory

In order to use item response theory to analyze test data, the statistical assumptions underlying the particular IRT model must first be adequately satisfied. Oftentimes, it is assumed that a test is unidimensional, i.e., that the test measures only one ability. If the test measures more than one trait but the responses are analyzed using a unidimensional IRT model, the resulting ability estimates and item parameter estimates may be highly inaccurate depending on the nature of the other variables being measured (Ansley & Forsyth, 1985; Reckase, 1985; Sireci, Thissen, & Wainer, 1991). An extension of item response theory—known as multidimensional item response theory (or "MIRT")—has been developed for use with multidimensional data (Ackerman, 1994; Ackerman, Gierl, & Walker, 2003; Reckase, 1985; Reckase, 2009).

Similar to unidimensional IRT, MIRT consists of a family of probabilistic models. MIRT models typically fall within one of two categories: compensatory and non-compensatory models (Ackerman, 1994; Ackerman, 1996; Ackerman et al., 2003;

Reckase, 2009). The mathematical form that compensatory MIRT models take is additive in nature, allowing examinees who have low proficiency on one trait to compensate for this weakness by having high proficiency on another trait. The non-compensatory MIRT model is multiplicative in nature and stands in contrast to the compensatory model: examinees that have low proficiency on one trait cannot compensate by having high proficiency on another trait. In comparison to unidimensional IRT models, which relate an examinee's proficiency level (θ) to the probability of correctly answering an item, both compensatory and non-compensatory MIRT models relate a vector of an examinee's proficiency levels ($\boldsymbol{\theta}$) on the specified traits to the probability of correctly answering an item. Each element in the proficiency vector ($\boldsymbol{\theta}$) corresponds to one of the specified dimensions on the exam (Ackerman, 1994; Ackerman, 1996; Ackerman et al., 2003; Reckase, 1985; Reckase, 2009).

Though several compensatory MIRT models exist, two of these models can be viewed as direct extensions of their unidimensional counterparts: the multidimensional compensatory logistic model and the multidimensional compensatory normal ogive model (Ackerman, 1994; Ackerman, 1996; Ackerman et al., 2003; Reckase, 1985; Reckase, 2009). Multidimensional logistic models are of the form:

$$p_{ij}(\boldsymbol{\theta}_i) = c_j + (1 - c_j)\left(\frac{\exp\left[1.7(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i + d_j)\right]}{1 + \exp\left[1.7(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i + d_j)\right]}\right) \tag{1.3}$$

Multidimensional normal ogive models are of form:

$$p_{ij}(\boldsymbol{\theta}_i) = c_j + (1 - c_j)\Phi(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i + d_j) \tag{1.4}$$

In both of these equations, $\mathbf{a}_j^{\mathrm{T}}$ represents the item discrimination vector, d_j represents a scalar location parameter related to multidimensional item difficulty, c_j represents the

lower asymptote parameter, $\boldsymbol{\theta}_i$ represents a vector of proficiency levels corresponding to each specified trait, $\mathbf{T}$ represents the transpose function, and, for the multidimensional normal ogive model, Φ represents the standard normal cumulative distribution function (CDF) (Ackerman, 1994; Ackerman, 1996; Ackerman et al., 2003; Reckase, 1985; Reckase, 2009).

Multidimensional IRT Equating

The success of an IRT equating is dependent in part upon the statistical assumptions associated with the chosen model being met. If a unidimensional IRT equating procedure is applied to multidimensional data, the resulting equating relationships will most likely be inaccurate. Rather than using unidimensional IRT equating methodology for multidimensional data, MIRT equating methodology should be developed to be used in conjunction with multidimensional data to help ensure equating accuracy (if it is desired to conduct equating within a specific psychometric framework such as IRT or MIRT).

To adjust for multidimensional data, several procedures have been developed to link scales within the MIRT framework (Davey et al., 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000; Thompson et al., 1997; Yon, 2006). Whereas unidimensional scale linking procedures estimate two coefficients to adjust for differences in (1) origin and (2) unit of measurement resulting from separate calibrations, MIRT scale linking procedures typically estimate scalar and matrix coefficients to adjust for differences in (1) rotation, (2) correlation, (3) translation (similar to “origin” in unidimensional IRT), and (4) dilation (similar to “unit of measurement” in unidimensional IRT). That is, separate calibrations within the MIRT framework may

place ability estimates and item parameter estimates on separate scales due to rotational indeterminacy (similar to factor analysis procedures), correlational indeterminacy (though solutions are often derived by fixing multidimensional traits to follow a multivariate standard normal distribution with no correlation between dimensions, i.e., $MVN(\mathbf{0},\mathbf{I})$), translation indeterminacy, and dilation indeterminacy.

Although various procedures have been developed to link scales in the MIRT framework, no procedures have yet been developed to equate number-correct scores within the MIRT framework (Reckase, 2009). Whereas scale linking *and* number-correct score equating have been developed for the unidimensional IRT framework, only scale linking has been developed for the MIRT framework. As a result, number-correct equating procedures should be developed for the MIRT framework to help ensure equating accuracy for multidimensional data.

Research Statements

Equating is an integral aspect of the test development process if scores are to be comparable across parallel forms. The equating procedure should estimate—as accurately as possible—the relationship between scores on two parallel forms. If a unidimensional IRT equating procedure is applied to multidimensional data, the resulting equating relationships will most likely contain a large amount of systematic error due to the violation of the unidimensionality assumption. Therefore, it is imperative that MIRT equating procedures are developed to be used in conjunction with the MIRT framework. The purpose of the present research is to:

(1) Develop theoretical foundations for conducting observed score and true score equating within the MIRT framework.

(2) Demonstrate how these procedures are conducted by applying these procedures to real test data.

(3) Compare the MIRT equating results to results produced by unidimensional IRT equating and traditional equipercentile equating using the same datasets.

CHAPTER 2

LITERATURE REVIEW

This chapter is comprised of seven main sections. First, a review of Item Response Theory and Multidimensional Item Response Theory will be presented, with special attention given to features related to equating methodology. Next, a discussion concerning unidimensionality and dimensionality assessment appears. Following that, current procedures for conducting IRT scale linking and MIRT scale linking will be presented. Lastly, current procedures for conducting IRT equating, followed by a discussion of existing methodology that will form the bases for conducting the proposed MIRT equating procedures, will be presented. These topics were selected to follow the equating sequence typically followed in practice, i.e., dimensionality assessment, linking scales (if necessary), and then conducting equating.

Item Response Theory

Item response theory consists of a family of probabilistic models which relate an examinee's proficiency level (θ) to the probability of correctly answering an item (Lord, 1980). For dichotomous items, the probability of correctly answering an item can be modeled mathematically using the logistic model or the normal ogive model. This relationship can also be represented graphically through the item characteristic curve (ICC) (see Figure 2-1 for a graphical representation of an ICC). The form of the three parameter logistic model is dictated by three parameters: the item discrimination parameter (a), the item difficulty parameter (b), and the lower asymptote parameter (c) (Lord, 1980).

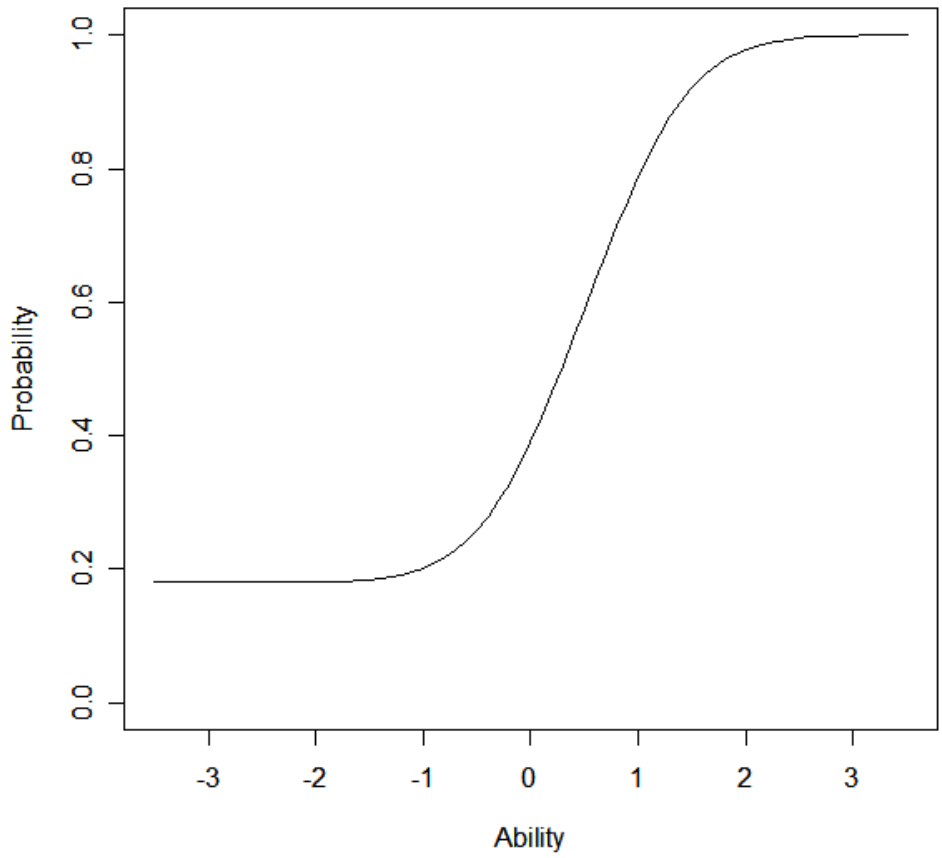


Figure 2-1. Example Item Characteristic Curve (ICC)

$a = 1.3$, $b = 0.5$, and $c = 0.18$

The item difficulty parameter provides an indication of the difficulty level of the item and primarily dictates the location of the ICC with respect to the ability (θ) scale. A larger difficulty parameter results in a more difficult item and shifts the ICC upscale in reference to the ability scale. The item discrimination parameter provides an indication of how well the item discriminates between examinees of similar ability level and primarily dictates the magnitude of the slope of the ICC. A larger discrimination parameter results in more discrimination power and yields a steeper ICC slope. Finally,

the lower asymptote parameter takes random responding (guessing) into account and provides an indication of how well an examinee of very low ability should perform on the item (Lord, 1980).

Two other concepts related to the form of the IRT model are the test characteristic curve (TCC) and the information function. The test characteristic curve is the sum of the item characteristic curves across all items and is conceptually viewed as the regression of the summed score responses on ability (Lord, 1980) (see Figure 2-2 for an example of a TCC). The information function can be viewed as the precision of the estimation procedure at a specific ability level (for MLE procedures) and is determined by the characteristics of the ICC. Information can be determined for each item individually, or item-level information functions can be summed across all items to determine information at the test level. Graphically, information for measuring a particular ability level from the summed score responses can be viewed as the slope of the TCC divided by the error variance at that particular ability level (Lord, 1980).

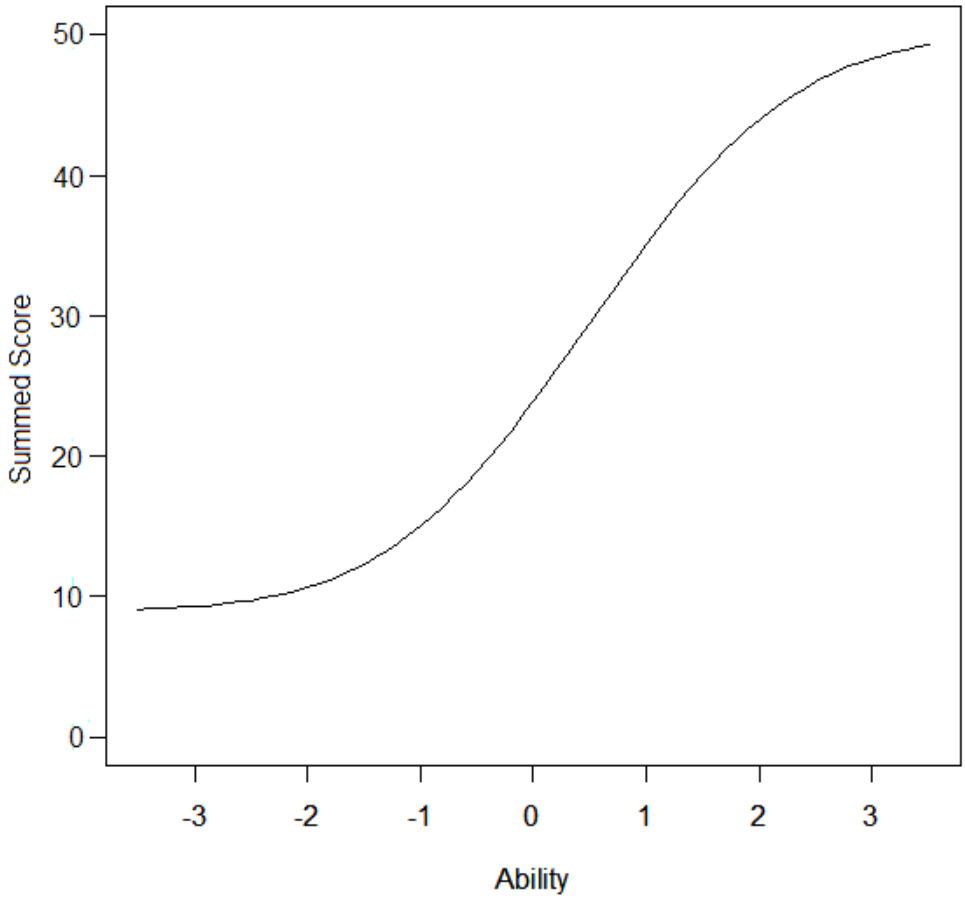


Figure 2-2. Example Test Characteristic Curve (TCC)

for a 50-item Test

Item response theory is, in essence, a scaling method. Estimation procedures attempt to locate both persons and items along a single continuum. In order to accurately locate persons and items along this continuum, it must be assumed that a test measures only one trait in the population of interest, and that items and examinees can be ordered along this continuum. If more than one dimension is present, yet the responses are modeled by a unidimensional IRT (UIRT) model, persons and items will still be scaled along a single continuum. In this situation, the UIRT procedure would collapse across

other relevant dimensions which discriminate between persons of different abilities, resulting in inaccurate relationships between examinee abilities and the probability of obtaining a correct response. Whereas examinees may be scaled in similar locations according to the single continuum, examinees may actually differ substantially on another dimension not accounted for in the unidimensional estimation procedure (Reckase, 2009).

Multidimensional Item Response Theory

Multidimensional item response theory (MIRT) was developed as an extension of IRT to provide a more accurate representation of persons and items in multidimensional space. As a result, the probability of obtaining a correct response can be more accurately modeled in the MIRT framework if, in fact, more than one dimension is being measured. Whereas UIRT models relate examinee ability to the probability of a correct response through the item characteristic curve, MIRT models relate examinee abilities (on two or more traits) to the probability of a correct response through the item characteristic surface (ICS) (see Figure 2-3 for a graphical representation of a two-dimensional ICS). Similar to the item characteristic curve, the item characteristic surface is dictated by parameters related to item discrimination, item difficulty, and a lower asymptote.

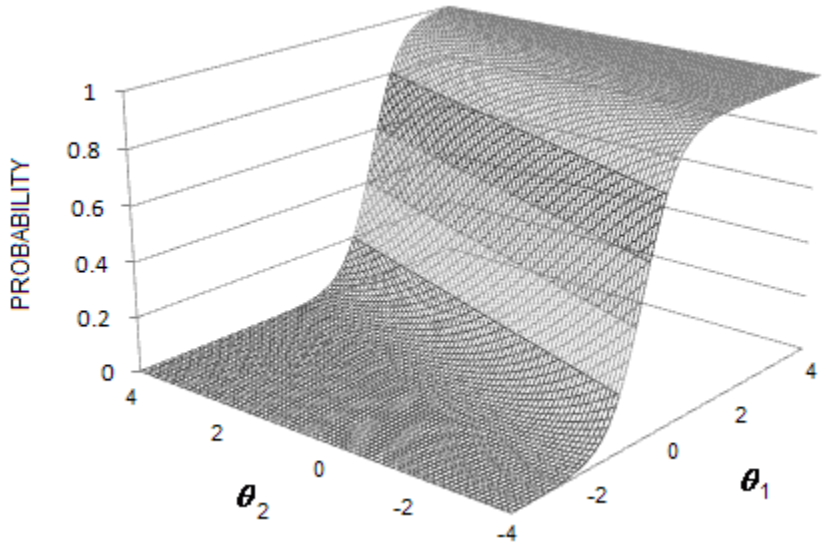


Figure 2-3. Example Item Characteristic Surface (ICS) plot

$a_1 = 1.5$, $a_2 = 0.5$, $c = 0$, and $d = -0.7$.

MIRT models typically contain one discrimination parameter (a) per dimension. Similar to UIRT, larger discrimination parameters are indicative of better discrimination power for the item. However, in contrast to UIRT—which only discriminates in one direction (the single trait being measured)—MIRT discrimination parameters provide an indication of how well the item discriminates on each dimension. This information provides an indication of which composite of skills are most precisely measured by the item, i.e., in which direction along the ICS the item best discriminates. The magnitude of the discrimination parameter on a particular dimension, relative to the magnitude of the discrimination parameters on other dimensions, indicates to what degree the item is measuring that particular trait. For example, if the first discrimination parameter is a positive value and all other discrimination parameters for the item are 0, then the item measures only the first dimension. In this situation, the item best discriminates in a

direction that falls parallel to the first dimension. If all discrimination parameters are equal, then the item equally measures all specified dimensions, and the item direction will be midway between all specified dimensions.

For comparative purposes, discrimination parameters were denoted by (a) for both UIRT and MIRT models. However, whereas UIRT models contain a difficulty parameter (b) directly related to the location of the ICC in reference to the ability scale, MIRT models are often parameterized using an index (d) that is indirectly related to the location of the ICS in reference to the ability axes. Consider the mathematical form of UIRT models which primarily dictates the location and the slope of the ICC, $a(\theta - b) = (a\theta - ab)$. The MIRT index ($d$) is the multidimensional equivalent to the unidimensional parameterization, ($-ab$). That is, $d = -ab$, and $b = \frac{-d}{a}$. As a result, the MIRT equivalent of the UIRT difficulty is computed as

$$b = \frac{-d}{\sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{a}}} \tag{2.1}$$

In this equation, d is the parameterized value for the model, and $\mathbf{a}^{\mathrm{T}}$ represents the vector of discrimination parameters.

The MIRT lower asymptote parameter (c) is directly comparable to the UIRT lower asymptote parameter. Both of these parameters account for random response (guessing) and dictate the lower bound for the ICC or ICS, respectively. In fact, the c-parameter resulting from a unidimensional calibration is often used as the c-parameter value for a MIRT parameterization of the same item (Reckase, 2009).

The direction in which the item best measures—described previously as being related to the item discrimination parameters—is of critical importance for understanding

the new equating methodology described in this research. This direction can be quantified by the vector of angles between the direction that the item best measures and each of the coordinate axes. For example, this vector may be displayed as $[\nu_{j1}, \nu_{j2}, \nu_{j3} ...]$, where ν_{jk} represents the angle between the direction that the item best measures and axis (dimension) k for item j. To determine these angles, the direction cosine corresponding to each dimension is first computed as

$$\cos(\nu_{jk}) = \frac{a_{jk}}{\sum_{k=1}^{d} a_{jk}^2} \tag{2.2}$$

The arccosine of $\cos(\nu_{jk})$ determines angle ν_{jk}. If the angle is 0 degrees between an item and a specific dimension, then the item measures only that dimension: in this case, the item does not measure a composite of the various traits being assessed by the test. As the angles between the item and the various dimensions become more similar, the item is interpreted as measuring a composite of the various traits.

An example in two-dimensional space appears in Figure 2-4. The angle between the direction of best measurement and the first coordinate axis is 80 degrees, implying that the angle between the direction of best measurement and the second coordinate axis is 10 degrees. This item provides an example where the second dimension is more precisely measured than the first dimension, yet the item does measure a composite of both traits.

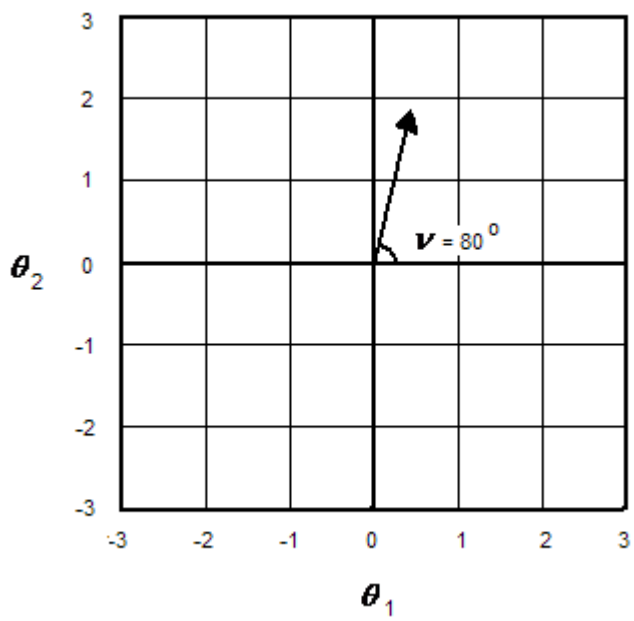


Figure 2-4. Angle between direction of best measurement and coordinate axes

Whereas these properties hold at the item level, it may be of interest to determine which direction (i.e., which composite of traits) is best measured at the test level. In order to do so, the multidimensional equivalent of the UIRT test characteristic curve and information function must first be reviewed.

A direct extension of the unidimensional test characteristic curve (TCC) is the multidimensional test characteristic surface (TCS). Similar to UIRT, the TCS is computed as the sum of the item characteristic surfaces across all items, and is conceptually viewed as the regression of the summed score responses on the vector of ability traits (Reckase, 2009) (see Figure 2-5 for an example TCC). However, unidimensional and multidimensional information functions are somewhat discrepant. Whereas in UIRT, information can be viewed as the slope of the TCC divided by the error variance at that particular ability level, in MIRT, the slope of the TCS is different

depending on the direction from the origin under consideration. As a result, information can be evaluated at each direction from the origin, and the information yielded in each direction will most likely be different.

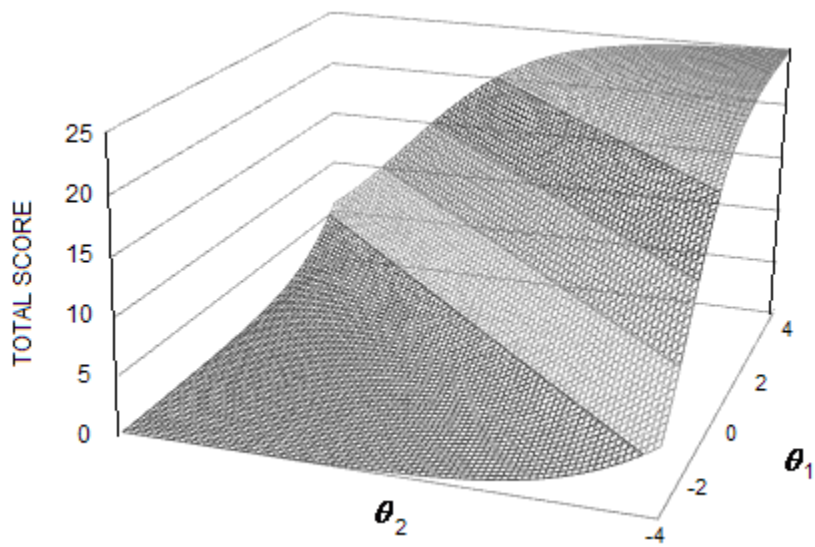


Figure 2-5. Example Test Characteristic Surface (TCS)

for a 25 Item Test

The properties of item discrimination and test information can be exploited to determine the unidimensional projection that would result from applying a UIRT model to the multidimensional data, and to determine which direction (i.e., which composite of traits) is best measured at the test level. Wang (1985, 1986) determined the relationship between the multidimensional ability space and the unidimensional projection that would result from applying a UIRT model to the multidimensional data. Defining the unidimensional scale as the "reference composite" for the test, she demonstrated that this unidimensional projection is equal to the first eigenvector of the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$, where the

matrix **A** contains discrimination parameters for each item on the exam. Each row in this matrix corresponds to an item, and each column in the matrix corresponds to a specific dimension. The reference composite contains one element per specified dimension, and each element is conceptually similar to the discrimination parameter on that dimension. To describe the direction of the reference composite, angles can be determined between each coordinate axis and the reference composite. A test that primarily measures the first dimension will yield a reference composite that falls nearly parallel to the first coordinate axis. A test that equally measures all dimensions will yield a reference composite that falls between each of the coordinate axes.

An example of a reference composite appears in Figure 2-6. Note that the test is primarily comprised of two clusters of items: a set of items that primarily measures the second dimension, and a set of items that primarily measures the first dimension (though both clusters measure a composite of the dimensions). As a result, the reference composite points in a direction midway between the directions that each of these clusters primarily point.

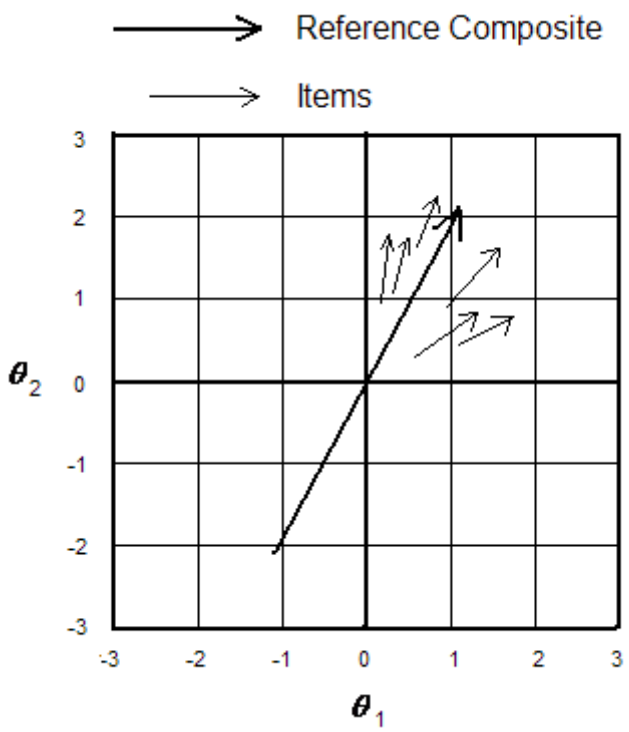


Figure 2-6. Example Reference Composite

Whereas the reference composite determines the unidimensional projection that would result from applying a UIRT model to the multidimensional data, Zhang and Stout (1999a) defined a similar concept, the "direction of best measurement," as the linear composite of abilities which the multidimensional test "best measures." Although these two concepts are similar in that a unidimensional projection is determined from the multidimensional space, the concepts are mathematically different. The reference composite yields a unidimensional scale that points in a direction solely determined by the item discrimination parameters. The direction of best measurement takes into account the multidimensional information yielded by the measurement procedure on each item and computes a unidimensional scale that represents that composite of traits best represented by a single observed score. Zhang and Stout (1999a) conceptually

determined the direction of best measurement by computing the average multidimensional information function across all directions. The result is a standardized vector of weights corresponding to each dimension. The mathematical expression for this direction will be presented later in this chapter.

Dimensionality Assessment

Several dimensionality assessment procedures have been created to help determine whether a unidimensional IRT model or a multidimensional IRT model should be used to analyze test data (Douglas, Kim, Habing, & Gao, 1996; Kim, 1994; Roussos & Stout, 1996; Stout, 1990; Zhang & Stout, 1999b). In order to adequately use UIRT to model examinee responses on an exam, the assumption of unidimensionality must not be seriously violated (Lord, 1980). A test that is unidimensional measures only one trait in the population of interest. If the test measures more than one trait but the responses are analyzed using a unidimensional IRT model, the resulting ability estimates and item parameter estimates may be highly inaccurate depending on the nature of the other variables being measured (Ansley & Forsyth, 1985; Reckase, 1985; Sireci, Thissen, and Wainer, 1991). Due to the potential seriousness of these violations, several dimensionality assessment procedures have been developed to estimate the number of dimensions measured by an exam, and to determine whether unidimensionality is a feasible assumption (Douglas, Kim, Habing, & Gao, 1996; Kim, 1994; Roussos & Stout, 1996; Stout, 1990; Zhang & Stout, 1999b).

One of these procedures—a nonparametric procedure originally developed by Zhang and Stout (1999b) and implemented in the computer program DETECT (Dimensionality Evaluation To Enumerate Contributing Traits)—has been widely used to

assess the feasibility of the unidimensionality assumption (Reckase 2009; Zhang & Stout, 1999b). The algorithm incorporated in this program essentially searches for homogeneous clusters of items that conform to either a pre-specified (confirmatory) or an unspecified (exploratory) simple structure solution. This procedure is accomplished by first determining the direction of best measurement at the test level, and then partitioning the test into homogeneous clusters such that the directions of best measurement at the cluster-level deviate as far as possible from the test-level direction of best measurement.

Included in the DETECT output are three statistics: the maximum DETECT value, the ratio r, and the IDN index value. The maximum DETECT value typically ranges between 0 and 1 and provides an indication of the degree of multidimensionality in the data (values less than 0.2 are viewed as "essentially unidimensional," values between 0.2 and 0.4 as "weak to moderate multidimensionality," values between 0.4 and 1.0 as "moderate to strong multidimensionality," and values above 1.0 as "large multidimensionality" (Kim, 1994; Zhang & Stout, 1999b). The ratio r ranges from 0 to 1 and incorporates a cross-validation technique (i.e., the procedure randomly divides the sample into two groups) to provide an indication of the stability of the solution. A value close to 1 is indicative of a stable solution, i.e., that a similar solution is likely to result if this procedure is conducted on a different (yet comparable) group of examinees. The IDN index value ranges from 0 to 1 and indicates how well the data conform to a simple structure model (as opposed to a "complex" structure model). Values close to 1 are indicative of good fit for a simple structure model.

Unidimensional IRT Scale Linking

Before IRT observed score or true score equating can be performed, item parameters and examinee abilities must first be estimated. Due to the scale

indeterminacy property of item response theory, any linear transformation of the ability scale will yield the same probabilistic relationships, assuming that both the ability scale and item parameters are transformed. "If an IRT model fits a set of data, then any linear transformation of the θ-scale also fits the set of data, provided that the item parameters also are transformed" (Kolen & Brennan, 2004, p. 161). To resolve this indeterminacy issue, IRT calibrations are typically specified to yield an ability distribution with mean of 0 and standard deviation of 1 (Kolen & Brennan, 2004).

When parallel forms are to be equated, parameters on the different forms can either be estimated at the same time (concurrent calibration) or in separate calibrations. If concurrent calibration is used, no scale linking methods need to be employed to ensure that parameter estimates are on the same scale. By nature of the single calibration, parameter estimates are already on the same scale. If separate calibrations are conducted under the random groups design, groups that were administered different forms are assumed to be equivalent on the measured trait. That is, groups are assumed to have equal means and standard deviations on the ability scale. If both calibrations are specified to yield an ability distribution with mean of 0 and standard deviation of 1, no scale linking method is necessary since means and standard deviations are specified to be equal across groups (Kolen & Brennan, 2004).

However, when item parameters and abilities are calibrated separately for groups of examinees who are different with respect to the measured trait (i.e., nonequivalent groups) the estimates will be placed on different—yet linearly related—ability scales. Therefore, a scale linking procedure must be conducted to place item parameter estimates and ability estimates from the Form A scale to that of the Form B scale before equating

can be conducted. "When the IRT model holds, the parameter estimates from different computer runs are on linearly related θ-scales. Thus, a linear equation can be used to convert IRT parameter estimates to the same scale" (Kolen & Brennan, 2004, p. 161).

When separate calibrations are performed under the nonequivalent groups design, the scale of Form A and the scale of Form B may differ in two aspects: the scales may differ in (1) origin (i.e., mean) and (2) unit of measurement (i.e., standard deviation). Since the scales are linearly related, two coefficients (K and L) must be determined to transform linking scale (I) parameters to the base scale (J) as follows:

$$\theta_{Ji} = K\theta_{Ii} + L \tag{2.3a}$$

$$a_{Jj} = \frac{a_{Ij}}{K} \tag{2.3b}$$

$$b_{Jj} = Kb_{Ij} + L \tag{2.3c}$$

$$c_{Jj} = c_{Ij} \tag{2.3d}$$

In this series of transformations, a is item discrimination, b is item difficulty, and c is the lower asymptote parameter for either scale I or J. The parameters K and L are two linking coefficients that account for differences in scale origin and unit of measurement resulting from separate calibrations. Several methods exist for estimating these parameters—namely, the mean/mean, mean/sigma, Haebara, and Stocking-Lord procedures (Kolen & Brennan, 2004).

Moment Transformation Methods

The mean/mean and mean/sigma transformation methods use moments of the discrimination and difficulty parameters over common items to estimate the two linking coefficients (K and L). Note that in the linear transformation of the ability scale (i.e.,

$\theta_{Ji} = K\theta_{Ii} + L$), K represents the slope coefficient and L represents the intercept coefficient. As a result, for any two individuals i and $i*$, or for any two items j and $j*$ (Kolen & Brennan, 2004, p. 163),

$$K = \frac{\theta_{Ji} - \theta_{Ji*}}{\theta_{Ii} - \theta_{Ii*}} = \frac{b_{Jj} - b_{Jj*}}{b_{Ij} - b_{Ij*}} = \frac{a_{Ij}}{a_{Jj}} \tag{2.4a}$$

and

$$L = b_{Jj} - Kb_{Ij} = \theta_{Ji} - K\theta_{Ii} \tag{2.4b}$$

As these equations express K and L in terms of two individuals or two items, the linking coefficients can also be expressed in terms of *populations* of items or *populations* of examinees as:

$$K = \frac{\sigma(b_J)}{\sigma(b_I)} = \frac{\mu(a_I)}{\mu(a_J)} = \frac{\sigma(\theta_J)}{\sigma(\theta_I)} \tag{2.5a}$$

$$L = \mu(b_J) - K\mu(b_I) = \mu(\theta_J) - K\mu(\theta_I) \tag{2.5b}$$

In this series of equations, μ represents the population mean, σ represents the population standard deviation, and θ, a, and b represent examinee ability, item discrimination, and item difficulty on common items for scale I or scale J, respectively.

When population values are used in Equations 2.5a and 2.5b, the relationships as defined in these equations hold perfectly. However, when statistics are substituted as estimates of population parameters, these relationships will not be equal. For example, when statistics are substituted for population parameters in Equation 2.5a, it is most likely that $\frac{\sigma(b_J)}{\sigma(b_I)} \neq \frac{\mu(a_I)}{\mu(a_J)}$. The mean/mean and mean/sigma differ with respect to which parameterization is used to estimate the K parameter (Kolen & Brennan, 2004).

Mean/Mean Method

The mean/mean method incorporates the means of the discrimination parameters and the means of the difficulty parameters over common items to estimate the linking coefficients as follows (Loyd & Hoover, 1980):

$$K = \frac{\mu(a_I)}{\mu(a_J)} \tag{2.6a}$$

$$L = \mu(b_J) - K\mu(b_I) \tag{2.6b}$$

An advantage of the mean/mean method (over the mean/sigma method) is that the sample mean tends to yield more stable estimates of the respective population value than the sample standard deviation (Baker & Al-Karni, 1991). However, item difficulty parameters tend to yield more stable estimates than item discrimination parameters (Kolen & Brennan, 2004).

Mean/Sigma Method

The mean/sigma method incorporates the means and the standard deviations of the difficulty parameters over common items to estimate the linking coefficients as follows (Marco, 1977):

$$K = \frac{\sigma(b_J)}{\sigma(b_I)} \tag{2.7a}$$

$$L = \mu(b_J) - K\mu(b_I) \tag{2.7b}$$

Again, item difficulty parameters tend to yield more stable estimates than item discrimination parameters (the mean/sigma method uses only item difficulty parameters), but sample means tend to yield more stable estimates than sample standard deviations (Baker & Al-Karni, 1991).

Characteristic Curve Transformation Methods

A weakness of the mean/mean and mean/sigma transformation methods is that these methods do not take into account all item parameter estimates simultaneously to estimate linking coefficients (Kolen & Brennan, 2004). In response to this inadequacy, Haebara (1980) and Stocking and Lord (1983) derived scale linking methods which focus on the item characteristic curve (ICC) or test characteristic curve (TCC), respectively, in order to estimate linking coefficients. The characteristic curve methods exploit the fact that, due to the scale indeterminacy property of IRT, for any set of item parameters,

$$p_{ij}(\theta_{Ji}; a_{Jj}, b_{Jj}, c_{Jj}) = p_{ij}(K\theta_{Ii} + L; \frac{a_{Ij}}{K}, Kb_{Ij} + L, c_{Ij}). \tag{2.8}$$

This equation is strictly true for item parameters. However, when item parameter estimates are substituted for population values, this equation most likely will not hold. The characteristic curve methods compute linking coefficients (K and L) which minimize differences in these probabilities over common items. The characteristic curve methods differ in which function is specified to derive differences in these probabilities.

Haebara Method

The Haebara (1980) method seeks to estimate linking coefficients (K and L) which minimize the squared differences in item characteristic curves across all common items. That is, the Haebara method seeks to minimize the function:

$$Hdiff(\theta_i) = \sum_{j:V} \left[p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - p_{ij}(K\theta_{Ii} + L; \frac{\hat{a}_{Ij}}{K}, K\hat{b}_{Ij} + L, \hat{c}_{Ij}) \right]^2 \tag{2.9}$$

The summation is taken over all common items ($j:V$). After this function is summed over common items, these differences are then summed across examinee abilities to derive the minimization function as:

$$Hcrit = \sum_i Hdiff(\theta_i) \tag{2.10}$$

Stocking and Lord Approach

The Stocking and Lord (1983) method seeks to estimate linking coefficients which minimize the squared differences in test characteristic curves across all common items. That is, the Stocking and Lord method seeks to minimize the function:

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - \sum_{j:V} p_{ij}(K\theta_{Ii} + L; \frac{\hat{a}_{Ij}}{K}, K\hat{b}_{Ij} + L, \hat{c}_{Ij}) \right]^2 \tag{2.11}$$

The summation is taken over all common items ($j:V$). After this function is summed over common items, these differences are then summed across examinee abilities to derive the minimization function as:

$$SLcrit = \sum_i SLdiff(\theta_i) \tag{2.12}$$

For the Haebara (1980) method and the Stocking and Lord (1983) method, the quadrature points and weights for $Hcrit$ and $SLcrit$ can be selected via a variety of methods (Kolen & Brennan, 2004).

MIRT Scale Linking

Several procedures have been developed to link scales within the multidimensional IRT (MIRT) framework (Davey et al., 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000; Thompson et al., 1997; Yon, 2006). Whereas unidimensional scale linking procedures estimate two coefficients (K and L) to adjust for differences in (1) origin and (2) unit of measurement from separate calibrations, MIRT scale linking procedures typically estimate scalar coefficients and matrices to adjust for differences in (1) rotation, (2) correlation, (3) translation (similar to

"origin" in unidimensional IRT), and (4) dilation (similar to "unit of measurement" in unidimensional IRT). That is, separate calibrations within the MIRT framework may place ability estimates and item parameter estimates on separate scales due to rotational indeterminacy (similar to factor analysis), correlation indeterminacy (though solutions are often derived by fixing multidimensional traits to be multivariate normally distributed and uncorrelated, i.e., distributed $MVN(\mathbf{0},\mathbf{I})$), and indeterminacy in origin and unit of measurement (see Figure 2-7 for a comparison between UIRT and MIRT linking procedures).

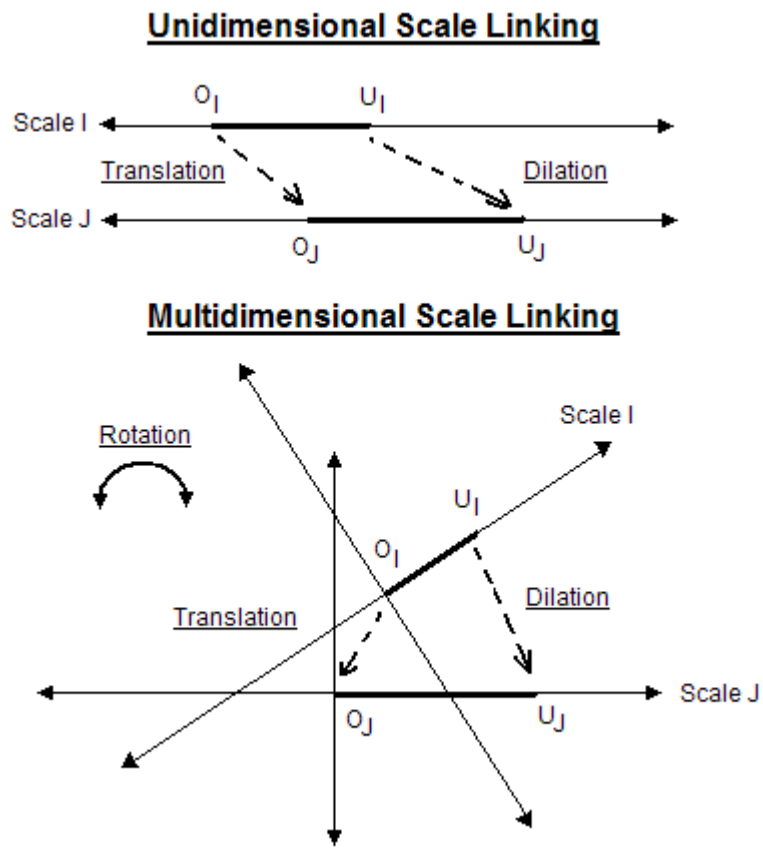


Figure 2-7. A comparison of UIRT and MIRT Linking Methods

O represents origin and U represents unit of measurement for

Scales I and J, respectively.

(Modified from Min, 2003)

Depending on the nature of the linking procedure and the assumptions associated with each procedure, several different mathematical expressions exist for transforming item parameter estimates and ability estimates from the linking scale (I) to the base scale (J). However, a generic expression is as follows:

$$\boldsymbol{\theta}_{Ji} = \mathbf{T}^{-1}\boldsymbol{\theta}_{Ii} + \boldsymbol{\beta} \tag{2.13a}$$

$$\mathbf{a}_{Jj}^{\mathrm{T}} = \mathbf{a}_{Ij}^{\mathrm{T}}\mathbf{T} \tag{2.13b}$$

$$d_{Jj} = d_{Ij} - \mathbf{a}_{Ij}^{\mathrm{T}}\mathbf{T}\boldsymbol{\beta} \tag{2.13c}$$

$$c_{Jj} = c_{Ij} \tag{2.13d}$$

(Davey et al., 1996; Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000; Thompson et al., 1997). In this series of equations, $\mathbf{T}$ represents an oblique or orthogonal matrix to account for rotational indeterminacy and dilation indeterminacy, and $\boldsymbol{\beta}$ represents a matrix to account for translation indeterminacy (Davey et al., 1996; Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000). Depending on the nature of the linking design, $\boldsymbol{\beta}$ might equal $\mathbf{0}$ (Thompson et al., 1997), $\mathbf{T}$ might be the product of a rotation matrix and a single dilation parameter (Li & Lissitz, 2000), $\mathbf{T}$ might be the product of a rotation matrix and a diagonal dilation matrix which accounts for separate dilation parameters on each dimension (Min, 2003), or $\mathbf{T}$ might simply be a rotation matrix. Note that the equations used to perform scale linking in the MIRT framework are direct extensions of their unidimensional counterparts, with matrices replacing scalar coefficients.

MIRT Scale Linking for Nonequivalent Groups

Most MIRT scale linking procedures that have been developed are used in conjunction with the nonequivalent groups design (Simon, 2008). In the MIRT framework, the nonequivalent groups design can either incorporate items which are common to both forms (common items), or examinees which are administered both forms (common examinees). The different scale linking procedures for the nonequivalent groups design typically differ in: (1) whether an orthogonal matrix or non-orthogonal (oblique) matrix is used to rotate parameter estimates to account for rotational indeterminacy; (2) whether dilation parameters are incorporated into the linking procedure to account for dilation (unit of measurement) indeterminacy; (3) for designs

which incorporate dilation parameters, whether one dilation parameter is applied to all dimensions or whether a separate dilation parameter is estimated to account for each dimension separately; and (4) whether transformation matrices and coefficients are estimated simultaneously or separately (Davey et al., 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000; Yon, 2006).

Hirsch (1989) developed a method to be used when there are common examinees, i.e., a set of examinees complete both Form B and Form A. This procedure contains an orthogonal rotation matrix to account for rotational indeterminacy. Furthermore, means and standard deviations are computed for common examinees on each dimension, which in turn are used to estimate translation and dilation coefficients. These coefficients are applied to make means and standard deviations equal for the common examinees across forms on each dimension.

