Comparison of MIRT observed score equating methods under the common-item nonequivalent groups design

Jiwon Choi
University of Iowa

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

Copyright © 2019 Jiwon Choi

This dissertation is available at Iowa Research Online: https://ir.uiowa.edu/etd/6716

Recommended Citation
Choi, Jiwon. "Comparison of MIRT observed score equating methods under the common-item nonequivalent groups design." PhD (Doctor of Philosophy) thesis, University of Iowa, 2019. https://doi.org/10.17077/etd.w4gp-27dt

Follow this and additional works at: https://ir.uiowa.edu/etd
COMPARISON OF MIRT OBSERVED SCORE EQUATING METHODS
UNDER THE COMMON-ITEM NONEQUIVALENT GROUPS DESIGN

by

Jiwon Choi

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

May 2019

Thesis Supervisor: Professor Won-Chan Lee
ACKNOWLEDGEMENTS

Firstly, I would like to thank Dr. Won-Chan Lee, my academic advisor and dissertation chair, for his guidance and support. He encouraged me to progress my doctoral study and provided me with essential psychometrics knowledge. I would also like to thank Dr. Jaime Malatesta, whose insight was critical to the progression of my dissertation. I am also thankful to the invaluable advice of Dr. Ariel Aloe, Dr. Brandon LeBeau, and Dr. Sanvesh Srivastava, all of whom helped make this dissertation possible.

I would like to thank my friends in the Educational Measurement and Statistics program, as they continuously encouraged me with their support and friendship. I would also like to express my gratitude for my colleagues and friends at ACT. It was a genuine pleasure to work with all of you. I would also like to thank my friends in Korea, whose emotional support means more than I can say.

Finally, I am grateful to my family. My accomplishment would not have been possible without your sacrifice, support, and love. I would especially like to thank my parents, my sister, my brother-in-law, and my little nephew, all of whom have shown me what it is to be unconditionally loved. Truly, I would not have reached this point had it not been for your steadfast love and encouragement.
ABSTRACT

For equating tests that measure several distinct proficiencies, procedures that reflect the multidimensional structure of the data are needed. Although there exist a few equating procedures developed under the multidimensional item response theory (MIRT) framework, there is a need for further research in this area. Therefore, the primary objectives of this dissertation are to consolidate and expand MIRT observed score equating research with a specific focus on the common-item nonequivalent groups (CINEG) design, which requires scale linking. Content areas and item types are two focal points of dimensionality. This dissertation uses two studies with different data types and comparison criteria to address the research objectives.

In general, a comparison between unidimensional item response theory (UIRT) and MIRT methods suggested a better performance of the MIRT methods over UIRT. The simple structure (SS) and full MIRT methods showed more accurate equating results compared to UIRT. In terms of calibration methods, concurrent calibration outperformed separate calibration for all equating methods under most of the studied conditions.
PUBLIC ABSTRACT

For equating tests that measure several distinct proficiencies, procedures that reflect the multidimensional structure of the data are needed. Although there exist a few equating procedures developed under the multidimensional item response theory (MIRT) framework, there is a need for further research in this area. Therefore, the primary objectives of this dissertation are to consolidate and expand MIRT observed score equating research with a specific focus on the common-item nonequivalent groups (CINEG) design, which requires scale linking. Content areas and item types are two focal points of dimensionality. This dissertation uses two studies with different data types and comparison criteria to address the research objectives.

In general, a comparison between unidimensional item response theory (UIRT) and MIRT methods suggested a better performance of the MIRT methods over UIRT. The simple structure (SS) and full MIRT methods showed more accurate equating results compared to UIRT. In terms of calibration methods, concurrent calibration outperformed separate calibration for all equating methods under most of the studied conditions.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................... vii

LIST OF FIGURES ......................................................................................................... ix

CHAPTER 1 INTRODUCTION ...................................................................................... 1
   Equating Designs ................................................................................................. 1
   Item Response Theory ......................................................................................... 2
   Multidimensional Item Response Theory ...................................................... 3
   Linking Methods ................................................................................................. 5
   IRT Equating Methods ....................................................................................... 8
   MIRT Equating Methods .................................................................................... 9
   Research Objectives ........................................................................................... 10