Oshima et al. (2000) derived four separate MIRT linking methods, three of which are direct extensions of unidimensional linking methods (the equated function method, derived as an extension of the mean/mean method; the test characteristic function method, derived as an extension of the Stocking-Lord method; and the item characteristic function method, derived as an extension of the Haebara method). For each of the four methods, a rotation matrix and a translation vector are determined simultaneously. The rotation matrix, as determined by Oshima et al. (2000), rotates parameter estimates as well as dilates parameters to account for rotational indeterminacy and dilation indeterminacy. When the correlation between traits is the same across forms, the rotation matrix is orthogonal; otherwise it is oblique.

Li and Lissitz (2000) derived a method to link MIRT scales which consists of an orthogonal rotation matrix, a translation vector, and one dilation parameter. The authors assert that only one dilation parameter is required (which is applied to all dimensions), since the process of estimating only one dilation parameter yields a more "tractable" solution, and since dispersion across specified dimensions is likely to be similar (Li & Lissitz, 2000, p. 116). Min (2003) extended the Li and Lissitz (2000) approach to incorporate a diagonal dilation matrix—containing one dilation parameter per dimension—rather than using one dilation parameter across all dimensions.

MIRT Scale Linking for Randomly Equivalent Groups

In UIRT, separate calibrations under the random groups design will yield item parameter estimates and ability estimates which are on the same scale, assuming that ability distributions are specified to be the same in both calibrations (typically specified to follow a standard normal distribution). Similar to UIRT, separate calibrations under the random groups design within the MIRT framework will yield item parameter estimates and ability estimates which are on the same scale in terms of translation (origin) and dilation (unit of measurement). However, separate calibrations under this design are still subject to rotational indeterminacy (Thompson et al., 1997). To account for rotational indeterminacy, Thompson et al. (1997) developed a procedure to estimate an orthogonal rotation matrix ($\mathbf{T}$) under the random groups design. The orthogonal Procrustes rotation is then applied to ability estimates and item discrimination estimates to place the estimates on the same scale.

The conceptual rationale underlying the Thompson et al. (1997) procedure is as follows. If the forms to be equated are in fact, parallel, then the forms should measure

the same composite construct, and the tables of specifications for the forms should be nearly identical. Based on the tables of specifications, and perhaps a further content analysis of what each item is intended to measure, items on each form can be grouped into homogeneous clusters. Items which were identified in the same cluster should logically be measuring the same composite of traits. Statistically, this means that the items within each cluster should discriminate in the same direction in multidimensional space, since they are intended to measure the same composite of traits. As a result, a reference composite can be computed for each cluster of items to determine the unidimensional projection that would result from applying a UIRT model to the cluster, since each cluster of items is expected to be internally homogeneous, i.e., the items discriminate in the same direction. The reference composite provides a stable indication of the direction of each cluster in multidimensional space.

The objective then is to determine the orthogonal rotation matrix such that, after the rotation is applied, corresponding reference composites on parallel forms will be pointing in the same direction in multidimensional space. Once the orthogonal rotation matrix has been determined for the reference composites (i.e., at the "cluster-level"), this matrix can be applied at the individual item level to ensure that item parameter estimates are on the same scale pending the rotational indeterminacy. The reason that the orthogonal rotation matrix could not originally be determined at the item level is because there are no common items in this data collection design.

Mathematically, the Thompson et al. (1997) procedure is conducted as follows. First, items are grouped into clusters based on content analysis of what each item is intended to measure. Separate matrices ($\mathbf{D}_C$) are created for each cluster which contain

item discrimination parameters for each item within the cluster. Each row of the matrix corresponds to a separate item, and each column of the matrix corresponds to a distinct dimension. A reference composite, defined as "the eigenvector associated with the largest eigenvalue of $\mathbf{D}_\mathbf{C}^\mathbf{T}\mathbf{D}_\mathbf{C}$, where $\mathbf{D}_\mathbf{C}$ is the matrix of item discriminations" (Thompson et al., 1997, p. 4) is then computed for each cluster of items to determine the unidimensional projection for this cluster. The total number of reference composites is equal to the total number of clusters identified on the test, and the number of elements in each reference composite is equal to the number of specified dimensions on the test.

The reference composites for each cluster on Form A and Form B are then concatenated to form matrices $\mathbf{M}$ and $\mathbf{N}$, corresponding to Form A and Form B, respectively. Each row in matrices $\mathbf{M}$ and $\mathbf{N}$ corresponds to a separate reference composite, and each column in matrices $\mathbf{M}$ and $\mathbf{N}$ corresponds to a distinct dimension. Corresponding reference composites on Form A and on Form B must be placed in the same row. The orthogonal rotation matrix $\mathbf{T}$ is the solution which minimizes $tr(\mathbf{E}^\mathbf{T}\mathbf{E})$ where $\mathbf{E} = \mathbf{N} - \mathbf{MT}$ and $tr(\cdot)$ represents the trace function. The solution to this matrix is obtained via the singular value decomposition, and is presented in Schonemann (1966). Specifically, the solution is obtained as:

$$\mathbf{S} = \mathbf{M}^\mathbf{T}\mathbf{N} \tag{2.14a}$$

$$\mathbf{S} = \mathbf{UQV}^\mathbf{T} \text{ (singular value decomposition)} \tag{2.14b}$$

$$\mathbf{T} = \mathbf{UV}^\mathbf{T} \tag{2.14c}$$

In this series of equations, $\mathbf{Q}$ is a diagonal matrix containing the square root of the eigenvalues of $\mathbf{S}$, and $\mathbf{U}$ and $\mathbf{V}$ are matrices containing the eigenvectors of $\mathbf{S}$ (Thompson et al., 1997). Once the orthogonal rotation matrix ($\mathbf{T}$) is computed, the

orthogonal Procrustes rotation is applied to the Form A discrimination parameters and ability estimates in order for the Form A parameter estimates to be placed on the same scale as the Form B parameter estimates as:

$$\boldsymbol{\theta}_{Ji} = \mathbf{T}^{-1}\boldsymbol{\theta}_{Ii} \tag{2.15a}$$

$$\mathbf{a}_{Jj}^{\mathbf{T}} = \mathbf{a}_{Ij}^{\mathbf{T}}\mathbf{T} \tag{2.15b}$$

$$d_{Jj} = d_{Ij} \tag{2.15c}$$

$$c_{Jj} = c_{Ij} \tag{2.15d}$$

If the covariance matrix ($\boldsymbol{\Sigma}$) for examinee abilities was computed prior to the rotation, this matrix may change as a result of the rotation. Let $\boldsymbol{\Sigma}_O$ represent the covariance matrix prior to the rotation, and $\boldsymbol{\Sigma}_N$ represent the covariance matrix after the rotation has been applied. From statistical theory of linear composites, $\boldsymbol{\Sigma}_N = Var(\mathbf{T}^{-1}\boldsymbol{\theta}) = \left(\mathbf{T}^{-1}\right)^{\mathbf{T}}\boldsymbol{\Sigma}_O\mathbf{T}^{-1}$ (Johnson & Wichern, 2007). However, often the correlational indeterminacy is solved by fixing multidimensional traits to follow a multivariate standard normal distribution with zero correlation between dimensions (i.e., $\boldsymbol{\theta} \sim MVN(\mathbf{0},\mathbf{I})$).

> The metric of MIRT item parameter estimates usually refers to reference axes that are orthogonal and of unit length, because most MIRT parameter estimation programs solve the identification problem (or result in a unique solution) by requiring that the multidimensional traits ($\boldsymbol{\theta}$) be distributed as a multivariate normal, $MVN(\mathbf{0},\mathbf{I})$. Although real traits are likely to be correlated, items in the bank can be tentatively defined with reference to orthogonal axes in order to make it easier for future MIRT equating or for additions to a MIRT item bank. The precalibrating parameters could be re-rotated obliquely, if necessary, for better interpretation (Li & Lissitz, 2000, p. 116).

In this situation (i.e., under an orthogonal solution), the covariance matrix prior to rotation will be the same as the covariance matrix after the rotation. Since $\mathbf{\Sigma}_O = \mathbf{I}$ (where $\mathbf{I}$ represents the identity matrix), $\mathbf{\Sigma}_N$ can be solved for as:

$$\mathbf{\Sigma}_N = \left(\mathbf{T}^{-1}\right)^{\mathbf{T}} \mathbf{\Sigma}_O \mathbf{T}^{-1} \tag{2.16a}$$

$$= \left(\mathbf{T}^{-1}\right)^{\mathbf{T}} \mathbf{I}\mathbf{T}^{-1} \tag{2.16b}$$

$$= \left(\mathbf{T}^{-1}\right)^{\mathbf{T}} \mathbf{T}^{-1} \tag{2.16c}$$

$$= \left(\mathbf{T}^{\mathbf{T}}\right)^{-1} \mathbf{T}^{-1} \tag{2.16d}$$

$$= \left(\mathbf{T}\mathbf{T}^{\mathbf{T}}\right)^{-1} \tag{2.16e}$$

$$= \mathbf{I}^{-1} \tag{2.16f}$$

$$= \mathbf{I} \tag{2.16g}$$

$$= \mathbf{\Sigma}_O \tag{2.16h}$$

Thus, the covariance matrix will only change under a non-orthogonal (oblique) solution.

Inherent in this procedure is the assumption that each form to be equated measures the same *composite* of traits. That is, it does not suffice to simply measure the same traits on each form, but the degree to which each trait contributes to the single reported score must also be the same across forms: "The main assumption made in these studies is that each test form measures exactly the same unidimensional reference composite. That is, not only must each test form measure the same constructs, but the composite construct formed must also be the same across test forms" (Thompson et al., 1997, p. 2). The assumptions associated with the scale linking procedure will be further addressed in the "Scale Linking and Equating Assumptions" section in Chapter 3.

Thompson et al. (1997) described several procedures for assessing the quality of the rotation. One method is to observe direction cosines (Miller & Hirsch, 1992) between each pair of reference composites to be aligned. Direction cosines—defined as the cosine of the angle between the vectors to be aligned—can be computed before the rotation is applied and after the rotation is applied. After the rotation is applied, the angle between corresponding reference composites should be smaller than the pre-rotation angle. Although no global criterion for assessing quality of the rotation based on direction cosines is presented, Thompson et al. (1997) noted that better fit is indicated by small post-rotation angles.

After the scale linking procedure has been conducted (i.e., item and ability parameters that have been estimated in separate calibrations have been placed on the same scale) and the linking procedure has been appropriately assessed, an equating procedure can be conducted to relate number-correct scores on the forms to be equated.

Unidimensional IRT Equating

If examinee scores are to be reported on the θ-scale or on a linear transformation of the θ-scale, then only an appropriate scale linking method is required. However, oftentimes it is desired to use a scale other than the θ-scale for score reporting. In this situation, a procedure must be conducted to equate observed scores or true scores across parallel forms of an exam. Two methods which have been derived for this purpose are IRT observed score equating and IRT true score equating. IRT observed score equating estimates observed score distributions on both forms based on the IRT model. Once the observed score distributions are estimated, equipercentile equating is conducted to determine an equating relationship between the forms (Kolen & Brennan, 2004). IRT

true score equating relates true scores on both forms using the IRT definition of true score—i.e. $\tau(\theta_i) = \sum_j p_{ij}(\theta_i; a_j, b_j, c_j)$. First, a true score on Form A is selected; the ability level (θ_i) associated with this true score is then estimated; finally, the true score on Form B associated with this ability level is computed (Kolen & Brennan, 2004).

IRT Observed Score Equating

To conduct unidimensional IRT observed score equating, conditional observed score distributions (i.e., $f(x \mid \theta_i)$) are first determined at each ability level (θ_i) using a recursion formula such as the Lord-Wingersky method (Lord & Wingersky, 1984). This algorithm computes conditional distributions across all specified ability levels as:

$$f_r(x \mid \theta_i) = f_{r-1}(x \mid \theta_i)(1 - p_{ir}) \qquad x = 0 \qquad (2.17a)$$

$$= f_{r-1}(x \mid \theta_i)(1 - p_{ir}) + f_{r-1}(x-1 \mid \theta_i) p_{ir} \qquad 0 < x < r \qquad (2.17b)$$

$$= f_{r-1}(x-1 \mid \theta_i) p_{ir} \qquad x = r \qquad (2.17c)$$

In this series of equations, $f_r(x \mid \theta_i)$ is the conditional observed score distribution over the first r items, $f_{r-1}(x \mid \theta_i)$ is the conditional observed score distribution over the first $r-1$ items, and p_{ir} is the probability of correctly answering item r for an examinee of ability level θ_i. In practice, this algorithm begins as:

$$f_r(x \mid \theta_i) = (1 - p_{ir}) \qquad x = 0 \qquad (2.18a)$$

$$= p_{ir} \qquad x = 1 \qquad (2.18b)$$

After this initial step, $r-1$ iterations are required to compute the conditional observed score distribution ($f_r(x \mid \theta_i)$) for r items.

Once the conditional observed score distribution is determined at each specified ability level, the conditional distribution is then multiplied by the ability density ($\psi(\theta)$) and either summed or integrated over all ability levels to determine an observed marginal distribution for each form. That is,

$$f(x) = \sum_{\theta} f(x \mid \theta)\psi(\theta) \tag{2.19a}$$

or

$$f(x) = \int_{\theta} f(x \mid \theta)\psi(\theta)d\theta \tag{2.19b}$$

depending on whether a discrete ability distribution or a continuous ability distribution is defined. Once marginal distributions ($f(x)$ and $f(y)$) are determined for each form, Form A and Form B are equated using traditional equipercentile methods. The equipercentile equating method relates observed scores on both forms with the same percentile rank (further discussion of this method will be presented in the next chapter).

IRT True Score Equating

IRT true score equating relates true scores on Form A with true scores on Form B using the IRT definition of true score. Although no theoretical justification exists for applying the true score relationships to observed scores, often this is conducted in practice (Kolen & Brennan, 2004). The unidimensional procedure is conducted in three stages: specifying a true score on Form A for which the corresponding true score on Form B is desired; determining the ability level (θ_i) which corresponds to the given true score on Form A; and determining the true score on Form B which corresponds to the ability level (θ_i). A detailed explanation appears below.

First, a true score on Form A (τ_A) is selected. Usually this value is between the sum of the lower asymptote parameters and the total number of items on Form A, i.e., $\sum_{j:A} c_j < \tau_A < N_A$, where N_A represents the total number of items on Form A. Recall that true score at a particular ability level (θ_i) is defined as the sum of the probabilities of obtaining a correct response for each item, i.e., $\tau_A(\theta_i) = \sum_{j:A} p_{ij}(\theta_i; a_j, b_j, c_j)$. In order to determine the θ_i associated with the particular true score on Form A, an iterative procedure—such as the Newton-Raphson method—is typically used. This root-finding algorithm essentially minimizes the difference $func(\theta_i) = \tau_A - \sum_{j:A} p_{ij}(\theta_i; a_j, b_j, c_j)$ by taking the derivative with respect to θ_i, setting this value equal to 0, and solving for the minimum. Once the particular θ_i is computed, this value is then substituted into the definition of true score on Form B, i.e., $\tau_B(\theta_i) = \sum_{j:B} p_{ij}(\theta_i; a_j, b_j, c_j)$. Typically in practice, corresponding true scores on Form B are derived for each Form A integer score.

Graphically, this procedure can be viewed as relating true scores through the test characteristic curve (TCC) for Form A and Form B (see Figure 2-8). Recall that the test characteristic curve can be viewed as the regression of observed score (X) on ability level (θ) (Lord, 1980). The expected value of the regression of observed score (X) on ability level (θ) is also known as the IRT true score (τ).

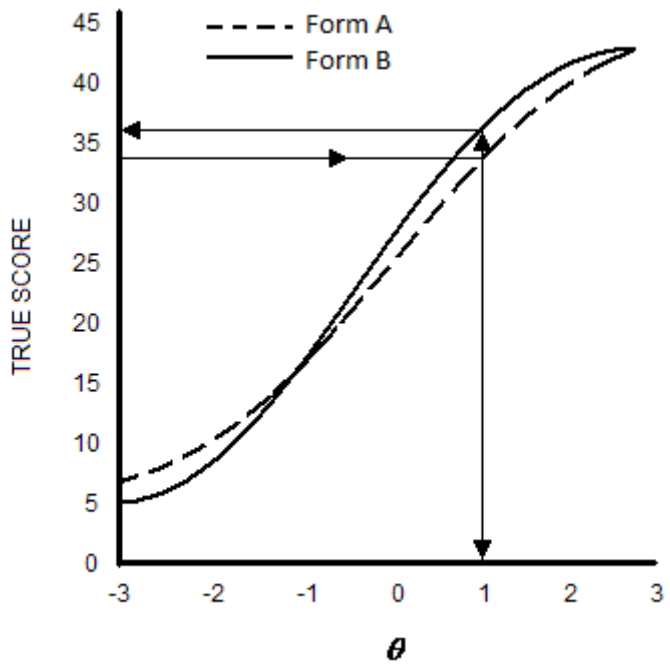


Figure 2-8. Graphical Representation of True Score Equating

Inherent in this procedure is the fact that the test characteristic curve (TCC) is a monotonically increasing function. If the TCC were not monotonically increasing, then a unique solution may not exist for the true score equating procedure.

Foundations for MIRT Equating

No procedures currently exist for conducting observed score or true score equating in the MIRT framework. As the purpose of the present research is to create such procedures, previously-existing literature which contributes toward the formation of this new methodology appears below.

Unidimensional Approximation

Zhang (1996), Zhang and Stout (1999a), and Zhang and Wang (1998) demonstrated that any set of item responses that can be adequately modeled by a multidimensional IRT model can be closely approximated by a unidimensional IRT

model with estimated unidimensional ability parameters and estimated unidimensional item parameters. Specifically, these authors demonstrated this to be true for any generalized δ-dimensional compensatory model. A generalized δ-dimensional compensatory model is defined as having form

$$p_{ij}(\boldsymbol{\theta}_i) = H_j(\mathbf{a}_j^{\mathbf{T}}\boldsymbol{\theta}_i + d_j) \equiv H_j\left(\sum_{k=1}^{\delta} a_{kj}\theta_k + d_j\right) \tag{2.20}$$

where "$\mathbf{a}_j^{\mathbf{T}} = (a_{j1}, a_{j2}, ..., a_{j\delta})$, $a_{j1}, a_{j2}, ..., a_{j\delta}$ are nonnegative and not all zero, and $H_j(x)$ is any non-decreasing function with $H_j^{'}(x) \geq 0$ for all x and $H_j^{'}(x)H_{j*}^{'}(x*)$ not being zero identically as $(x, x*)$ varies for $j, j* = 1,2,...,n$ (mathematically rigorously, not being zero almost everywhere with respect to Lebesque measure suffices)" (Zhang & Stout, 1999a, p. 133). For this family of models, $\mathbf{a}_j^{\mathbf{T}}$ is the discrimination parameter vector, d_j is an index related to the difficulty parameter, and $H_j(\cdot)$ is a link function (Zhang, 1996, Zhang & Stout, 1999a, and Zhang & Wang, 1998).

Two commonly used generalized δ-dimensional compensatory models include the multidimensional compensatory three parameter logistic model (M3PL) and the multidimensional compensatory normal ogive model. The M3PL can be written as:

$$p_{ij}(\boldsymbol{\theta}_i) = c_j + (1 - c_j)\left(\frac{\exp\left[1.7(\mathbf{a}_j^{\mathbf{T}}\boldsymbol{\theta}_i + d_j)\right]}{1 + \exp\left[1.7(\mathbf{a}_j^{\mathbf{T}}\boldsymbol{\theta}_i + d_j)\right]}\right) \tag{2.21}$$

with corresponding link function

$$H_j(\theta_i) = c_j + (1 - c_j)\left(\frac{\exp[1.7\theta]}{1 + \exp[1.7\theta]}\right) \tag{2.22}$$

The multidimensional compensatory normal ogive model can be written as:

$$p_{ij}(\boldsymbol{\theta}_i) = c_j + (1 - c_j)\Phi(\mathbf{a}_j^{\mathbf{T}}\boldsymbol{\theta}_i + d_j) \tag{2.23}$$

with corresponding link function

$$H_j(\theta_i) = c_j + (1 - c_j)\Phi(\cdot) \tag{2.24}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

Again, any generalized δ-dimensional compensatory model which adequately models the probability of a correct response for examinees of ability vector $\boldsymbol{\theta}_i$ can be closely approximated by a unidimensional model with estimated unidimensional ability parameters and unidimensional item parameters. Zhang (1996), Zhang and Stout (1999a), and Zhang and Wang (1998) define the estimated unidimensional ability (θ_α, where α denotes a unidimensional approximation) as a standardized linear composite of the latent variables ($\theta_1, \theta_2, ..., \theta_\delta$). The authors estimate this composite as:

$$\hat{\theta}_\alpha = \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\hat{\boldsymbol{\theta}} = \sum_{k=1}^{\delta} \hat{\alpha}_k \hat{\theta}_k \tag{2.25}$$

where $\sum_{k=1}^{\delta} \hat{\alpha}_k^2 = 1$. The coefficients for the linear composite are estimated as

$$\hat{\alpha}_k = \omega \sum_{j=1}^{N} w_j E\left\{ H_j^{'}\left(\hat{\mathbf{a}}_j^{\mathrm{T}}\hat{\boldsymbol{\Theta}}\right)\left[\sum_{j=1}^{N} w_j^2 H_j\left(\hat{\mathbf{a}}_j^{\mathrm{T}}\hat{\boldsymbol{\Theta}}\right)\left[1 - H_j\left(\hat{\mathbf{a}}_j^{\mathrm{T}}\hat{\boldsymbol{\Theta}}\right)\right]\right]^{-\frac{1}{2}}\right\}\hat{a}_{jk} \tag{2.26}$$

$$= \omega \sum_{j=1}^{N} w_j E\left[\frac{H_j^{'}\left(\hat{\mathbf{a}}_j^{\mathrm{T}}\hat{\boldsymbol{\Theta}}\right)}{\sqrt{Var(Y \mid \hat{\boldsymbol{\Theta}})}}\right]\hat{a}_{jk}, \;\; k = 1,2,...,\delta \tag{2.27}$$

where $\hat{\boldsymbol{\Theta}}$ is the complete latent trait vector, w_j is the score weight, ω is a positive constant such that $\sum_{k=1}^{\delta} \hat{\alpha}_k^2 = 1$, and E is the expectation operator. Under the assumption that all terms

$$w_j E\left[\frac{H_j^{'}\left(\hat{\mathbf{a}}_j^{\mathbf{T}}\hat{\mathbf{\Theta}}\right)}{\sqrt{Var(Y \mid \hat{\mathbf{\Theta}})}}\right], \quad j = 1,2,...,N \tag{2.28}$$

are equal, this becomes

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{a}_{jk}}{\sqrt{\sum_{k=1}^{\delta}\left(\sum_{j=1}^{N} \hat{a}_{jk}\right)^2}} \tag{2.29}$$

where N is the total number of items on the test. The terms $w_j E\left[\frac{H_j^{'}\left(\hat{\mathbf{a}}_j^{\mathbf{T}}\hat{\mathbf{\Theta}}\right)}{\sqrt{Var(Y \mid \hat{\mathbf{\Theta}})}}\right]$ "may be considered to be 'compound weights' for items contributing to the test direction; each term is completely determined by the score weight w_j and the item model (theoretical) weight $E\left[\frac{H_j^{'}\left(\hat{\mathbf{a}}_j^{\mathbf{T}}\hat{\mathbf{\Theta}}\right)}{\sqrt{Var(Y \mid \hat{\mathbf{\Theta}})}}\right]$ which is mainly determined by the derivative of the link function and the item discrimination parameter vector" (Zhang, 1996, p. 25).

Furthermore, if the exam demonstrates approximate simple structure (i.e., under the oblique solution, each item on the exam loads on only one dimension), then the oblique solution for the jth item discrimination parameter vector in the kth cluster yields

$$\mathbf{a}_{jk}^{\mathbf{T}} = (\overbrace{0,...,0}^{k-1}, a_{jk}, \overbrace{0,...,0}^{\delta-k}), \quad a_{jk} > 0 \tag{2.30}$$

As a result,

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{n_k} \hat{a}_{jk}}{\sqrt{\sum_{k=1}^{\delta}\left(\sum_{j=1}^{n_k} \hat{a}_{jk}\right)^2}} \tag{2.31}$$

In this equation, a_{jk} represents the non-zero discrimination parameter for the jth item in the kth cluster, where $j = 1,2,...,n_k$, $k = 1,2,...,\delta$, n_k is the number of items in the kth

cluster, and δ is the number of dimensions. It should further be noted that in the special case that all item discriminations are equal (i.e., under the Rasch model),

$$\hat{\alpha}_k = \frac{n_k}{\sqrt{\sum_{k=1}^{\delta} n_k^2}} \tag{2.32}$$

When using the multidimensional normal ogive model, corresponding unidimensional item parameter estimates are derived as a function of the linear composite coefficients ($\boldsymbol{\alpha}$), the multidimensional item discrimination vector ($\mathbf{a}_j^{\mathrm{T}}$), the index related to multidimensional item difficulty (d_j), and the multidimensional ability covariance matrix ($\boldsymbol{\Sigma}$) in the population of interest as follows:

$$P(U_j = 1 \mid \hat{\theta}_\alpha) = E(U_j \mid \hat{\theta}_\alpha) = \Phi(\hat{a}_{\alpha j}\hat{\theta}_\alpha + \hat{d}_{\alpha j}) \tag{2.33}$$

where Φ is the standard normal cumulative distribution function (CDF) and

$$\hat{a}_{\alpha j} = (1 + \hat{\sigma}_{\alpha j}{}^2)^{-\frac{1}{2}} \hat{\mathbf{a}}_j^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\alpha}} \tag{2.34a}$$

$$\hat{d}_{\alpha j} = (1 + \hat{\sigma}_{\alpha j}{}^2)^{-\frac{1}{2}} \hat{d}_j \tag{2.34b}$$

and

$$\hat{\sigma}_{\alpha j}{}^2 = \hat{\mathbf{a}}_j^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \hat{\mathbf{a}}_j - (\hat{\mathbf{a}}_j^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\alpha}})^2 \tag{2.34c}$$

The authors continue by noting that the IRT true score (T_α) associated with the linear composite (θ_α)—which is the sum of the probabilities of obtaining correct responses over all items at each composite ability level—is expressed as

$$T_\alpha = \xi(\theta_\alpha) \equiv \xi(\sum_{k=1}^{\delta} \alpha_k \theta_k) \tag{2.35}$$

This expression preserves the property of unidimensional IRT true scores in that the function $\xi(\cdot)$ is strictly increasing.

As the focus of this paper is not on the *methods* Zhang (1996), Zhang and Stout (1999a), and Zhang and Wang (1998) used to derive unidimensional linear composites and corresponding item parameter estimates, it should be noted in passing that the basis for their procedure was provided from MIRT theory and multivariate normal distribution theory. Specifically, Zhang and colleagues used the definition of the "direction of best measurement" described earlier in this chapter—which is the direction corresponding to the average multidimensional information function evaluated in all directions—to determine which unidimensional composite would best represent a single reported score. Using the direction of best measurement as the unidimensional ability scale, Zhang and colleagues then proceeded to use multivariate normal distribution theory to determine unidimensional (marginal) item parameters corresponding to the unidimensional ability scale. In essence, the coefficients for the linear composites (i.e., $\boldsymbol{\alpha}$) as defined by Zhang and colleagues provide a unidimensional ability estimate which can be conceptually viewed as an approximation of the combination of the proficiencies that the test most precisely measures.

CHAPTER 3

METHODOLOGY

This chapter consists of seven main sections. First, a description of the data used in this study appears, accompanied by a list of the procedures that were conducted. Second, a description of how the unidimensional IRT equating procedures were conducted appears. Third, a detailed description of the proposed new methodology for conducting MIRT observed score and true score equating is presented, followed by a description of how the MIRT equating procedures were conducted. Next, a discussion of the assumptions associated with the MIRT equating procedures is presented. Finally, a description of the equipercentile procedures that were conducted, followed by a framework for evaluating the procedures, is presented.

Data and Procedures

The data used in this study were collected under the random groups equating design and came from two forms of the Iowa Tests of Educational Development (ITED) (Forsyth, Ansley, Feldt, & Alnot, 2001), Level 17/18 battery. Specifically, each of the following tests within these batteries were equated: (a) Mathematics: Concepts and Problem Solving, (b) Analysis of Science Materials, and (c) Analysis of Social Studies Materials. The sample size for each form was 2,500.

Each of the three tests consisted of several clusters of items as indicated by the ITED content classification tables. For example, each item on the Mathematics: Concepts and Problem Solving test was classified as either Numbers and Operations on Numbers, Data Analysis/Probability/Statistics, Geometry/Measurement, or Algebraic

Concepts. These clusters formed the bases by which the MIRT linking procedures were conducted.

Each of the following procedures were conducted to equate Form A and Form B tests: (1) unidimensional IRT observed score equating, (2) unidimensional IRT true score equating, (3) full MIRT observed score equating, (4) unidimensional approximation of MIRT observed score equating, (5) unidimensional approximation of MIRT true score equating, and (6) equipercentile equating. The equipercentile equating procedure was conducted for the purpose of comparison with the other procedures.

Unidimensional Equating Procedures

BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) was originally used to estimate unidimensional item parameters. Several problems arose when calibrating the sets of item responses using BILOG-MG, however. Specifically, both discrimination and difficulty parameters for each form were very high and did not appear to be within the bounds typical for item parameters. Provided that the medians for each form were relatively low, it was hypothesized that BILOG-MG may have had difficulty estimating the lower asymptote parameters for these difficult tests. Therefore, various conditions were manipulated in BILOG-MG and the item parameters were recalibrated to see what the effects were. Specifically, a series of BILOG-MG calibrations was conducted with various fixed lower asymptote parameters provided for each item (as opposed to estimating lower asymptote parameters during the calibration). The results revealed that as the fixed lower asymptote parameters decreased in value, both discrimination parameter estimates and difficulty parameter estimates appeared more reasonable. Discrimination and difficulty parameters appeared the “best” when lower asymptote

parameters were all fixed at zero. Therefore, all of the data analyses presented in this research were conducted under the two parameter logistic (2PL) model—rather than the three parameter logistic (3PL) model—in order to enhance the quality of the item parameter estimates.

Furthermore, after the BILOG-MG calibrations were conducted, it seemed prudent to calibrate the unidimensional item parameters using TESTFACT (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, 2003), given that the multidimensional item parameters were also calibrated using this program. As a result, any differences between unidimensional equating procedures and multidimensional equating procedures could be more confidently attributed to differences between the equating procedures and not to differences between item parameter estimation programs. Therefore, the unidimensional procedures incorporated in this study were conducted using the TESTFACT program rather than BILOG-MG (it should be noted that although fixed lower asymptotes can be supplied for TESTFACT calibrations, this program does not estimate lower asymptote parameters).

Since the data used in this study were collected under the random groups design, no scale linking procedures were required for the unidimensional procedures: the item parameter estimates were assumed to be on the same scale. After item parameters were estimated, the unidimensional item parameters were then substituted into the computer program PIE (a computer Program for IRT Equating) (Hanson & Zeng, n.d.) to conduct both unidimensional observed score and true score equating procedures.

Multidimensional Equating Methodology

No procedures currently exist for conducting observed score or true score equating in the MIRT framework. As the purpose of the present research is to create such procedures, three new equating procedures—two observed score procedures and one true score procedure—are detailed below. First, an observed score equating procedure is presented as a direct extension of UIRT observed score equating. This procedure is referred to as the "Full MIRT Observed Score Equating" procedure. The true score equating procedure and the second observed score equating procedure use the unidimensional approximation methods detailed by Zhang and colleagues (Zhang, 1996; Zhang & Stout, 1999a; and Zhang & Wang, 1998), and are referred to as the "Unidimensional Approximation of MIRT True Score Equating" procedure and the "Unidimensional Approximation of MIRT Observed Score Equating" procedure, respectively.

Full MIRT Observed Score Equating

Relatively straightforward extensions can be implemented to conduct observed score equating in the MIRT framework. In the unidimensional IRT framework, conditional observed score distributions (i.e., $f(x \mid \theta)$) are first determined at each ability level (θ_i). The MIRT analog is to derive conditional distributions for each *combination* of ability levels (i.e., $f(x \mid \boldsymbol{\theta})$), where $\boldsymbol{\theta}$ denotes the entire ability space (Kolen & Wang, 2007). Similar to the UIRT framework, these can be computed using the Lord-Wingersky algorithm. To implement this algorithm, however, a vector of ability levels is used in place of a single ability level as follows:

$$f_r(x \mid \boldsymbol{\theta}_i) \quad = f_{r-1}(x \mid \boldsymbol{\theta}_i)(1 - p_{ir}) \qquad x = 0 \qquad (3.1a)$$

$$= f_{r-1}(x \mid \boldsymbol{\theta}_i)(1 - p_{ir}) + f_{r-1}(x-1 \mid \boldsymbol{\theta}_i)p_{ir} \qquad 0 < x < r \qquad (3.1b)$$

$$= f_{r-1}(x-1 \mid \boldsymbol{\theta}_i)p_{ir} \qquad x = r \qquad (3.1c)$$

In this series of equations, $f_r(x \mid \boldsymbol{\theta}_i)$ is the conditional observed score distribution over the first r items, $f_{r-1}(x \mid \boldsymbol{\theta}_i)$ is the conditional observed score distribution over the first $r-1$ items, and p_{ir} is the probability of correctly answering item r for an examinee of ability vector $\boldsymbol{\theta}_i$. In practice, this algorithm begins as:

$$f_r(x \mid \boldsymbol{\theta}_i) \qquad = (1 - p_{ir}) \qquad x = 0 \qquad (3.2a)$$

$$= p_{ir} \qquad x = 1 \qquad (3.2b)$$

After this initial step, $r-1$ iterations are required to compute the conditional observed score distribution ($f_r(x \mid \boldsymbol{\theta}_i)$) for r items.

In the unidimensional IRT framework, these conditional distributions are then multiplied by the ability density ($\psi(\theta)$) and either summed or integrated over all ability levels to obtain a marginal observed score distribution for each form. The MIRT analog is to multiply the conditional distributions by the multivariate ability density ($\psi(\boldsymbol{\theta})$) and sum (or integrate) over all ability spaces as,

$$f(x) = \sum_1 \sum_2 \ldots \sum_\delta f(x \mid \boldsymbol{\theta})\psi(\boldsymbol{\theta}) \qquad (3.3a)$$

or

$$f(x) = \int_1 \int_2 \ldots \int_\delta f(x \mid \boldsymbol{\theta})\psi(\boldsymbol{\theta})d\boldsymbol{\theta} \qquad (3.3b)$$

In these equations, δ represents the number of dimensions. Similar to the unidimensional case, traditional equipercentile methods can then be used to equate the two forms.

Unidimensional Approximation

Both the unidimensional approximation of the MIRT true score equating procedure and the unidimensional approximation of the MIRT observed score equating procedure are conducted by estimating unidimensional item parameters and unidimensional ability distributions from the multidimensional data. Motivation for using unidimensional estimates, along with the unidimensional estimation procedures, appears below.

Several problems arise when trying to extend unidimensional IRT true score equating to the MIRT framework. In the UIRT framework, true score equating was conducted by relating the forms to be equated through the respective test characteristic curves (TCCs). Recall that the TCC relates the IRT definition of true score ($\tau(\theta_i) = \sum_j p_{ij}(\theta_i; a_j, b_j, c_j)$) to ability level ($\theta$). In MIRT, true scores are related to ability levels through the test characteristic surface (TCS), which is the multidimensional equivalent of the test characteristic curve. For each *combination* of ability levels (corresponding to each dimension), the probabilities of obtaining correct responses to each item are summed to form true scores (i.e., $\tau(\boldsymbol{\theta}) = \sum p(\boldsymbol{\theta})$). As such, an infinite number of combinations of ability levels is associated with a particular true score. The problem arises in that, when the test characteristic surface is computed for Form B, different combinations of ability levels corresponding to the Form A true score may map to *different* true scores on Form B. That is, there is no *unique* solution for mapping true scores on Form A to true scores on Form B.

The problem discussed above can be bypassed using the results presented in Zhang (1996), Zhang and Stout (1999a), and Zhang and Wang (1998). These authors

demonstrated that any set of item responses that can be adequately modeled by a multidimensional compensatory IRT model can be closely approximated by a unidimensional IRT model with estimated unidimensional ability and item parameters.

First, the vector of weights corresponding to the test-level direction of best measurement can be estimated as

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{a}_{jk}}{\sqrt{\sum_{k=1}^{\delta} \left(\sum_{j=1}^{N} \hat{a}_{jk} \right)^2}} \tag{3.4}$$

where α_k represents the kth standardized coefficient in the vector of linear composite coefficients ($\boldsymbol{\alpha}$), a_{jk} represents the discrimination parameter for the jth item on the kth dimension, $j = 1,2,...,N$, $k = 1,2,...,\delta$, N is the total number of items on the test, and δ is the number of dimensions (Zhang, 1996). Under the oblique (simple structure) solution, this reduces to

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{n_k} \hat{a}_{jk}}{\sqrt{\sum_{k=1}^{\delta} \left(\sum_{j=1}^{n_k} \hat{a}_{jk} \right)^2}} \tag{3.5}$$

where n_k represents the number of items in the kth cluster. (The number of clusters is assumed to equal the number of dimensions in this study). Next, corresponding unidimensional item parameters can be estimated as:

$$\hat{a}_{\alpha j} = (1 + \hat{\sigma}_{\alpha j}{}^2)^{-\frac{1}{2}} \hat{\mathbf{a}}_j^{\mathbf{T}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\alpha}} \tag{3.6a}$$

$$\hat{d}_{\alpha j} = (1 + \hat{\sigma}_{\alpha j}{}^2)^{-\frac{1}{2}} \hat{d}_j \tag{3.6b}$$

$$\hat{b}_{\alpha j} = \frac{-\hat{d}_{\alpha j}}{\hat{a}_{\alpha j}} \tag{3.6c}$$

$$\hat{c}_{\alpha j} = \hat{c}_j \tag{3.6d}$$

and

$$\hat{\sigma}_{\alpha j}{}^2 = \hat{\mathbf{a}}_j^{\mathrm{T}} \hat{\mathbf{\Sigma}} \hat{\mathbf{a}}_j - (\hat{\mathbf{a}}_j^{\mathrm{T}} \hat{\mathbf{\Sigma}} \hat{\boldsymbol{\alpha}})^2 \tag{3.6e}$$

In this series of equations, the vector $\boldsymbol{\alpha}$ contains the standardized linear composite coefficients, $\mathbf{a}_j^{\mathrm{T}}$ is the multidimensional discrimination vector, d_j is an index related to the multidimensional difficulty parameter, and $\mathbf{\Sigma}$ is the multidimensional ability covariance matrix in the population. Whereas this parameterization is in accordance with the normal ogive model, these parameters can be substituted as logistic model parameters by noting the fact that these models yield nearly identical probabilities. That is,

$$p_{\alpha ij}(\theta_{\alpha i}) = \hat{c}_j + (1-\hat{c}_j)\Phi\left(\hat{a}_{\alpha j}(\theta_{\alpha i} - \hat{b}_{\alpha j})\right) \approx \hat{c}_j + (1-\hat{c}_j)\frac{e^{1.7\hat{a}_{\alpha j}(\theta_{\alpha i}-\hat{b}_{\alpha j})}}{1+e^{1.7\hat{a}_{\alpha j}(\theta_{\alpha i}-\hat{b}_{\alpha j})}} \tag{3.7}$$

where the first probability is parameterized according to the normal ogive model and the second probability is parameterized according to the logistic model. The differences in these probabilities are less than 0.01 across the entire ability space (Lord, 1980). Although this change in parameterization will result in minor loss of precision, this should not have a significant effect on the results. Thus, the probability of a correct response can be modeled as:

$$p_{\alpha ij}(\theta_{\alpha i}) = \hat{c}_j + (1-\hat{c}_j)\frac{e^{1.7\hat{a}_{\alpha j}(\theta_{\alpha i}-\hat{b}_{\alpha j})}}{1+e^{1.7\hat{a}_{\alpha j}(\theta_{\alpha i}-\hat{b}_{\alpha j})}} \tag{3.8}$$

See Appendix C for an example of how these procedures are conducted. It should be noted that the IRT true score (T_α) associated with the linear composite (θ_α)—which is the sum of the probabilities of obtaining correct responses over all items at each composite ability level—is expressed as

$$T_\alpha = \xi(\theta_\alpha) \equiv \xi(\sum_{k=1}^{\delta} \alpha_k \theta_k) \tag{3.9}$$

This expression preserves the property of unidimensional IRT true scores in that the function $\xi(\cdot)$ is strictly increasing.

The previous results imply that 1) a unidimensional composite ability (θ_α) which approximates the ability which the test most closely measures can be constructed from any generalized δ-dimensional compensatory model and that 2) unidimensional item parameters corresponding to the unidimensional composite ability can be produced. Consequently, a unidimensional test characteristic curve can be constructed which relates unidimensional composite ability level (θ_α) to composite true score (T_α).

Unidimensional Approximation of MIRT True Score Equating

Using the unidimensional approximations provided above, unidimensional IRT true score equating procedures can be conducted to relate composite true scores (T_α) on the multidimensional test forms. First, a true score on Form A ($\tau_{\alpha A}$) can be selected for which the corresponding composite true score on Form B is desired. Using an iterative procedure (specifically, the Newton-Raphson method), the corresponding unidimensional composite ability level (θ_α) associated with this Form A true score can be obtained by minimizing the difference $func(\theta_{\alpha i}) = \tau_{\alpha A} - \sum_{j:A} p_{ij}(\theta_{\alpha i}; a_{\alpha j}, b_{\alpha j}, c_j)$. Finally, using the IRT definition of true score, the composite true score on Form B associated with the Form A composite true score can be computed as $\tau_{\alpha B}(\theta_{\alpha i}) = \sum_{j:B} p_{ij}(\theta_{\alpha i}; a_{\alpha j}, b_{\alpha j}, c_j)$.

Unidimensional Approximation of MIRT Observed Score Equating

To conduct unidimensional approximation of MIRT observed score equating, unidimensional item parameters and abilities can first be estimated using the methodology described above. Conditional distributions $f(x \mid \theta_\alpha)$ can then be determined at each composite ability level (θ_α) using the Lord-Wingersky recursion formula. The conditional distributions can then be multiplied by the estimated unidimensional ability distribution in the population of examinees and summed (or integrated) across the estimated unidimensional ability space as:

$$f(x) = \sum_{\theta_\alpha} f(x \mid \theta_{\alpha i})\psi(\theta_{\alpha i}) \tag{3.10a}$$

or

$$f(x) = \int_{\theta_\alpha} f(x \mid \theta_{\alpha i})\psi(\theta_{\alpha i})d\theta_\alpha \tag{3.10b}$$

Once marginal distributions are computed for each form, the forms can be equated using traditional equipercentile procedures.

Multidimensional Procedures

TESTFACT (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, 2003) was used to estimate multidimensional item parameters. The TESTFACT program essentially conducts a factor analysis procedure on inter-item tetrachoric correlations using the marginal maximum likelihood procedure (Bock & Aitken, 1981). The resulting parameterization is in accordance with the multidimensional normal ogive model. Each TESTFACT calibration was specified to yield an orthogonal solution, with the number of specified dimensions equal to the number of clusters identified on the ITED form.

Item parameter estimates on Form A were then rotated to the scale of Form B to account for rotational indeterminacy (Thompson et al., 1997). First, the orthogonal rotation matrix ($\mathbf{T}$) was determined using the methods discussed in the previous chapter. The matrix of item discrimination parameters on Form A was then post-multiplied by the orthogonal rotation matrix to rotate the estimated discrimination parameters as follows:

$$\mathbf{a}_{Jj}^{\mathbf{T}} = \mathbf{a}_{Ij}^{\mathbf{T}}\mathbf{T} \tag{3.11}$$

In TESTFACT, translation and dilation indeterminacies for exploratory solutions are resolved by specifying axes to be orthogonal and of unit length (i.e., $MVN(\mathbf{0},\mathbf{I})$). Therefore, quadrature points and weights were determined from the multivariate normal $MVN(\mathbf{0},\mathbf{I})$ distribution.