CHAPTER 2 LITERATURE REVIEW ......................................................................... 13
   Unidimensional Item Response Theory ........................................................... 13
   Multidimensional Item Response Theory ....................................................... 15
      The M3PL Model and its Characteristics .................................................... 15
      MIRT Item Parameter Recovery ................................................................. 17
      Multidimensional Graded Response Model ................................................ 19
      The Bifactor Model ....................................................................................... 20
      The Simple Structure Model ..................................................................... 22
   Dimensionality Assessment ............................................................................ 22
   UIRT Scale Linking ............................................................................................ 26
      Scale Linking Methods .................................................................................. 27
      Concurrent Calibration .................................................................................. 30
      Comparison Studies ...................................................................................... 31
   MIRT Scale Linking ............................................................................................ 35
      Translation of the Origin ............................................................................... 36
      Rotation of the Coordinate Axes .................................................................. 37
      Change of the Unit of Measurement .............................................................. 38
      Comparison Studies ...................................................................................... 39
   IRT Equating Methods ....................................................................................... 44
      IRT True Score Equating .............................................................................. 45
      UIRT Observed Score Equating ................................................................. 46
      Simple Structure MIRT Observed Score Equating .................................... 48
      Bifactor Observed Score Equating ............................................................... 49
      Full MIRT Observed Score Equating ......................................................... 51
   Summary ............................................................................................................. 54

CHAPTER 3 METHODOLOGY ............................................................................... 55
   MIRT Observed Score Equating Under the CINEG Design .............................. 55
      Test Forms ..................................................................................................... 56
      Dimensionality and Examinee Groups ......................................................... 56
      Correlations between Dimensions and Calibration Options ....................... 58
LIST OF TABLES

Table 3-1. Study 1: Characteristics of the Spanish and French exams ................. 80
Table 3-2. Study 1: Descriptive statistics of samples for Spanish and French .......... 80
Table 3-3. Study 1: Raw-to-scale score conversion table for Spanish .................... 81
Table 3-4. Study 1: Raw-to-scale score conversion table for French ...................... 82
Table 3-5. Study 2: Characteristics of Spanish and French ................................. 83
Table 3-6. Study 2: Descriptive statistics of intact forms for Spanish and French ...... 83
Table 3-7. Study 2: Descriptive statistics of pseudo forms for Spanish and French .... 84
Table 3-8. Study 2: Study conditions ..................................................................... 85
Table 3-9. Study 2: Raw-to-scale score conversion table for Spanish .................... 86
Table 3-10. Study 2: Raw-to-scale score conversion table for French ................... 87
Table 3-11. Differences between Study 1 and Study 2 ......................................... 88
Table 4-1. Preliminary study: Comparison of true item parameters with three types of estimated item parameters (r=0.2) .............................................................. 114
Table 4-2. Preliminary study: Comparison of true item parameters with three types of estimated item parameters (r=0.75) .............................................................. 114
Table 4-3. Preliminary study: Comparison of true item parameters with three types of estimated item parameters (r=0.8) .............................................................. 115
Table 4-4. Preliminary study: Comparison of true item parameters with three types of estimated item parameters (r=0.95) .............................................................. 115
Table 4-5. Preliminary study: Values of ABIAS, ASE, and ARMSE for all three conditions ......................................................................................................................... 116
Table 4-6. Study 1: DETECT results for Spanish and French ............................... 117
Table 4-7. Study 1: CFA results for Spanish and French ....................................... 117
Table 4-8. Study 1: Disattenuated correlation, observed correlation, and Cronbach’s alpha for Spanish and French ..........................................................................................................................118

Table 4-9. Study 2: Poly-DETECT results for Spanish and French .................................................................118

Table 4-10. Study 2: CFA results for Spanish and French .................................................................................119

Table 4-11. Study 2: Disattenuated correlation, observed correlation, and Cronbach’s alpha for Spanish and French ..........................................................................................................................119

Table 4-12. Study 2: Unweighted ABIAS, ASE, and ARSME for Spanish raw score .............................. 120

Table 4-13. Study 2: Weighted ABIAS, ASE, and ARSME for Spanish raw score .............................. 121

Table 4-14. Study 2: Unweighted ABIAS, ASE, and ARSME for Spanish unrounded scale score .................................................................................................................................................. 122

Table 4-15. Study 2: Weighted ABIAS, ASE, and ARSME for Spanish unrounded scale score .................................................................................................................................................. 123

Table 4-16. Study 2: Unweighted ABIAS, ASE, and ARSME for French raw score ............................................. 124