DETECT (Zhang & Stout, 1999b) was then used to assess dimensionality for each form, as well as to observe which clusters of items on each form were the most statistically similar (homogeneous). This procedure was conducted twice for each form: the procedure was conducted in both an exploratory and a confirmatory manner. For the confirmatory procedure, items on each form were clustered according to ITED content classification tables. For example, each item on the Mathematics: Concepts and Problem Solving test was specified to load on either Numbers and Operations on Numbers, Data Analysis/Probability/Statistics, Geometry/Measurement, or Algebraic Concepts. The exploratory procedure was conducted to determine how well the original ITED classification scheme was recovered by the DETECT procedure. Ideally, the exploratory solution would yield as many clusters as there are classifications according to the ITED content classification tables, with each item clustering within its respective content classification. Realistically, however, this is not likely to be the case. It is more likely

that the original ITED classification scheme will not be perfectly recovered by the DETECT procedure.

The discrepancy between the ITED classification tables and the clusters recovered under the exploratory DETECT procedure was then used as an indicator of how well the MIRT linking procedure can perform. Recall that to perform this procedure, items were first clustered according to the content specifications. The reference composite—conceptually viewed as the average direction that the individual items in the cluster are pointing in—was then determined for each cluster. If all items in a cluster are pointing in exactly the same direction (and thus measuring the same composite of traits), then the exploratory DETECT results will yield the same clustering scheme as the ITED classification tables. In this situation, the reference composite can be used as a stable indicator of the direction that these items are pointing in. To the degree that the items within a cluster point in different directions, it is likely that the DETECT exploratory results will yield a different clustering scheme than the ITED classification tables. In this situation, the reference composite will still point in the average direction that the individual items within a cluster are pointing, yet the items will empirically appear to be measuring different traits. Since the MIRT linking procedure is performed by aligning reference composites on Form A to be pointing in the same direction as the corresponding reference composites on Form B, a comparison of the exploratory DETECT solution and the ITED classification tables was used to evaluate how well the MIRT linking procedure may have worked.

The maximum DETECT value was used to gauge the extent to which each form is multidimensional, and the IDN index value was used to gauge the extent to which the

specified solution (the ITED content classification scheme for the confirmatory analysis, and the homogeneous clusters dictated by the exploratory analysis) conforms to simple structure. The ratio r was used to gauge the stability of the estimation procedure.

After MIRT parameters were placed on the same scale, the MIRT true score equating procedure and both MIRT observed score equating procedures were conducted. To conduct the unidimensional approximation of the MIRT true score equating procedure and the unidimensional approximation of the MIRT observed score equating procedure, the unidimensional item parameters were first estimated using the methodology described above. The unidimensional ability distributions were determined as follows.

First, quadrature weights were defined in accordance with the multivariate standard normal distribution. The number of quadrature points was equal to n^{δ}, where n represents the number of quadrature points per dimension and δ represents the number of dimensions. For example, the Science exams were calibrated with respect to two dimensions and forty quadrature points per dimension (i.e., $n = 40$ and $\delta = 2$). Thus, there were $40^2 = 1600$ total quadrature points, and the multidimensional ability density was represented as,

$$\begin{bmatrix} \underline{\theta_1} & \underline{\theta_2} & \underline{Density} \\ -4.0 & -4.0 & [density] \\ -4.0 & -3.8 & [density] \\ \vdots & \vdots & \vdots \\ -4.0 & 4.0 & [density] \\ -3.8 & -4.0 & [density] \\ -3.8 & -3.8 & [density] \\ \vdots & \vdots & \vdots \\ 4.0 & 3.8 & [density] \\ 4.0 & 4.0 & [density] \end{bmatrix}$$

To obtain unidimensional quadrature points, each vector of multidimensional quadrature points (i.e., $[\theta_1 = -4.0\,,\ \theta_2 = -4.0]$, $[\theta_1 = -4.0\,,\ \theta_2 = -3.8]$, etc.) was multiplied by the vector of standardized linear composite coefficients ($\boldsymbol{\alpha}$) in accordance with the test-level direction of best measurement. For example, the unidimensional quadrature point was equal to $\theta_1\alpha_1 + \theta_2\alpha_2$, where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]$. The density associated with each unidimensional quadrature point remained the same, and was equal to the multivariate normal density at the specific vector, $[\theta_1, \theta_2]$.

The resulting unidimensional density still contained n^{δ} quadrature points (which was the same number of points as the multivariate quadrature density), which proved to be quite unwieldy. Therefore, the quadrature points were then rank ordered according to magnitude. For example, the unidimensional ability density after rank ordering was represented as,

$$\begin{bmatrix} \underline{\theta_\alpha} & \underline{Density} \\ \theta_{11}\alpha_1 + \theta_{12}\alpha_2 & [density] \\ \theta_{21}\alpha_1 + \theta_{22}\alpha_2 & [density] \\ \vdots & \vdots \\ \theta_{n^{\delta}1}\alpha_1 + \theta_{n^{\delta}2}\alpha_2 & [density] \end{bmatrix}$$

where $\theta_{11}\alpha_1 + \theta_{12}\alpha_2$ is the smallest quadrature point, $\theta_{21}\alpha_1 + \theta_{22}\alpha_2$ is the second smallest quadrature point,…, and $\theta_{n^{\delta}1}\alpha_1 + \theta_{n^{\delta}2}\alpha_2$ is the largest quadrature point. The n^{δ} quadrature points and weights were then collapsed into forty points and weights in order to yield a more feasible solution. Specifically, equal-sized intervals were created from the n^{δ} quadrature points by dividing n^{δ} by 40 to determine how many of the n^{δ} points should appear within each interval. The mean quadrature point for each interval was

used as the final quadrature point, and the sum of the quadrature weights within each interval was used as the final quadrature weight.

After this procedure was conducted, however, the resulting ability density was revealed to be rather ragged (as opposed to being a smooth distribution). To investigate this phenomenon, additional analyses which incorporated larger numbers of quadrature points per dimension were conducted. The results revealed that as the number of quadrature points per dimension increased, the quadrature distribution converged to a standard normal distribution (see Figure A-22 in Appendix A, which reveals how the unidimensional ability distribution changes by incorporating more quadrature points per dimension). Therefore, instead of using the "unsmoothed" ability distributions for the unidimensional approximation of MIRT observed score equating procedures, standard normal ability distributions were substituted. These values were incorporated in the computer program PIE (Hanson & Zeng, n.d.) to conduct both the observed score and the true score equating procedure.

To conduct the full MIRT observed score equating procedure, conditional observed score distributions were first determined for each combination of θ-values (using multivariate normal quadrature points) using a modified version of the Lord-Wingersky algorithm for the MIRT framework. The multivariate normal quadrature weights ($\psi(\boldsymbol{\theta})$) were then multiplied by the conditional observed score distributions ($f(x|\boldsymbol{\theta})$) and summed over the ability space to determine an estimated marginal observed score distribution. After this procedure was conducted for both forms, traditional equipercentile equating was conducted to equate the forms. The entire full MIRT observed score equating procedure was conducted using the computer programs R

(R Development Core Team, 2008) and RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui, & Chien, 2004). Specifically, R code was created to determine the conditional observed score distributions and the marginal observed score distribution for each form, and RAGE-RGEQUATE was used to conduct the equipercentile equating.

Scale Linking and Equating Assumptions

Several assumptions were made in order to conduct the scale linking and equating procedures described in this study. Recall that the data used in this research were collected under the random groups equating design. Therefore, examinees who were administered Form A and examinees who were administered Form B were assumed to be equivalent in both origin (translation) and unit of measurement (dilation) on each of the measured traits. Thus, only an orthogonal rotation matrix was required to account for rotational indeterminacy for the scale linking procedures (Thompson et al., 1997).

It is evident that the scale linking procedure used in this study requires that the same traits be measured on both forms to be equated. Furthermore, this procedure also requires that the same *composite* of traits be measured by each form (as opposed to each trait simply being measured by each form). Conceptually, this implies that the extent to which each trait contributes to the total score be the same across each form. Mathematically, this implies that the test-level reference composite must be the same for both forms to be equated: "Not only must each test form measure the same constructs, but the composite construct formed must also be the same across test forms. The composite construct formed by the test items may be thought of as analogous to an alloy composed of various metals. Just as differing proportions of copper and tin can be used to form different bronze alloys, different mixtures of geometry, trigonometry, and algebra items

(for example) can be used to form different overall math achievement reference composites" (Thompson et al., 1997, p. 2).

Furthermore, the scale linking procedure only performs well when the angle between any two reference composites on *one form* to be equated is similar to the angle between the corresponding reference composites on the *other form* to be equated (recall that the angle between any two reference composites on the same form is related to the correlation between the reference composites). The objective of the scale linking procedure is to determine the orthogonal rotation matrix ($\mathbf{T}$) such that, after the rotation is applied, corresponding reference composites on *both forms* to be equated are pointing in the same direction. Mathematically, this implies that $\mathbf{T}$ is determined by minimizing the function $tr(\mathbf{E}^{\mathbf{T}}\mathbf{E})$ where $\mathbf{E} = \mathbf{N} - \mathbf{MT}$, $tr(\cdot)$ represents the trace function, and $\mathbf{N}$ and $\mathbf{M}$ contain reference composites on Form B and Form A, respectively. Conceptually, this implies that $\mathbf{T}$ is the orthogonal rotation matrix such that after the rotation is applied, corresponding rows on $\mathbf{N}$ and $\mathbf{M}$ are as similar as possible. The matrices $\mathbf{N}$ and $\mathbf{M}$ display the following pattern:

$$\mathbf{N} = \begin{bmatrix} REFERENCE & COMPOSITE & 1_B \\ REFERENCE & COMPOSITE & 2_B \\ & \vdots & \\ REFERENCE & COMPOSITE & \delta_B \end{bmatrix} \mathbf{M} = \begin{bmatrix} REFERENCE & COMPOSITE & 1_A \\ REFERENCE & COMPOSITE & 2_A \\ & \vdots & \\ REFERENCE & COMPOSITE & \delta_A \end{bmatrix}$$

Recall that the angle between any two eigenvectors on the *same form* does not change under an orthogonal rotation, and that a reference composite is, in essence, an eigenvector. Under an orthogonal rotation, each eigenvector will rotate in relation to the coordinate axes, but the angle between the eigenvectors on the *same form* does not change. Therefore, in order for each pair of corresponding reference composites to be

pointing in identical directions (or, equivalently, in order for corresponding rows in $\mathbf{N}$ and $\mathbf{M}$ to be as similar as possible), the angle between any two reference composites on one form must necessarily be similar to the angle between the corresponding reference composites on the other form.

After MIRT parameters were placed on the same scale, equating was conducted within the MIRT framework. Two of the MIRT equating procedures—the unidimensional approximation of the MIRT observed score equating procedure and the unidimensional approximation of the MIRT true score equating procedure—are based on the derivations provided by Zhang (1996), Zhang and Stout (1999a), and Zhang and Wang (1998). These authors demonstrated that any set of item responses that can be adequately modeled by a multidimensional compensatory IRT model can be closely approximated by a unidimensional IRT model with estimated unidimensional ability and item parameters.

To compute unidimensional item parameter estimates, the multidimensional ability distribution in the population of examinees was assumed to follow a multivariate normal distribution (i.e., $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$). Furthermore, to estimate the unidimensional ability parameters, an assumption was made that each term $w_j E\left[\frac{H_j^{'}\left(\hat{\mathbf{a}}_j^{\mathrm{T}}\hat{\boldsymbol{\Theta}}\right)}{\sqrt{Var(Y \mid \hat{\boldsymbol{\Theta}})}}\right]$, $j = 1,2,...,N$ is equal across all items. The terms $w_j E\left[\frac{H_j^{'}\left(\hat{\mathbf{a}}_j^{\mathrm{T}}\hat{\boldsymbol{\Theta}}\right)}{\sqrt{Var(Y \mid \hat{\boldsymbol{\Theta}})}}\right]$ "may be considered to be 'compound weights' for items contributing to the test direction; each term is completely determined by the score weight w_j and the item model (theoretical) weight

$E\left[\frac{H_j^{'}\left(\hat{\mathbf{a}}_j^{\mathrm{T}}\hat{\mathbf{\Theta}}\right)}{\sqrt{Var(Y \mid \hat{\mathbf{\Theta}})}}\right]$ which is mainly determined by the derivative of the link function and the item discrimination parameter vector" (Zhang, 1996, p. 25).

For the full MIRT observed score equating procedure, the multivariate quadrature distribution was specified to follow the multivariate standard normal distribution with uncorrelated abilities (i.e., $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \mathbf{I})$. Since the MIRT parameter estimation procedure resolves translation and dilation indeterminacies by fixing the coordinate axes to yield a mean of 0 and standard deviation of 1 on each dimension, the ability density ($\boldsymbol{\theta} \sim MVN(\mathbf{0}, \mathbf{I})$) should be correct with respect to mean and standard deviation. However, the ability estimates are likely to be correlated in practice. That is, it is likely that $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \mathbf{\Sigma})$, where the off-diagonal elements of $\mathbf{\Sigma}$ are positive values between 0 and 1. In this study, however, correlations between ability estimates on different dimensions were not estimated, as item parameter estimates and ability estimates were calibrated under a "complex structure" model as opposed to a "simple structure" model (thus resulting in orthogonal axes as opposed to correlated axes). Although the correlations between ability estimates on different dimensions could have been obtained via several different methods, these methods would not be empirically justified, as the exams did not demonstrate simple structure in accordance with the ITED classification tables (see the results pertaining to Dimensionality Assessment).

For the unidimensional approximation of the MIRT observed score equating procedure, the unidimensional quadrature distribution was specified to follow a standard normal distribution. This parameterization should be accurate, as the unidimensional quadrature distribution was empirically verified as following the standard normal

distribution (see the discussion concerning the quadrature distribution under the "Multidimensional Procedures" section above).

Other Procedures for Conducting MIRT Equating

It should be noted that the MIRT equating methods could have been conducted several different ways. For example, another method for conducting the MIRT equating procedures would be to obtain the MIRT estimates of the lower asymptote parameter (c) by calibrating each cluster of items separately under a UIRT model. Since each cluster of items is expected to be internally homogeneous, this method might yield better estimates of the lower asymptote parameter for each item. In IRT estimation procedures, parameter estimates are highly dependent upon one another. In this study, all items within a specific test were calibrated simultaneously using a UIRT model to obtain lower asymptote parameters for each item (despite the multidimensional structure of the exam). These unidimensional lower asymptote parameters were originally substituted as the lower asymptote parameters for the MIRT model (before it was decided to fix all lower asymptote parameters to 0). Since the unidimensional calibration is expected to yield biased estimates of item and person parameters due to the multidimensional structure of the exam, estimates for the lower asymptote may also be inaccurate (given that parameter estimates are highly dependent upon one another). To obtain more accurate estimates, each cluster of items could be calibrated separately under a UIRT model. In this particular study, however, this methodology did not seem feasible provided that some clusters of items contained very few items. Furthermore, this methodology was not required, as all lower asymptote parameters were fixed to 0.

Continuing, although the multidimensional procedures were conducted under a "complex structure" solution (i.e., items were allowed to measure more than one trait), the multidimensional procedures could have been conducted under a "simple structure" solution. In this situation, each item would be specified to measure only one trait (as dictated by the ITED classification tables), and a correlation coefficient would be estimated for each pair of traits. For example, all Math items within the Mathematics: Concepts and Problem Solving domain would be specified to measure only this domain, all Math items within the Numbers and Operations on Numbers domain would be specified to measure only this domain, and a correlation coefficient would provide an indication of the relationship between the Mathematics: Concepts and Problem Solving domain and the Numbers and Operations on Numbers domain.

Under the "simple-structure" solution, the discrimination parameters for each item *not* associated with a particular domain would be fixed to 0, resulting in only one positive discrimination parameter for each item. Specifically, each row in the matrix of item discrimination parameters ($\mathbf{A}$) would contain only one non-zero element, which would appear in the column corresponding to the dimension that the specific item measures. For this matrix, each row corresponds to an item and each column corresponds to a dimension. Thus, the pattern for the entire item discrimination matrix ($\mathbf{A}$) would be:

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ \vdots & 0 & 0 & \cdots & 0 \\ a_{n_1 1} & 0 & 0 & \cdots & 0 \\ 0 & a_{12} & 0 & \cdots & 0 \\ 0 & \vdots & 0 & \cdots & 0 \\ 0 & a_{n_2 2} & 0 & \cdots & 0 \\ 0 & 0 & a_{13} & 0 & 0 \\ 0 & 0 & \vdots & 0 & 0 \\ 0 & 0 & a_{n_3 3} & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & a_{1\delta} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & a_{n_d \delta} \end{pmatrix} \tag{3.11}$$

In this series of equations, n_k represents the number of items in the kth cluster, and δ represents the number of dimensions. For each item, the simple structure solution constrains $\delta - 1$ discrimination parameters to 0. The simple structure solution assumes that the item responses adequately conform to simple structure, and that the matrix of item discrimination parameters ($\mathbf{A}$) can be adequately modeled with only one non-zero element per item. This implies that each item measures only one trait (specifically, the trait as identified by the ITED classification tables) and that the traits may be correlated. However, this solution may not adequately model the data, as was the case in this study. If the matrix of item responses demonstrates complex structure—as opposed to simple structure—this assumption may not be feasible. A better estimate of the item discrimination matrix ($\mathbf{A}$) may be obtained under an unconstrained (complex structure) solution. The complex structure solution frees the previously constrained (to 0) $\delta - 1$ discrimination parameters for each item. As a result, the complex structure solution may yield a better estimate of the item discrimination matrix ($\mathbf{A}$).

The simple-structure method did not appear feasible in this study, as empirical evidence from the DETECT procedure revealed that the ITED exams did not demonstrate simple structure (see the discussion of the Dimensionality Assessment procedures). That is, the items within each ITED domain did not necessarily measure the exact same trait, and some items across the various ITED domains empirically appeared to measure similar traits.

Lastly, it should be noted that the MIRT equating procedures presented in this research could be conducted in conjunction with any of the previously described equating designs (i.e., single group, random groups, or common item nonequivalent groups design). Whereas the linking procedures would differ depending on the equating design used, the MIRT procedure to equate number-correct scores would remain the same.

Equipercentile Equating

Equipercentile equating was also conducted to equate each pair of forms, and a post-smoothing method was incorporated to reduce random error associated with this procedure. Specifically, cubic splines were fit to each equipercentile equating relationship to determine an appropriate smoothing parameter. After examining the effects of various post-smoothing parameters, a value of $S = 0.20$ was applied to each of the three equipercentile equating procedures. The results from both the unsmoothed equipercentile equating procedure and the smoothed equipercentile equating procedure were then compared with the results from the other equating procedures, since the assumptions associated with this procedure are not expected to be violated in this study (for example, the UIRT procedures are expected to violate the unidimensionality assumption, given the multidimensional nature of the exams). The equipercentile

equating procedure essentially relates observed scores on both forms with the same percentile rank. Since the data used were collected under the random groups design, examinees that were administered different forms were assumed to have identical distributions on the traits being measured. These procedures were conducted using the computer program RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui, & Chien, 2004) and Equating Error (Hanson, n.d.).

Evaluation Procedures

The following six procedures were conducted to equate Form A and Form B for each set of tests: (1) unidimensional IRT observed score equating, (2) unidimensional IRT true score equating, (3) full MIRT observed score equating, (4) unidimensional approximation of MIRT observed score equating, (5) unidimensional approximation of MIRT true score equating, and (6) equipercentile equating.

Several methods were used to determine the adequacy of the MIRT scale linking procedure for each set of forms. Furthermore, dimensionality assessment methods were employed to determine the extent to which each form is multidimensional and to determine how well the ITED classification schemes were empirically recovered. The standard error of equating (SEE) was used to determine how well the equipercentile procedure worked to equate each set of forms. Discrepancies between each IRT or MIRT equating procedure and the equipercentile procedure were evaluated according to the Differences That Matter (DTM) standard (Dorans, Holland, Thayer, & Tateneni, 2003) and the standard error of equating (SEE). A description of each procedure follows.

Evaluation of Linking Procedures

The accuracy of each linking rotation was determined by using a procedure described by Thompson, et al. (1997). Recall that the orthogonal rotation matrix $\mathbf{T}$ is the solution which minimizes $tr(\mathbf{E}^{\mathrm{T}}\mathbf{E})$, where $\mathbf{E} = \mathbf{N} - \mathbf{MT}$, $tr(\cdot)$ represents the trace function, and the matrices $\mathbf{M}$ and $\mathbf{N}$ contain reference composites for Form A and Form B, respectively. The purpose of the rotation is to align reference composites to be pointing in exactly the same direction in the δ-dimensional space after the rotation. If the rotation works perfectly, then $\mathbf{E} = \mathbf{0}$, meaning that the reference composites are pointing in exactly the same direction. As a result, the angles between corresponding reference composites were compared prior to the rotation and after the rotation was applied. A small post-rotation angle between corresponding reference composites is indicative of an accurate rotation. Thompson, et al. (1997) warned that there is no global criterion based on pre- and post-rotation angles for assessing quality of rotation.

The cluster scheme provided by the exploratory DETECT procedure was also compared to the ITED classification tables to determine how well the MIRT linking procedure may have worked. If the DETECT procedure yields the exact same clusters as the ITED classification tables, then each item within the respective cluster is pointing in nearly the same direction (i.e., each item is measuring the same composite of traits), and the corresponding reference composites provide a solid basis on which to conduct the MIRT linking procedure. To the degree that the DETECT procedure yields clusters that are inconsistent with the ITED classification tables, the reference composites for each cluster might not provide a clear indication of which direction the items are intended to point (i.e., which composite of skills the items are intended to measure). In this case, the

reference composites may still be used as the basis to conduct the MIRT scale linking procedure, but the solution for the reference composite is less stable.

Standard Error of Equating

All equating procedures are susceptible to both systematic error and random error (Kolen & Brennan, 2004). The goal of an equating procedure is to compute the most accurate equating relationship possible so that examinees can be compared on parallel forms. This goal is achieved by minimizing both systematic error and random error. Systematic error may result by performing the equating procedure in a manner that violates the assumptions associated with the procedure. For example, if the data to be equated were collected under the random equivalent groups design, yet the groups were actually quite different on the measured trait, the equating results may be highly inaccurate due to the violation of this assumption. In this particular study, performing the UIRT equating procedures on multidimensional data may result in systematic error due to the violation of the unidimensionality assumption.

Whereas the amount of systematic error that affects an equating procedure may be difficult to quantify, procedures are available to help determine the amount of random error in an equating procedure. Random error may result from performing the equating procedure on a random sample of examinees as opposed to the entire population.

The standard error of equating (SEE) was computed to help determine the amount of random error for the equipercentile procedure. Specifically, bootstrap standard errors were used for this purpose. Bootstrap methods, in general, incorporate a resampling procedure to estimate a population parameter a large number of times. The standard

deviation of the estimated population parameter then serves as an estimate of the standard error of the sampling distribution for that particular statistic.

In this situation, the raw scores for Form A and Form B were both sampled with replacement 2500 times (which is the same sample size as the number of examinees that completed each form). The equipercentile equating procedure was then conducted to equate Form A and Form B using the resampled scores. After this entire process was completed 1,000 times (i.e., resampling with replacement 2500 raw scores per form and then conducting equipercentile equating using the resampled scores), the standard deviation of the estimated equipercentile equating relationships at each raw score served as the estimated standard error of equating at each score. These procedures were conducted using the computer program Equating Error (Hanson, n.d.).

The bootstrap standard error of equating was used to generate confidence bands around the estimated equipercentile equating relationship. The equipercentile equating relationship and corresponding confidence bands were displayed graphically, along with the equating relationships for the other five procedures. Equating relationships that fall outside of the error band are noted as differing significantly from the equipercentile equating relationships; equating relationships that fall within the error band are described as being similar to the equipercentile equating relationship. Furthermore, the estimated SEE at each raw score was used as a cautionary device. If the difference between the equipercentile equating procedure and another equating procedure is large at a particular raw score, yet the SEE is also large at that raw score, the interpretation should be restricted: the equipercentile procedure is subject to much random error at this score. At scores for which the SEE is smaller, the interpretations do not need to be as restricted.

Differences That Matter

Several problems arise when trying to determine an appropriate criterion by which the performance of each equating procedure can be evaluated. First, the "true" equating relationship between each pair of forms is unknown. However, even if the true equating relationship could be specified (via a simulation study), this relationship would most likely be different for observed score procedures and true score procedures. Specifically, IRT observed score procedures and IRT true score procedures are not necessarily expected to yield the same results, given that they are defined differently. True score equating relates true scores on both forms to be equated; although no theoretical justification exists for applying the true score equating results to observed scores, often this is conducted in practice. Observed score equating, on the contrary, provides a statistical adjustment such that the observed score distributions on each form are as similar as possible.

Since no perfect criterion exists for evaluating the performance of each equating procedure in this study, the equipercentile equating procedure was used as a benchmark for comparison for the UIRT and MIRT equating procedures. Since the assumptions associated with this procedure are not expected to be violated in this study (as the UIRT procedures are expected to violate the unidimensionality assumption), this procedure might provide a better indication of how well each of the other equating procedures performs. The UIRT equating procedures are expected to contain a large amount of systematic error, considering that these procedures do not take into account the multidimensional structure of the exams. The MIRT equating procedures, on the contrary, are expected to perform more similarly to the equipercentile equating procedure

given that the MIRT procedures take into account the multidimensionality in the item responses.

The criteria for this study included differences and absolute differences between equated scores for the equipercentile procedure and each of the other five procedures. The differences and absolute differences were also used to compute averages of the differences across all score points to provide a single summary statistic for each procedure. Both tables (including numeric values) and plots were generated to reveal the magnitude of these differences. Furthermore, because there are flaws in using the equipercentile equating procedure as the benchmark for comparison (described above), general trends for each of the equating procedure were also described qualitatively.

Each difference was evaluated against the Difference That Matters (DTM) criterion (Dorans, Holland, Thayer, & Tateneni, 2003). Although Dorans, et al. (2003) defined the DTM in terms of test linking procedures (i.e., placing scores on the same scale for tests that are not intended to be strictly parallel) and in terms of subgroups of examinees (for example, comparing linking results for males with linking results for females), the DTM concept was extended in this study to serve as a criterion for strict parallel form equating across different methods of equating. Specifically, Dorans, et al. define the DTM as a 0.5 raw score difference point between linking results. In this study, the absolute differences between equated scores for the equipercentile procedure and the other five procedures were compared to the 0.5 criterion.

CHAPTER 4

RESULTS

This chapter contains the results from the data analyses and is comprised of five main sections. First, descriptive statistics for each test used in this study are presented. Second, descriptive statistics pertaining to the dimensionality assessment procedures are presented, followed by the results from the multidimensional scale linking procedures. Finally, standard errors of equating for the equipercentile equating procedures are presented, followed by the equating results for all methods.

Descriptive Statistics for Each Form

The data used in this study came from Form A and Form B of the Iowa Tests of Educational Development (ITED) (Forsyth, Ansley, Feldt, & Alnot, 2001), Level 17/18. Specifically, the scores used in this study consisted of a subset of the data that were collected for the purposes of national standardization of the ITED. In the standardization, the exams were administered under uniform conditions to groups selected to reflect the relative breakdown of various minority and socioeconomic groups (Forsyth, Ansley, Feldt, & Alnot, 2001). The sample size for each form used in this study was 2500.

Descriptive statistics for each form used in this study are presented in Table A-1, and the distribution of scores on each form is presented in Figures A-1, A-2, and A-3. For each test, both the mean and the median scores were relatively low in comparison to the length of the scale (and in comparison to the standardization data as a whole), and the distribution of scores on each form was positively skewed. Reliability for each form ranged between 0.88 and 0.90.

Dimensionality Assessment

To investigate the dimensional structure of each form, the DETECT procedure was conducted in both an exploratory and a confirmatory fashion. The exploratory procedure recovers internally homogeneous clusters of items via empirical investigation (i.e., items that point in nearly identical directions in multidimensional space are clustered together) and provides summary statistics in accordance with the recovered cluster scheme. In contrast, the confirmatory procedure computes the summary statistics under a pre-specified solution (i.e., the user indicates how many clusters should appear on each form and the items to be associated with each cluster). The program output for both the exploratory and the confirmatory analyses contains the DETECT value, the IDN index value, and the ratio r. The DETECT value provides an indication of the extent to which the data are multidimensional (values less than 0.2 are viewed as "essentially unidimensional," values between 0.2 and 0.4 as "weak to moderate multidimensionality," values between 0.4 and 1.0 as "moderate to strong multidimensionality," and values above 1.0 as "large multidimensionality" (Kim, 1994; Zhang & Stout, 1999b). The IDN index value ranges from 0 to 1 and indicates how well the data conform to a simple structure model (as opposed to a "complex" structure model). Values close to 1 are indicative of good fit for a simple structure model. The ratio r ranges from 0 to 1 and provides an indication of the stability of the solution. A value close to 1 is indicative of a stable solution, i.e., a similar solution is likely to result if this procedure is conducted on a different (yet comparable) group of examinees.

The summary statistics yielded by the DETECT analyses are presented in Tables A-2, A-3, and A-4. For the unconstrained (exploratory) solution, the DETECT value

ranged between 0.223 and 0.300 across all subjects (Math, Science, and Social Studies) and both forms (Form A and Form B). These results indicate that the scores on each form are "weak to moderately" multidimensional. Under the constrained (confirmatory) solution, these values range between 0.073 and 0.108 and are higher on both Math forms than for the Science and Social Studies forms. These results imply that the ITED domains are highly correlated within each test, resulting in a unidimensional interpretation under the simple structure model.

Across all subjects and both forms, the IDN index values ranged between 0.662 and 0.722 under the exploratory solution. These values imply that—according to the clustering scheme derived by the DETECT procedure—the data moderately adhere to a simple structure solution. Under the confirmatory solution, the IDN index values were smaller (ranging from 0.528 to 0.578). The IDN index values are expected to be smaller under the confirmatory solution, as the clustering schemes recovered under the exploratory DETECT are formed by empirical evidence concerning which items point in the same direction. Since items within each cluster as specified by the ITED classification tables do not necessarily point in the same direction, the degree to which the data conform to a simple structure model would necessarily be smaller. Furthermore, as opposed to the DETECT values—which are higher for the Math forms than for the Science or Social Studies forms—the IDN index values do not appear to vary by subject.

Under the exploratory solution, the ratio r ranges from 0.488 to 0.600 and does not appear to vary by subject: the Social Studies Form A yielded the highest ratio r value, whereas the Social Studies Form B yielded the lowest ratio r value. The ratio r values obtained for each form indicate that a moderately stable solution was obtained.

These results imply that a different cluster scheme may be recovered under a different (yet comparable) group of examinees. Similar to the other two statistics, the ratio r is smaller under the confirmatory solution, and ranges between 0.149 and 0.227. In contrast to the exploratory results, this statistic does appear to vary by subject, however: the confirmatory solution appears to yield a more stable result for both Math forms than for either of the Science or Social Studies forms.

Linking Results

After the study was completed, it was discovered that multidimensional scale linking procedures under the random groups design (which only consists of an orthogonal rotation) were not required to conduct observed score equating or true score equating within the MIRT framework. That is, the same equating relationships would result regardless of whether rotational indeterminacy was accounted for (recall that the linking procedure did not take into account translation or dilation indeterminacies, as examinees who were administered Form A and examinees who were administered Form B are assumed to have the same mean (translation) and unit of measurement (dilation) under the random groups equating design). This will only hold under certain conditions, however. This topic will be further addressed in the next chapter. Regardless, the scale linking results are still presented below, as the results provide additional information concerning the nature of each test used in this study.

To perform the scale linking procedures, items were first grouped into clusters which are intended to be internally homogeneous as dictated by the ITED classification tables. Given that each item within a given cluster is expected to measure the same composite of traits, statistically these items should point in the same direction in

multidimensional space. As a result, a reference composite can be determined for each cluster of items to provide a stable indication of the average direction for these items. The objective of the scale linking procedure is then to determine the orthogonal rotation such that, after the rotation is applied, corresponding reference composites on the forms to be equated are pointing in the same direction. Therefore, angles between corresponding reference composites can be computed prior to the rotation and after the rotation is applied to provide an indication of how well the scale linking procedure may have worked: smaller post-rotation angles are indicative of a more accurate linking procedure (Thompson, Nering, & Davey, 1997).

Furthermore, the item clusters recovered by the exploratory DETECT procedure (Zhang & Stout, 1999b) can be compared with the ITED classification tables to provide an indication of how well the scale linking procedure may have worked. This procedure determines the direction of best measurement for the overall test, and then proceeds to empirically determine homogeneous clusters of items such that the direction of best measurement for each cluster of items deviates as far as possible from the overall direction of best measurement. The resulting cluster scheme provides empirical information concerning which items measure the same composites of traits. Therefore, the DETECT classification scheme can be compared with the ITED classification tables to determine the extent to which reference composites should be used to align the orientation of the axes: if the empirical clustering scheme (i.e., the results from the exploratory DETECT procedure) is perfectly congruent with the ITED classification tables, then each cluster as specified by the ITED classification tables is internally homogeneous, and each reference composite provides a stable indication of the average

direction of the items. As the empirical cluster scheme and the ITED classification tables become more disparate, each reference composite will still point in the average direction of the items within each cluster, yet this estimate of the cluster-level direction may be less stable.

Table A-5 in Appendix A presents pre-rotation angles and post-rotation angles between corresponding reference composites on Forms A and B for each exam, respectively. Nearly all of the post-rotation angles are smaller than pre-rotation angles, which indicate that the scale linking procedures were successful to some extent (there is currently no method available for estimating the degree of success, as no global criterion exists which indicates the extent to which post-rotation angles should be reduced). However, the angles between corresponding reference composites on two of the four Math composites *increased* as a result of the orthogonal rotation, which indicates that the procedure used to link scales on the Math exams may not have been as successful as the procedures used to link scales on the Science and Social Studies exams.

Tables A-6, A-7, and A-8 present the cluster scheme recovered by the exploratory DETECT procedure for each of the six tests (Math Forms A and B; Science Forms A and B; and Social Studies Forms A and B, respectively). For each form, the original ITED classification tables are presented, along with the classification scheme recovered under the exploratory DETECT procedure. Furthermore, a column denoted "DETECT Pattern" was added for ease of interpretation. This column provides a letter corresponding to each item in the "DETECT" column, indicating which ITED domain the item is associated with. For example, each item corresponding to the Math table is coded as either "N," "D," "G," or "A," denoting whether the item was classified, respectively, as Numbers and

Operations on Numbers, Data Analysis/Probability/Statistics, Geometry/Measurement, or Algebraic Concepts according to the original ITED classification tables. This information may provide a visually appealing method of determining the degree to which items within each specific content category empirically clustered together (for example, the degree to which items in the Numbers and Operations on Numbers category empirically clustered together, the degree to which items in the Data Analysis/Probability/Statistics category empirically clustered together, the degree to which items in the Geometry/Measurement category empirically clustered together, and the degree to which items in the Algebraic Concepts category empirically clustered together).

Overall, the original ITED classification scheme was partially—but not perfectly—recovered. For the Math exams, the Numbers and Operations on Numbers items tended to blend with the Data Analysis/Probability/Statistics items. Geometry/Measurement items and Algebraic Concepts items were interspersed throughout the empirical clusters. Furthermore, whereas the DETECT procedure recovered four clusters on the Form B Math exam (as there were four ITED domains), the DETECT procedure recovered five clusters on the Form A Math exam.

The exploratory DETECT procedure did not perfectly recover the ITED classification tables for the Science exams. Though some clusters recovered by the DETECT procedure resembled the ITED classification tables (i.e., all of the items in the empirical cluster were intended to measure the same trait as specified by the ITED tables), other clusters recovered by the DETECT procedure contained a mixture of items from different domains. For example, the second and third clusters on Form A and the

second cluster on Form B all contained items which were intended to measure the Physical Sciences/Earth and Environmental Science domain, and the fifth cluster on Form A contained items which were all intended to measure the Biological Science/Life Science domain. All of the other empirical clusters contained a mixture of Biological Science/Life Science items and Physical Sciences/Earth and Environmental Science items. Whereas only two domains were specified by the ITED classification tables, the DETECT procedure recovered five clusters on Form A and four clusters on Form B.

Similarly, the ITED classification tables for the Social Studies exams were not perfectly recovered by the exploratory DETECT procedure. The second empirical cluster formed by the DETECT procedure for both Forms A and B primarily contained Economics items, which align with the ITED classification tables. Aside from these two clusters, however, items from each ITED domain were interspersed throughout the empirical clusters. The exploratory DETECT procedure did recover the appropriate number of clusters, though: four clusters were recovered for both Form A and Form B Social Studies exams, just as the ITED classification tables specified that four domains were to be measured.

Standard Error of Equating

The bootstrap standard error of equating (SEE) was computed for each of the equipercentile equating procedures (corresponding to the Math, Science, and Social Studies exams) to estimate how much random error (i.e., error related to the sampling of examinees) may have affected this procedure. The SEE values for the unsmoothed equipercentile equating procedures are presented in Table A-9, and the SEE values for the smoothed equipercentile equating procedures are presented in Table A-10. The last

row in each table presents the average SEE across all raw score points for each respective exam.

Overall, the equipercentile equating procedure for the Math exams appears to contain the least amount of random error. The average SEE for the Math exams were 0.300 and 0.265 (for the unsmoothed and smoothed procedures, respectively), whereas the average SEE's for the Science and Social Studies exams were 0.349 and 0.296 (unsmoothed and smoothed, respectively) and 0.363 and 0.320 (unsmoothed and smoothed, respectively). Given that the score scales are of different lengths, however, comparisons at each raw score point cannot be made. (For example, a raw score of 20 on the Math exam is in the exact middle of the score scale, whereas a raw score of 20 on the Social Studies exam is at the mid-to-lower end of the score scale; equating error is expected to vary at different points along the score scale).

These values imply that—although the equipercentile equating procedure is used as the benchmark for comparison in this study—this procedure is not "error-free." Furthermore, the magnitude of the SEE value at each specific raw score must be taken into account when evaluating the differences between the equipercentile equating procedure and each of the other respective procedures. At scores where the SEE is large, interpretations concerning differences between the equipercentile equating procedure and another procedure must be tempered. At scores where the SEE is smaller, interpretations do not need to be as restricted.

Equating Results

Lastly, equating results for each of the six procedures (equipercentile equating, UIRT observed score equating, unidimensional approximation of MIRT observed score

equating, full MIRT observed score equating, UIRT true score equating, and unidimensional approximation of MIRT true score equating) on each of the three subjects (Math, Science, and Social Studies) are presented in Tables A-11, A-12, and A-13. Each table contains the raw score on Form A in the first column. The Form B equivalents (i.e., the equated scores) appear in the next six columns and are primarily distinguished by procedure type (i.e., observed score versus true score). These tables provide a useful resource for determining the equated scores for each of the equating procedures under each set of exams. From these tables, it can be seen that both unidimensional equating procedures performed similarly and that all three of the multidimensional procedures performed similarly. Also, within each type of equating (i.e., unidimensional and multidimensional), the observed score procedures and the true score procedure performed similarly.

However, based on these tables alone, it is difficult to visually gauge exactly how similarly the unidimensional procedures performed and how similarly the multidimensional procedures performed. Furthermore, it is also difficult to gauge how differently the unidimensional procedures and the multidimensional procedures performed. Therefore, plots containing the differences between the identity equating procedure (i.e., the equating that would result if forms did not differ at all in difficulty) and each of the other six equating procedures appear in Figures A-4 through A-9. Figures A-4 and A-7 correspond to the Math exams, Figures A-5 and A-8 correspond to the Science exams, and Figures A-6 and A-9 correspond to the Social Studies exams. The difference between Figures A-4 through A-6 and Figures A-7 through A-9 pertains to the equipercentile equating procedure. Specifically, Figures A-4, A-5, and A-6 contain

differences between the identity equating procedure and the unsmoothed equipercentile equating procedure, whereas Figures A-7, A-8, and A-9 contain differences between the identity equating procedure and the smoothed equipercentile equating procedure. (Differences between the identity equating procedure and each of the other equating procedures remain the same across both sets of figures). A standard error band—formed by adding and subtracting the standard error of equating at each raw score point—is presented around the equipercentile equating results on each figure.

Figures A-4 through A-9 reveal that although the unidimensional equating procedures and the multidimensional equating procedures performed differently, equating trends were very similar for the two sets of procedures. Specifically, as the unidimensional procedures revealed larger differences between Form A and Form B, the multidimensional procedures also tended to reveal greater differences between Form A and Form B. The magnitude of the equated scores tended to be dissimilar for the unidimensional and the multidimensional procedures, however.

Whereas the aforementioned tables and figures primarily focus on similarities and differences between the unidimensional procedures and the multidimensional procedures, it is also of interest to determine how each set of equating procedures performed in relation to the equipercentile equating procedure. Tables A-14, A-15, and A-16 contain differences between equated scores on each of the five procedures and the unsmoothed equipercentile equating procedure. Differences between equated scores on each of the five procedures and the smoothed equipercentile procedure appear in Tables A-17, A-18, and A-19. These differences are presented at each raw score on Form A. Furthermore, average differences between the equipercentile equating procedure and each of the

respective equating procedures were computed and appear in the last four rows of each table. Specifically, the first of these rows presents the average difference across all raw score points. The absolute value of each difference was also computed and the average of these values appears in the next row. The last two rows present average differences and average absolute differences using a weighted means procedure. Specifically, the relative frequency for each Form A raw score was used as the weight for the difference at each particular score to compute average weighted differences. The conceptual difference between the weighted and the unweighted means procedures is that the weighted means procedure gives more weight to differences where examinees score more frequently and less weight to differences where examinees do not score as frequently; the unweighted means procedure weights each difference equally.

From these tables (A-14 through A-19), it can be seen that the multidimensional equating procedures tended to perform more similarly to the equipercentile equating procedure than the unidimensional equating procedures for both the Math and the Social Studies exams. For the Science exams, the multidimensional equating procedures performed more similarly to the equipercentile equating procedure in an absolute sense, but the unidimensional equating procedures performed more similarly to the equipercentile equating procedure overall (using the unweighted means criterion). For the weighted means criterion, the multidimensional equating procedures performed more similarly to the equipercentile equating procedure for the Science exams. These results imply that for the Science exams, differences between the equipercentile equating procedure and the multidimensional procedures were smaller where examinees scored more frequently; differences between the equipercentile equating procedure and the

unidimensional procedures were smaller where examinees scored less frequently. Furthermore, both the unidimensional procedures and the multidimensional procedures tended to yield lower equated scores than the equipercentile equating procedure (given that most of the differences are negative).

Plots containing the differences between the unsmoothed equipercentile procedure and each of the other five equating procedures appear in Figures A-10, A-11, and A-12 as a means of visually representing the data presented in Tables A-14 through A-19. Plots containing the differences between the smoothed equipercentile procedure and each of the other five equating procedures appear in Figures A-13, A-14, and A-15. Horizontal lines appear at values of -0.5 and 0.5 to elucidate the Difference That Matters (DTM) criterion (Dorans, Holland, Thayer, & Tateneni, 2003). Furthermore, plots containing the absolute values of the differences between the unsmoothed equipercentile procedure and each of the other five equating procedures appear in Figures A-16, A-17, and A-18. Plots containing the absolute values of the differences between the smoothed equipercentile procedure and each of the other five equating procedures appear in Figures A-19, A-20, and A-21. A horizontal line appears at a value of 0.5 to elucidate the Difference That Matters (DTM) criterion. These figures reveal that—as previously mentioned—the multidimensional equating procedures tended to perform more similarly to the equipercentile equating procedure than the unidimensional equating procedures for both the Math and the Social Studies exams. For the Science exams, the multidimensional procedures performed more similarly to the equipercentile equating procedure in an absolute sense, though the unidimensional procedures performed more similarly to the equipercentile equating procedure according to the unweighted mean difference.

Overall Trends

Across all three exams that were equated, the most significant difference among the various equating procedures was not necessarily by procedure type (i.e., observed score versus true score), but rather by psychometric framework (i.e., unidimensional versus multidimensional). These trends are most visible in Figures A-4 through A-9. The UIRT observed score equating procedure and the UIRT true score equating procedure performed quite similarly, and the unidimensional approximation of the MIRT observed score equating procedure, the full MIRT observed score equating procedure, and the unidimensional approximation of the MIRT true score equating procedure performed quite similarly. Although the unidimensional procedures and the multidimensional procedures performed differently, each set of procedures did reveal similar trends across the score scale. For example, as the unidimensional procedures revealed greater differences between Form A and Form B, the multidimensional procedures also tended to reveal greater differences between Form A and Form B (see Figures A-10, A-11, and A-12). Furthermore, although both sets of equating procedures performed somewhat differently from the equipercentile equating procedure at various points along the score scale, the multidimensional procedures tended to perform more similarly to the equipercentile equating procedures than the unidimensional procedures (see Figures A-16 through A-21).

For both the Math and the Social Studies exams, Form B appeared to be more difficult than Form A at all points along the score scale (see Figures A-4, A-6, A-7, and A-9). For the Science exams, Form B appeared to be more difficult than Form A at the upper end of the score scale, whereas Form A appeared to be more difficult than Form B

at the lower end of the score scale (see Figures A-5 and A-8). In general, however, both the unidimensional procedures and the multidimensional procedures tended to reveal greater differences between Form A and Form B (in terms of difficulty) than the equipercentile procedure. Further discussion of the equating results for each set of exams appears below.

Equating of the Math Exams

Equating trends for both the unidimensional procedures and the multidimensional procedures on the Math exams were very similar, though the unidimensional procedures tended to reveal greater differences between Form A and Form B than the multidimensional procedures (Figures A-4 and A-7). Furthermore, the unidimensional procedures revealed greater differences from the equipercentile equating procedure than the multidimensional procedures at nearly every raw score (Figures A-16 and A-19).

For the most part, differences between the unidimensional procedures and the equipercentile equating procedure remained around the 0.5 DTM criterion along most of the score scale (Figures A-16 and A-19). At both the lower end of the score scale (raw scores of 10 and below) and at the upper end of the score scale (raw scores of 30 and above), the unidimensional procedures performed more similarly to the equipercentile procedure than in the middle of the score scale. Furthermore, the equated scores for the unidimensional procedures were almost always less than the equated scores for the equipercentile procedure (Figures A-10 and A-13).

As previously noted, trends for the unidimensional procedures and the multidimensional procedures were very similar for the Math exams. Whereas differences between the unidimensional procedures and the equipercentile equating procedure tended

to remain around the 0.5 DTM criterion, differences between the multidimensional procedures and the equipercentile equating procedure were nearly all less than the 0.5 DTM criterion (Figures A-16 and A-19).

Equating of the Science Exams

Similar to the Math exams, the unidimensional procedures and the multidimensional procedures also revealed similar equating trends for the Science exams (Figures A-5 and A-8). At the region of the score scale where these procedures differed the most (at the upper end of the score scale), the unidimensional procedures revealed greater differences between Form A and Form B (in terms of difficulty) than the multidimensional procedures. Unlike the equating for the Math exams, however, both the unidimensional procedures and the multidimensional procedures yielded higher equated scores than the equipercentile procedure at the lower end of the scale score and at the upper end of the scale score; both sets of procedures yielded lower equated scores than the equipercentile procedure in the middle of the score scale (Figures A-11 and A-14). Furthermore, the unidimensional procedures performed more similarly to the equipercentile equating procedure across all score points according to the unweighted mean difference criterion (Tables A-15 and A-18). In contrast, the multidimensional procedures performed more similarly to the equipercentile equating procedure according to the absolute unweighted mean difference criterion and according to both weighted difference criteria (Tables A-15 and A-18). The unweighted results indicate that although absolute differences between the unidimensional procedures and the equipercentile procedure were larger than the absolute differences between the multidimensional procedures and the equipercentile procedure, the differences between

the unidimensional procedures and the equipercentile procedure were both positive and negative, thus balancing out to yield a lower mean difference. The weighted results indicate that differences between the equipercentile equating procedure and the multidimensional procedures were smaller where examinees scored more frequently, whereas differences between the equipercentile equating procedure and the unidimensional procedures were smaller where examinees scored less frequently.

Both the unidimensional procedures and the multidimensional procedures nearly all exceeded that 0.5 DTM criterion at the lower end of the score scale (between raw scores of 0 and 9) for the Science exams (Figures A-17 and A-20). Within this range, both sets of equating procedures yielded higher equated scores than the equipercentile equating procedure (Figures A-11 and A-14). In the middle of the score scale, both sets of equating procedures tended to yield lower equated scores than the equipercentile equating procedure. Within this range (roughly raw scores of 10 through 40), most differences with the equipercentile equating procedure were less than the 0.5 DTM criterion for both types of procedure; the exception is that the unidimensional procedures tended to exceed this criterion between raw scores of 28 through 37. At the upper end of the score scale—between and including raw scores of 40 through 50—both the unidimensional and the multidimensional procedures yielded higher equated scores than the equipercentile procedure. Within this range, most differences with the equipercentile procedure (for both unidimensional and multidimensional procedures) exceeded the 0.5 DTM criterion.

Equating of the Social Studies Exams

Similar to both the Math and the Science exams, the equating trends for the Social Studies exams were similar for the unidimensional procedures and the multidimensional procedures (Figures A-6 and A-9). Furthermore, the unidimensional procedures tended to reveal greater differences between Form A and Form B (in terms of difficulty) than the multidimensional procedures, which is also similar to the results for the Math and Science exams. Lastly, the multidimensional procedures performed more similarly to the equipercentile equating procedure than the unidimensional procedures on the Social Studies exams (Figures A-18 and A-21).

At nearly all points along the score scale, the differences between the unidimensional procedures and the equipercentile equating procedure exceeded the 0.5 DTM criterion, with the unidimensional procedures nearly always yielding lower equated scores than the equipercentile equating procedure (Figures A-12 and A-15). The multidimensional procedures, on the contrary, tended only to exceed the 0.5 DTM criterion at the upper end of the score scale (between raw scores of 37 through 50). Similar to the unidimensional procedures, the multidimensional procedures tended to yield lower equated scores than the equipercentile equating procedure, with the exception of the range between scores of 18 through 31 (Figures A-12 and A-15).

Summary of Equating Results

Six procedures (equipercentile equating, UIRT observed score equating, unidimensional approximation of MIRT observed score equating, full MIRT observed score equating, UIRT true score equating, and unidimensional approximation of MIRT true score equating) were used to equate Forms A and B of the Math, Science, and Social

Studies Levels 17/18 ITED exams. Differences between the equipercentile equating procedure and each of the five other equating procedures were evaluated according to the Difference That Matters (DTM) criterion (Dorans, Holland, Thayer, & Tateneni, 2003) and the standard error of equating (SEE) values.

The most significant difference among the various equating procedures was not necessarily by procedure type (i.e., observed score versus true score), but rather by psychometric framework (i.e., unidimensional versus multidimensional). The equated scores were very similar for both the UIRT observed score equating procedure and the UIRT true score equating procedure for all three subjects (Math, Science, and Social Studies). Similarly, the equated scores were very similar for the unidimensional approximation of MIRT observed score equating procedure, the full MIRT observed score equating procedure, and the unidimensional approximation of MIRT true score equating procedure for all three subjects. Given that two of the MIRT procedures were formed by approximating a UIRT model and one of the MIRT procedures was based on a full MIRT model, the fact that all three MIRT procedures performed very similarly is especially of interest. Furthermore, the equated scores for the unidimensional procedures and the multidimensional procedures revealed very similar trends, though the magnitude of the equated scores tended to differ by type of equating procedure (the unidimensional procedures tended to yield lower equated scores than the multidimensional procedures).

Across all three subjects (Math, Science, and Social Studies), the multidimensional procedures tended to perform more similarly to the equipercentile equating procedure than the unidimensional procedures. Within the multidimensional procedures, the unidimensional approximation of the MIRT true score equating

procedure performed most similarly to the equipercentile equating procedure. In general, both the unidimensional procedures and the multidimensional procedures tended to yield lower equated scores than the equipercentile equating procedure. The only exceptions are at both ends of the score scale for the Science exams, and at several score points for the Social Studies exams.

Whereas this chapter only provided the results for the data analyses, the next chapter explores these results in more detail and discusses possible explanations as to why these results were obtained. Furthermore, implications of this research, as well as limitations, are discussed.

CHAPTER 5

DISCUSSION AND CONCLUSION

This chapter consists of six main sections. First, results and discussions are presented for each of the four main statistical analyses (the dimensionality assessment procedures, the scale linking procedures, the standard error of equating for the equipercentile equating procedures, and the equating procedures). Next, limitations to this study are addressed. Finally, a summary and conclusion are presented.

Dimensionality Assessment

DETECT (Zhang & Stout, 1999b) was used to assess dimensionality for each form, as well as to observe which clusters of items on each form were the most statistically similar (homogeneous). The maximum DETECT value obtained under the exploratory solutions ranged between 0.223 and 0.300 across all test forms (Tables A-2, A-3, and A-4). Therefore, each form used in this study was classified as "weak to moderately" multidimensional according to the DETECT classification scheme (Zhang & Stout, 1999b). The maximum DETECT values obtained under the confirmatory solution are not as useful as the values obtained under the exploratory solution in this investigation, as the equating procedures were only conducted under an exploratory (i.e., "complex structure") MIRT model.

Given the "weak to moderately" multidimensional structure of each exam, the unidimensional equating procedures are expected to contain more systematic error than the multidimensional equating procedures due to the violation of the unidimensionality assumption. However, the extent to which systematic error is expected to affect these results is restricted. In general, forms that are "strongly" multidimensional would

logically be expected to contain more systematic error, whereas forms that are “weak” or “moderately” multidimensional would logically be expected to contain less systematic error. Since each form used in this investigation was only classified as “weak to moderately” multidimensional, the unidimensional equating results are not expected to contain much systematic error.

The unidimensional procedures and the multidimensional procedures performed as expected, given that each form was classified as “weak to moderately” multidimensional. Specifically, as test forms become increasingly unidimensional, the unidimensional procedures and the multidimensional procedures would logically be expected to perform very similarly. As test forms become increasingly multidimensional, the unidimensional procedures and the multidimensional procedures would logically be expected to perform differently, as the unidimensional procedures are expected to be greatly affected by systematic error. In this investigation, unidimensional procedures and multidimensional procedures did perform differently, yet the equated scores for both sets of procedures revealed very similar trends. This finding would most likely be expected from forms that are “weak to moderately” multidimensional.

As the DETECT procedure classified each form as “weak to moderately” multidimensional, the DETECT procedure also produced either four or five homogeneous clusters of items for each form (Tables A-6, A-7, and A-8). These results are congruent with the Math and Social Studies blueprints, given that the Math and Social Studies exams were intended to measure four distinct domains (though the empirical clusters derived by the DETECT procedure did not perfectly match the ITED classification tables). The Science exams were only specified to measure two domains

(according to the ITED classification tables), despite the fact that five clusters were empirically recovered for Form A and four clusters were empirically recovered for Form B.

As a result, the Science item parameters could have been calibrated under a four- or five-dimensional solution, as opposed to a two-dimensional solution (both the Math and the Social Studies exams were calibrated under a four-dimensional solution, as dictated by the ITED classification tables). In conformity with the ITED classification tables, however, the Science exams were calibrated with respect to two dimensions. The differences that would result from equating the exams under a two-dimensional solution as opposed to a four- or five-dimensional solution are currently unknown. In general, the hypothesis is that after an adequate number of dimensions have been accounted for, equating results under even higher dimensional solutions would most likely be similar. For example, if dimensionality assessment procedures revealed that a two-dimensional solution would adequately account for multidimensionality in the item responses, and the equating procedures were also conducted under a four-dimensional solution, the equating results under the four-dimensional solution would most likely be similar to the equating results under the two-dimensional solution. Equating results under lower dimensional solutions that do not adequately account for multidimensionality in the item responses would most likely be different, presumably due to increases in systematic error (specifically, not specifying an adequate number of dimensions).

This logic is consistent with the results presented in Reckase (2009). Although the author did not investigate equating results under varying number of dimensions, the author demonstrated that—after an adequate number of dimensions have been accounted

for—item parameter estimates and ability estimates would maintain the same *relative* structure under even higher dimensional solutions. For example, consider the situation in which a two-dimensional solution is deemed sufficient to represent the item response pattern. Geometrically, parameter estimates would vary with respect to two reference axes, i.e., the parameter estimates would be geometrically located along a plane. If a three-dimensional solution were obtained for the same dataset, but examinees only varied according to two dimensions, the resulting parameter estimates would reveal differences in examinee abilities along a plane, though this plane would be embedded within a cube (i.e., with respect to three axes). Thus, item parameter estimates and ability estimates would maintain the same *relative* structure under higher order solutions, which might be extended to imply that MIRT equating results should be similar under higher order solutions (after an adequate number of dimensions have been accounted for).

Whereas the DETECT value provides an indication of the degree to which each form is multidimensional, the ratio r value provides an indication of the stability of the solution (i.e., whether a similar solution is likely to result if the procedure is conducted on a different—yet comparable—group of examinees). The ratio r value obtained under the confirmatory solution was small, ranging between 0.149 and 0.227. These values imply that a confirmatory solution that adhered to the ITED classification tables was not very stable. Similarly, under the exploratory solution, the ratio r values ranged between 0.488 and 0.600—which imply that moderate stability was attained.

Given that the ratio r values obtained in this investigation were not very large, perhaps greater stability would be obtained if more examinees were administered each form (recall that in this investigation, 2500 examinees completed each form).

Furthermore, these results indicate that the cluster schemes recovered under the exploratory DETECT procedures were susceptible to random error. If each exam used in this investigation were administered to a different (yet comparable) group of examinees, it is quite possible that the ITED classification scheme would be more closely recovered. Regardless, the clusters recovered by the exploratory DETECT procedure did provide empirical evidence concerning which items pointed in the same direction in multidimensional space for this sample.

Linking Results

As a result of the research conducted in this study, it was discovered that the MIRT scale linking procedures under the random groups design (which only consists of an orthogonal rotation) are not required in order to conduct the MIRT observed score and true score equating procedures. That is, the same equating relationships would result regardless of whether the orthogonal rotation was first incorporated to account for rotational indeterminacy. This will only hold under certain conditions, however. First, item parameters and ability estimates must be calibrated with respect to orthogonal reference axes (i.e., the solution must be orthogonal as opposed oblique). Second, the specified variance-covariance matrix for the ability estimates must be the identity matrix (i.e., each measured trait must be specified as uncorrelated with other measured traits and of unit length). If either of these conditions does not hold, then the scale linking procedures under the random groups design must first be conducted. An explanation as to why the scale linking procedures are not required under these conditions appears below.

To conduct the full MIRT observed score equating procedure, conditional observed score distributions are first determined at each vector of ability level (i.e., $f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the ability vector). These conditional distributions are then multiplied by the multivariate ability density ($\psi(\boldsymbol{\theta})$) and summed over the ability space as,

$$f(x) = \sum_{1}\sum_{2}\cdots\sum_{\delta} f(x|\boldsymbol{\theta})\psi(\boldsymbol{\theta}) \tag{5.1}$$

After marginal distributions are determined for each form using Equation 5.1, the marginal distributions are then equated using traditional equipercentile equating methods.

Under an orthogonal rotation, the MIRT difficulty parameter (d) will remain the same, though the discrimination parameters will change. However, the overall discrimination power for each item—which is related to the geometric length of the item as represented in multidimensional space—will not change since each item is only rotated orthogonally as opposed to being dilated as well. Furthermore, under the orthogonal rotation, the direction of best measurement for each item will be rotated the exact same amount (see Figure 5-1 for a two-item example in two-dimensional space). Given that (1) the MIRT difficulty parameters will not change under an orthogonal rotation and (2) the geometric length of each item remains the same in multidimensional space (i.e., the discrimination parameters are rotated but not dilated), the conditional distribution ($f(x|\boldsymbol{\theta})$) at any set of quadrature points prior to rotation will be identical to the conditional distribution at the *rotated set of quadrature points* after the rotation is applied. This is the result of rotational indeterminacy in the MIRT framework, or specifically,

$$p(X=1|\boldsymbol{\theta},\mathbf{a},d,c) = p(X=1|\boldsymbol{\theta}^*,\mathbf{a}^*,d,c), \tag{5.2a}$$

where

$$\boldsymbol{\theta}^* = \mathbf{T}^{-1}\boldsymbol{\theta} \tag{5.2b}$$

and

$$\mathbf{a}^{*\mathbf{T}} = \mathbf{a}^{\mathbf{T}}\mathbf{T} \tag{5.2c}$$

Furthermore, because the ability density ($\psi(\boldsymbol{\theta})$) is specified to follow a multivariate standard normal distribution with zero correlation between dimensions, all vectors of quadrature points which maintain the same Euclidean distance from the origin of the reference axes also have the same probability density (represented by the circles in Figure 5-1). Therefore, after the orthogonal rotation is applied, the ability density associated with the set of quadrature points prior to rotation will be identical to the ability density associated with the *rotated set of quadrature points* after the rotation is applied.

Given that each conditional distribution is the same prior to rotation and after the rotation is applied (though each conditional distribution now appears at the rotated vector of quadrature points), and given that the ability density corresponding to each conditional distribution is the same prior to rotation and after the rotation is applied, the term $f(x \mid \boldsymbol{\theta})\psi(\boldsymbol{\theta})$ will be identical for the set of quadrature points before the rotation is applied and for the rotated set of quadrature points. Therefore, the marginal distribution for each form—which is computed using Equation 5.1—will be identical for each form. In conclusion, when item parameters are calibrated with respect to orthogonal reference axes, and when the ability covariance matrix is specified as the identity matrix, the full MIRT procedure will perform the same regardless of whether rotational indeterminacy is taken into account.

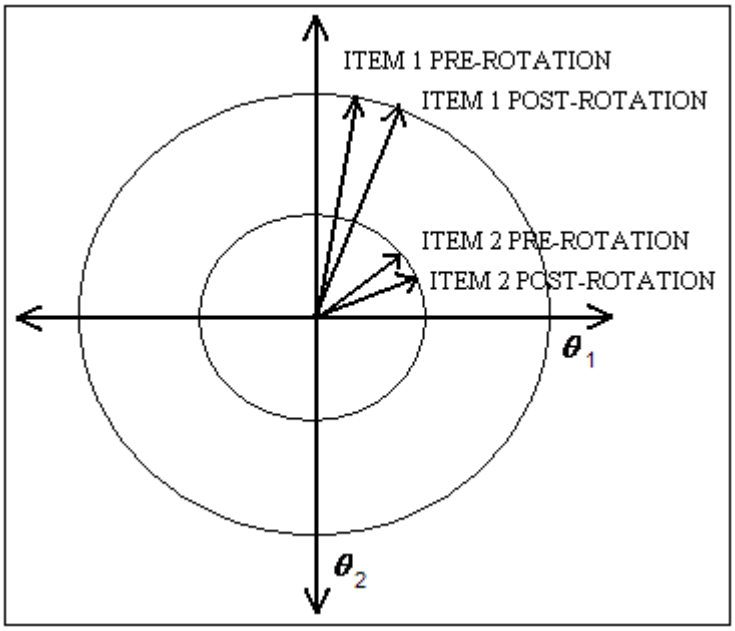


Figure 5-1. Comparison of Pre-Rotation and Post-Rotation Items

Similar logic can be used to explain why MIRT scale linking procedures are not required for the unidimensional approximation of the MIRT observed score equating procedure and the unidimensional approximation of the MIRT true score equating procedure under the random groups design. Recall that to perform these procedures, unidimensional item parameters are first estimated given the multidimensional parameter estimates. To compute unidimensional item parameters, the direction of best measurement is first determined at the test-level using Equation 3.4. Unidimensional discrimination and difficulty parameters are then computed using Equations 3.6a—3.6e.

Under an orthogonal rotation, the direction of best measurement for each item will be rotated by exactly the same angle. Furthermore, the test-level direction of best measurement will be rotated by the same angle as the item-level directions of best measurement. The unidimensional item parameter estimates are primarily governed by the item discrimination vector, the test-level direction of best measurement, and the

covariance matrix of ability estimates. Since both the test-level direction of best measurement and the item discrimination vector are rotated by the same angle, and the covariance matrix of ability estimates is the identity matrix in this situation, the resulting unidimensional item parameters will be the same prior to an orthogonal rotation and after an orthogonal rotation. Thus, both the unidimensional approximation of the MIRT observed score equating procedure and the unidimensional approximation of the MIRT true score equating procedure will yield the same results regardless of whether rotational indeterminacy is taken into account, as the estimated unidimensional item parameters prior to rotation are the same as the estimated unidimensional item parameters after the rotation.

Standard Error of Equating

The standard error of equating (SEE) was computed to determine the extent to which random error might have affected the equipercentile equating procedures. Across all score points, the SEE was smaller (on average) for the Math exams than for the Science and Social Studies exams. For the Math exams, the (smoothed) average SEE was 0.265 across all score points, whereas for the Science and Social Studies exams, the (smoothed) average SEE was 0.296 and 0.320 across all score points, respectively. These values imply that the equipercentile equating results may be less susceptible to random error for the Math equating than for the Science or Social Studies equatings. It should be noted that the SEE values obtained in this study are similar to the SEE values obtained for comparable exams (Kolen & Brennan, 2004).

Standard error bands were constructed around the equipercentile equating results by adding and subtracting one SEE at each raw score point. These bands are plotted

around the equipercentile results in Figures A-4 through A-9. Each band can be interpreted as a 68% confidence interval for the true equipercentile equating relationship at each raw score point, providing an indication of the true equipercentile equating relationship in the population of examinees. This methodology is consistent with the methods typically used in practice (Kolen & Brennan, 2004).

The standard error of equating is quite useful in the interpretation of each of the equating results. For the most part, differences and absolute differences between the equipercentile equating procedure and each of the other equating procedures form the bases by which to evaluate the performance of each equating procedure. However, the equipercentile equating procedure was not "error-free," as indicated by the standard error of equating. At several points along the score scale (for each set of exams), the SEE was near to or exceeds 0.5. Despite the fact that many of these values were near to or greater than 0.5, differences between the equipercentile equating procedure and each of the other equating procedures were still evaluated according to the 0.5 DTM criterion (Dorans, Holland, Thayer, & Tateneni, 2003). Absolute differences that exceeded 0.5 were classified as "significantly different," whereas absolute differences that were smaller than 0.5 were classified as insignificant. The magnitude of each SEE value should guide the interpretation of each difference between the equipercentile equating procedure and the other respective equating procedures: when interpreting the magnitude of each difference, the SEE value at each raw score must also be taken into account. Provided this limitation, perhaps general trends for each of the equating procedures—as opposed to strict differences and absolute differences between each equating procedure and the

equipercentile equating procedure—should be used for investigating the performance of each procedure.

Equating Results

After investigating the performance of each equating procedure, four themes were identified and will be described in detail in the next two sections:

(1) Both unidimensional procedures performed similarly, and all three multidimensional procedures performed similarly.

(2) Unidimensional procedures and multidimensional procedures performed differently, though the pattern of equated scores was similar for both types of procedures.

(3) Multidimensional procedures tended to perform more similar to the equipercentile equating procedure than the unidimensional procedures.

(4) Both the unidimensional procedures and the multidimensional procedures tended to indicate greater differences (in terms of difficulty) between Forms A and B than the equipercentile equating procedure.

Differences Between Unidimensional Procedures and Multidimensional Procedures

The most significant differences between the five equating procedures (unidimensional IRT observed score equating, unidimensional IRT true score equating, full MIRT observed score equating, unidimensional approximation of MIRT observed score equating, and unidimensional approximation of MIRT true score equating) were not necessarily by procedure type (i.e., observed score versus true score), but rather by psychometric framework (i.e., unidimensional versus multidimensional). Both of the unidimensional procedures performed similarly, and each of the three multidimensional

procedures performed similarly. The unidimensional procedures and the multidimensional procedures performed differently, though equating trends were similar for both types of procedures. The fact that UIRT observed score and true score equating procedures performed similarly is consistent with previous research comparing these two methods (Han, Kolen, & Pohlmann, 1997).

The fact that all three MIRT procedures performed very similarly—despite the fact that one of these procedures was based on a full MIRT model and the other two procedures incorporate multidimensional parameter estimates to approximate a UIRT model—is especially noteworthy. It might appear as if the unidimensional approximation methods derived by Zhang and colleagues (Zhang, 1996; Zhang & Stout, 1999a; Zhang & Wang, 1998) performed very well at approximating a MIRT model using unidimensional item parameters. However, this conclusion may be premature. That is, it is currently unknown as to how well these procedures approximate a full MIRT model, or why the full MIRT procedure and the unidimensional approximation procedures performed so similarly.

From a practical perspective, if future research in this area reveals that the full MIRT observed score equating procedure and the unidimensional approximation equating procedures always perform similarly under a specified set of conditions, then the unidimensional approximation methods may be the procedure of choice in practice. The full MIRT observed score equating procedure takes a substantially longer amount of computing time and resources than the unidimensional approximation procedures. However, this should be investigated in more detail in order to understand how each procedure performs under varying conditions.

A second noticeable theme in this study was that although the unidimensional equating procedures and the multidimensional equating procedures performed differently, equating trends were similar for both types of procedures. For example, at points along the score scale where the unidimensional procedures revealed greater discrepancies between Form A and Form B exams (in terms of difficulty), the multidimensional procedures also tended to reveal greater discrepancies between the two forms.

To investigate this occurrence, unidimensional item parameter estimates were compared with the unidimensional approximation item parameter estimates. Recall that the unidimensional approximation procedures incorporate the statistical definition of the "direction of best measurement" used in conjunction with the multidimensional framework in order to estimate unidimensional item parameters for each item. On the other hand, the unidimensional procedures simply estimate unidimensional item parameters regardless of the dimensional structure of the exam. Therefore, in this study, three sets of item parameters were obtained for each item: unidimensional item parameter estimates, multidimensional item parameter estimates, and unidimensional approximation item parameter estimates. Direct comparisons between unidimensional item parameter estimates and multidimensional item parameter estimates are difficult to evaluate, given that the item parameters are used in conjunction with different psychometric models. However, the unidimensional item parameter estimates and the unidimensional approximation item parameter estimates can be directly compared given that both sets of parameters are used in conjunction with the unidimensional logistic model. Table A-20 in Appendix A provides the mean, median, and standard deviation of discrimination parameters and difficulty parameters for both the unidimensional and the unidimensional

approximation procedures. Table A-21 provides correlations between each set of estimated item parameters.

Overall, both sets of item parameter estimates (unidimensional item parameter estimates and unidimensional approximation item parameter estimates) do appear to be very similar. Correlations between unidimensional and unidimensional approximation discrimination parameter estimates ranged between 0.964 and 0.990, and correlations between unidimensional and unidimensional approximation difficulty parameter estimates ranged between 0.997 and 0.999. Although there was a strong linear relationship between both sets of item parameter estimates, the magnitude of the parameter estimates was slightly different. Specifically, unidimensional approximation difficulty parameter estimates tended to be slightly lower than unidimensional difficulty parameter estimates. Differences in these parameter estimates were greater for both the Math and the Social Studies exams than for the Science exams (it should also be noted that the unidimensional procedures and the unidimensional approximation procedures performed more similarly for the Science exams than for the Math and the Social Studies exams). Furthermore, although the mean discrimination parameter estimates were very similar, these values also tended to be slightly lower for the unidimensional approximation procedures than for the unidimensional procedures.

The similarity between both sets of item parameter estimates may help to explain why the unidimensional procedures and the multidimensional procedures yielded similar trends across the score scale. However, the fact that the unidimensional procedures and the unidimensional approximation procedures performed similarly at some locations along the score scale—whereas the two sets of procedures performed differently at other

locations along the score scale—may be explained via a generic discussion of how the dimensional structure of each exam may effect interpretations at various points along the scale.

When more than one trait is measured by an exam, differences between scores at one part of the scale may have an entirely different meaning than differences between scores at another part of the scale. For example, on the Science exams, differences between scores at the lower end of the scale may be primarily due to differences between examinees on the Biological Science/Life Science trait, whereas differences between scores at the upper end of the scale may be primarily due to differences between examinees on the Physical Sciences/Earth and Environmental Science trait. Centroid plots (Reckase, 2009) are often used to determine which trait(s) contribute the most towards differences in scores at various points along the scale.

In this study, there may be points along the scale that are more "unidimensional" than other points along the scale. That is, at some points along the score scale, differences between scores may be the result of differences on only one trait, whereas at other points along the score scale, differences between scores may be the result of differences on more than one trait. Therefore, the unidimensional and multidimensional equating procedures may perform more similarly where differences in scores are primarily due to differences on only one trait. Continuing, the unidimensional and multidimensional equating procedures may perform less similarly where differences in scores are primarily due to differences on more than one trait. However, this explanation remains a hypothesis at this point: no empirical investigations have been conducted in order to examine this phenomenon.

Differences with the Equipercentile Equating Procedure

A third noticeable theme in this study was that the multidimensional procedures performed more similarly to the equipercentile equating procedure than the unidimensional procedures. These results were expected, given that the unidimensional procedures were expected to contain more systematic error than the multidimensional procedures due to the violation of the unidimensionality assumption. Furthermore, it might be expected that the multidimensional observed score procedures would perform more similarly to the equipercentile equating procedure than the multidimensional true score procedure, given that the equipercentile equating procedure is an observed score procedure (as opposed to a true score procedure). On the contrary, the unidimensional approximation of the MIRT true score equating procedure performed more similarly to the equipercentile equating procedure than either of the multidimensional observed score equating procedures. However, the differences between the MIRT observed score equating procedures and the MIRT true score equating procedure were minimal.

The last equating theme identified in this study was that both the unidimensional procedures and the multidimensional procedures tended to reveal greater discrepancies (in terms of difficulty) between Form A and Form B than the equipercentile procedure; between the unidimensional procedures and the multidimensional procedures, the unidimensional procedures tended to reveal greater discrepancies between Form A and Form B.

It is currently unknown as to why both the unidimensional procedures and the multidimensional procedures yielded lower equated scores than the equipercentile equating procedure. As previously noted, the unidimensional equating procedures were

expected to contain more systematic error than the multidimensional equating procedures and the equipercentile equating procedure due to the violation of the unidimensionality assumption, which might help to explain why the unidimensional procedures performed differently than the equipercentile equating procedure. As all equating methods are prone to systematic error, the multidimensional procedures were not expected to contain a large amount of systematic error, given that the multidimensional structure of each exam was accounted for in these procedures. Most likely, the multidimensional procedures did contain *some* systematic error—though not as much as the unidimensional procedures—which caused the multidimensional procedures to yield lower equated scores than the equipercentile equating procedure.

Limitations and Future Studies

There are several limitations of this study that should be addressed. Each limitation serves as a disclaimer that the equating procedures conducted in this study were not performed in a perfectly controlled setting, and therefore the conclusions should be restricted accordingly. Each limitation can be primarily classified in one of two categories: limitations associated with using the equipercentile equating procedure as the benchmark for comparison, and limitations associated with using "real" test data to perform the equating procedures (as opposed to incorporating a simulation study). Each set of limitations is addressed below.

Limitations Associated with using the Equipercentile Equating Procedure as the Benchmark for Comparison

A major limitation of this study is that there was no global criterion to serve as a benchmark for comparison with each of the equating procedures. In this study, the

results obtained for the equipercentile equating procedures served as the benchmark by which the results from the other procedures were evaluated since this method does not explicitly violate any statistical assumptions. However, several limitations result from using the equipercentile equating procedure as the standard for comparison.

First, the equipercentile equating procedure is still subject to random error, which is why the standard error of equating (SEE) was estimated for each equipercentile equating procedure. The fact that random error is still present in the equipercentile equating results implies that—although this procedure might be a good alternative to a criterion since the assumptions associated with this procedure are not expected to be violated in this study—the estimated equating relationship for this sample of examinees is still different than it would be for the population of examinees. An estimate of how much the sample equating relationships might deviate from the population relationships is quantified by the SEE values.

Furthermore, the estimated SEE for each equipercentile equating was large at several raw score points. This result limits the utility of the 0.5 Difference That Matters (DTM) criteria. Whereas the difference between the equipercentile equating procedure and another equating procedure may have exceeded 0.5 at a given raw score—thus resulting in a "significant difference" according to the DTM criteria—if the SEE was large at that score, this may not be a fair interpretation.

Secondly—and perhaps more importantly—each type of equating procedure is not necessarily expected to yield the same equating relationships, despite the fact that one benchmark (the equipercentile equating procedure) was used for comparison with all procedures. Each equating procedure is expected to perform in accordance with how that

procedure is defined. For example, true score equating procedures are defined as relating true scores on both forms to be equated. In practice, this relationship is then applied to the observed scores on each form, although there is no theoretical justification for proceeding in this manner. Observed score equating procedures, on the other hand, estimate a statistical adjustment such that the observed score distributions on both forms to be equated are as similar as possible. Therefore, true score equating procedures and observed score equating procedures are not necessarily expected to perform identically, even under ideal conditions.

Fundamentally, observed score equating procedures should be compared with an observed score equating criterion, and true score equating procedures should be compared with a true score equating criterion. However, even this approach would be difficult to incorporate in a research setting: the equating procedure that is *defined* most similarly to the "true" equating relationship would most likely *perform* most similarly to the "true" equating relationship. For example, consider the situation in which the "true" parameters are known, and the "true" model is specified as a multidimensional IRT model. Then, if the "true" equating relationships were computed using observed score methods, then the observed score equating procedures would be expected to perform most similarly to the true equating relationship. On the other hand, if the "true" equating relationship were computed using true score methods, then the true score equating procedures would be expected to perform most similarly to the true equating relationship. Thus, there should be a different criterion for observed score equating procedures and true score equating procedures.

Limitations of using "Real" Data

The fact that "real" test data were used in this study (as opposed to incorporating a simulation study) limits the interpretations and conclusions that can be generated from this research. First and foremost, the "true" equating relationships for each set of procedures were not known since authentic data were used. These values could only be known by specifying the true parameters via a simulation study. Therefore, the equipercentile equating procedure served as a benchmark for comparison with the other equating procedures.

By incorporating a series of simulation studies—as opposed to demonstrating how these procedures work by using authentic data—several limitations in this study would be addressed. First, the forms used in this study were demonstrated to be "weak to moderately" multidimensional. A series of simulation studies could incorporate varying levels of multidimensionality (i.e., unidimensional, weak multidimensionality, moderate multidimensionality, strong multidimensionality) in order to determine how these procedures perform across varying levels of multidimensionality. Secondly, each equating procedure was only conducted three times in this study (to equate the Math, Science, and Social Studies exams). Therefore, the interpretations yielded in this study were generated from a small sample of equating results. A simulation study would replicate each equating procedure many times (perhaps 100 or 1,000 replications) and therefore interpretations would be based on a much larger sample.

Future research could incorporate a series of simulation studies to investigate how each of these procedures performs under a variety of settings. Furthermore, it may also be of interest to determine how robust each of the MIRT equating procedures is to

violations of the assumptions that were required to perform them (see the Scale Linking and Equating Assumptions section in Chapter 3). This research—in part—would require that the assumptions associated with both the MIRT scale linking procedures and the MIRT equating procedures be explicitly stated. In general, the body of literature that comprises statistical assumptions associated with MIRT scale linking is incomplete.

Oftentimes, discussions pertaining to scale linking within the MIRT framework are written from a purely mathematical perspective. These discussions present methods for determining scale linking coefficients related to translation, dilation, rotation, and correlation indeterminacies, but often the practical assumptions related to the measurement of constructs across diverse populations are neglected. Under the random groups design, this may not be as significant of an issue, as the same constructs are expected to be measured in the populations who were administered each form. Under the nonequivalent groups design, however—where populations of examinees are not assumed to be identical in terms of the traits being measured—this limitation in the MIRT literature may be of much more consequence.

For example, future research in this area should address the extent to which diverse populations under the nonequivalent groups design must be identical on the measured traits in order for the scale linking and equating procedures to be successful. This research might address whether the same constructs must be measured across nonequivalent populations, and the degree to which the relationships between the measured constructs must be identical in each population. Although forms to be equated are typically designed to be the same in terms of content specifications, the fact that populations of examinees are different under the nonequivalent groups design implies

that various dimensions might carry different weight in constituting the total score for each population. For example, some populations might not vary at all on specific measured traits, whereas other populations might vary a significant amount on measured traits. Correlations between measured traits might also be expected to differ under the nonequivalent groups design. In general, the assumptions related to both item parameter and ability estimation procedures—as well as scale linking and equating procedures—must be explored and documented in order for these procedures to be implemented in practice.

Summary and Conclusion

The purposes of this research were to create equating procedures that can be used in conjunction with the MIRT framework and to demonstrate how these procedures were conducted using data from the Iowa Tests of Educational Development (Forsyth, Ansley, Feldt, & Alnot, 2001). Six equating procedures were conducted and evaluated in this study: (1) unidimensional IRT observed score equating, (2) unidimensional IRT true score equating, (3) full MIRT observed score equating, (4) unidimensional approximation of MIRT observed score equating, (5) unidimensional approximation of MIRT true score equating, and (6) equipercentile equating.

Both unidimensional equating procedures performed similarly and all three multidimensional equating procedures performed similarly. This is especially noteworthy provided that two of the multidimensional procedures were formed by approximating a unidimensional IRT model given multidimensional parameter estimates, whereas the other procedure was based on a full MIRT model. The unidimensional procedures and the multidimensional procedures tended to perform differently, though

both sets of procedures did yield similar trends. Also, the multidimensional procedures performed more similarly to the equipercentile procedure than the unidimensional procedures; among the multidimensional procedures, the unidimensional approximation of the MIRT true score equating procedure performed most similarly to the equipercentile procedure. Lastly, both the unidimensional procedures and the multidimensional procedures tended to reveal greater differences between Form A and Form B (in terms of difficulty) than the equipercentile equating procedure.

Although the equipercentile equating procedure was used as the benchmark for comparison in this study, this procedure still provided a less-than-perfect benchmark for comparison. This procedure was selected since the assumptions associated with this procedure are not expected to be violated in this study, and therefore systematic error should be limited under this procedure. However, the equipercentile equating procedure was still subject to random error, as indicated by the standard error of equating (SEE). This implies that differences between the equipercentile equating procedure and each of the other equating procedures must interpreted with caution based on the magnitude of the SEE at each score point.

Furthermore, the observed score and the true score equating procedures were not expected to perform identically even under ideal conditions, as each procedure was defined differently. Observed score procedures provided a statistical adjustment in attempt to make observed score distributions on both forms as similar as possible, whereas true score procedures attempted to link true scores on both forms to be equated. Although no theoretical justification exists for applying the true score equating relationship to observed scores, often this is conducted in practice. Therefore, perhaps a

description of the general trends for each equating procedure proves to be the most useful method for interpreting the equating methods.

In conclusion, it appears as if the multidimensional equating procedures presented in this research may provide an adequate alternative to unidimensional IRT equating when the data are not strictly unidimensional. Provided that the dimensional structure of each form is taken into account by the multidimensional equating procedures, results under these procedures may contain less systematic error than results under the unidimensional equating procedures when the data are not strictly unidimensional. However, the performance of these procedures should be empirically verified via a simulation study before these procedures are fully implemented in practice.

REFERENCES

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*, 311-329.

Ackerman, T. A. (1997). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255-278.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues & Practice*, *22*, 37–53.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D, Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

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

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three advanced placement program exams. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to advanced placement program examinations* (p. 79-118), Research Report 03-27. Princeton, NJ: Educational Testing Service.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigation local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*, 129-151.

Forsyth, R. A., Ansley, T. A., Feldt, L. S., & Alnot, S. D. (2001). *The Iowa Tests of Educational Development. Forms A and B. Levels 15-17/18*. Itasca, IL: Riverside Publishing.

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

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT True- and Observed Score Equatings and Traditional Equipercentile Equating. *Applied Measurement in Education, 10*, 105-121.

Hanson, B. (n.d.). *Equating Error: A program for computing equating error using the bootstrap. (Windows Console Version, Revised by Z. Cui, May 18, 2004)* [Manual]. Unpublished manuscript, College of Education, University of Iowa, Iowa City, Iowa.

Hanson, B., & Zeng, L. (n.d.). *PIE: A computer program for IRT equating. (Windows Console Version, Revised by Z. Cui, May 20, 2004*) [Manual]. Unpublished manuscript, College of Education, University of Iowa, Iowa City, Iowa .

Hattie, J. (1985). Methodology Review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

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

Johnson, R. A. & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (Sixth ed.). Upper Saddle River, New Jersey: Pearson.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. Unpublished doctoral dissertation, Department of Statistics, University of Illinois at Urbana-Champaign.

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

Kolen, M. J., & Wang, T.-Y. (1998, April). *Conditional standard errors of measurement for composite scores using IRT*. Paper presented at the National Council on Measurement in Education, San Diego, CA.

Kolen, M. J., & Wang, T.-Y. (2007). *Conditional standard errors of measurement for composite scores using IRT*. (Unpublished manuscript).

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

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

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics, 7*, 165-174.

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.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139-160.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometrika Monographs, No. 15.*

Miller, T. R. & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education, 5*(3), 193-211.

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

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

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

R Development Core Team (2008). 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. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D., & Martineau, J. (2004). Vertical scaling of science achievement tests. Paper commissioned by the Committee on Test Design for K-12 Science Achievement, Center for Education, National Research Council, Washington, DC, October.

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

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Schonemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika, 31*, 1-16.

Simon, M. K. Comparison of concurrent and separate multidimensional IRT linking of item parameters. Unpublished doctoral dissertation. University of Minnesota.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

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

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55*, 293-326.

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

Wang, M. (1985). Fitting a unidimensional model to multidimensional item response data: the effect of latent space misspecification on the application of IRT (Research Report MW: 6-24-85). University of Iowa, Iowa City, IA

Wang, M. (1986). Fitting a unidimensional model to multidimensional item response data. Paper presented at the Office of Naval Research Contractors Meeting, Gatlinburg, TN.

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

Zeng, L., Kolen, M. J., Hanson, B. A., Cui, Z., & Chien, Y. (2004). RAGE-RGEQUATE [Computer software]. Iowa City: University of Iowa.

Zhang, J. (1996). Some fundamental issues in item response theory with applications. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Department of Statistics.

Zhang, J., & Stout, W. F., (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129-152.

Zhang, J., & Stout, W. F., (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.

Zhang, J., & Wang, M. (1998, April). Relating reported scores to latent traits in a multidimensional test. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R.D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

APPENDIX A. TABLES AND FIGURES

Table A-1. Descriptive Statistics for Each Form

	Math		Science		Social Studies	
	Form A	Form B	Form A	Form B	Form A	Form B
Scale	0-40	0-40	0-48	0-48	0-50	0-50
Guessing	8.0	8.0	9.6	9.6	10.0	10.0
Mean	16.35	14.85	19.27	19.42	22.87	20.31
SD	8.01	7.70	9.65	9.03	9.90	9.33
Minimum	0	2	1	0	3	2
Median	15	13	16	17	21	18
Maximum	40	39	47	47	49	48
KR-20	0.88	0.88	0.90	0.88	0.90	0.89

"Scale" represents the length of the scale for each form (for example, raw scores on the Math forms are between (and include) raw scores of 0 through 40). "Guessing" represents the score that an individual can be expected to obtain if the individual responds randomly or "guesses" on each item.

Table A-2. Confirmatory and Exploratory DETECT Statistics for Math Forms

	Form A		Form B	
	Confirmatory	Exploratory	Confirmatory	Exploratory
DETECT Value	0.1080	0.2755	0.0950	0.2696
IDN Index Value	0.5564	0.6936	0.5782	0.7218
Ratio r	0.2268	0.5783	0.2075	0.5890

Table A-3. Confirmatory and Exploratory DETECT Statistics for Science Forms

	Form A		Form B	
	Confirmatory	Exploratory	Confirmatory	Exploratory
DETECT Value	0.0733	0.2861	0.0735	0.2590
IDN Index Value	0.5550	0.7057	0.5284	0.6622
Ratio r	0.1502	0.5860	0.1500	0.5285

Table A-4. Confirmatory and Exploratory DETECT Statistics for Social Studies Forms

	Form A		Form B	
	Confirmatory	Exploratory	Confirmatory	Exploratory
DETECT Value	0.0745	0.2998	0.0798	0.2227
IDN Index Value	0.5380	0.6955	0.5641	0.6743
Ratio r	0.1491	0.6000	0.1747	0.4878

Table A-5. Pre-rotation and Post-rotation Angles Between Corresponding Reference Composites

	Math		Science		Social Studies	
Reference Composite	Pre-Angle	Post-Angle	Pre-Angle	Post-Angle	Pre-Angle	Post-Angle
1	11.08	57.17	11.85	1.47	18.11	15.66
2	11.45	33.57	22.71	1.47	46.09	6.02
3	14.39	13.78			36.31	9.85
4	29.14	5.12			35.92	31.78

Table A-6. Comparison of Math ITED Classification Tables and DETECT Procedure

	Form A			Form B		
CLUSTER (DOMAIN)	ITED	DETECT	DETECT PATTERN	ITED	DETECT	DETECT PATTERN
Cluster 1 (N)	1 2 3 4 6 14 15 16 17 19 21 23 25 26 28 29 31 32 38	1 2 3 4 5 6 7 8 9 10 14 15 16 17 19	N N N N D N D D D D N N N N N	1 2 8 9 10 12 14 16 17 21 22 25 26 27 28 39	1 5 8 16 17 18 19 21 22 23 24 25 26	N D N N N D D N N A G N N
Cluster 2 (D)	5 7 8 9 10 35 36 37 39 40	18 26 27 30 31 32 33 34 35 36 37 39 40	A N G A N N G A D D D D D	3 4 5 6 7 18 19 33 34 35	2 3 4 6 7 9 10 11 12 13 14 20	N D D D D N N G N G N A
Cluster 3 (G)	11 12 13 22 24 27 33	11 12 13 20	G G G A	11 13 24 30 31 32 36 37 38	15 35 36 38 39 40	A D G G N A
Cluster 4 (A)	18 20 30 34	21 22 23 24 25	N G N G N	15 20 23 29 40	27 28 29 30 31 32 33 34 37	N N A G G G D D G
Cluster 5		28 29 38	N N N			

N = Numbers and Operations on Numbers
D = Data Analysis/Probability/Statistics
G = Geometry/Measurement
A = Algebraic Concepts

Table A-7. Comparison of Science ITED Classification Tables and DETECT Procedure

	Form A			Form B		
CLUSTER (DOMAIN)	ITED	DETECT PATTERN	DETECT	ITED	DETECT	DETECT PATTERN
Cluster 1 (B)	1 5 9 10 19 20 21 22 23 24 25 26 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48	1 2 3 4 5 6 7 9 10 11 12 13 14 16 17 19 20 24 25 26	B P P P B P P B B P P P P P P B B B B B	2 3 4 6 8 9 10 12 13 15 16 17 18 19 20 21 22 23 24 31 32 33 34 35 36 37	1 2 3 4 5 6 7 8 9 10 12 13 14 15 16 17 18 20	P B B B P B P B B B B B P B B B B B
Cluster 2 (P)	2 3 4 6 7 8 11 12 13 14 15 16 17 18 27 28 29 30 31 32 33	8 15 32	P P P	1 5 7 11 14 25 26 27 28 29 30 38 39 40 41 42 43 44 45 46 47 48	11 42 43	P P P
Cluster 3		18	P		19 21 22 23 24 25 26 27 28 29 30	B B B B B P P P P P P
Cluster 4		21 22 23 27 28 29 30 31 33 34	B B B P P P P P P B		31 32 33 34 35 36 37 38 39 40 41 44 45 46 47 48	B B B B B B B P P P P P P P P P
Cluster 5		35 36 37 38 39 40 41 42 43 44 45 46 47 48	B B B B B B B B B B B B B B			

B = Biological Science/Life Science
P = Physical Sciences/Earth and Environmental Science

Table A-8. Comparison of Social Studies ITED Classification Tables and DETECT Procedure

	Form A			Form B		
CLUSTER (DOMAIN)	ITED	DETECT	DETECT PATTERN	ITED	DETECT	DETECT PATTERN
Cluster 1 (H)	2 9 17 18 19 23 24 25	1 2 3 4 5 6 7 8 9 11 12 13 14 15 16 18 34	P H S E S S P P H P P E S P P H S	4 13 14 15 16 17 34 35 39 40	1 2 3 4 6 7 8 9 10 11 12 13 14 15 16 17 23	E P P H P P P S S P S H H H H H E
Cluster 2 (E)	4 10 13 21 30 31 32 40 41 43 44 45 46 48 49 50	40 42 43 44 45 46 47 48 49 50	E P E E E E P E E E	1 5 18 19 20 21 22 23 24 25 36 37 38 41 42 49 50	5 18 19 20 21 22 24 27 32	E E E E E E E P P
Cluster 3 (P)	1 7 8 11 12 15 16 26 27 28 29 35 36 38 42 47	19 20 21 22 23 25 26 27 30 31 32 33 35 36 37	H S E S H H P P E E E S P P S	2 3 6 7 8 11 26 27 28 29 30 31 32 33	25 26 28 29 30 33 34 36 37 39 40	E P P P P P H E E H H
Cluster 4 (S)	3 5 6 14 20 22 33 34 37 39	10 17 24 28 29 38 39 41	E H H P P P S E	9 10 12 43 44 45 46 47 48	31 35 38 41 42 43 44 45 46 47 48 49 50	P H E E E S S S S S S E E

H = History/Geography
E = Economics
P = Political Science
S = Psychology/Sociology/Education/Anthropology

Table A-9. Standard Error of Equating for Unsmoothed Equipercentile Procedure

Score	Math	Science	Social	Score	Math	Science	Social
0	0.0341	0.0394	0.0266	26	0.4896	0.5853	0.4577
1	0.1024	0.1182	0.0797	27	0.6015	0.6149	0.4968
2	0.1707	0.1970	0.1328	28	0.5631	0.6520	0.4902
3	0.2390	0.2759	0.1859	29	0.5266	0.6543	0.5340
4	0.2102	0.3547	0.2390	30	0.5852	0.6207	0.5314
5	0.1769	0.4202	0.2921	31	0.6316	0.5944	0.5535
6	0.1705	0.4262	0.3262	32	0.6387	0.5481	0.6415
7	0.1581	0.3413	0.3501	33	0.5777	0.5103	0.6089
8	0.1493	0.2175	0.2494	34	0.6194	0.5307	0.6138
9	0.1641	0.1871	0.2331	35	0.5309	0.5694	0.6546
10	0.1533	0.1680	0.2160	36	0.4752	0.5755	0.5847
11	0.1572	0.1638	0.1818	37	0.4752	0.5920	0.5203
12	0.1741	0.1751	0.1734	38	0.4691	0.5910	0.6637
13	0.1935	0.1802	0.1812	39	0.2918	0.5403	0.6338
14	0.2266	0.1923	0.1778	40	0.0973	0.6498	0.5572
15	0.3001	0.2194	0.1954	41		0.6730	0.5548
16	0.3047	0.2688	0.2272	42		0.5586	0.5172
17	0.2931	0.3047	0.1968	43		0.4647	0.6290
18	0.2963	0.3465	0.2009	44		0.4446	0.8071
19	0.3646	0.4231	0.2330	45		0.5216	0.6475
20	0.4131	0.4852	0.2905	46		0.7206	0.6376
21	0.3877	0.5529	0.3136	47		0.4697	0.5610
22	0.3999	0.5517	0.3519	48		0.1566	0.3923
23	0.4561	0.5631	0.3937	49			0.2354
24	0.5253	0.5414	0.4058	50			0.0785
25	0.4641	0.4988	0.3925	Mean	0.3001	0.3490	0.3625

Table A-10. Standard Error of Equating for Smoothed Equipercentile Procedure

Score	Math	Science	Social	Score	Math	Science	Social
0	0.0260	0.0256	0.0170	26	0.4300	0.4367	0.4141
1	0.0781	0.0767	0.0509	27	0.4380	0.4434	0.4337
2	0.1301	0.1278	0.0848	28	0.4451	0.4490	0.4508
3	0.1814	0.1789	0.1187	29	0.4513	0.4538	0.4652
4	0.1761	0.2300	0.1526	30	0.4568	0.4580	0.4767
5	0.1585	0.2783	0.1865	31	0.4617	0.4617	0.4854
6	0.1443	0.2860	0.2157	32	0.4658	0.4651	0.4915
7	0.1359	0.2433	0.2208	33	0.4696	0.4683	0.4952
8	0.1334	0.2039	0.2028	34	0.4751	0.4711	0.4968
9	0.1365	0.1740	0.1874	35	0.4864	0.4735	0.4968
10	0.1439	0.1572	0.1762	36	0.5073	0.4752	0.4957
11	0.1553	0.1551	0.1691	37	0.5075	0.4762	0.4940
12	0.1715	0.1650	0.1659	38	0.4218	0.4764	0.4925
13	0.1924	0.1829	0.1664	39	0.2599	0.4760	0.4920
14	0.2165	0.2052	0.1705	40	0.0866	0.4753	0.4935
15	0.2418	0.2300	0.1779	41		0.4749	0.4981
16	0.2668	0.2561	0.1887	42		0.4758	0.5067
17	0.2907	0.2825	0.2028	43		0.4789	0.5199
18	0.3133	0.3084	0.2206	44		0.4863	0.5379
19	0.3340	0.3328	0.2417	45		0.5239	0.5611
20	0.3529	0.3551	0.2656	46		0.6462	0.5891
21	0.3697	0.3750	0.2911	47		0.4489	0.6046
22	0.3848	0.3921	0.3173	48		0.1508	0.4629
23	0.3983	0.4066	0.3433	49			0.2781
24	0.4102	0.4186	0.3684	50			0.0927
25	0.4207	0.4285	0.3922	Mean	0.2652	0.2956	0.3198

Table A-11. Equating Results for Math Exams

		Observed Score			True Score	
Raw Score	Equi	UIRT	Approx	Full MIRT	UIRT	Approx
0	1.625	-0.047	-0.054	-0.039	0.000	0.000
1	1.750	0.873	0.873	0.878	0.924	0.962
2	2.000	1.750	1.772	1.749	1.816	1.839
3	3.269	2.587	2.654	2.604	2.655	2.710
4	3.927	3.397	3.527	3.453	3.458	3.578
5	4.640	4.196	4.399	4.312	4.244	4.443
6	5.412	4.988	5.269	5.183	5.024	5.306
7	6.321	5.778	6.137	6.063	5.807	6.168
8	7.222	6.572	7.004	6.950	6.597	7.030
9	8.201	7.381	7.871	7.837	7.398	7.893
10	9.104	8.203	8.740	8.725	8.212	8.759
11	9.891	9.037	9.611	9.617	9.040	9.628
12	10.613	9.884	10.488	10.511	9.883	10.501
13	11.350	10.746	11.373	11.409	10.742	11.381
14	12.119	11.623	12.263	12.309	11.616	12.267
15	13.019	12.515	13.160	13.217	12.507	13.160
16	14.046	13.423	14.065	14.133	13.413	14.062
17	14.934	14.344	14.978	15.055	14.334	14.973
18	15.719	15.278	15.901	15.980	15.268	15.894
19	16.639	16.222	16.834	16.910	16.214	16.825
20	17.793	17.177	17.778	17.850	17.171	17.768
21	18.782	18.143	18.733	18.800	18.139	18.722
22	19.659	19.118	19.700	19.759	19.117	19.689
23	20.690	20.105	20.679	20.732	20.106	20.668
24	21.886	21.104	21.671	21.713	21.108	21.662
25	22.960	22.117	22.677	22.705	22.125	22.669
26	23.900	23.147	23.698	23.714	23.160	23.692
27	25.056	24.196	24.733	24.738	24.218	24.731
28	26.523	25.266	25.784	25.779	25.302	25.787
29	27.453	26.360	26.851	26.838	26.417	26.860

Table A-11—continued

30	28.453	27.476	27.935	27.914	27.566	27.952
31	29.419	28.630	29.034	29.011	28.750	29.064
32	30.455	29.810	30.150	30.125	29.967	30.195
33	31.411	31.008	31.278	31.258	31.211	31.346
34	32.735	32.209	32.417	32.404	32.470	32.515
35	34.225	33.397	33.569	33.563	33.728	33.701
36	35.643	34.584	34.740	34.738	34.964	34.901
37	36.429	35.788	35.914	35.919	36.159	36.110
38	37.111	36.963	37.084	37.090	37.310	37.321
39	37.875	38.120	38.241	38.250	38.464	38.535
40	38.750	39.285	39.379	39.389	40.000	40.000

Equi = Equipercentile Equating Procedure
UIRT = Unidimensional IRT Procedures (Observed Score or True Score)
Approx = Unidimensional Approximation of MIRT Procedures (Observed Score or True Score)
Full MIRT = Full MIRT Observed Score Procedure

Table A-12. Equating Results for Science Exams

		Observed Score			True Score	
Raw Score	Equi	UIRT	Approx	Full MIRT	UIRT	Approx
0	-0.500	0.248	0.337	0.346	0.000	0.000
1	0.000	1.385	1.476	1.516	1.337	1.376
2	0.500	2.477	2.561	2.622	2.457	2.539
3	1.500	3.546	3.627	3.703	3.538	3.629
4	3.625	4.603	4.676	4.767	4.598	4.684
5	4.813	5.650	5.709	5.823	5.643	5.718
6	5.794	6.689	6.729	6.857	6.680	6.737
7	7.000	7.722	7.739	7.859	7.712	7.745
8	8.044	8.752	8.739	8.826	8.740	8.745
9	9.208	9.776	9.732	9.765	9.767	9.737
10	10.311	10.795	10.719	10.681	10.790	10.724
11	11.526	11.808	11.700	11.592	11.808	11.705
12	12.539	12.812	12.675	12.532	12.818	12.681
13	13.628	13.805	13.645	13.542	13.818	13.652
14	14.606	14.787	14.609	14.596	14.806	14.618
15	15.571	15.755	15.568	15.633	15.779	15.578
16	16.489	16.710	16.520	16.620	16.737	16.533
17	17.642	17.650	17.467	17.568	17.680	17.481
18	18.766	18.576	18.408	18.483	18.606	18.422
19	19.793	19.488	19.343	19.378	19.517	19.357
20	20.758	20.388	20.270	20.274	20.414	20.285
21	21.635	21.275	21.192	21.172	21.297	21.206
22	22.687	22.151	22.107	22.076	22.168	22.121
23	23.534	23.018	23.016	22.988	23.029	23.029
24	24.514	23.876	23.921	23.903	23.881	23.931
25	25.214	24.728	24.820	24.811	24.725	24.828
26	26.220	25.574	25.715	25.704	25.564	25.720
27	27.061	26.418	26.607	26.592	26.400	26.608

Table A-12—continued

28	27.838	27.262	27.497	27.483	27.234	27.493
29	28.780	28.106	28.386	28.365	28.069	28.377
30	29.690	28.952	29.275	29.271	28.906	29.260
31	30.476	29.802	30.165	30.196	29.747	30.144
32	31.239	30.659	31.057	31.136	30.596	31.030
33	31.969	31.522	31.954	32.079	31.454	31.920
34	32.737	32.401	32.856	33.006	32.324	32.816
35	33.757	33.295	33.765	33.898	33.209	33.719
36	34.603	34.204	34.684	34.751	34.112	34.631
37	35.294	35.133	35.614	35.608	35.036	35.556
38	36.074	36.083	36.557	36.514	35.984	36.497
39	36.833	37.058	37.517	37.463	36.961	37.455
40	37.735	38.061	38.497	38.449	37.972	38.437
41	39.167	39.095	39.501	39.461	39.023	39.447
42	40.175	40.160	40.534	40.486	40.120	40.492
43	40.958	41.257	41.601	41.538	41.272	41.582
44	41.875	42.381	42.709	42.636	42.493	42.729
45	42.808	43.531	43.862	43.793	43.806	43.958
46	43.423	44.809	45.056	45.014	45.264	45.311
47	45.625	46.121	46.268	46.278	47.100	46.914
48	48.000	47.397	47.440	47.429	48.000	48.000

Equi = Equipercentile Equating Procedure
UIRT = Unidimensional IRT Procedures (Observed Score or True Score)
Approx = Unidimensional Approximation of MIRT Procedures (Observed Score or True Score)
Full MIRT = Full MIRT Observed Score Procedure

Table A-13. Equating Results for Social Studies Exams

		Observed Score			True Score	
Raw Score	Equi	UIRT	Approx	Full MIRT	UIRT	Approx
0	0.000	-0.192	0.009	0.004	0.000	0.000
1	1.000	0.604	1.012	1.022	0.752	1.034
2	1.500	1.358	1.993	2.018	1.536	2.085
3	3.000	2.114	2.941	2.968	2.283	3.050
4	4.800	2.868	3.859	3.878	3.013	3.967
5	5.050	3.614	4.752	4.759	3.738	4.854
6	5.350	4.360	5.626	5.625	4.466	5.720
7	6.500	5.112	6.485	6.479	5.199	6.571
8	7.444	5.869	7.335	7.326	5.942	7.413
9	8.272	6.634	8.177	8.169	6.696	8.247
10	9.202	7.411	9.014	9.008	7.462	9.076
11	9.976	8.201	9.847	9.849	8.241	9.902
12	10.708	9.002	10.677	10.690	9.034	10.726
13	11.563	9.816	11.506	11.533	9.840	11.550
14	12.449	10.641	12.338	12.378	10.660	12.375
15	13.264	11.480	13.171	13.227	11.494	13.201
16	14.194	12.334	14.006	14.080	12.341	14.030
17	14.918	13.200	14.844	14.936	13.203	14.863
18	15.526	14.079	15.685	15.794	14.078	15.700
19	16.235	14.971	16.532	16.654	14.966	16.543
20	16.979	15.875	17.385	17.513	15.867	17.391
21	17.825	16.791	18.245	18.378	16.781	18.246
22	18.725	17.718	19.111	19.251	17.707	19.108
23	19.489	18.657	19.984	20.128	18.645	19.978
24	20.434	19.606	20.865	21.010	19.595	20.856
25	21.250	20.566	21.753	21.901	20.555	21.743
26	22.162	21.535	22.651	22.800	21.526	22.638
27	23.186	22.513	23.558	23.707	22.506	23.543

Table A-13—continued

28	24.090	23.499	24.474	24.624	23.495	24.458
29	25.075	24.493	25.399	25.552	24.492	25.383
30	26.009	25.495	26.335	26.489	25.497	26.318
31	27.000	26.503	27.281	27.430	26.508	27.264
32	28.333	27.518	28.238	28.378	27.527	28.220
33	29.500	28.539	29.205	29.332	28.552	29.188
34	30.722	29.566	30.184	30.296	29.582	30.168
35	31.710	30.600	31.176	31.268	30.619	31.161
36	32.771	31.640	32.180	32.247	31.664	32.166
37	33.634	32.689	33.198	33.235	32.716	33.186
38	34.769	33.747	34.231	34.235	33.777	34.222
39	36.339	34.817	35.281	35.250	34.851	35.276
40	37.473	35.901	36.349	36.281	35.940	36.349
41	38.548	37.003	37.438	37.332	37.051	37.446
42	39.387	38.127	38.553	38.406	38.190	38.572
43	40.423	39.277	39.702	39.514	39.369	39.733
44	42.100	40.457	40.887	40.671	40.604	40.939
45	43.471	41.708	42.113	41.881	41.922	42.206
46	44.667	43.022	43.380	43.151	43.365	43.558
47	45.857	44.370	44.727	44.489	45.022	45.040
48	47.063	45.840	46.146	45.991	47.068	46.742
49	47.500	47.324	47.552	47.487	49.415	48.813
50	50.000	48.900	49.131	49.200	50.000	50.000

Equi = Equipercentile Equating Procedure
UIRT = Unidimensional IRT Procedures (Observed Score or True Score)
Approx = Unidimensional Approximation of MIRT Procedures (Observed Score or True Score)
Full MIRT = Full MIRT Observed Score Procedure

Table A-14. Differences Between Equating Results and Unsmoothed Equipercentile Results for Math Exams

	Observed Score			True Score	
Raw Score	UIRT	Approx	Full MIRT	UIRT	Approx
0	-1.672	-1.679	-1.664	-1.625	-1.625
1	-0.877	-0.877	-0.872	-0.826	-0.788
2	-0.250	-0.228	-0.251	-0.184	-0.161
3	-0.682	-0.615	-0.665	-0.614	-0.559
4	-0.530	-0.400	-0.474	-0.469	-0.349
5	-0.444	-0.241	-0.328	-0.396	-0.197
6	-0.424	-0.143	-0.229	-0.388	-0.106
7	-0.543	-0.184	-0.258	-0.514	-0.153
8	-0.650	-0.218	-0.272	-0.625	-0.192
9	-0.820	-0.330	-0.364	-0.803	-0.308
10	-0.901	-0.364	-0.379	-0.892	-0.346
11	-0.854	-0.280	-0.274	-0.851	-0.263
12	-0.729	-0.125	-0.102	-0.730	-0.112
13	-0.604	0.023	0.059	-0.609	0.031
14	-0.496	0.144	0.190	-0.503	0.148
15	-0.504	0.141	0.198	-0.512	0.141
16	-0.623	0.019	0.087	-0.633	0.016
17	-0.590	0.044	0.121	-0.600	0.039
18	-0.442	0.182	0.261	-0.452	0.175
19	-0.417	0.195	0.271	-0.425	0.186
20	-0.616	-0.015	0.057	-0.622	-0.025
21	-0.639	-0.049	0.018	-0.643	-0.060
22	-0.541	0.041	0.100	-0.542	0.030
23	-0.585	-0.011	0.042	-0.584	-0.022
24	-0.782	-0.215	-0.173	-0.778	-0.225
25	-0.843	-0.283	-0.255	-0.835	-0.291
26	-0.753	-0.202	-0.186	-0.740	-0.208
27	-0.860	-0.323	-0.318	-0.839	-0.325

Table A-14—continued

28	-1.257	-0.739	-0.744	-1.221	-0.737
29	-1.094	-0.602	-0.615	-1.037	-0.593
30	-0.977	-0.518	-0.539	-0.888	-0.501
31	-0.790	-0.385	-0.408	-0.669	-0.355
32	-0.645	-0.305	-0.330	-0.488	-0.260
33	-0.403	-0.133	-0.153	-0.200	-0.065
34	-0.526	-0.318	-0.331	-0.265	-0.220
35	-0.828	-0.656	-0.662	-0.497	-0.524
36	-1.059	-0.903	-0.905	-0.679	-0.742
37	-0.641	-0.515	-0.510	-0.270	-0.319
38	-0.148	-0.027	-0.021	0.199	0.210
39	0.245	0.366	0.375	0.589	0.660
40	0.535	0.629	0.639	1.250	1.250
Mean	-0.640	-0.246	-0.241	-0.547	-0.189
Abs Mean	0.678	0.333	0.359	0.646	0.330
Wt. Mean	-0.669	-0.150	-0.136	-0.643	-0.134
Wt. Abs	0.671	0.215	0.247	0.649	0.202

"Mean" represents the unweighted mean difference across all score points
"Abs Mean" represents the unweighted mean absolute difference across all score points
"Wt. Mean" represents the weighted mean difference across all score points
"Wt. Abs" represents the weighted mean absolute difference across all score points

Table A-15. Differences Between Equating Results and Unsmoothed Equipercentile Results for Science Exams

	Observed Score			True Score	
Raw Score	UIRT	Approx	Full MIRT	UIRT	Approx
0	0.748	0.837	0.846	0.500	0.500
1	1.385	1.476	1.516	1.337	1.376
2	1.977	2.061	2.122	1.957	2.039
3	2.046	2.127	2.203	2.038	2.129
4	0.978	1.051	1.142	0.973	1.059
5	0.837	0.896	1.010	0.830	0.905
6	0.895	0.935	1.063	0.886	0.943
7	0.722	0.739	0.859	0.712	0.745
8	0.708	0.695	0.782	0.696	0.701
9	0.568	0.524	0.557	0.559	0.529
10	0.484	0.408	0.370	0.479	0.413
11	0.282	0.174	0.066	0.282	0.179
12	0.273	0.136	-0.007	0.279	0.142
13	0.177	0.017	-0.086	0.190	0.024
14	0.181	0.003	-0.010	0.200	0.012
15	0.184	-0.003	0.062	0.208	0.007
16	0.221	0.031	0.131	0.248	0.044
17	0.008	-0.175	-0.074	0.038	-0.161
18	-0.190	-0.358	-0.283	-0.160	-0.344
19	-0.305	-0.451	-0.415	-0.276	-0.436
20	-0.371	-0.488	-0.484	-0.344	-0.473
21	-0.360	-0.443	-0.463	-0.338	-0.429
22	-0.536	-0.580	-0.611	-0.519	-0.566
23	-0.516	-0.518	-0.546	-0.505	-0.505
24	-0.638	-0.594	-0.611	-0.633	-0.583
25	-0.486	-0.394	-0.403	-0.489	-0.387
26	-0.646	-0.505	-0.516	-0.656	-0.501
27	-0.643	-0.454	-0.469	-0.661	-0.453

Table A-15—continued

28	-0.576	-0.341	-0.355	-0.604	-0.345
29	-0.674	-0.394	-0.415	-0.712	-0.403
30	-0.738	-0.415	-0.419	-0.785	-0.430
31	-0.674	-0.311	-0.280	-0.729	-0.332
32	-0.581	-0.182	-0.103	-0.643	-0.209
33	-0.447	-0.015	0.110	-0.515	-0.049
34	-0.336	0.119	0.269	-0.413	0.079
35	-0.462	0.008	0.141	-0.548	-0.038
36	-0.399	0.081	0.148	-0.491	0.028
37	-0.161	0.320	0.314	-0.259	0.262
38	0.009	0.483	0.440	-0.090	0.423
39	0.225	0.684	0.630	0.128	0.622
40	0.326	0.762	0.714	0.237	0.702
41	-0.072	0.334	0.294	-0.144	0.280
42	-0.015	0.359	0.311	-0.055	0.317
43	0.299	0.643	0.580	0.314	0.624
44	0.506	0.834	0.761	0.618	0.854
45	0.723	1.054	0.985	0.998	1.150
46	1.386	1.633	1.591	1.841	1.888
47	0.496	0.643	0.653	1.475	1.289
48	-0.603	-0.561	-0.571	0.000	0.000
Mean	0.127	0.263	0.277	0.152	0.278
Abs Mean	0.552	0.556	0.567	0.584	0.549
Wt. Mean	0.024	0.020	0.018	0.028	0.024
Wt. Abs	0.356	0.295	0.290	0.373	0.296

"Mean" represents the unweighted mean difference across all score points
"Abs Mean" represents the unweighted mean absolute difference across all score points
"Wt. Mean" represents the weighted mean difference across all score points
"Wt. Abs" represents the weighted mean absolute difference across all score points

Table A-16. Differences Between Equating Results and Unsmoothed Equipercentile Results for Social Studies Exams

	Observed Score			True Score	
Raw Score	UIRT	Approx	Full MIRT	UIRT	Approx
0	-0.192	0.009	0.004	0.000	0.000
1	-0.396	0.012	0.022	-0.248	0.034
2	-0.142	0.493	0.518	0.036	0.585
3	-0.886	-0.059	-0.032	-0.717	0.050
4	-1.932	-0.941	-0.922	-1.787	-0.833
5	-1.436	-0.298	-0.291	-1.312	-0.196
6	-0.990	0.276	0.275	-0.884	0.370
7	-1.388	-0.015	-0.021	-1.301	0.071
8	-1.575	-0.110	-0.118	-1.502	-0.031
9	-1.638	-0.095	-0.103	-1.576	-0.025
10	-1.791	-0.188	-0.194	-1.740	-0.126
11	-1.775	-0.129	-0.127	-1.735	-0.074
12	-1.706	-0.031	-0.018	-1.674	0.018
13	-1.748	-0.057	-0.030	-1.723	-0.013
14	-1.808	-0.111	-0.071	-1.789	-0.075
15	-1.784	-0.093	-0.037	-1.770	-0.063
16	-1.860	-0.188	-0.114	-1.853	-0.164
17	-1.718	-0.074	0.018	-1.715	-0.055
18	-1.447	0.159	0.268	-1.449	0.174
19	-1.264	0.297	0.419	-1.269	0.308
20	-1.104	0.406	0.534	-1.112	0.412
21	-1.035	0.420	0.553	-1.044	0.421
22	-1.007	0.386	0.526	-1.018	0.383
23	-0.832	0.495	0.639	-0.844	0.489
24	-0.828	0.431	0.576	-0.839	0.422
25	-0.685	0.503	0.651	-0.695	0.493
26	-0.627	0.489	0.638	-0.636	0.476
27	-0.673	0.372	0.521	-0.680	0.357

Table A-16—continued

28	-0.591	0.384	0.534	-0.595	0.368
29	-0.582	0.324	0.477	-0.583	0.308
30	-0.514	0.326	0.480	-0.512	0.309
31	-0.497	0.281	0.430	-0.492	0.264
32	-0.815	-0.095	0.045	-0.806	-0.113
33	-0.961	-0.295	-0.168	-0.949	-0.312
34	-1.156	-0.538	-0.426	-1.140	-0.554
35	-1.110	-0.534	-0.442	-1.091	-0.549
36	-1.131	-0.591	-0.524	-1.108	-0.605
37	-0.945	-0.436	-0.399	-0.919	-0.448
38	-1.022	-0.538	-0.534	-0.992	-0.547
39	-1.522	-1.058	-1.089	-1.488	-1.064
40	-1.572	-1.124	-1.192	-1.533	-1.124
41	-1.545	-1.110	-1.216	-1.497	-1.102
42	-1.260	-0.834	-0.981	-1.197	-0.815
43	-1.146	-0.721	-0.909	-1.054	-0.690
44	-1.643	-1.213	-1.429	-1.496	-1.161
45	-1.763	-1.358	-1.590	-1.550	-1.265
46	-1.646	-1.287	-1.516	-1.302	-1.109
47	-1.487	-1.130	-1.368	-0.836	-0.817
48	-1.224	-0.917	-1.072	0.005	-0.321
49	-0.176	0.052	-0.013	1.915	1.313
50	-1.101	-0.869	-0.800	0.000	0.000
Mean	-1.170	-0.214	-0.189	-1.021	-0.130
Abs Mean	1.170	0.454	0.507	1.098	0.429
Wt. Mean	-1.280	-0.033	0.036	-1.254	-0.014
Wt. Abs	1.280	0.341	0.389	1.260	0.325

"Mean" represents the unweighted mean difference across all score points
"Abs Mean" represents the unweighted mean absolute difference across all score points
"Wt. Mean" represents the weighted mean difference across all score points
"Wt. Abs" represents the weighted mean absolute difference across all score points

Table A-17. Differences Between Equating Results and Smoothed Equipercentile Results for Math Exams

	Observed Score			True Score	
Raw Score	UIRT	Approx	Full MIRT	UIRT	Approx
0	-0.033	-0.040	-0.025	0.014	0.014
1	-0.086	-0.086	-0.081	-0.035	0.003
2	-0.182	-0.160	-0.183	-0.117	-0.093
3	-0.318	-0.251	-0.301	-0.250	-0.195
4	-0.390	-0.260	-0.334	-0.328	-0.209
5	-0.440	-0.237	-0.324	-0.392	-0.193
6	-0.497	-0.216	-0.302	-0.460	-0.179
7	-0.560	-0.201	-0.275	-0.532	-0.171
8	-0.624	-0.192	-0.246	-0.599	-0.166
9	-0.672	-0.182	-0.216	-0.655	-0.160
10	-0.705	-0.168	-0.183	-0.697	-0.150
11	-0.723	-0.149	-0.143	-0.720	-0.132
12	-0.727	-0.123	-0.100	-0.728	-0.109
13	-0.720	-0.094	-0.057	-0.724	-0.085
14	-0.706	-0.066	-0.020	-0.713	-0.063
15	-0.690	-0.045	0.012	-0.698	-0.045
16	-0.673	-0.031	0.037	-0.683	-0.034
17	-0.658	-0.024	0.053	-0.669	-0.030
18	-0.649	-0.025	0.054	-0.658	-0.032
19	-0.646	-0.034	0.042	-0.654	-0.043
20	-0.652	-0.051	0.021	-0.658	-0.061
21	-0.665	-0.075	-0.008	-0.669	-0.086
22	-0.688	-0.106	-0.047	-0.689	-0.117
23	-0.716	-0.142	-0.089	-0.715	-0.153
24	-0.750	-0.183	-0.141	-0.746	-0.193
25	-0.785	-0.225	-0.197	-0.777	-0.233
26	-0.817	-0.266	-0.250	-0.804	-0.272
27	-0.842	-0.305	-0.300	-0.820	-0.307
28	-0.856	-0.338	-0.343	-0.820	-0.335
29	-0.855	-0.363	-0.376	-0.798	-0.354
30	-0.839	-0.380	-0.401	-0.750	-0.363
31	-0.796	-0.391	-0.414	-0.676	-0.362
32	-0.736	-0.396	-0.421	-0.579	-0.351
33	-0.670	-0.400	-0.420	-0.467	-0.333
34	-0.615	-0.407	-0.420	-0.353	-0.308

Table A-17—continued

35	-0.583	-0.411	-0.417	-0.251	-0.278
36	-0.557	-0.401	-0.403	-0.177	-0.240
37	-0.491	-0.365	-0.360	-0.120	-0.169
38	-0.486	-0.365	-0.359	-0.140	-0.128
39	-0.548	-0.427	-0.418	-0.205	-0.134
40	-0.604	-0.510	-0.500	0.111	0.111
Mean	-0.616	-0.222	-0.216	-0.522	-0.164
Abs Mean	0.616	0.222	0.227	0.528	0.171
Wt. Mean	-0.678	-0.159	-0.146	-0.653	-0.144
Wt. Abs	0.678	0.159	0.162	0.653	0.144

"Mean" represents the unweighted mean difference across all score points
"Abs Mean" represents the unweighted mean absolute difference across all score points
"Wt. Mean" represents the weighted mean difference across all score points
"Wt. Abs" represents the weighted mean absolute difference across all score points

Table A-18. Differences Between Equating Results and Smoothed Equipercentile Results for Science Exams

	Observed Score			True Score	
Raw Score	UIRT	Approx	Full MIRT	UIRT	Approx
0	0.244	0.333	0.342	-0.004	-0.004
1	0.373	0.464	0.504	0.325	0.364
2	0.457	0.541	0.602	0.437	0.519
3	0.518	0.599	0.675	0.510	0.601
4	0.567	0.640	0.731	0.562	0.648
5	0.607	0.666	0.780	0.600	0.675
6	0.613	0.653	0.781	0.604	0.661
7	0.577	0.594	0.714	0.567	0.600
8	0.537	0.524	0.611	0.525	0.529
9	0.492	0.448	0.481	0.483	0.453
10	0.446	0.370	0.332	0.441	0.375
11	0.401	0.293	0.185	0.401	0.297
12	0.355	0.218	0.075	0.361	0.224
13	0.308	0.148	0.045	0.321	0.155
14	0.261	0.083	0.070	0.279	0.091
15	0.210	0.023	0.088	0.234	0.033
16	0.158	-0.032	0.068	0.185	-0.019
17	0.102	-0.081	0.020	0.131	-0.067
18	0.043	-0.125	-0.050	0.073	-0.110
19	-0.017	-0.163	-0.127	0.012	-0.147
20	-0.077	-0.194	-0.190	-0.051	-0.179
21	-0.137	-0.220	-0.240	-0.115	-0.205
22	-0.197	-0.241	-0.272	-0.179	-0.227
23	-0.254	-0.256	-0.284	-0.243	-0.243
24	-0.311	-0.267	-0.284	-0.306	-0.256
25	-0.363	-0.271	-0.280	-0.366	-0.264
26	-0.414	-0.273	-0.284	-0.423	-0.268
27	-0.459	-0.270	-0.285	-0.477	-0.269
28	-0.497	-0.262	-0.276	-0.525	-0.266
29	-0.530	-0.250	-0.271	-0.567	-0.259
30	-0.557	-0.234	-0.238	-0.603	-0.248
31	-0.576	-0.213	-0.182	-0.631	-0.234
32	-0.587	-0.188	-0.109	-0.649	-0.215
33	-0.589	-0.157	-0.032	-0.657	-0.191
34	-0.576	-0.121	0.029	-0.653	-0.162

Table A-18—continued

35	-0.549	-0.079	0.054	-0.635	-0.126
36	-0.509	-0.029	0.038	-0.602	-0.082
37	-0.452	0.029	0.023	-0.549	-0.029
38	-0.377	0.097	0.054	-0.476	0.037
39	-0.281	0.178	0.124	-0.377	0.117
40	-0.160	0.276	0.228	-0.249	0.216
41	-0.012	0.394	0.354	-0.084	0.340
42	0.164	0.538	0.490	0.124	0.496
43	0.370	0.714	0.651	0.386	0.695
44	0.602	0.930	0.857	0.714	0.950
45	0.830	1.161	1.092	1.105	1.258
46	0.839	1.086	1.044	1.293	1.341
47	0.381	0.528	0.538	1.360	1.174
48	-0.182	-0.140	-0.150	0.421	0.421
Mean	0.037	0.173	0.186	0.062	0.188
Abs Mean	0.390	0.339	0.331	0.446	0.354
Wt. Mean	0.081	0.076	0.075	0.085	0.080
Wt. Abs	0.323	0.227	0.199	0.347	0.230

"Mean" represents the unweighted mean difference across all score points
"Abs Mean" represents the unweighted mean absolute difference across all score points
"Wt. Mean" represents the weighted mean difference across all score points
"Wt. Abs" represents the weighted mean absolute difference across all score points

Table A-19. Differences Between Equating Results and Smoothed Equipercentile Results for Social Studies Exams

	Observed Score			True Score	
Raw Score	UIRT	Approx	Full MIRT	UIRT	Approx
0	-0.166	0.035	0.030	0.026	0.026
1	-0.317	0.091	0.101	-0.169	0.113
2	-0.511	0.124	0.149	-0.333	0.216
3	-0.702	0.125	0.152	-0.534	0.233
4	-0.896	0.095	0.114	-0.751	0.203
5	-1.097	0.041	0.048	-0.973	0.142
6	-1.296	-0.030	-0.031	-1.191	0.063
7	-1.451	-0.078	-0.084	-1.364	0.008
8	-1.527	-0.062	-0.070	-1.454	0.017
9	-1.590	-0.047	-0.055	-1.528	0.022
10	-1.639	-0.036	-0.042	-1.588	0.025
11	-1.672	-0.026	-0.024	-1.632	0.028
12	-1.691	-0.016	-0.003	-1.660	0.033
13	-1.695	-0.004	0.023	-1.670	0.040
14	-1.683	0.014	0.054	-1.664	0.050
15	-1.656	0.035	0.091	-1.642	0.065
16	-1.611	0.061	0.135	-1.604	0.085
17	-1.554	0.090	0.182	-1.552	0.109
18	-1.486	0.120	0.229	-1.487	0.135
19	-1.410	0.151	0.273	-1.415	0.162
20	-1.329	0.181	0.309	-1.337	0.187
21	-1.248	0.207	0.340	-1.258	0.208
22	-1.168	0.225	0.365	-1.178	0.223
23	-1.090	0.237	0.381	-1.102	0.231
24	-1.019	0.240	0.385	-1.030	0.231
25	-0.955	0.233	0.381	-0.965	0.222
26	-0.899	0.217	0.366	-0.908	0.204
27	-0.855	0.190	0.339	-0.862	0.176
28	-0.822	0.153	0.303	-0.827	0.137
29	-0.803	0.103	0.256	-0.804	0.087
30	-0.796	0.044	0.198	-0.794	0.027
31	-0.804	-0.026	0.123	-0.798	-0.043
32	-0.824	-0.104	0.036	-0.815	-0.122
33	-0.857	-0.191	-0.064	-0.845	-0.208
34	-0.902	-0.284	-0.172	-0.885	-0.300

Table A-19—continued

35	-0.955	-0.379	-0.287	-0.935	-0.394
36	-1.016	-0.476	-0.409	-0.993	-0.490
37	-1.082	-0.573	-0.536	-1.056	-0.585
38	-1.152	-0.668	-0.664	-1.122	-0.677
39	-1.221	-0.757	-0.788	-1.188	-0.763
40	-1.286	-0.838	-0.906	-1.247	-0.838
41	-1.342	-0.907	-1.013	-1.294	-0.899
42	-1.385	-0.959	-1.106	-1.321	-0.939
43	-1.409	-0.984	-1.172	-1.317	-0.953
44	-1.411	-0.981	-1.197	-1.264	-0.929
45	-1.348	-0.943	-1.175	-1.134	-0.850
46	-1.225	-0.866	-1.095	-0.881	-0.688
47	-1.096	-0.739	-0.977	-0.444	-0.426
48	-1.044	-0.737	-0.892	0.186	-0.141
49	-1.005	-0.777	-0.842	1.086	0.483
50	-0.877	-0.645	-0.576	0.224	0.224
Mean	-1.154	-0.199	-0.173	-1.006	-0.114
Abs Mean	1.154	0.317	0.383	1.065	0.287
Wt. Mean	-1.299	-0.052	0.017	-1.273	-0.033
Wt. Abs	1.299	0.211	0.282	1.277	0.212

"Mean" represents the unweighted mean difference across all score points
"Abs Mean" represents the unweighted mean absolute difference across all score points
"Wt. Mean" represents the weighted mean difference across all score points
"Wt. Abs" represents the weighted mean absolute difference across all score points

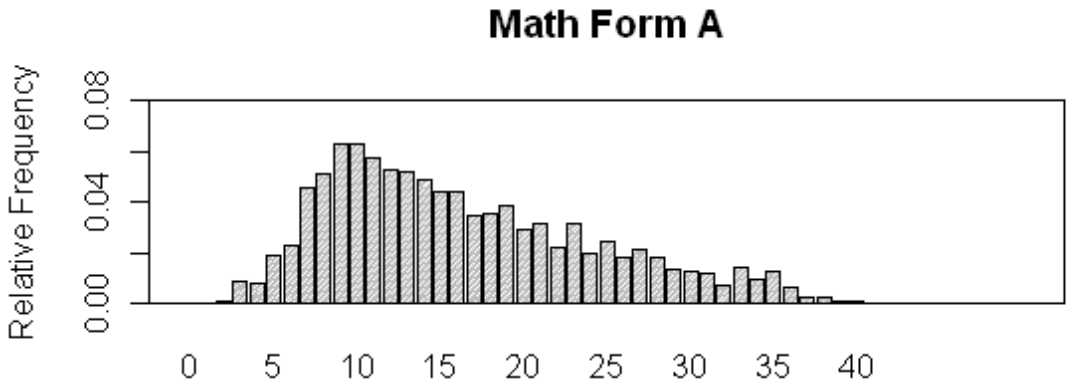
Table A-20. Statistics for Item Parameter Estimates

	Discrimination			Difficulty		
	Median	Mean	SD	Median	Mean	SD
UIRT Math A	0.989	1.004	0.412	0.570	0.804	1.156
MIRT Math A	0.999	1.009	0.380	0.505	0.709	1.062
UIRT Math B	0.967	0.978	0.405	0.877	1.145	1.115
MIRT Math B	0.954	0.966	0.365	0.845	1.096	1.126
UIRT Science A	1.015	0.986	0.369	0.449	0.977	1.410
MIRT Science A	1.003	0.957	0.335	0.434	0.953	1.334
UIRT Science B	0.935	0.913	0.402	0.535	0.867	1.112
MIRT Science B	0.924	0.886	0.374	0.521	0.850	1.093
UIRT Social A	0.929	0.975	0.361	0.357	0.538	1.004
MIRT Social A	0.972	0.984	0.331	0.290	0.426	1.141
UIRT Social B	0.837	0.896	0.344	0.508	0.809	1.089
MIRT Social B	0.849	0.865	0.314	0.482	0.773	1.051

Table A-21. Correlations Between Unidimensional Item Parameter Estimates and Unidimensional Approximation Item Parameter Estimates

	Discrimination	Difficulty
Math A	0.965	0.999
Math B	0.971	0.999
Science A	0.990	0.999
Science B	0.985	0.999
Social A	0.964	0.997
Social B	0.981	0.999

Figure A-1. Observed Score Distributions for Math Forms



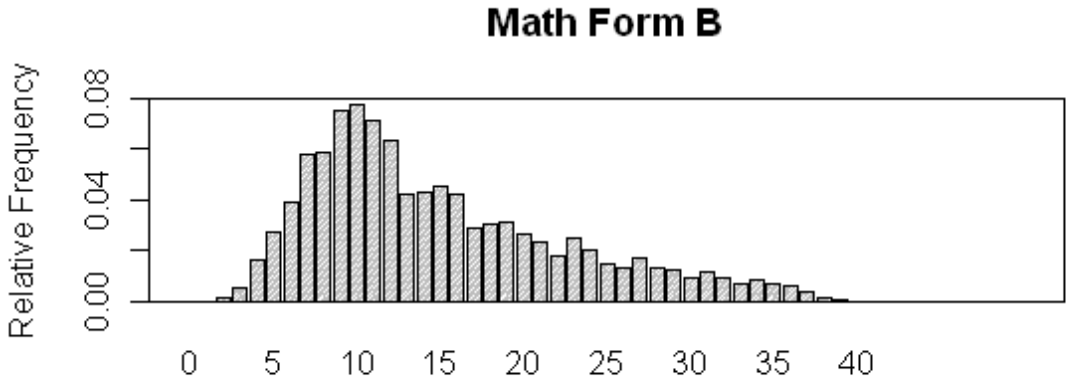


Figure A-2. Observed Score Distributions for Science Forms

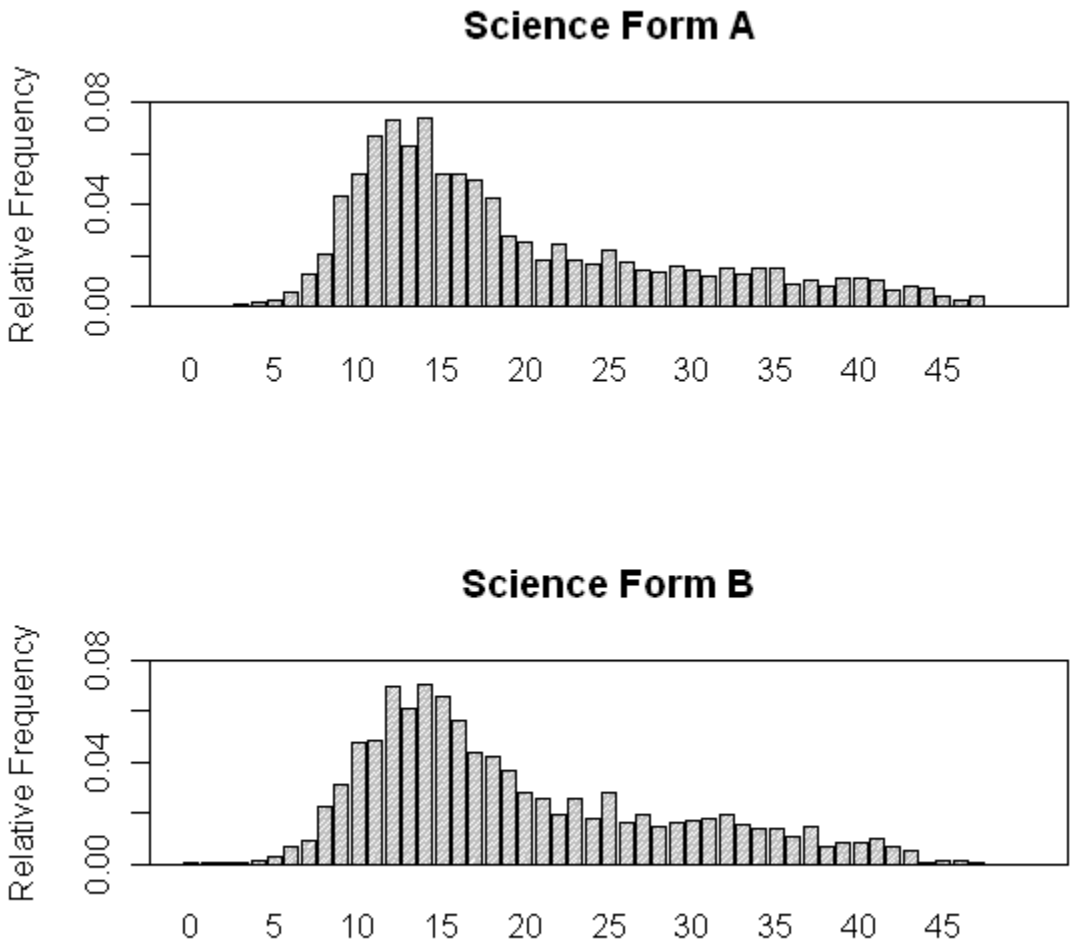
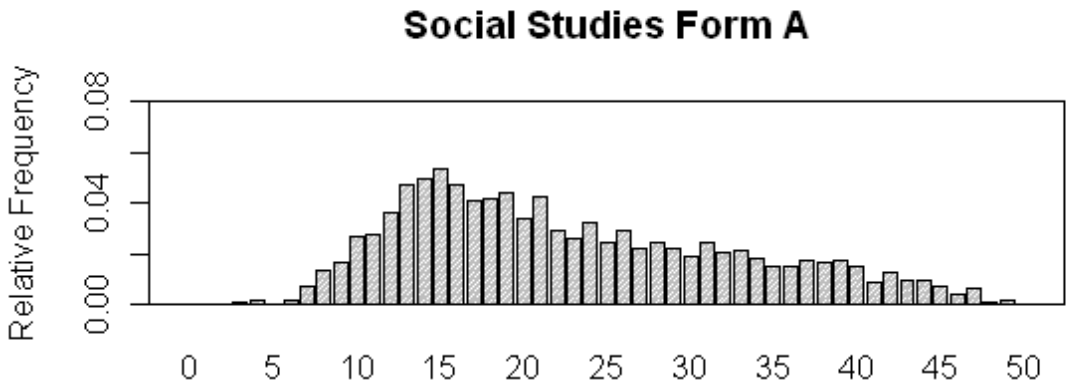


Figure A-3. Observed Score Distributions for Social Studies Forms



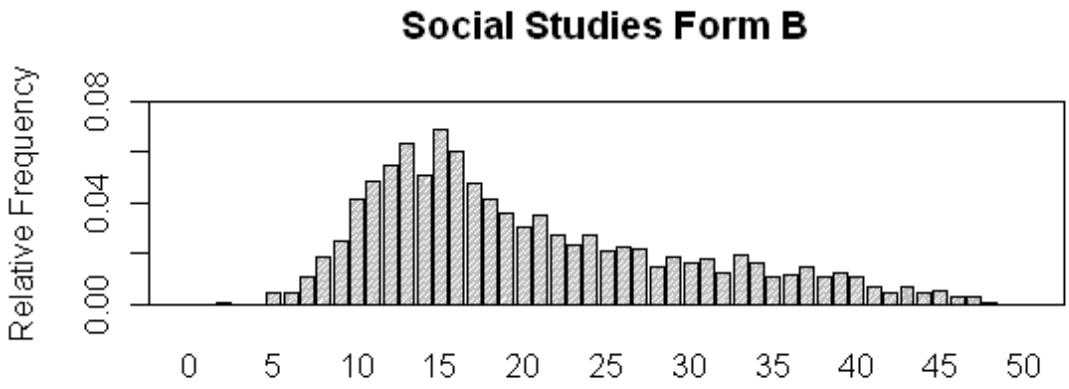


Figure A-4. Differences Between Unsmoothed Equating Results and Identity Equating for Math Exams

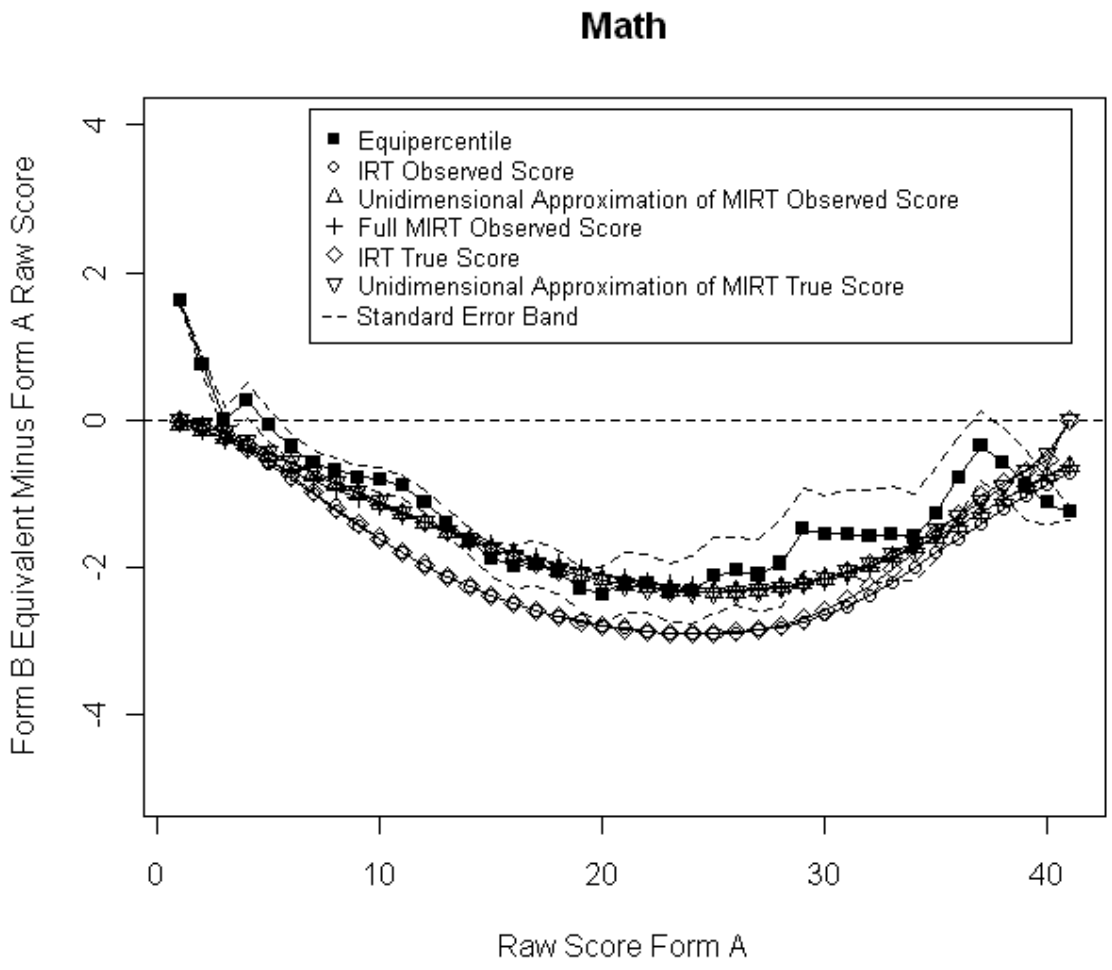


Figure A-5. Differences Between Unsmoothed Equating Results and Identity Equating for Science Exams

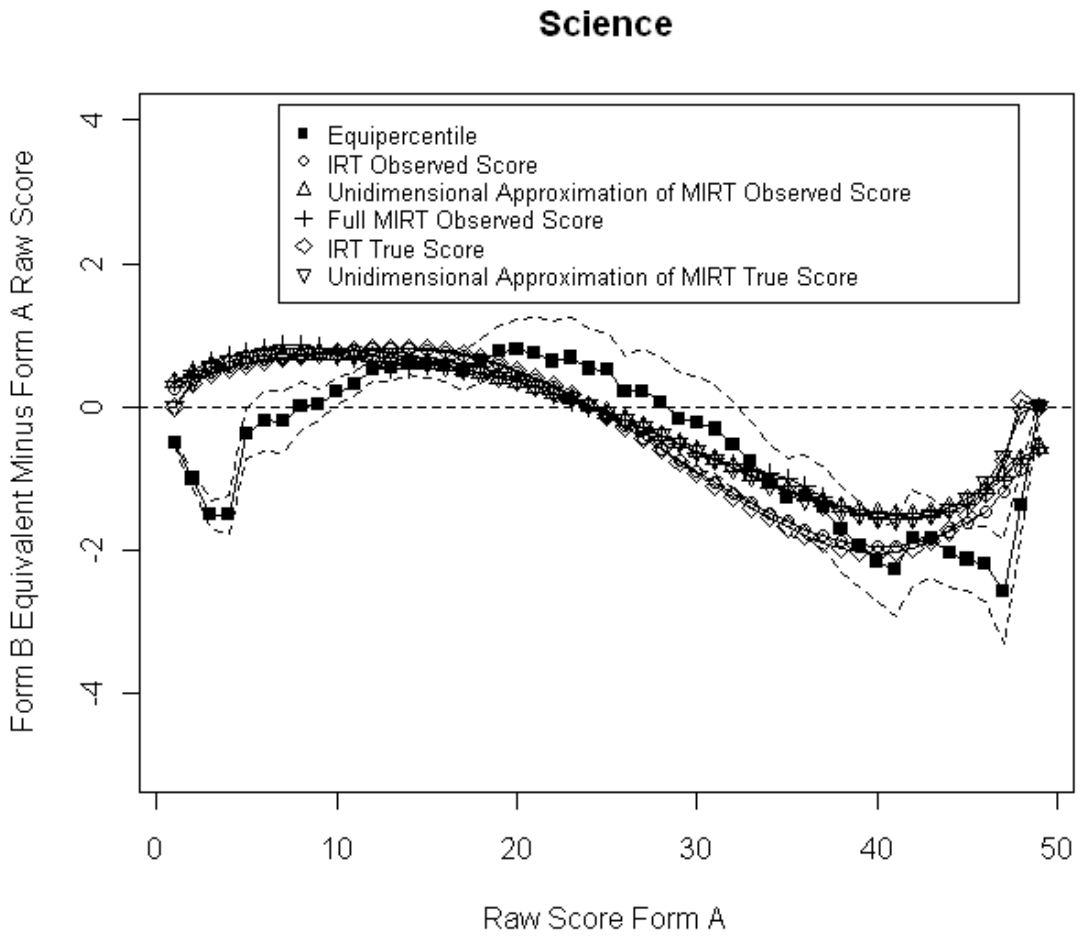


Figure A-6. Differences Between Unsmoothed Equating Results and Identity Equating for Social Studies Exams

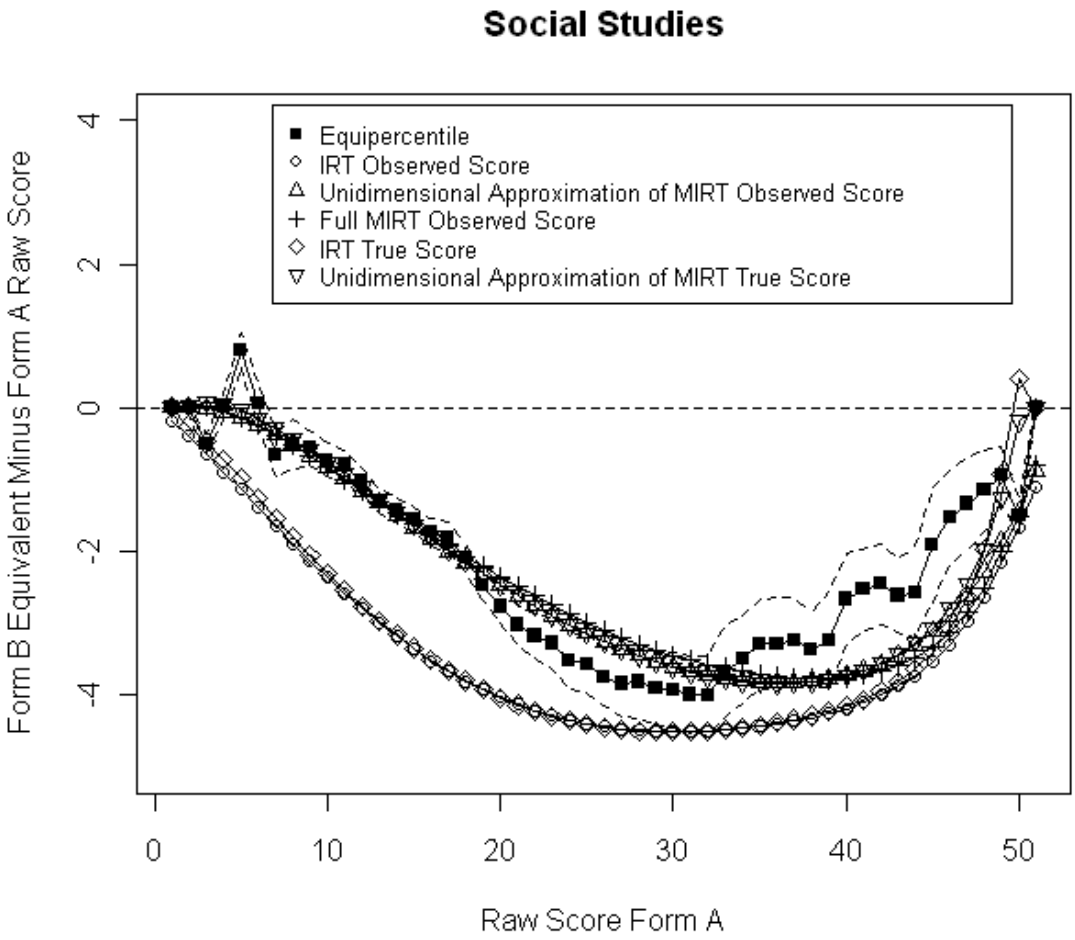


Figure A-7. Differences Between Smoothed Equating Results and Identity Equating for Math Exams

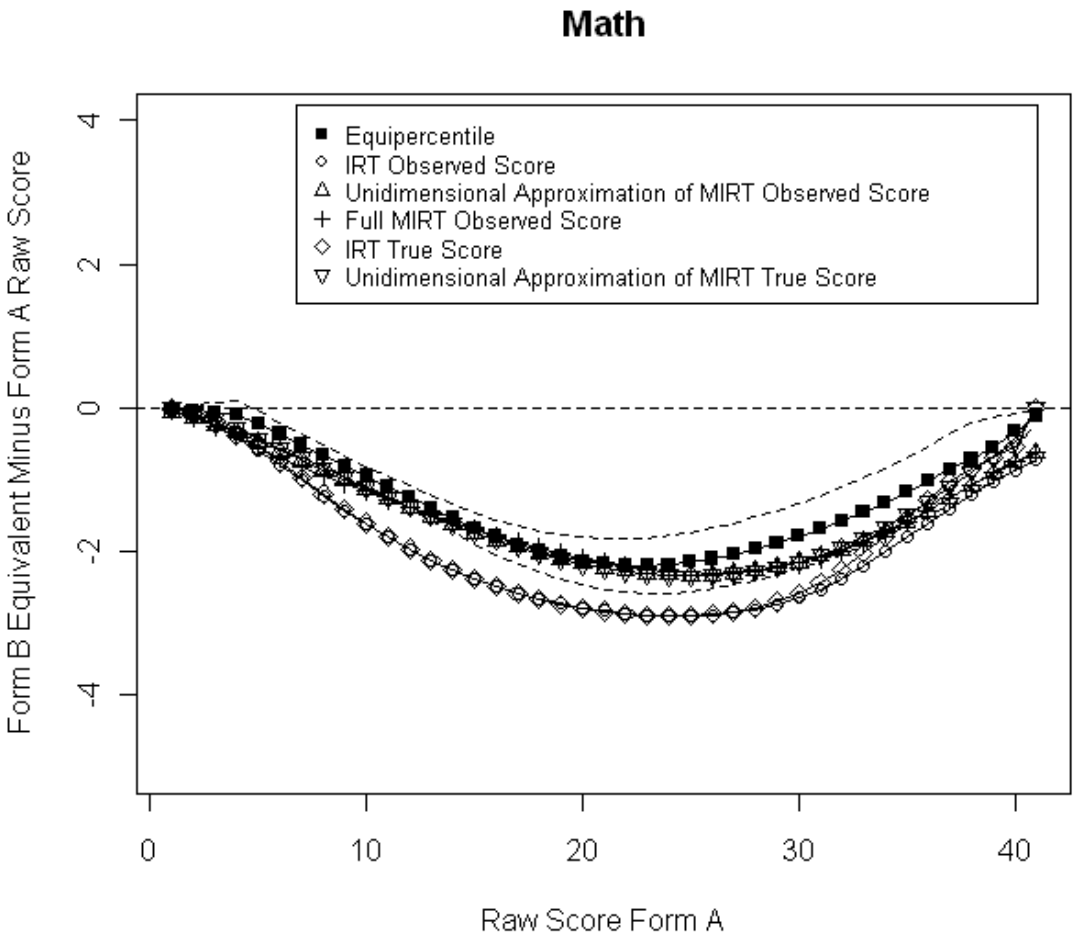


Figure A-8. Differences Between Smoothed Equating Results and Identity Equating for Science Exams

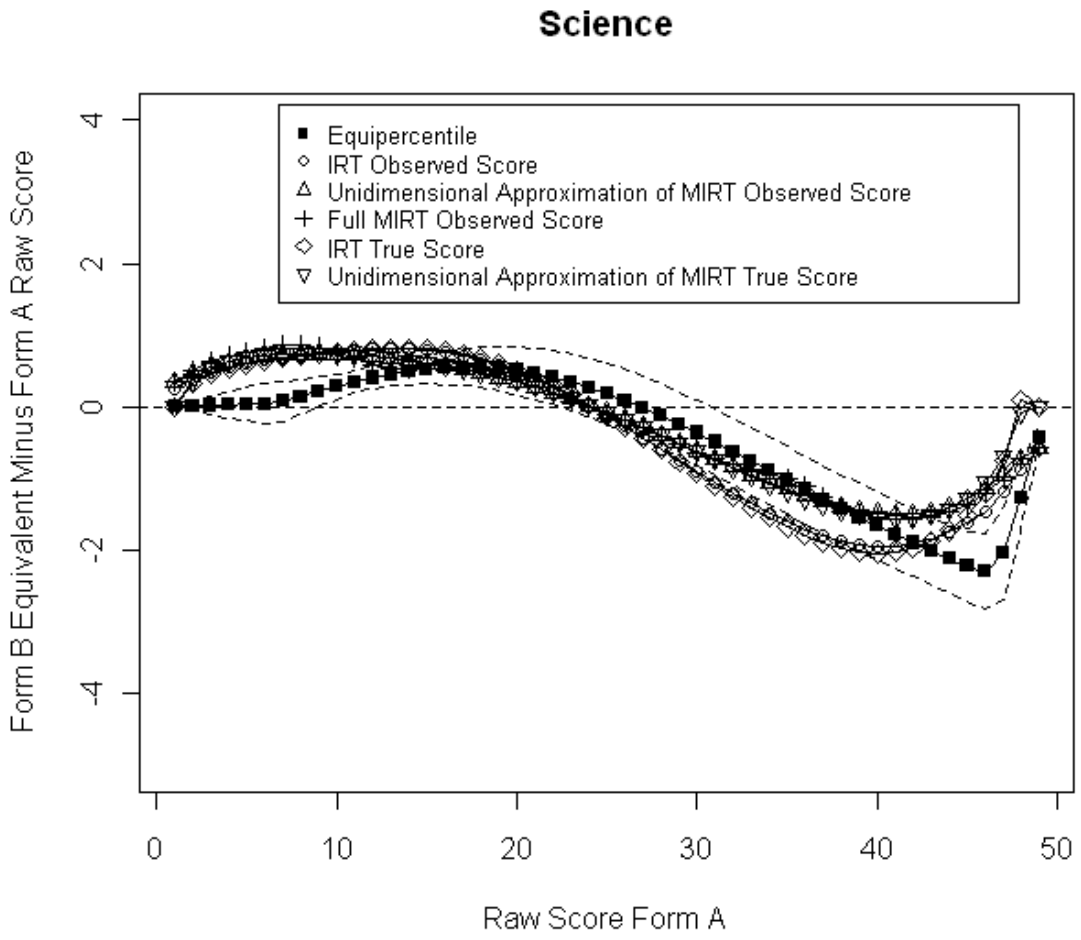


Figure A-9. Differences Between Smoothed Equating Results and Identity Equating for Social Studies Exams

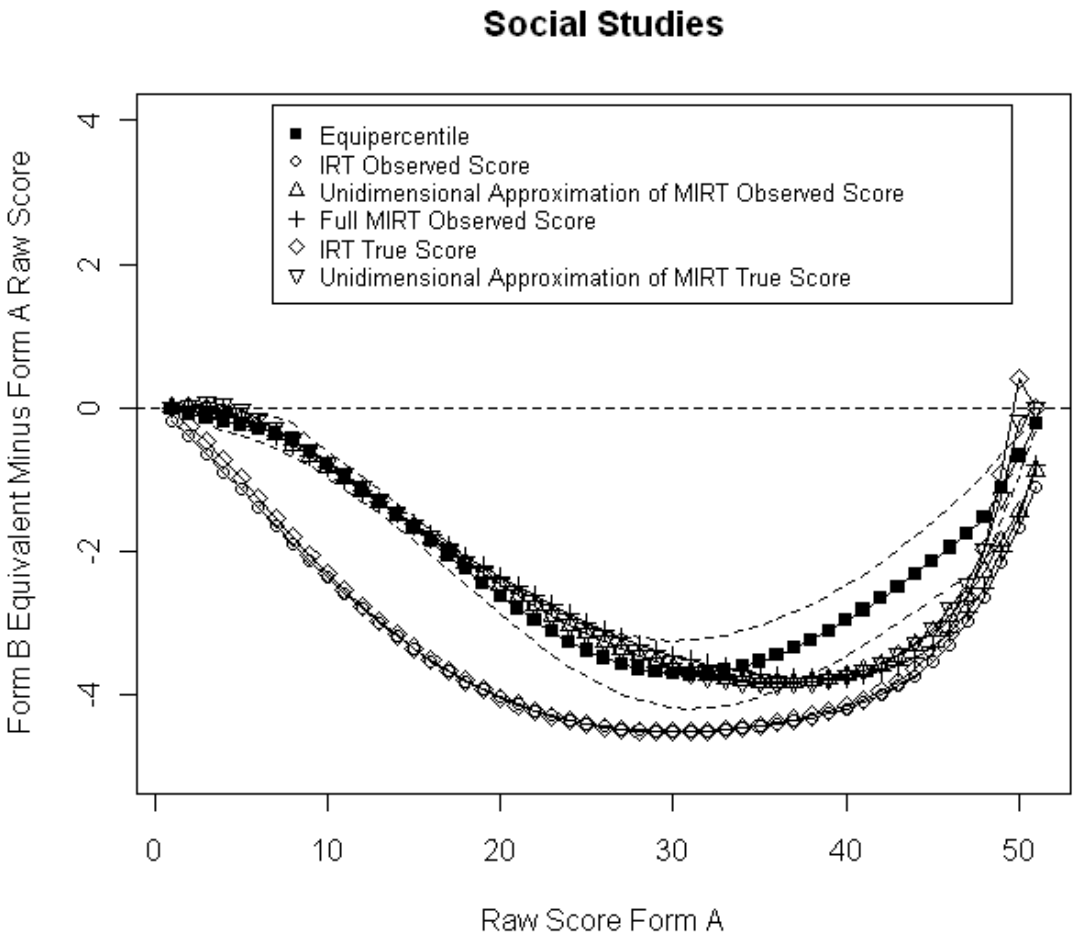


Figure A-10. Differences Between Equating Results and Unsmoothed Equipercentile Results for Math Exams

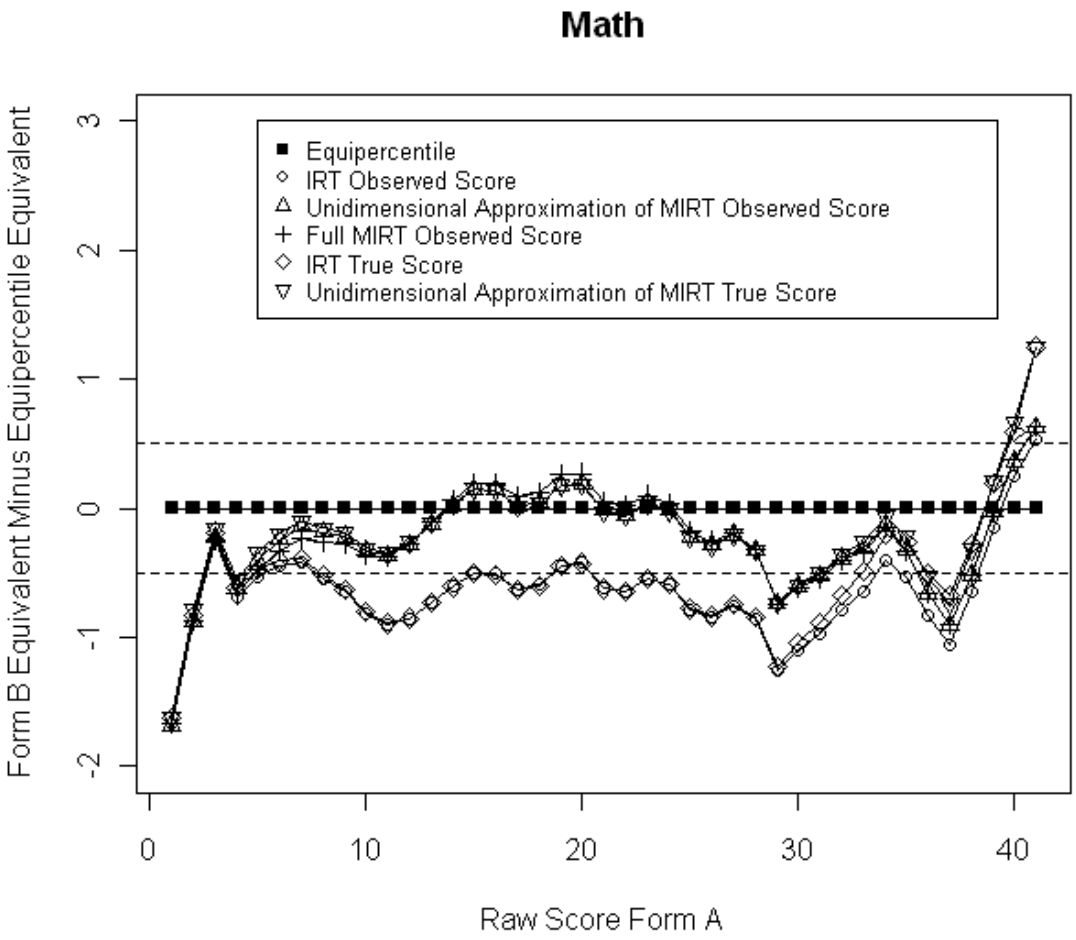


Figure A-11. Differences Between Equating Results and Unsmoothed Equipercentile Results for Science Exams

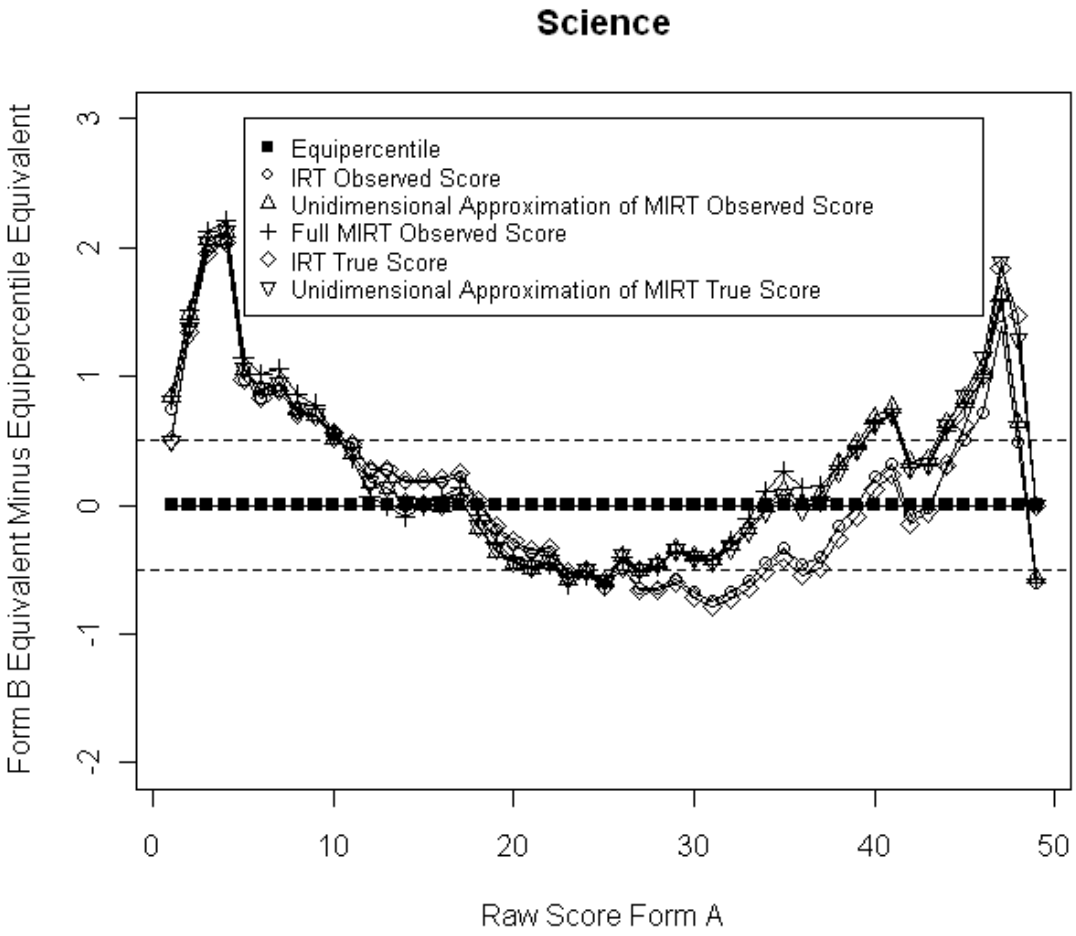


Figure A-12. Differences Between Equating Results and Unsmoothed Equipercentile Results for Social Studies Exams

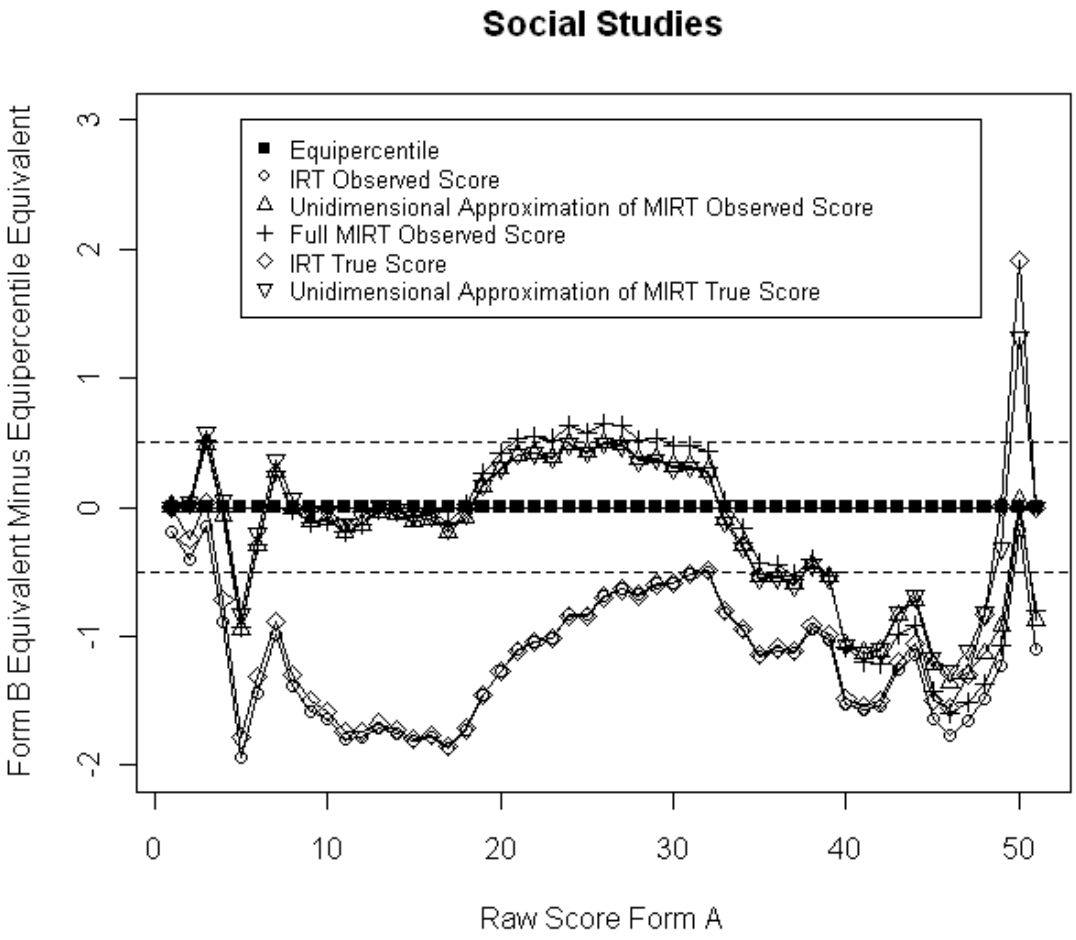


Figure A-13. Differences Between Equating Results and Smoothed Equipercentile Results for Math Exams

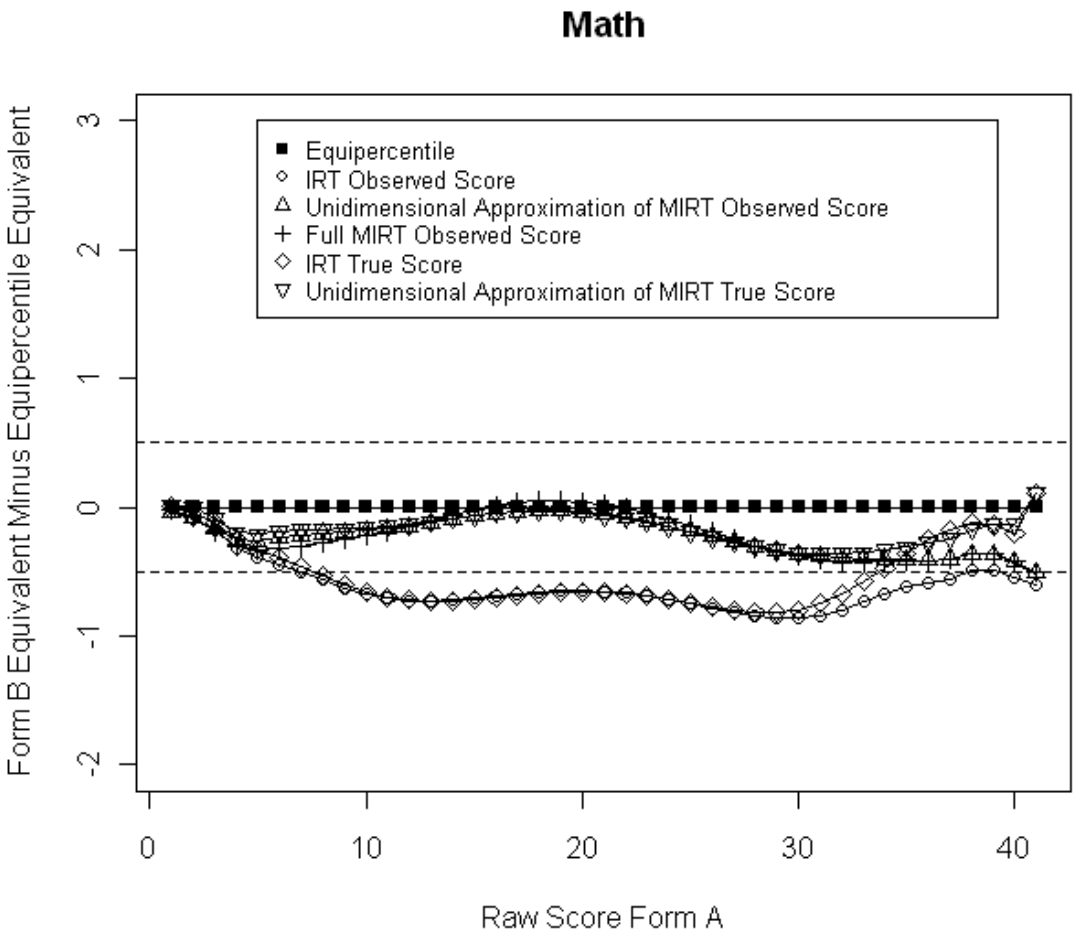


Figure A-14. Differences Between Equating Results and Smoothed Equipercentile Results for Science Exams

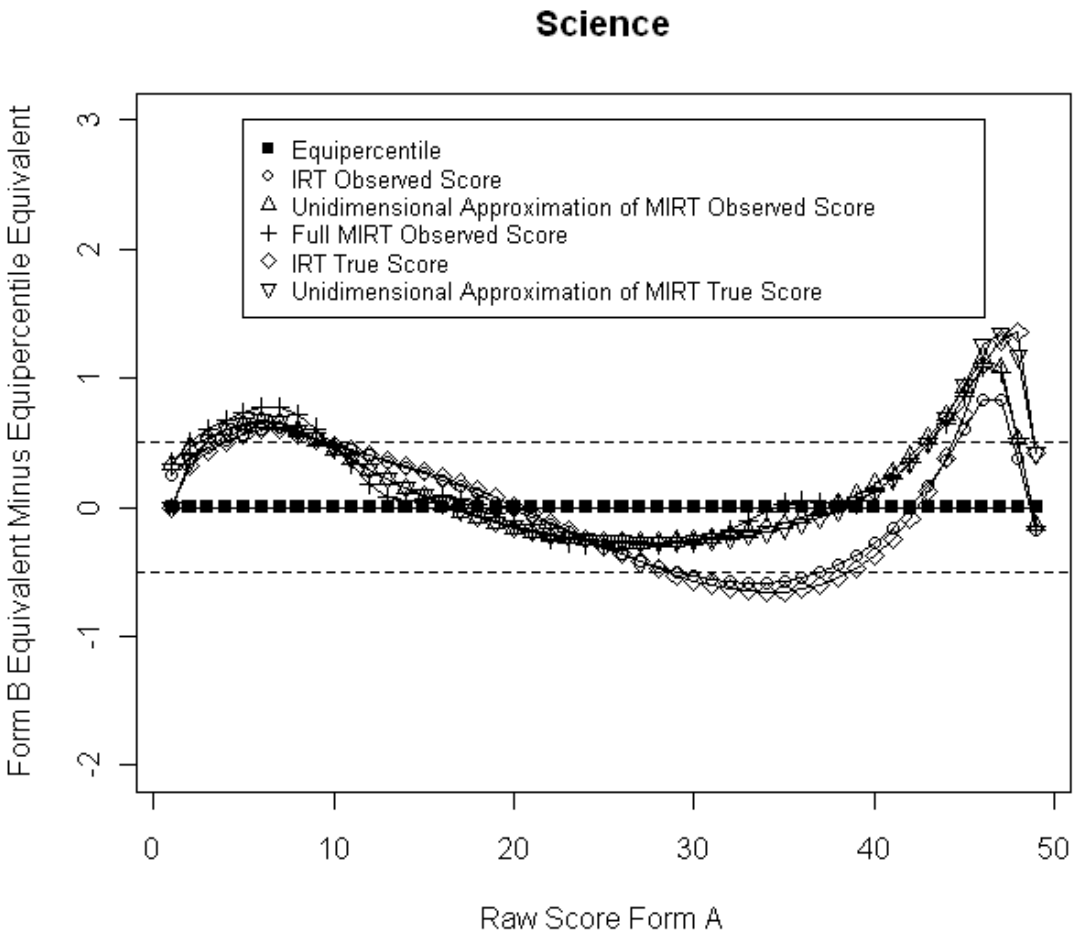


Figure A-15. Differences Between Equating Results and Smoothed Equipercentile Equating Results for Social Studies Exams

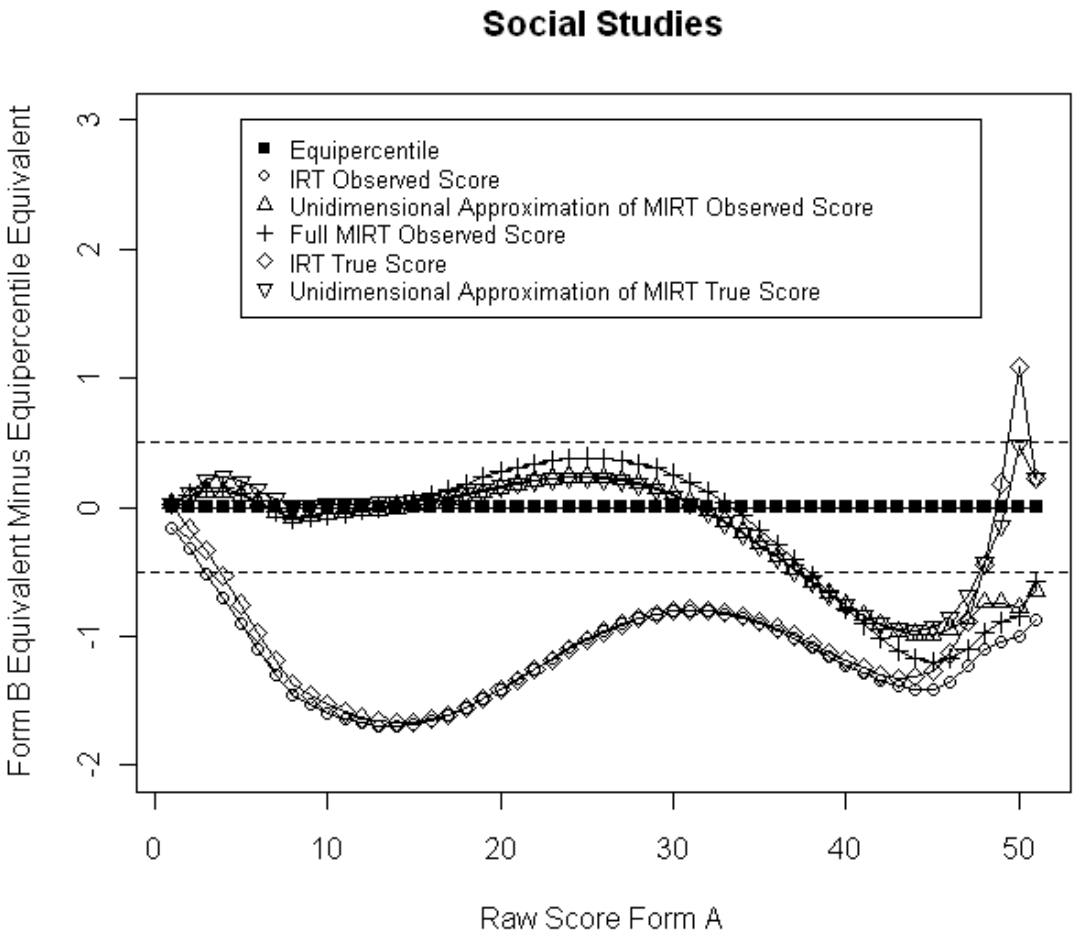


Figure A-16. Absolute Differences Between Equating Results and Unsmoothed Equipercentile Results for Math Exams

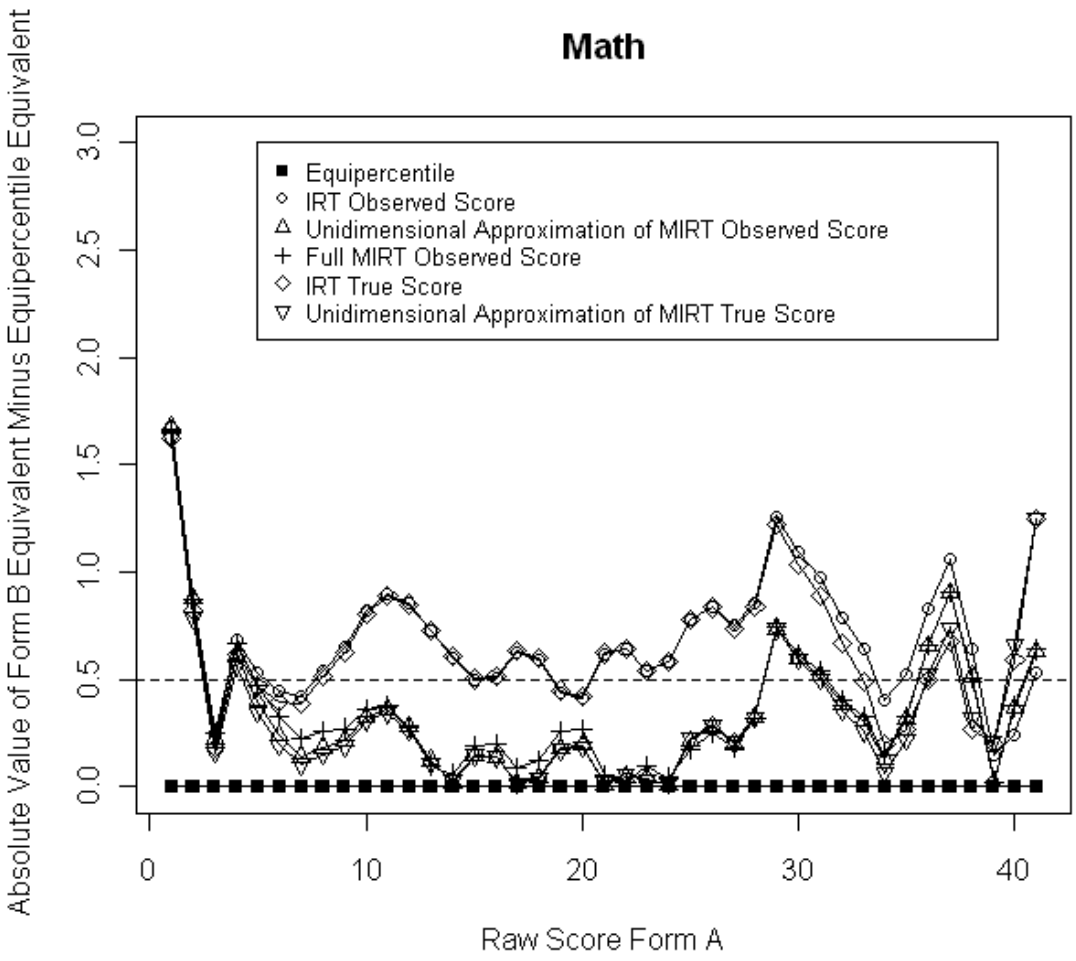


Figure A-17. Absolute Differences Between Equating Results and Unsmoothed Equipercentile Results for Science Exams

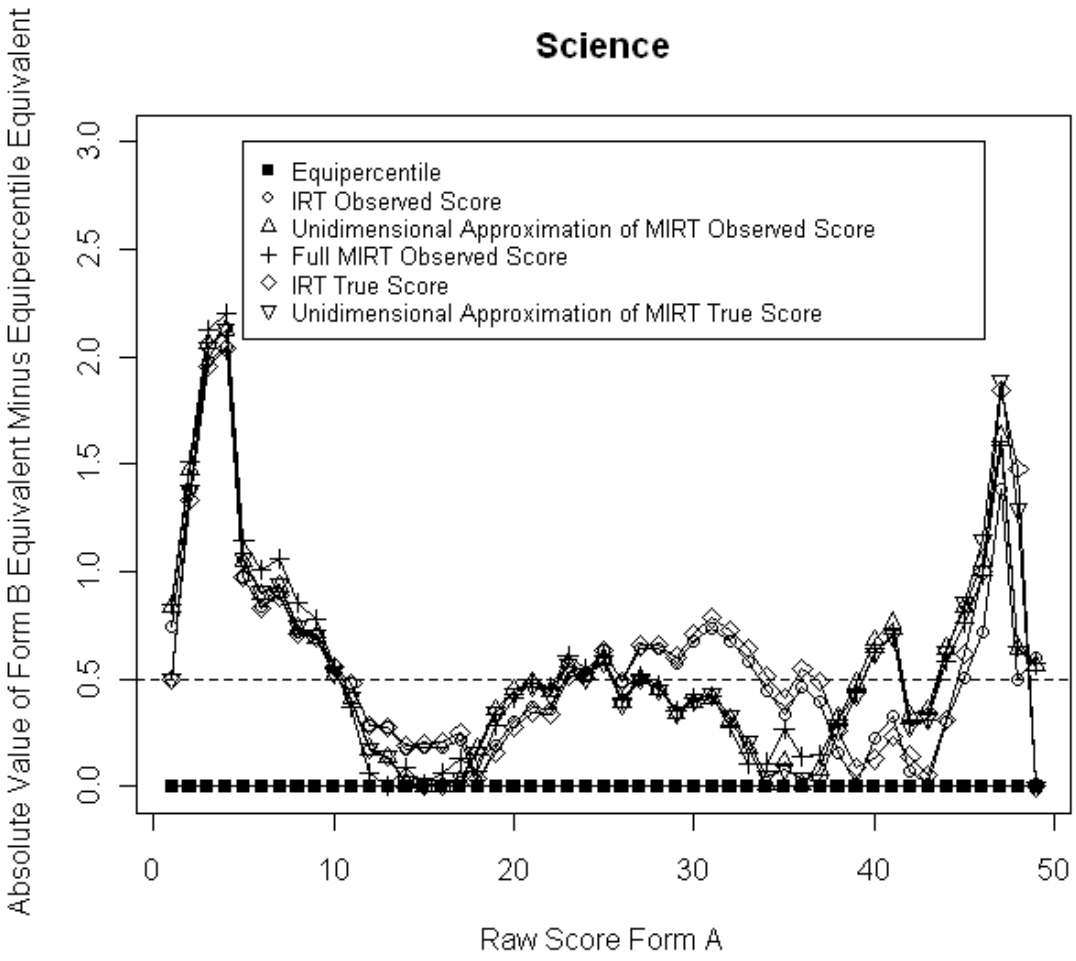


Figure A-18. Absolute Differences Between Equating Results and Unsmoothed Equipercentile Results for Social Studies Exams

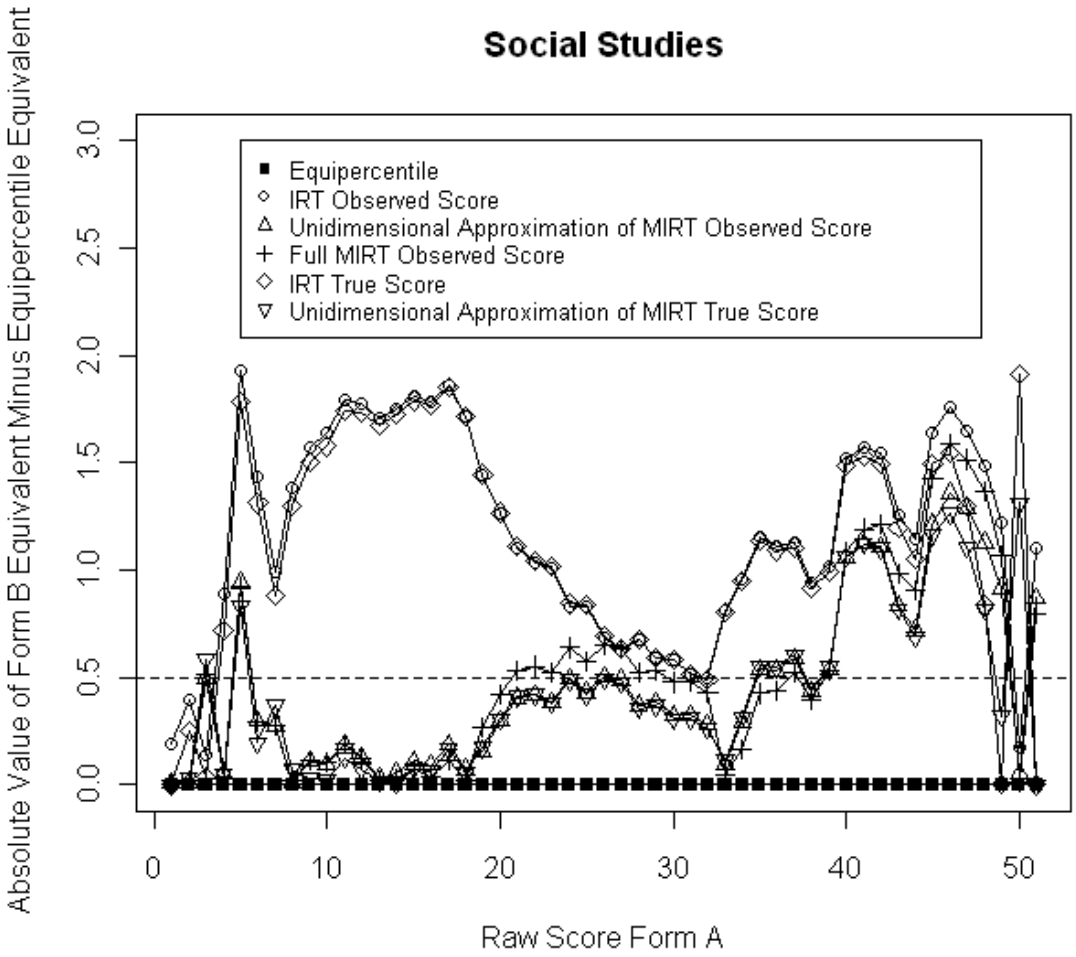


Figure A-19. Absolute Differences Between Equating Results and Smoothed Equipercentile Results for Math Exams

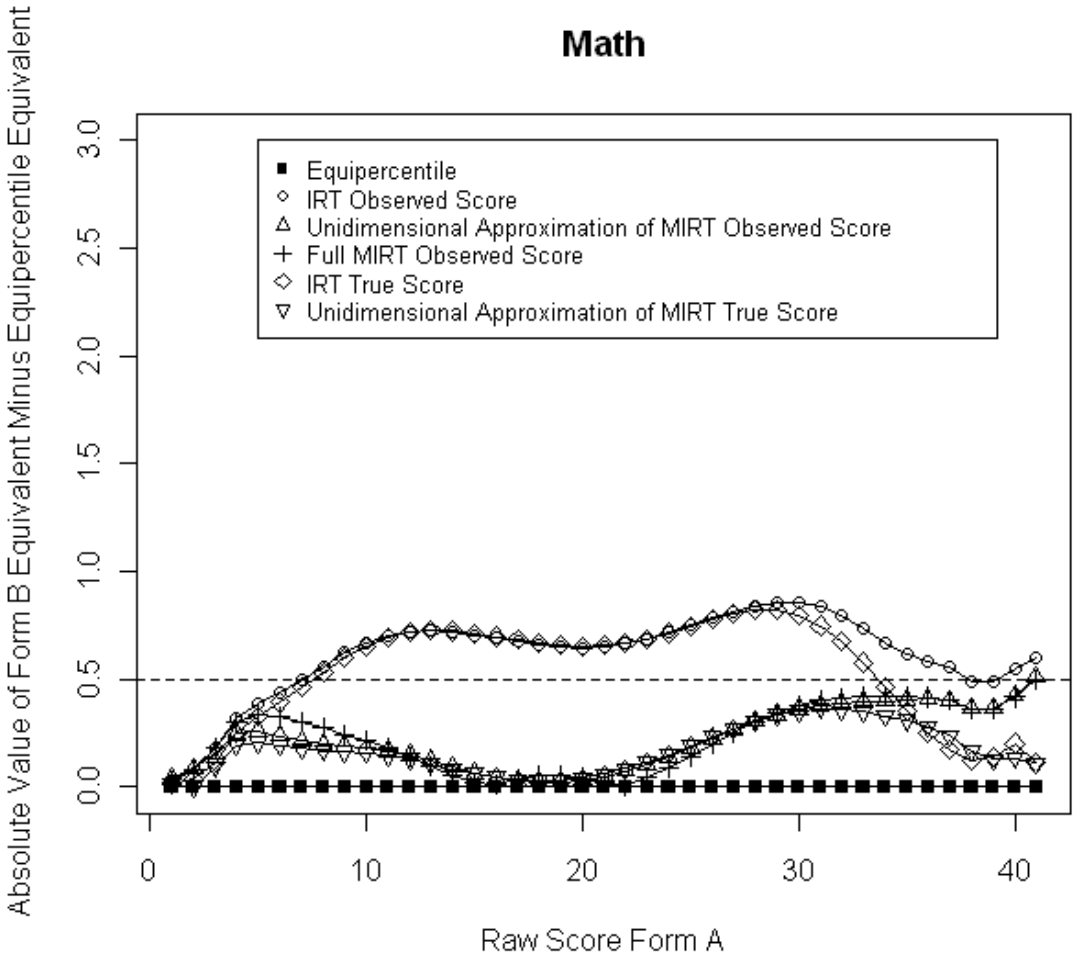


Figure A-20. Absolute Differences Between Equating Results and Smoothed Equipercentile Results for Science Exams

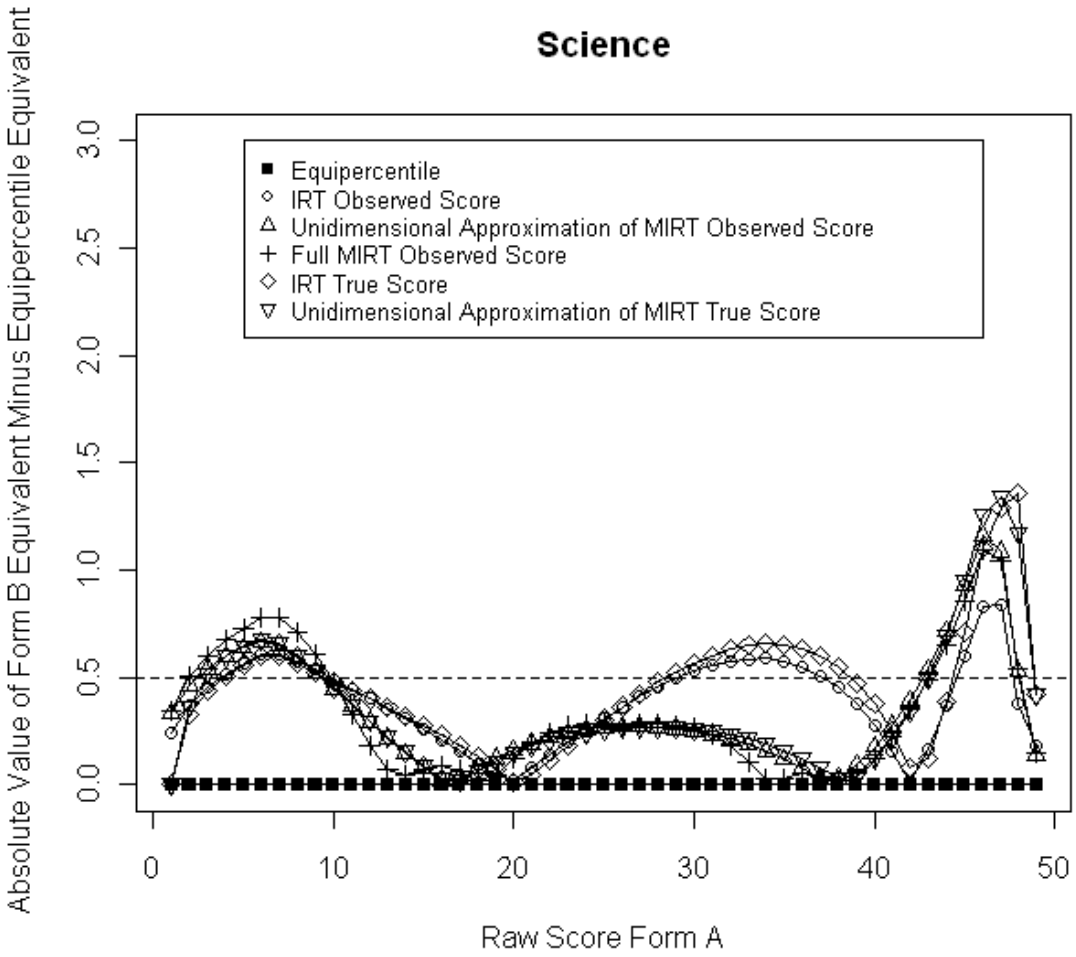


Figure A-21. Abolute Differences Between Equating Results and Smoothed Equipercentile Results for Social Studies Exams

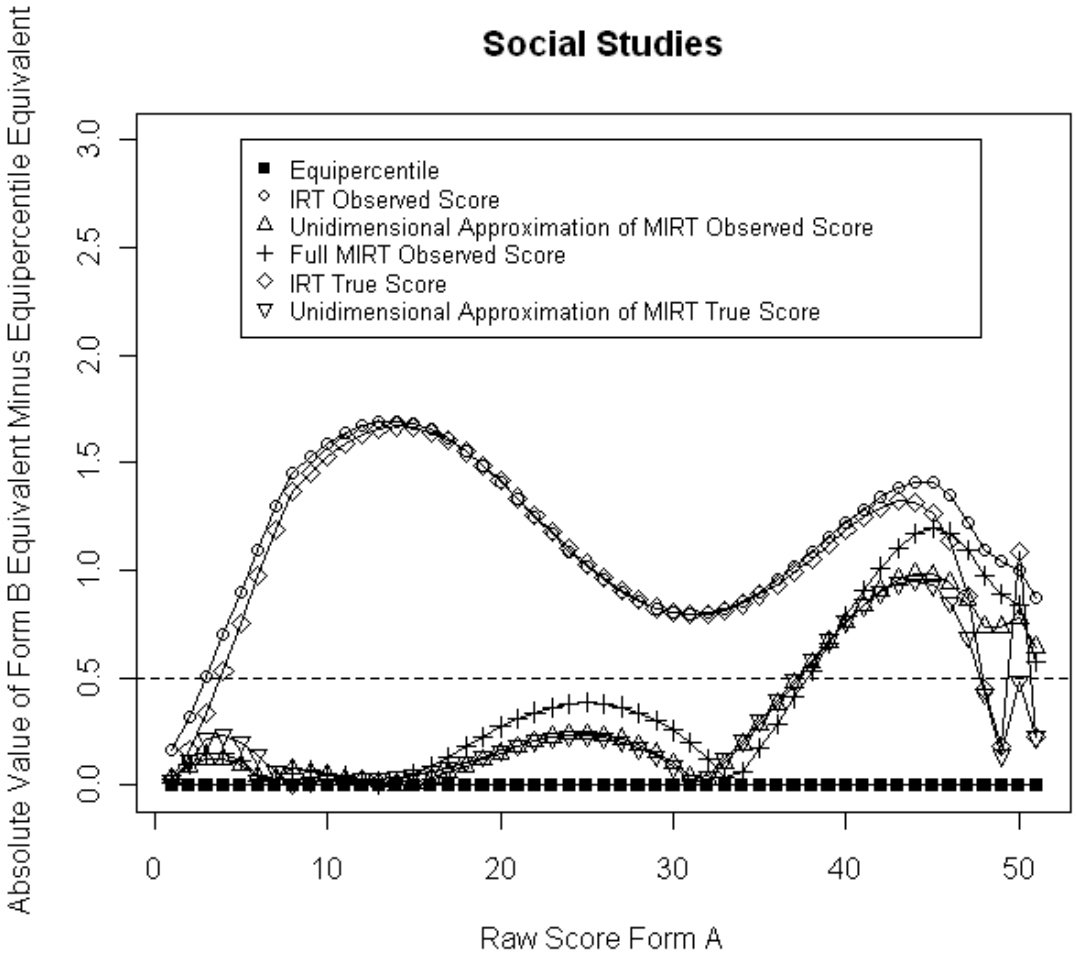
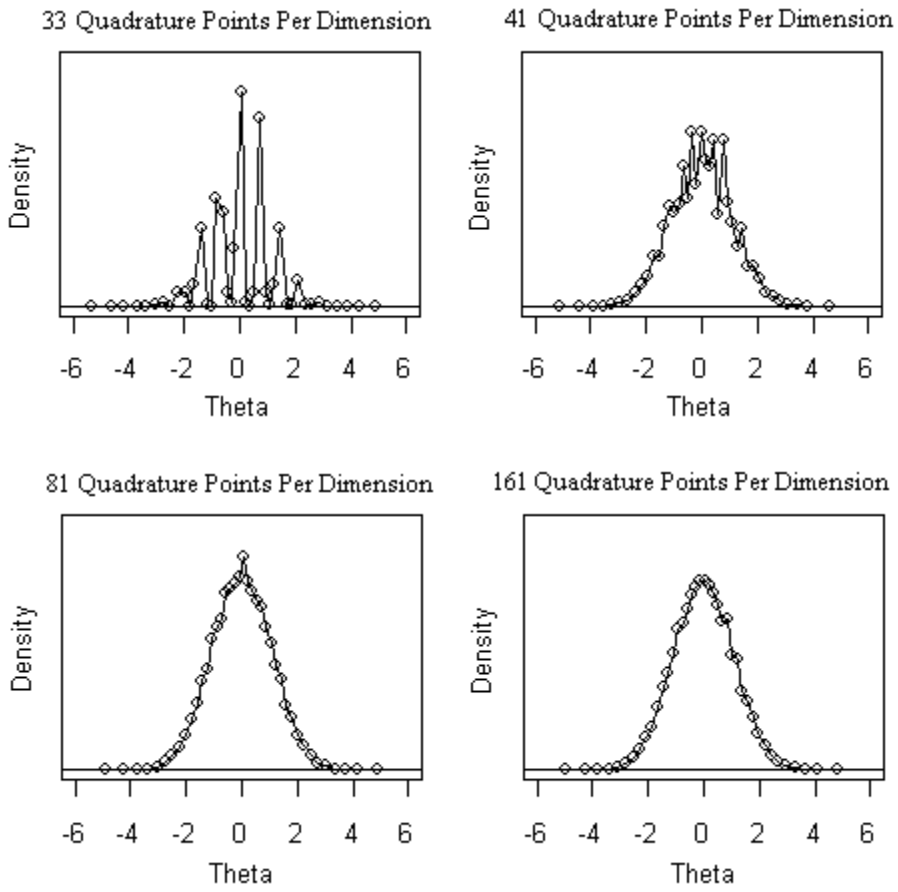


Figure A-22. Ability Distributions as the Number of Quadrature Points Increases



APPENDIX B. SAMPLE COMPUTER CODE

Figure B-1. TESTFACT Code for Form A Math Exam

```
>TITLE
MATHA.TSF-TRIAL RUN
        Trial Run
>PROBLEM NITEMS=40,RESPONSES=3;
>NAMES      I01,I02,I03,I04,I05,I06,I07,I08,I09,I10,I11,I12,I13,I14,I15,
I16,I17,I18,I19,I20,I21,I22,I23,I24,I25,I26,I27,I28,I29,I30,I31,I32,I33,I34,I35,
I36,I37,I38,I39,I40;
>RESPONSE  ' ', '0', '1';
>KEY  1111111111111111111111111111111111111111;
>TETRACHORIC    LIST,NDEC=3;
>FACTOR     NFAC=4,NROOT=4,ROTATE=VARIMAX;
>FULL               CYCLES=80;
>TECHNICAL          PRECISION=0.005;
>SCORE      METHOD=2,LIST=20;
>SAVE               PARM,FSCORE,ROTATE;
>INPUT      NIDCHAR=5,SCORES,FILE='MATHA.TXT';
(5A1,4X,40A1)
>STOP
```

Figure B-2. R Code

```
#COVARIANCE MATRICES FOR ABILITY DISTRIBUTIONS

MATHA_SIGMA <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=4,byrow=T)
MATHB_SIGMA <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=4,byrow=T)
SCIENCEA_SIGMA <- matrix(c(1,0,0,1),ncol=2,byrow=T)
SCIENCEB_SIGMA <- matrix(c(1,0,0,1),ncol=2,byrow=T)
SOCIALA_SIGMA <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=4,byrow=T)
SOCIALB_SIGMA <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=4,byrow=T)

#DETERMINE DIRECTION OF BEST MEASUREMENT

MATHA_alpha_1 <- sum(FINAL_MATHA[,1]) / sqrt((sum(FINAL_MATHA[,1]))^2 +
(sum(FINAL_MATHA[,2]))^2 + (sum(FINAL_MATHA[,3]))^2 +
(sum(FINAL_MATHA[,4]))^2)

MATHA_alpha_2 <- sum(FINAL_MATHA[,2]) / sqrt((sum(FINAL_MATHA[,1]))^2 +
(sum(FINAL_MATHA[,2]))^2 + (sum(FINAL_MATHA[,3]))^2 +
(sum(FINAL_MATHA[,4]))^2)

MATHA_alpha_3 <- sum(FINAL_MATHA[,3]) / sqrt((sum(FINAL_MATHA[,1]))^2 +
(sum(FINAL_MATHA[,2]))^2 + (sum(FINAL_MATHA[,3]))^2 +
(sum(FINAL_MATHA[,4]))^2)

MATHA_alpha_4 <- sum(FINAL_MATHA[,4]) / sqrt((sum(FINAL_MATHA[,1]))^2 +
(sum(FINAL_MATHA[,2]))^2 + (sum(FINAL_MATHA[,3]))^2 +
(sum(FINAL_MATHA[,4]))^2)

MATHB_alpha_1 <- sum(FINAL_MATHB[,1]) / sqrt((sum(FINAL_MATHB[,1]))^2 +
(sum(FINAL_MATHB[,2]))^2 + (sum(FINAL_MATHB[,3]))^2 +
(sum(FINAL_MATHB[,4]))^2)
```

Figure B-2—continued

```
MATHB_alpha_2 <- sum(FINAL_MATHB[,2]) / sqrt((sum(FINAL_MATHB[,1]))^2 +
(sum(FINAL_MATHB[,2]))^2 + (sum(FINAL_MATHB[,3]))^2 +
(sum(FINAL_MATHB[,4]))^2)

MATHB_alpha_3 <- sum(FINAL_MATHB[,3]) / sqrt((sum(FINAL_MATHB[,1]))^2 +
(sum(FINAL_MATHB[,2]))^2 + (sum(FINAL_MATHB[,3]))^2 +
(sum(FINAL_MATHB[,4]))^2)

MATHB_alpha_4 <- sum(FINAL_MATHB[,4]) / sqrt((sum(FINAL_MATHB[,1]))^2 +
(sum(FINAL_MATHB[,2]))^2 + (sum(FINAL_MATHB[,3]))^2 +
(sum(FINAL_MATHB[,4]))^2)

SCIENCEA_alpha_1 <- sum(FINAL_SCIENCEA[,1]) / sqrt((sum(FINAL_SCIENCEA[,1]))^2
+ (sum(FINAL_SCIENCEA[,2]))^2)

SCIENCEA_alpha_2 <- sum(FINAL_SCIENCEA[,2]) / sqrt((sum(FINAL_SCIENCEA[,1]))^2
+ (sum(FINAL_SCIENCEA[,2]))^2)

SCIENCEB_alpha_1 <- sum(FINAL_SCIENCEB[,1]) / sqrt((sum(FINAL_SCIENCEB[,1]))^2
+ (sum(FINAL_SCIENCEB[,2]))^2)

SCIENCEB_alpha_2 <- sum(FINAL_SCIENCEB[,2]) / sqrt((sum(FINAL_SCIENCEB[,1]))^2
+ (sum(FINAL_SCIENCEB[,2]))^2)

SOCIALA_alpha_1 <- sum(FINAL_SOCIALA[,1]) / sqrt((sum(FINAL_SOCIALA[,1]))^2 +
(sum(FINAL_SOCIALA[,2]))^2 + (sum(FINAL_SOCIALA[,3]))^2 +
(sum(FINAL_SOCIALA[,4]))^2)

SOCIALA_alpha_2 <- sum(FINAL_SOCIALA[,2]) / sqrt((sum(FINAL_SOCIALA[,1]))^2 +
(sum(FINAL_SOCIALA[,2]))^2 + (sum(FINAL_SOCIALA[,3]))^2 +
(sum(FINAL_SOCIALA[,4]))^2)

SOCIALA_alpha_3 <- sum(FINAL_SOCIALA[,3]) / sqrt((sum(FINAL_SOCIALA[,1]))^2 +
(sum(FINAL_SOCIALA[,2]))^2 + (sum(FINAL_SOCIALA[,3]))^2 +
(sum(FINAL_SOCIALA[,4]))^2)

SOCIALA_alpha_4 <- sum(FINAL_SOCIALA[,4]) / sqrt((sum(FINAL_SOCIALA[,1]))^2 +
(sum(FINAL_SOCIALA[,2]))^2 + (sum(FINAL_SOCIALA[,3]))^2 +
(sum(FINAL_SOCIALA[,4]))^2)
```

Figure B-2—continued

```
SOCIALB_alpha_1 <- sum(FINAL_SOCIALB[,1]) / sqrt((sum(FINAL_SOCIALB[,1]))^2 +
(sum(FINAL_SOCIALB[,2]))^2 + (sum(FINAL_SOCIALB[,3]))^2 +
(sum(FINAL_SOCIALB[,4]))^2)

SOCIALB_alpha_2 <- sum(FINAL_SOCIALB[,2]) / sqrt((sum(FINAL_SOCIALB[,1]))^2 +
(sum(FINAL_SOCIALB[,2]))^2 + (sum(FINAL_SOCIALB[,3]))^2 +
(sum(FINAL_SOCIALB[,4]))^2)

SOCIALB_alpha_3 <- sum(FINAL_SOCIALB[,3]) / sqrt((sum(FINAL_SOCIALB[,1]))^2 +
(sum(FINAL_SOCIALB[,2]))^2 + (sum(FINAL_SOCIALB[,3]))^2 +
(sum(FINAL_SOCIALB[,4]))^2)

SOCIALB_alpha_4 <- sum(FINAL_SOCIALB[,4]) / sqrt((sum(FINAL_SOCIALB[,1]))^2 +
(sum(FINAL_SOCIALB[,2]))^2 + (sum(FINAL_SOCIALB[,3]))^2 +
(sum(FINAL_SOCIALB[,4]))^2)

MATHA_alpha <- matrix(c(MATHA_alpha_1, MATHA_alpha_2, MATHA_alpha_3,
MATHA_alpha_4),ncol=1,byrow=T)

MATHB_alpha <- matrix(c(MATHB_alpha_1, MATHB_alpha_2, MATHB_alpha_3,
MATHB_alpha_4),ncol=1,byrow=T)

SCIENCEA_alpha <- matrix(c(SCIENCEA_alpha_1, SCIENCEA_alpha_2), ncol=1,byrow=T)

SCIENCEB_alpha <- matrix(c(SCIENCEB_alpha_1, SCIENCEB_alpha_2), ncol=1,byrow=T)

SOCIALA_alpha <- matrix(c(SOCIALA_alpha_1, SOCIALA_alpha_2,
SOCIALA_alpha_3,SOCIALA_alpha_4),ncol=1,byrow=T)

SOCIALB_alpha <- matrix(c(SOCIALB_alpha_1, SOCIALB_alpha_2,
SOCIALB_alpha_3,SOCIALB_alpha_4),ncol=1,byrow=T)

MATHA_alpha_repeat1 <- rep(MATHA_alpha,each=40)
MATHA_alpha_repeat2 <- matrix(MATHA_alpha_repeat1,nrow=4,byrow=T)

MATHB_alpha_repeat1 <- rep(MATHB_alpha,each=40)
```

Figure B-2—continued

```
MATHB_alpha_repeat2 <- matrix(MATHB_alpha_repeat1,nrow=4,byrow=T)

SCIENCEA_alpha_repeat1 <- rep(SCIENCEA_alpha,each=48)
SCIENCEA_alpha_repeat2 <- matrix(SCIENCEA_alpha_repeat1,nrow=2,byrow=T)

SCIENCEB_alpha_repeat1 <- rep(SCIENCEB_alpha,each=48)
SCIENCEB_alpha_repeat2 <- matrix(SCIENCEB_alpha_repeat1,nrow=2,byrow=T)

SOCIALA_alpha_repeat1 <- rep(SOCIALA_alpha,each=50)
SOCIALA_alpha_repeat2 <- matrix(SOCIALA_alpha_repeat1,nrow=4,byrow=T)

SOCIALB_alpha_repeat1 <- rep(SOCIALB_alpha,each=50)
SOCIALB_alpha_repeat2 <- matrix(SOCIALB_alpha_repeat1,nrow=4,byrow=T)

#CREATE MATRICES THAT ONLY CONTAIN DISCRIMINATION PARAMETERS

rotated_MATHA <- MATHA[,-6]
rotated_MATHA <- rotated_MATHA[,-5]
rotated_MATHB <- MATHB[,-6]
rotated_MATHB <- rotated_MATHB[,-5]

rotated_SCIENCEA <- SCIENCEA[,-4]
rotated_SCIENCEA <- rotated_SCIENCEA[,-3]
rotated_SCIENCEB <- SCIENCEB[,-4]
rotated_SCIENCEB <- rotated_SCIENCEB[,-3]

rotated_SOCIALA <- SOCIALA[,-6]
rotated_SOCIALA <- rotated_SOCIALA[,-5]
rotated_SOCIALB <- SOCIALB[,-6]
rotated_SOCIALB <- rotated_SOCIALB[,-5]

#DETERMINE UNIDIMENSIONAL ITEM PARAMETERS
```

Figure B-2—continued

```
MATHA_sigma_star_sq_matrix <- rotated_MATHA %*% MATHA_SIGMA %*%
t(rotated_MATHA) - (rotated_MATHA %*% MATHA_SIGMA %*%
MATHA_alpha_repeat2)^2


MATHA_sigma_star_sq <- diag(MATHA_sigma_star_sq_matrix)


MATHA_a_star <- (1+(MATHA_sigma_star_sq))^(-1/2) * diag(rotated_MATHA %*%
MATHA_SIGMA %*% MATHA_alpha_repeat2)


MATHA_d_star <- (1+(MATHA_sigma_star_sq))^(-1/2) * FINAL_MATHA[,5]


MATHB_sigma_star_sq_matrix <- rotated_MATHB %*% MATHB_SIGMA %*%
t(rotated_MATHB) - (rotated_MATHB %*% MATHB_SIGMA %*%
MATHB_alpha_repeat2)^2


MATHB_sigma_star_sq <- diag(MATHB_sigma_star_sq_matrix)


MATHB_a_star <- (1+(MATHB_sigma_star_sq))^(-1/2) * diag(rotated_MATHB %*%
MATHB_SIGMA %*% MATHB_alpha_repeat2)


MATHB_d_star <- (1+(MATHB_sigma_star_sq))^(-1/2) * FINAL_MATHB[,5]


SCIENCEA_sigma_star_sq_matrix <- rotated_SCIENCEA %*% SCIENCEA_SIGMA %*%
t(rotated_SCIENCEA) - (rotated_SCIENCEA %*% SCIENCEA_SIGMA %*%
SCIENCEA_alpha_repeat2)^2


SCIENCEA_sigma_star_sq <- diag(SCIENCEA_sigma_star_sq_matrix)


SCIENCEA_a_star <- (1+(SCIENCEA_sigma_star_sq))^(-1/2) * diag(rotated_SCIENCEA
%*% SCIENCEA_SIGMA %*% SCIENCEA_alpha_repeat2)


SCIENCEA_d_star <- (1+(SCIENCEA_sigma_star_sq))^(-1/2) * FINAL_SCIENCEA[,3]


SCIENCEB_sigma_star_sq_matrix <- rotated_SCIENCEB %*% SCIENCEB_SIGMA %*%
t(rotated_SCIENCEB) - (rotated_SCIENCEB %*% SCIENCEB_SIGMA %*%
SCIENCEB_alpha_repeat2)^2
```

Figure B-2—continued

```
SCIENCEB_sigma_star_sq <- diag(SCIENCEB_sigma_star_sq_matrix)

SCIENCEB_a_star <- (1+(SCIENCEB_sigma_star_sq))^(-1/2) * diag(rotated_SCIENCEB
%*% SCIENCEB_SIGMA %*% SCIENCEB_alpha_repeat2)

SCIENCEB_d_star <- (1+(SCIENCEB_sigma_star_sq))^(-1/2) * FINAL_SCIENCEB[,3]

SOCIALA_sigma_star_sq_matrix <- rotated_SOCIALA %*% SOCIALA_SIGMA %*%
t(rotated_SOCIALA) - (rotated_SOCIALA %*% SOCIALA_SIGMA %*%
SOCIALA_alpha_repeat2)^2

SOCIALA_sigma_star_sq <- diag(SOCIALA_sigma_star_sq_matrix)

SOCIALA_a_star <- (1+(SOCIALA_sigma_star_sq))^(-1/2) * diag(rotated_SOCIALA %*%
SOCIALA_SIGMA %*% SOCIALA_alpha_repeat2)

SOCIALA_d_star <- (1+(SOCIALA_sigma_star_sq))^(-1/2) * FINAL_SOCIALA[,5]

SOCIALB_sigma_star_sq_matrix <- rotated_SOCIALB %*% SOCIALB_SIGMA %*%
t(rotated_SOCIALB) - (rotated_SOCIALB %*% SOCIALB_SIGMA %*%
SOCIALB_alpha_repeat2)^2

SOCIALB_sigma_star_sq <- diag(SOCIALB_sigma_star_sq_matrix)

SOCIALB_a_star <- (1+(SOCIALB_sigma_star_sq))^(-1/2) * diag(rotated_SOCIALB %*%
SOCIALB_SIGMA %*% SOCIALB_alpha_repeat2)

SOCIALB_d_star <- (1+(SOCIALB_sigma_star_sq))^(-1/2) * FINAL_SOCIALB[,5]

#NORMAL OGIVE PARAMETERS

MATHA_parm <- cbind(MATHA_a_star,MATHA_d_star,MATHA[,6])
MATHB_parm <- cbind(MATHB_a_star,MATHB_d_star,MATHB[,6])
```

Figure B-2—continued

```
SCIENCEA_parm <- cbind(SCIENCEA_a_star,SCIENCEA_d_star,SCIENCEA[,4])
SCIENCEB_parm <- cbind(SCIENCEB_a_star,SCIENCEB_d_star,SCIENCEB[,4])

SOCIALA_parm <- cbind(SOCIALA_a_star,SOCIALA_d_star,SOCIALA[,6])
SOCIALB_parm <- cbind(SOCIALB_a_star,SOCIALB_d_star,SOCIALB[,6])

#CONVERT NORMAL OGIVE PARAMATERS TO 3PL PARAMETERS
#NOTE: 3PL PARAMETERIZATION IS: a(theta-b), WE HAVE a*theta+d

MATHA_parm_3PL <- cbind(MATHA_a_star,-MATHA_d_star/MATHA_a_star,
MATHA[,6])

MATHB_parm_3PL <- cbind(MATHB_a_star,-MATHB_d_star/MATHB_a_star, MATHB[,6])

SCIENCEA_parm_3PL <- cbind(SCIENCEA_a_star,
-SCIENCEA_d_star/SCIENCEA_a_star,SCIENCEA[,4])

SCIENCEB_parm_3PL <- cbind(SCIENCEB_a_star,
-SCIENCEB_d_star/SCIENCEB_a_star,SCIENCEB[,4])

SOCIALA_parm_3PL <- cbind(SOCIALA_a_star,
-SOCIALA_d_star/SOCIALA_a_star,SOCIALA[,6])

SOCIALB_parm_3PL <- cbind(SOCIALB_a_star,
-SOCIALB_d_star/SOCIALB_a_star,SOCIALB[,6])

#CREATE ITEM NUMBERS FOR PIE INPUT

ITEM_NUM_MATH <- matrix(c(1:40),ncol=1)
ITEM_NUM_SCIENCE <- matrix(c(1:48),ncol=1)
ITEM_NUM_SOCIAL <- matrix(c(1:50),ncol=1)

#CREATE PIE INPUT
```

Figure B-2—continued

```
MATHA_PIE <- cbind(ITEM_NUM_MATH, MATHA_parm_3PL[,1],
MATHA_parm_3PL[,2],MATHA_parm_3PL[,3])

MATHB_PIE <- cbind(ITEM_NUM_MATH, MATHB_parm_3PL[,1],
MATHB_parm_3PL[,2],MATHB_parm_3PL[,3])

SCIENCEA_PIE <- cbind(ITEM_NUM_SCIENCE, SCIENCEA_parm_3PL[,1],
SCIENCEA_parm_3PL[,2],SCIENCEA_parm_3PL[,3])

SCIENCEB_PIE <- cbind(ITEM_NUM_SCIENCE, SCIENCEB_parm_3PL[,1],
SCIENCEB_parm_3PL[,2],SCIENCEB_parm_3PL[,3])

SOCIALA_PIE <- cbind(ITEM_NUM_SOCIAL, SOCIALA_parm_3PL[,1],
SOCIALA_parm_3PL[,2],SOCIALA_parm_3PL[,3])

SOCIALB_PIE <- cbind(ITEM_NUM_SOCIAL, SOCIALB_parm_3PL[,1],
SOCIALB_parm_3PL[,2],SOCIALB_parm_3PL[,3])

MATH_PIE_IN <- rbind(MATHA_PIE,MATHB_PIE)
SCIENCE_PIE_IN <- rbind(SCIENCEA_PIE,SCIENCEB_PIE)
SOCIAL_PIE_IN <- rbind(SOCIALA_PIE,SOCIALB_PIE)

#write.table(MATH_PIE_IN, file="I:\\FINAL_MMATH_PIE_IN.txt", sep="\t",
row.names=FALSE, col.names=FALSE)

#write.table(SCIENCE_PIE_IN, file="I:\\FINAL_MSCIENCE_PIE_IN.txt", sep="\t",
row.names=FALSE, col.names=FALSE)

#write.table(SOCIAL_PIE_IN, file="I:\\FINAL_MSOCIAL_PIE_IN.txt", sep="\t",
row.names=FALSE, col.names=FALSE)

#QUADRATURE POINTS AND WEIGHTS FOR PIE INPUT

#install.packages("MASS")
#install.packages("mvtnorm")
```

Figure B-2—continued

```
library(MASS)
library(mvtnorm)

theta_2d <- expand.grid(theta1=c(seq(-4,4,by=0.20)), theta2=c(seq(-4,4,by=0.20)))
theta_2d <- data.matrix(theta_2d)

theta_4d <- expand.grid(theta1=c(seq(-4,4,by=0.20)), theta2=c(seq(-4,4,by=0.20)),
theta3=c(seq(-4,4,by=0.20)), theta4=c(seq(-4,4,by=0.20)))
theta_4d <- data.matrix(theta_4d)

#MATH

quadwts_MATH <- matrix(0,nrow=dim(theta_4d)[1],ncol=1)
for(j in 1:dim(theta_4d)[1])
    {
quadwts_MATH[j,1] <- dmvnorm(theta_4d[j,], mean = rep(0, dim(theta_4d)[2]),
MATHA_SIGMA, log = FALSE)
    }
quadpts_MATH <- theta_4d %*% MATHA_alpha
QUAD_MATH <- cbind(quadpts_MATH,quadwts_MATH)
QUAD_MATH=QUAD_MATH[order(QUAD_MATH[,1]), ]
QUAD_MATH=QUAD_MATH[-dim(QUAD_MATH)[1],]
QUAD_MATH <- data.frame(QUAD_MATH)
QUAD_MATH$INDICATOR=rep(1:40,each=(dim(QUAD_MATH)[1])/40)
FINAL_QUAD_WTS_MATH <- tapply(QUAD_MATH[,2], QUAD_MATH$INDICATOR,
sum)
FINAL_QUAD_PTS_MATH <- tapply(QUAD_MATH[,1], QUAD_MATH$INDICATOR,
mean)
FINAL_QUAD_MATH <- cbind(FINAL_QUAD_PTS_MATH,
FINAL_QUAD_WTS_MATH)

#SCIENCE

quadwts_SCIENCE <- matrix(0,nrow=dim(theta_2d)[1],ncol=1)
```

Figure B-2—continued

```
for(j in 1:dim(theta_2d)[1])
    {
quadwts_SCIENCE[j,1] <- dmvnorm(theta_2d[j,], mean = rep(0, dim(theta_2d)[2]),
SCIENCEA_SIGMA, log = FALSE)
    }
quadpts_SCIENCE <- theta_2d %*% SCIENCEA_alpha
QUAD_SCIENCE <- cbind(quadpts_SCIENCE,quadwts_SCIENCE)
QUAD_SCIENCE=QUAD_SCIENCE[order(QUAD_SCIENCE[,1]), ]
QUAD_SCIENCE=QUAD_SCIENCE[-dim(QUAD_SCIENCE)[1],]
QUAD_SCIENCE <- data.frame(QUAD_SCIENCE)
QUAD_SCIENCE$INDICATOR=rep(1:40,each=(dim(QUAD_SCIENCE)[1])/40)
FINAL_QUAD_WTS_SCIENCE <- tapply(QUAD_SCIENCE[,2],
QUAD_SCIENCE$INDICATOR, sum)
FINAL_QUAD_PTS_SCIENCE <- tapply(QUAD_SCIENCE[,1],
QUAD_SCIENCE$INDICATOR, mean)
FINAL_QUAD_SCIENCE <- cbind(FINAL_QUAD_PTS_SCIENCE,
FINAL_QUAD_WTS_SCIENCE)

#SOCIAL STUDIES

quadwts_SOCIAL <- matrix(0,nrow=dim(theta_4d)[1],ncol=1)
for(j in 1:dim(theta_4d)[1])
    {
quadwts_SOCIAL[j,1] <- dmvnorm(theta_4d[j,], mean = rep(0, dim(theta_4d)[2]),
SOCIALA_SIGMA, log = FALSE)
    }
quadpts_SOCIAL <- theta_4d %*% SOCIALA_alpha
QUAD_SOCIAL <- cbind(quadpts_SOCIAL,quadwts_SOCIAL)
QUAD_SOCIAL=QUAD_SOCIAL[order(QUAD_SOCIAL[,1]), ]
QUAD_SOCIAL=QUAD_SOCIAL[-dim(QUAD_SOCIAL)[1],]
QUAD_SOCIAL <- data.frame(QUAD_SOCIAL)
QUAD_SOCIAL$INDICATOR=rep(1:40,each=(dim(QUAD_SOCIAL)[1])/40)
FINAL_QUAD_WTS_SOCIAL <- tapply(QUAD_SOCIAL[,2],
QUAD_SOCIAL$INDICATOR, sum)
```

Figure B-2—continued

```
FINAL_QUAD_PTS_SOCIAL <- tapply(QUAD_SOCIAL[,1],
QUAD_SOCIAL$INDICATOR, mean)

FINAL_QUAD_SOCIAL <- cbind(FINAL_QUAD_PTS_SOCIAL,
FINAL_QUAD_WTS_SOCIAL)


#write.table(FINAL_QUAD_MATH, file="I:\\QUAD_MATH.txt", sep="\t",
row.names=FALSE, col.names=FALSE)

#write.table(FINAL_QUAD_SCIENCE, file="I:\\QUAD_SCIENCE.txt", sep="\t",
row.names=FALSE, col.names=FALSE)

#write.table(FINAL_QUAD_SOCIAL, file="I:\\QUAD_SOCIAL.txt", sep="\t",
row.names=FALSE, col.names=FALSE)


#FULL MIRT PROCEDURE


MATHA_SIGMA <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=4,byrow=T)

MATHB_SIGMA <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=4,byrow=T)

SCIENCEA_SIGMA <- matrix(c(1,0,0,1),ncol=2,byrow=T)

SCIENCEB_SIGMA <- matrix(c(1,0,0,1),ncol=2,byrow=T)

SOCIALA_SIGMA <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=4,byrow=T)

SOCIALB_SIGMA <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=4,byrow=T)


#PROBS WILL BE A (40)^d x NUMBER_ITEMS MATRIX CONTAINING
PROBABILITIES FOR EACH THETA COMBINATION (ROW) BY EACH ITEM
(COLUMNS)


theta_2d <- expand.grid(theta1=c(seq(-4,4,by=0.20)), theta2=c(seq(-4,4,by=0.20)))

theta_4d <- expand.grid(theta1=c(seq(-4,4,by=0.20)), theta2=c(seq(-4,4,by=0.20)),


#install.packages("MASS")

#install.packages("mvtnorm")


library(MASS)

library(mvtnorm)


#MATHA
```

Figure B-2—continued

```
probs <- matrix(0,nrow=dim(theta_4d)[1],ncol=dim(MATHA)[1])
for(j in 1:dim(theta_4d)[1]) #LOOPS FOR ALL ROWS OF THETA COMBINATIONS
    {
for(i in 1:dim(MATHA)[1])
    {
probs[j,i] <- MATHA[i,6] + (1-(MATHA[i,6])) *
pnorm((MATHA[i,1])*(theta_4d[j,1]) + (MATHA[i,2])*(theta_4d[j,2]) +
(MATHA[i,3])*(theta_4d[j,3]) + (MATHA[i,4])*(theta_4d[j,4]) + (MATHA[i,5]),
mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
    }
    }
#WE HAVE PROBABILITIES (PROBS), NOW WE NEED TO CALCULATE LORD-
WINGERSKY
#px = Prob X=x|theta, nrows = number of combinations of theta (quad points), ncols = number
of items + 1 (could score 0)
px <- matrix(0,nrow=dim(theta_4d),ncol=((dim(MATHA)[1])+1))
for(j in 1:dim(theta_4d)[1])
    {
for(i in 1:dim(MATHA)[1])
    {
if(i == 1)
        {
px[j,1] = 1 - probs[j,i]
px[j,2] = probs[j,i]
        }
else
        {
vector=c(rep(0,i+1))
vector[1]       =px[j,i-(i-1)]*(1-probs[j,i])
vector[i+1]     =px[j,i]*probs[j,i]
for(x in 2:i)
                {
vector[x] = px[j,x]*(1-probs[j,i])+px[j,x-1]*probs[j,i]
```

Figure B-2—continued

```
                    }
px[j,1:(i+1)] = vector
          }
      }
      }
#DEFINE THETA DISTRIBUTION
quadwts_MATH <- matrix(0,nrow=dim(theta_4d)[1],ncol=1)
for(j in 1:dim(theta_4d)[1])
      {
quadwts_MATH[j,1] <- dmvnorm(theta_4d[j,], mean = rep(0, dim(theta_4d)[2]),
MATHA_SIGMA, log = FALSE)
      }
#MULTIPLY THETA DISTRIBUTION (quadwts) AND CONDITIONAL DISTRIBUTIONS
(px) TO FORM JOINT DISTRIBUTION (joint_dist)
joint_dist <- matrix(0,nrow=dim(theta_4d),ncol=((dim(MATHA)[1])+1))
joint_dist = px*quadwts_MATH[,]
#OBTAIN MARGINAL OBSERVED SCORE DISTRIBUTION (marginal)
marginal_MATHA <- matrix(0,nrow=1,ncol=((dim(MATHA)[1])+1))
for(j in 1:dim(joint_dist)[1])
      {
for(i in 1:dim(joint_dist)[2])
      {
marginal_MATHA[,i]=sum(joint_dist[,i])
      }
      }
marginal_MATHA <- round(10000*marginal_MATHA)

#MATHB

probs <- matrix(0,nrow=dim(theta_4d)[1],ncol=dim(MATHB)[1])
for(j in 1:dim(theta_4d)[1]) #LOOPS FOR ALL ROWS OF THETA COMBINATIONS
      {
for(i in 1:dim(MATHB)[1])
      {
```

Figure B-2—continued

```
probs[j,i] <- MATHB[i,6] + (1-(MATHB[i,6])) *
pnorm((MATHB[i,1])*(theta_4d[j,1]) + (MATHB[i,2])*(theta_4d[j,2]) +
(MATHB[i,3])*(theta_4d[j,3]) + (MATHB[i,4])*(theta_4d[j,4]) + (MATHB[i,5]),
mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
    }
    }
#WE HAVE PROBABILITIES (PROBS), NOW WE NEED TO CALCULATE LORD-
WINGERSKY
#px = Prob X=x|theta, nrows = number of combinations of theta (quad points), ncols = number
of items + 1 (could score 0)
px <- matrix(0,nrow=dim(theta_4d),ncol=((dim(MATHB)[1])+1))
for(j in 1:dim(theta_4d)[1])
    {
for(i in 1:dim(MATHB)[1])
    {
if(i == 1)
        {
px[j,1] = 1 - probs[j,i]
px[j,2] = probs[j,i]
        }
else
        {
vector=c(rep(0,i+1))
vector[1]       =px[j,i-(i-1)]*(1-probs[j,i])
vector[i+1]     =px[j,i]*probs[j,i]
for(x in 2:i)
                {
vector[x] = px[j,x]*(1-probs[j,i])+px[j,x-1]*probs[j,i]
                }
px[j,1:(i+1)] = vector
        }
    }
    }
#DEFINE THETA DISTRIBUTION
```

Figure B-2—continued

```
quadwts_MATH <- matrix(0,nrow=dim(theta_4d)[1],ncol=1)
for(j in 1:dim(theta_4d)[1])
    {
quadwts_MATH[j,1] <- dmvnorm(theta_4d[j,], mean = rep(0, dim(theta_4d)[2]),
MATHB_SIGMA, log = FALSE)
    }
#MULTIPLY THETA DISTRIBUTION (quadwts) AND CONDITIONAL DISTRIBUTIONS
(px) TO FORM JOINT DISTRIBUTION (joint_dist)
joint_dist <- matrix(0,nrow=dim(theta_4d),ncol=((dim(MATHB)[1])+1))
joint_dist = px*quadwts_MATH[,]
#OBTAIN MARGINAL OBSERVED SCORE DISTRIBUTION (marginal)
marginal_MATHB <- matrix(0,nrow=1,ncol=((dim(MATHB)[1])+1))
for(j in 1:dim(joint_dist)[1])
    {
for(i in 1:dim(joint_dist)[2])
    {
marginal_MATHB[,i]=sum(joint_dist[,i])
    }
    }
marginal_MATHB <- round(10000*marginal_MATHB)

#SCIENCEA

probs <- matrix(0,nrow=dim(theta_2d)[1],ncol=dim(SCIENCEA)[1])
for(j in 1:dim(theta_2d)[1]) #LOOPS FOR ALL ROWS OF THETA COMBINATIONS
    {
for(i in 1:dim(SCIENCEA)[1])
    {
probs[j,i] <- SCIENCEA[i,4] + (1-(SCIENCEA[i,4])) *
pnorm((SCIENCEA[i,1])*(theta_2d[j,1]) + (SCIENCEA[i,2])*(theta_2d[j,2]) +
(SCIENCEA[i,3]),
mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
    }
    }
```

Figure B-2—continued

```
#WE HAVE PROBABILITIES (PROBS), NOW WE NEED TO CALCULATE LORD-
WINGERSKY
#px = Prob X=x|theta, nrows = number of combinations of theta (quad points), ncols = number
of items + 1 (could score 0)
px <- matrix(0,nrow=dim(theta_2d),ncol=((dim(SCIENCEA)[1])+1))
for(j in 1:dim(theta_2d)[1])
    {
for(i in 1:dim(SCIENCEA)[1])
    {
if(i == 1)
        {
px[j,1] = 1 - probs[j,i]
px[j,2] = probs[j,i]
        }
else
        {
vector=c(rep(0,i+1))
vector[1]       =px[j,i-(i-1)]*(1-probs[j,i])
vector[i+1]     =px[j,i]*probs[j,i]
for(x in 2:i)
                {
vector[x] = px[j,x]*(1-probs[j,i])+px[j,x-1]*probs[j,i]
                }
px[j,1:(i+1)] = vector
        }
    }
    }
#DEFINE THETA DISTRIBUTION
quadwts_SCIENCE <- matrix(0,nrow=dim(theta_2d)[1],ncol=1)
for(j in 1:dim(theta_2d)[1])
    {
quadwts_SCIENCE[j,1] <- dmvnorm(theta_2d[j,], mean = rep(0, dim(theta_2d)[2]),
SCIENCEA_SIGMA, log = FALSE)
    }
```

Figure B-2—continued

```
#MULTIPLY THETA DISTRIBUTION (quadwts) AND CONDITIONAL DISTRIBUTIONS
(px) TO FORM JOINT DISTRIBUTION (joint_dist)
joint_dist <- matrix(0,nrow=dim(theta_2d),ncol=((dim(SCIENCEA)[1])+1))
joint_dist = px*quadwts_SCIENCE[,]
#OBTAIN MARGINAL OBSERVED SCORE DISTRIBUTION (marginal)
marginal_SCIENCEA <- matrix(0,nrow=1,ncol=((dim(SCIENCEA)[1])+1))
for(j in 1:dim(joint_dist)[1])
    {
for(i in 1:dim(joint_dist)[2])
    {
marginal_SCIENCEA[,i]=sum(joint_dist[,i])
    }
    }
marginal_SCIENCEA <- round(10000*marginal_SCIENCEA)

#SCIENCEB

probs <- matrix(0,nrow=dim(theta_2d)[1],ncol=dim(SCIENCEB)[1])
for(j in 1:dim(theta_2d)[1]) #LOOPS FOR ALL ROWS OF THETA COMBINATIONS
    {
for(i in 1:dim(SCIENCEB)[1])
    {
probs[j,i] <- SCIENCEB[i,4] + (1-(SCIENCEB[i,4])) *
pnorm((SCIENCEB[i,1])*(theta_2d[j,1]) + (SCIENCEB[i,2])*(theta_2d[j,2]) +
(SCIENCEB[i,3]),
mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
    }
    }
#WE HAVE PROBABILITIES (PROBS), NOW WE NEED TO CALCULATE LORD-
WINGERSKY
#px = Prob X=x|theta, nrows = number of combinations of theta (quad points), ncols = number
of items + 1 (could score 0)
px <- matrix(0,nrow=dim(theta_2d),ncol=((dim(SCIENCEB)[1])+1))
for(j in 1:dim(theta_2d)[1])
```

Figure B-2—continued

```
    {
for(i in 1:dim(SCIENCEB)[1])
    {
if(i == 1)
        {
px[j,1] = 1 - probs[j,i]
px[j,2] = probs[j,i]
        }
else
        {
vector=c(rep(0,i+1))
vector[1]       =px[j,i-(i-1)]*(1-probs[j,i])
vector[i+1]     =px[j,i]*probs[j,i]
for(x in 2:i)
                {
vector[x] = px[j,x]*(1-probs[j,i])+px[j,x-1]*probs[j,i]
                }
px[j,1:(i+1)] = vector
        }
    }
    }
#DEFINE THETA DISTRIBUTION
quadwts_SCIENCE <- matrix(0,nrow=dim(theta_2d)[1],ncol=1)
for(j in 1:dim(theta_2d)[1])
    {
quadwts_SCIENCE[j,1] <- dmvnorm(theta_2d[j,], mean = rep(0, dim(theta_2d)[2]),
SCIENCEB_SIGMA, log = FALSE)
    }
#MULTIPLY THETA DISTRIBUTION (quadwts) AND CONDITIONAL DISTRIBUTIONS
(px) TO FORM JOINT DISTRIBUTION (joint_dist)
joint_dist <- matrix(0,nrow=dim(theta_2d),ncol=((dim(SCIENCEB)[1])+1))
joint_dist = px*quadwts_SCIENCE[,]
#OBTAIN MARGINAL OBSERVED SCORE DISTRIBUTION (marginal)
marginal_SCIENCEB <- matrix(0,nrow=1,ncol=((dim(SCIENCEB)[1])+1))
```

Figure B-2—continued

```
for(j in 1:dim(joint_dist)[1])
    {
for(i in 1:dim(joint_dist)[2])
    {
marginal_SCIENCEB[,i]=sum(joint_dist[,i])
    }
    }
marginal_SCIENCEB <- round(10000*marginal_SCIENCEB)

#SOCIALA

probs <- matrix(0,nrow=dim(theta_4d)[1],ncol=dim(SOCIALA)[1])
for(j in 1:dim(theta_4d)[1]) #LOOPS FOR ALL ROWS OF THETA COMBINATIONS
    {
for(i in 1:dim(SOCIALA)[1])
    {
probs[j,i] <- SOCIALA[i,6] + (1-(SOCIALA[i,6])) *
pnorm((SOCIALA[i,1])*(theta_4d[j,1]) + (SOCIALA[i,2])*(theta_4d[j,2]) +
(SOCIALA[i,3])*(theta_4d[j,3]) + (SOCIALA[i,4])*(theta_4d[j,4]) + (SOCIALA[i,5]),
mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
    }
    }
#WE HAVE PROBABILITIES (PROBS), NOW WE NEED TO CALCULATE LORD-
WINGERSKY
#px = Prob X=x|theta, nrows = number of combinations of theta (quad points), ncols = number
of items + 1 (could score 0)
px <- matrix(0,nrow=dim(theta_4d),ncol=((dim(SOCIALA)[1])+1))
for(j in 1:dim(theta_4d)[1])
    {
for(i in 1:dim(SOCIALA)[1])
    {
if(i == 1)
        {
px[j,1] = 1 - probs[j,i]
```

Figure B-2—continued

```
px[j,2] = probs[j,i]
        }
else
        {
vector=c(rep(0,i+1))
vector[1]       =px[j,i-(i-1)]*(1-probs[j,i])
vector[i+1]     =px[j,i]*probs[j,i]
for(x in 2:i)
                {
vector[x] = px[j,x]*(1-probs[j,i])+px[j,x-1]*probs[j,i]
                }
px[j,1:(i+1)] = vector
        }
    }
    }
#DEFINE THETA DISTRIBUTION
quadwts_SOCIAL <- matrix(0,nrow=dim(theta_4d)[1],ncol=1)
for(j in 1:dim(theta_4d)[1])
    {
quadwts_SOCIAL[j,1] <- dmvnorm(theta_4d[j,], mean = rep(0, dim(theta_4d)[2]),
SOCIALA_SIGMA, log = FALSE)
    }
#MULTIPLY THETA DISTRIBUTION (quadwts) AND CONDITIONAL DISTRIBUTIONS
(px) TO FORM JOINT DISTRIBUTION (joint_dist)
joint_dist <- matrix(0,nrow=dim(theta_4d),ncol=((dim(SOCIALA)[1])+1))
joint_dist = px*quadwts_SOCIAL[,]
#OBTAIN MARGINAL OBSERVED SCORE DISTRIBUTION (marginal)
marginal_SOCIALA <- matrix(0,nrow=1,ncol=((dim(SOCIALA)[1])+1))
for(j in 1:dim(joint_dist)[1])
    {
for(i in 1:dim(joint_dist)[2])
    {
marginal_SOCIALA[,i]=sum(joint_dist[,i])
    }
```

Figure B-2—continued

```
    }
marginal_SOCIALA <- round(10000*marginal_SOCIALA)


#SOCIALB


probs <- matrix(0,nrow=dim(theta_4d)[1],ncol=dim(SOCIALB)[1])
for(j in 1:dim(theta_4d)[1]) #LOOPS FOR ALL ROWS OF THETA COMBINATIONS
    {
for(i in 1:dim(SOCIALB)[1])
    {
probs[j,i] <- SOCIALB[i,6] + (1-(SOCIALB[i,6])) *
pnorm((SOCIALB[i,1])*(theta_4d[j,1]) + (SOCIALB[i,2])*(theta_4d[j,2]) +
(SOCIALB[i,3])*(theta_4d[j,3]) + (SOCIALB[i,4])*(theta_4d[j,4]) + (SOCIALB[i,5]),
mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
    }
    }
#WE HAVE PROBABILITIES (PROBS), NOW WE NEED TO CALCULATE LORD-
WINGERSKY
#px = Prob X=x|theta, nrows = number of combinations of theta (quad points), ncols = number
of items + 1 (could score 0)
px <- matrix(0,nrow=dim(theta_4d),ncol=((dim(SOCIALB)[1])+1))
for(j in 1:dim(theta_4d)[1])
    {
for(i in 1:dim(SOCIALB)[1])
    {
if(i == 1)
        {
px[j,1] = 1 - probs[j,i]
px[j,2] = probs[j,i]
        }
else
        {
vector=c(rep(0,i+1))
vector[1]        =px[j,i-(i-1)]*(1-probs[j,i])
```

Figure B-2—continued

```
vector[i+1]      =px[j,i]*probs[j,i]
for(x in 2:i)
                 {
vector[x] = px[j,x]*(1-probs[j,i])+px[j,x-1]*probs[j,i]
                 }
px[j,1:(i+1)] = vector
        }
    }
    }
#DEFINE THETA DISTRIBUTION
quadwts_SOCIAL <- matrix(0,nrow=dim(theta_4d)[1],ncol=1)
for(j in 1:dim(theta_4d)[1])
    {
quadwts_SOCIAL[j,1] <- dmvnorm(theta_4d[j,], mean = rep(0, dim(theta_4d)[2]),
SOCIALB_SIGMA, log = FALSE)
    }
#MULTIPLY THETA DISTRIBUTION (quadwts) AND CONDITIONAL DISTRIBUTIONS
(px) TO FORM JOINT DISTRIBUTION (joint_dist)
joint_dist <- matrix(0,nrow=dim(theta_4d),ncol=((dim(SOCIALB)[1])+1))
joint_dist = px*quadwts_SOCIAL[,]
#OBTAIN MARGINAL OBSERVED SCORE DISTRIBUTION (marginal)
marginal_SOCIALB <- matrix(0,nrow=1,ncol=((dim(SOCIALB)[1])+1))
for(j in 1:dim(joint_dist)[1])
    {
for(i in 1:dim(joint_dist)[2])
    {
marginal_SOCIALB[,i]=sum(joint_dist[,i])
    }
    }
marginal_SOCIALB <- round(10000*marginal_SOCIALB)

MATH_ITEM_ID <- matrix(rep(0:40),ncol=1)
SCIENCE_ITEM_ID <- matrix(rep(0:48),ncol=1)
SOCIAL_ITEM_ID <- matrix(rep(0:50),ncol=1)
```

Figure B-2—continued

```
FULLMATH <- cbind(MATH_ITEM_ID, t(marginal_MATHA), t(marginal_MATHB),
MATH_ITEM_ID)

FULLSCIENCE <- cbind(SCIENCE_ITEM_ID, t(marginal_SCIENCEA),
t(marginal_SCIENCEB), SCIENCE_ITEM_ID)

FULLSOCIAL <- cbind(SOCIAL_ITEM_ID, t(marginal_SOCIALA), t(marginal_SOCIALB),
SOCIAL_ITEM_ID)


#write.table(FULLMATH, file="I:\\FINAL_FULLMATH.txt", sep="\t", row.names=FALSE,
col.names=FALSE)

#write.table(FULLSCIENCE, file="I:\\FINAL_FULLSCIENCE.txt", sep="\t",
row.names=FALSE, col.names=FALSE)

#write.table(FULLSOCIAL, file="I:\\FINAL_FULLSOCIAL.txt", sep="\t",
row.names=FALSE, col.names=FALSE)


#PLOTS


#install.packages("fields")


data <- read.csv(file="E:\\IOWA\\1DIS.Equating Results.csv", header=TRUE)


attach(data)

library(fields)


#par(mfrow=c(3,2))


#UNSMOOTHED IDENTITY DIFFERENCE PLOTS


plot(MATHRAW,IMATHEQUI,type="o",main="Math",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Form A Raw Score",xlim=c(0,40),ylim=c(-5,4),pch=15)

lines(IMATHOBS,type="o",lty=1,pch=1)

lines(IMATHUAOBS,type="o",lty=1,pch=2)

lines(IMATHFULL,type="o",lty=1,pch=3)

lines(IMATHTRUE,type="o",lty=1,pch=5)

lines(IMATHUATRUE,type="o",lty=1,pch=6)
```

Figure B-2—continued

```
lines(IMATHBELOW,lty=2)

lines(IMATHABOVE,lty=2)

legend(6.8, 4.2, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(0,lty=2)


plot(SCIENCERAW,ISCIENCEEQUI,type="o",main="Science",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Form A Raw Score",xlim=c(0,48),ylim=c(-5,4),pch=15)

lines(ISCIENCEOBS,type="o",lty=1,pch=1)

lines(ISCIENCEUAOBS,type="o",lty=1,pch=2)

lines(ISCIENCEFULL,type="o",lty=1,pch=3)

lines(ISCIENCETRUE,type="o",lty=1,pch=5)

lines(ISCIENCEUATRUE,type="o",lty=1,pch=6)

lines(ISCIENCEBELOW,lty=2)

lines(ISCIENCEABOVE,lty=2)

legend(6.8, 4.2, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(0,lty=2)


plot(SOCIALRAW,ISOCIALEQUI,type="o",main="Social Studies",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Form A Raw Score",xlim=c(0,50),ylim=c(-5,4),pch=15)

lines(ISOCIALOBS,type="o",lty=1,pch=1)

lines(ISOCIALUAOBS,type="o",lty=1,pch=2)

lines(ISOCIALFULL,type="o",lty=1,pch=3)

lines(ISOCIALTRUE,type="o",lty=1,pch=5)

lines(ISOCIALUATRUE,type="o",lty=1,pch=6)

lines(ISOCIALBELOW,lty=2)

lines(ISOCIALABOVE,lty=2)

legend(6.8, 4.2, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(0,lty=2)
```

Figure B-2—continued

```
#SMOOTHED IDENTITY DIFFERENCE PLOTS

plot(MATHRAW,IMATHEQUISMOOTH,type="o",main="Math",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Form A Raw Score",xlim=c(0,40),ylim=c(-5,4),pch=15)
lines(IMATHOBS,type="o",lty=1,pch=1)
lines(IMATHUAOBS,type="o",lty=1,pch=2)
lines(IMATHFULL,type="o",lty=1,pch=3)
lines(IMATHTRUE,type="o",lty=1,pch=5)
lines(IMATHUATRUE,type="o",lty=1,pch=6)
lines(IMATHBELOWSMOOTH,lty=2)
lines(IMATHABOVESMOOTH,lty=2)
legend(6.8, 4.2, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));
yline(0,lty=2)

plot(SCIENCERAW,ISCIENCEEQUISMOOTH,type="o",main="Science",xlab="Raw Score
Form A",ylab="Form B Equivalent Minus Form A Raw Score", xlim=c(0,48), ylim=c(-
5,4),pch=15)
lines(ISCIENCEOBS,type="o",lty=1,pch=1)
lines(ISCIENCEUAOBS,type="o",lty=1,pch=2)
lines(ISCIENCEFULL,type="o",lty=1,pch=3)
lines(ISCIENCETRUE,type="o",lty=1,pch=5)
lines(ISCIENCEUATRUE,type="o",lty=1,pch=6)
lines(ISCIENCEBELOWSMOOTH,lty=2)
lines(ISCIENCEABOVESMOOTH,lty=2)
legend(6.8, 4.2, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));
yline(0,lty=2)

plot(SOCIALRAW,ISOCIALEQUISMOOTH,type="o",main="Social Studies", xlab="Raw
Score Form A",ylab="Form B Equivalent Minus Form A Raw Score", xlim=c(0,50),ylim=c(-
5,4),pch=15)
lines(ISOCIALOBS,type="o",lty=1,pch=1)
lines(ISOCIALUAOBS,type="o",lty=1,pch=2)
```

Figure B-2—continued

```
lines(ISOCIALFULL,type="o",lty=1,pch=3)

lines(ISOCIALTRUE,type="o",lty=1,pch=5)

lines(ISOCIALUATRUE,type="o",lty=1,pch=6)

lines(ISOCIALBELOWSMOOTH,lty=2)

lines(ISOCIALABOVESMOOTH,lty=2)

legend(6.8, 4.2, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(0,lty=2)


#UNSMOOTHED EQUIPERCENTILE DIFFERENCE PLOTS


plot(MATHRAW,MATHZERO,type="o",main="Math",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Equipercentile Equivalent",xlim=c(0,40),ylim=c(-
2,3),pch=15)

lines(EMATHOBS,type="o",lty=1,pch=1)

lines(EMATHUAOBS,type="o",lty=1,pch=2)

lines(EMATHFULL,type="o",lty=1,pch=3)

lines(EMATHTRUE,type="o",lty=1,pch=5)

lines(EMATHUATRUE,type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(c(-0.5,0.5),lty=2)


plot(MATHRAW,MATHZERO,type="o",main="Math",xlab="Raw Score Form A",
ylab="Absolute Value of Form B Equivalent Minus Equipercentile Equivalent",
xlim=c(0,40),ylim=c(0,3),pch=15)

lines(abs(EMATHOBS),type="o",lty=1,pch=1)

lines(abs(EMATHUAOBS),type="o",lty=1,pch=2)

lines(abs(EMATHFULL),type="o",lty=1,pch=3)

lines(abs(EMATHTRUE),type="o",lty=1,pch=5)

lines(abs(EMATHUATRUE),type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));
```

Figure B-2—continued

```
yline(0.5,lty=2)

plot(SCIENCERAW,SCIENCEZERO,type="o",main="Science",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Equipercentile Equivalent",xlim=c(0,48),ylim=c(-
2,3),pch=15)
lines(ESCIENCEOBS,type="o",lty=1,pch=1)
lines(ESCIENCEUAOBS,type="o",lty=1,pch=2)
lines(ESCIENCEFULL,type="o",lty=1,pch=3)
lines(ESCIENCETRUE,type="o",lty=1,pch=5)
lines(ESCIENCEUATRUE,type="o",lty=1,pch=6)
legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));
yline(c(-0.5,0.5),lty=2)

plot(SCIENCERAW,SCIENCEZERO,type="o",main="Science",xlab="Raw Score Form
A",ylab="Absolute Value of Form B Equivalent Minus Equipercentile
Equivalent",xlim=c(0,48),ylim=c(0,3),pch=15)
lines(abs(ESCIENCEOBS),type="o",lty=1,pch=1)
lines(abs(ESCIENCEUAOBS),type="o",lty=1,pch=2)
lines(abs(ESCIENCEFULL),type="o",lty=1,pch=3)
lines(abs(ESCIENCETRUE),type="o",lty=1,pch=5)
lines(abs(ESCIENCEUATRUE),type="o",lty=1,pch=6)
legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));
yline(0.5,lty=2)

plot(SOCIALRAW,SOCIALZERO,type="o",main="Social Studies",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Equipercentile Equivalent",xlim=c(0,50),ylim=c(-
2,3),pch=15)
lines(ESOCIALOBS,type="o",lty=1,pch=1)
lines(ESOCIALUAOBS,type="o",lty=1,pch=2)
lines(ESOCIALFULL,type="o",lty=1,pch=3)
lines(ESOCIALTRUE,type="o",lty=1,pch=5)
lines(ESOCIALUATRUE,type="o",lty=1,pch=6)
```

Figure B-2—continued

```
legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(c(-0.5,0.5),lty=2)


plot(SOCIALRAW,SOCIALZERO,type="o",main="Social Studies",xlab="Raw Score Form A",ylab="Absolute Value of Form B Equivalent Minus Equipercentile Equivalent", xlim=c(0,50),ylim=c(0,3),pch=15)

lines(abs(ESOCIALOBS),type="o",lty=1,pch=1)

lines(abs(ESOCIALUAOBS),type="o",lty=1,pch=2)

lines(abs(ESOCIALFULL),type="o",lty=1,pch=3)

lines(abs(ESOCIALTRUE),type="o",lty=1,pch=5)

lines(abs(ESOCIALUATRUE),type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(0.5,lty=2)


#SMOOTHED EQUIPERCENTILE DIFFERENCE PLOTS


plot(MATHRAW,MATHZERO,type="o",main="Math",xlab="Raw Score Form A",ylab="Form B Equivalent Minus Equipercentile Equivalent",xlim=c(0,40),ylim=c(-2,3),pch=15)

lines(EMATHOBSSMOOTH,type="o",lty=1,pch=1)

lines(EMATHUAOBSSMOOTH,type="o",lty=1,pch=2)

lines(EMATHFULLSMOOTH,type="o",lty=1,pch=3)

lines(EMATHTRUESMOOTH,type="o",lty=1,pch=5)

lines(EMATHUATRUESMOOTH,type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(c(-0.5,0.5),lty=2)


plot(MATHRAW,MATHZERO,type="o",main="Math",xlab="Raw Score Form A",ylab="Absolute Value of Form B Equivalent Minus Equipercentile Equivalent", xlim=c(0,41),ylim=c(0,3),pch=15)

lines(abs(EMATHOBSSMOOTH),type="o",lty=1,pch=1)
```

Figure B-2—continued

```
lines(abs(EMATHUAOBSSMOOTH),type="o",lty=1,pch=2)

lines(abs(EMATHFULLSMOOTH),type="o",lty=1,pch=3)

lines(abs(EMATHTRUESMOOTH),type="o",lty=1,pch=5)

lines(abs(EMATHUATRUESMOOTH),type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(0.5,lty=2)


plot(SCIENCERAW,SCIENCEZERO,type="o",main="Science",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Equipercentile Equivalent",xlim=c(0,48),ylim=c(-
2,3),pch=15)

lines(ESCIENCEOBSSMOOTH,type="o",lty=1,pch=1)

lines(ESCIENCEUAOBSSMOOTH,type="o",lty=1,pch=2)

lines(ESCIENCEFULLSMOOTH,type="o",lty=1,pch=3)

lines(ESCIENCETRUESMOOTH,type="o",lty=1,pch=5)

lines(ESCIENCEUATRUESMOOTH,type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(c(-0.5,0.5),lty=2)


plot(SCIENCERAW,SCIENCEZERO,type="o",main="Science",xlab="Raw Score Form
A",ylab="Absolute Value of Form B Equivalent Minus Equipercentile
Equivalent",xlim=c(0,48),ylim=c(0,3),pch=15)

lines(abs(ESCIENCEOBSSMOOTH),type="o",lty=1,pch=1)

lines(abs(ESCIENCEUAOBSSMOOTH),type="o",lty=1,pch=2)

lines(abs(ESCIENCEFULLSMOOTH),type="o",lty=1,pch=3)

lines(abs(ESCIENCETRUESMOOTH),type="o",lty=1,pch=5)

lines(abs(ESCIENCEUATRUESMOOTH),type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(0.5,lty=2)
```

Figure B-2—continued

```
plot(SOCIALRAW,SOCIALZERO,type="o",main="Social Studies",xlab="Raw Score Form
A",ylab="Form B Equivalent Minus Equipercentile Equivalent", xlim=c(0,50),ylim=c(-
2,3),pch=15)

lines(ESOCIALOBSSMOOTH,type="o",lty=1,pch=1)

lines(ESOCIALUAOBSSMOOTH,type="o",lty=1,pch=2)

lines(ESOCIALFULLSMOOTH,type="o",lty=1,pch=3)

lines(ESOCIALTRUESMOOTH,type="o",lty=1,pch=5)

lines(ESOCIALUATRUESMOOTH,type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(c(-0.5,0.5),lty=2)


plot(SOCIALRAW,SOCIALZERO,type="o",main="Social Studies",xlab="Raw Score Form
A",ylab="Absolute Value of Form B Equivalent Minus Equipercentile Equivalent",
xlim=c(0,50),ylim=c(0,3),pch=15)

lines(abs(ESOCIALOBSSMOOTH),type="o",lty=1,pch=1)

lines(abs(ESOCIALUAOBSSMOOTH),type="o",lty=1,pch=2)

lines(abs(ESOCIALFULLSMOOTH),type="o",lty=1,pch=3)

lines(abs(ESOCIALTRUESMOOTH),type="o",lty=1,pch=5)

lines(abs(ESOCIALUATRUESMOOTH),type="o",lty=1,pch=6)

legend(5, 3.0, c("Equipercentile","IRT Observed Score","Unidimensional Approximation of
MIRT Observed Score","Full MIRT Observed Score","IRT True Score","Unidimensional
Approximation of MIRT True Score"), cex=0.8, pch=c(15,1,2,3,5,6));

yline(0.5,lty=2)


#HISTOGRAMS


par(mfrow=c(2,1))

barplot(data$AMFREQ,ylim=c(0,0.08),density=c(rep(100,50)),names.arg=data$RAWM,main=
"Math Form A",xlab=" ",ylab="Relative Frequency")

box()

barplot(data$BMFREQ,ylim=c(0,0.08),density=c(rep(100,50)),names.arg=data$RAWM,main=
"Math Form B",xlab=" ",ylab="Relative Frequency")

box()
```

Figure B-2—continued

```
par(mfrow=c(2,1))
barplot(data$ASCIFREQ,ylim=c(0,0.08),density=c(rep(100,50)),names.arg=data$RAWSCI,ma
in="Science Form A",xlab=" ",ylab="Relative Frequency")
box()
barplot(data$BSCIFREQ,ylim=c(0,0.08),density=c(rep(100,50)),names.arg=data$RAWSCI,mai
n="Science Form B",xlab=" ",ylab="Relative Frequency")
box()

par(mfrow=c(2,1))
barplot(data$ASSFREQ,ylim=c(0,0.08),density=c(rep(100,50)),names.arg=data$RAWSS,main
="Social Studies Form A",xlab=" ",ylab="Relative Frequency")
box()
barplot(data$BSSFREQ,ylim=c(0,0.08),density=c(rep(100,50)),names.arg=data$RAWSS,main
="Social Studies Form B",xlab=" ",ylab="Relative Frequency")
box()

#QUADRATURE PLOTS

#install.packages("fields")

data <- read.csv(file="E:\\IOWA\\1DIS.Equating Results.csv", header=TRUE)

attach(data)
library(fields)

par(mfrow=c(1,2))
plot(UMATHPT,UMATHWT,type="o",main="Unidimensional Math Ability
Distribution",xlab="Theta",ylab="Density",xlim=c(-6,6),ylim=c(0,0.10),pch=1)
yline(0,lty=1)
plot(MMATHPT,MMATHWT,type="o",main="Unidimensional Approximation Math Ability
Distribution",xlab="Theta",ylab="Density",xlim=c(-6,6),ylim=c(0,0.10),pch=1)
yline(0,lty=1)

par(mfrow=c(1,2))
```

Figure B-2—continued

```
plot(USCIENCEPT,USCIENCEWT,type="o",main="Unidimensional Science Ability
Distribution",xlab="Theta",ylab="Density",xlim=c(-6,6),ylim=c(0,0.25),pch=1)

yline(0,lty=1)

plot(MSCIENCEPT,MSCIENCEWT,type="o",main="Unidimensional Approximation Science
Ability Distribution",xlab="Theta",ylab="Density",xlim=c(-6,6),ylim=c(0,0.25),pch=1)

yline(0,lty=1)


par(mfrow=c(1,2))

plot(USOCIALPT,USOCIALWT,type="o",main="Unidimensional Social Studies Ability
Distribution",xlab="Theta",ylab="Density",xlim=c(-6,6),ylim=c(0,0.10),pch=1)

yline(0,lty=1)

plot(MSOCIALPT,MSOCIALWT,type="o",main="Unidimensional Approximation Social
Studies Ability Distribution",xlab="Theta",ylab="Density",xlim=c(-6,6),ylim=c(0,0.10),pch=1)

yline(0,lty=1)
```

APPENDIX C. MULTIDIMENSIONAL EQUATING EXAMPLE

The procedures below provide a step-by-step example of how the unidimensional approximation equating procedures were conducted. The example below uses the first five items for each Science exam.

STEP I: Estimate Multidimensional Item Parameters using TESTFACT

Form A Parameters:

	a_1	a_2	d
Item 1	0.318	1.126	0.391
Item 2	0.404	0.562	-0.350
Item 3	0.990	0.230	0.304
Item 4	0.427	0.233	-0.134
Item 5	0.765	0.370	0.126

Form B Parameters:

	a_1	a_2	d
Item 1	0.415	0.305	-0.704
Item 2	0.952	0.333	0.418
Item 3	1.174	0.115	0.475
Item 4	0.351	0.301	-0.353
Item 5	0.241	1.317	-0.663

STEP II: Estimate Test-Level Direction of Best Measurement for Form A and Form B

$$\alpha_A = \begin{bmatrix} \dfrac{0.318 + 0.404 + 0.990 + 0.427 + 0.765}{\sqrt{(0.318 + \ldots + 0.765)^2 + (1.126 + \ldots + 0.370)^2}} \\ \\ \dfrac{1.126 + 0.562 + 0.230 + 0.233 + 0.370}{\sqrt{(0.318 + \ldots + 0.765)^2 + (1.126 + \ldots + 0.370)^2}} \end{bmatrix} = \begin{bmatrix} 0.755 \\ \\ 0.656 \end{bmatrix}$$

$$\alpha_B = \begin{bmatrix} \dfrac{0.415 + 0.952 + 1.174 + 0.351 + 0.241}{\sqrt{(0.415 + \ldots + 0.241)^2 + (0.305 + \ldots + 1.317)^2}} \\ \\ \dfrac{0.305 + 0.333 + 0.115 + 0.301 + 1.317}{\sqrt{(0.415 + \ldots + 0.241)^2 + (0.305 + \ldots + 1.317)^2}} \end{bmatrix} = \begin{bmatrix} 0.797 \\ \\ 0.603 \end{bmatrix}$$

STEP III: Estimate Unidimensional Item Parameters

Sigma-Squared Values:

$$\hat{\sigma}^2_{A1} = \begin{bmatrix} 0.318 & 1.126 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.318 \\ 1.126 \end{bmatrix} - \left(\begin{bmatrix} 0.318 & 1.126 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.755 \\ 0.656 \end{bmatrix} \right)^2 = 0.411$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\hat{\sigma}^2_{A5} = \begin{bmatrix} 0.765 & 0.370 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.765 \\ 0.370 \end{bmatrix} - \left(\begin{bmatrix} 0.765 & 0.370 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.755 \\ 0.656 \end{bmatrix} \right)^2 = 0.049$$

$$\hat{\sigma}^2_{B1} = \begin{bmatrix} 0.415 & 0.305 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.415 \\ 0.305 \end{bmatrix} - \left(\begin{bmatrix} 0.415 & 0.305 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.797 \\ 0.603 \end{bmatrix} \right)^2 = 0.000$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\hat{\sigma}^2_{B5} = \begin{bmatrix} 0.241 & 1.317 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.241 \\ 1.317 \end{bmatrix} - \left(\begin{bmatrix} 0.241 & 1.317 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.797 \\ 0.603 \end{bmatrix} \right)^2 = 0.820$$

Unidimensional Discrimination Parameters:

$$\hat{a}_{A1} = (1 + 0.411)^{-\frac{1}{2}} \begin{bmatrix} 0.318 & 1.126 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.755 \\ 0.656 \end{bmatrix} = 0.824$$

$$\vdots \qquad \vdots$$

$$\hat{a}_{A5} = (1 + 0.049)^{-\frac{1}{2}} \begin{bmatrix} 0.765 & 0.370 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.755 \\ 0.656 \end{bmatrix} = 0.801$$

$$\hat{a}_{B1} = (1 + 0.000)^{-\frac{1}{2}} \begin{bmatrix} 0.415 & 0.305 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.797 \\ 0.603 \end{bmatrix} = 0.515$$

$$\vdots \qquad \vdots$$

$$\hat{a}_{B5} = (1 + 0.820)^{-\frac{1}{2}} \begin{bmatrix} 0.241 & 1.317 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.797 \\ 0.603 \end{bmatrix} = 0.731$$

Unidimensional Difficulty-"Related" Parameters:

$$d_{A1} = (1 + 0.411)^{-\frac{1}{2}} (0.391) = 0.329$$

$$\vdots$$

$$d_{A5} = (1 + 0.049)^{-\frac{1}{2}} (0.126) = 0.123$$

$$d_{B1} = (1 + 0.000)^{-\frac{1}{2}}(-0.704) = -0.704$$
$$\vdots$$
$$d_{B5} = (1 + 0.820)^{-\frac{1}{2}}(-0.663) = -0.491$$

Unidimensional Difficulty Parameters:

$$b_{A1} = \frac{-0.329}{0.824} = -0.399$$
$$\vdots$$
$$b_{A5} = \frac{-0.123}{0.801} = -0.154$$

$$b_{B1} = \frac{0.704}{0.515} = 1.367$$
$$\vdots$$
$$b_{B5} = \frac{0.491}{0.731} = 0.672$$

STEP IV: Convert Normal Ogive Parameters to Logistic Parameters
(Note: this step may differ depending on whether the logistic model incorporates that constant 1.7 to make the logistic function similar to the normal ogive).

STEP V: Conduct Observed Score or True Score Equating Given Unidimensional Approximation Item Parameters