Table 4-17. Study 2: Weighted ABIAS, ASE, and ARSME for French raw score ............................................. 125

Table 4-18. Study 2: Unweighted ABIAS, ASE, and ARSME for French unrounded scale score .................................................................................................................................................. 126

Table 4-19. Study 2: Weighted ABIAS, ASE, and ARSME for French unrounded scale score .................................................................................................................................................. 127
**LIST OF FIGURES**

Figure 3-1. Pseudo-form construction for the Spanish exam ..................................... 88

Figure 4-1. Item vectors for the Spanish language 2011 form and common items ....... 128

Figure 4-2. Item vectors for the Spanish language 2012 form and common items ...... 129

Figure 4-3. Item vectors for the French language 2012 form and common items ....... 130

Figure 4-4. Item vectors for the French language 2013 form and common items ....... 131

Figure 4-5. Study 1: Scree plots for Spanish and French ............................................. 132

Figure 4-6. Study 1: Spanish observed and fitted distributions ................................. 133

Figure 4-7. Study 1: French observed and fitted distributions ................................. 134

Figure 4-8. Study 1: Spanish raw-to-raw score equivalents for all equating methods using separate and concurrent calibration ......................................................... 135

Figure 4-9. Study 1: French raw-to-raw score equivalents for all equating methods using separate and concurrent calibration ......................................................... 136

Figure 4-10. Study 1: Differences between (M)IRT and CE equated raw and unrounded scale scores for Spanish ................................................................. 137

Figure 4-11. Study 1: Differences between (M)IRT and CE equated raw and unrounded scale scores for French ................................................................. 138

Figure 4-12. Study 2: Scree plots for Spanish and French ............................................. 139

Figure 4-13. Study 2: Spanish observed and fitted distributions ................................. 140

Figure 4-14. Study 2: French observed and fitted distributions ................................. 141

Figure 4-15. Study 2: Conditional bias for Spanish raw scores ................................. 142

Figure 4-16. Study 2: Conditional bias for Spanish unrounded scale scores .......... 143

Figure 4-17. Study 2: Conditional SE for Spanish raw scores ................................. 144

Figure 4-18. Study 2: Conditional SE for Spanish unrounded scale scores .......... 145
Figure 4-19. Study 2: Conditional RMSE for Spanish raw scores ........................................ 146
Figure 4-20. Study 2: Conditional RMSE for Spanish unrounded scale scores ............ 147
Figure 4-21. Study 2: Conditional bias for French raw scores ........................................ 148
Figure 4-22. Study 2: Conditional bias for French unrounded scale scores ............... 149
Figure 4-23. Study 2: Conditional SE for French raw scores ........................................ 150
Figure 4-24. Study 2: Conditional SE for French unrounded scale scores ............... 151
Figure 4-25. Study 2: Conditional RMSE for French raw scores ........................................ 152
Figure 4-26. Study 2: Conditional RMSE for French unrounded scale scores ............ 153
CHAPTER 1

INTRODUCTION

It is common for large-scale testing programs to construct and administer alternate test forms. The testing industry uses alternate forms to accommodate different testing dates and item exposure issues. Ideally, the alternate forms should be parallel to the original form in terms of content and statistical specifications (Kolen & Brennan, 2014); however, alternate forms are likely to vary in regard to difficulty. In order for the alternate forms to be interchangeable across multiple occasions, a statistical procedure called equating is required to adjust differences in difficulty between forms (Kolen & Brennan, 2014). Since the form differences are adjusted in the equating procedure, examinees can be compared regardless of which forms they take.

Equating Designs

Equating is a decision-making process, and the first step of the decision is to choose the data collection design. The three most commonly utilized equating designs are random groups (RG) design, single group (SG) design, and common-item nonequivalent groups (CINEG) design. In the RG design, examinees are randomly assigned to alternate forms. The groups taking each form are assumed to be randomly equivalent; therefore, the differences in the examinees’ scores are considered to be due only to form differences. In the SG design, the same examinees take both forms of a test. Since the same examinees are administered two forms, the SG design is often used with counterbalancing the order of form administration.
In the CINEG design, which is the focus of this dissertation, different forms are administered to separate groups of examinees that differ in ability and each form contains common items present on both forms. These common items should be a mini-version of the full test in terms of content and statistical specifications (Kolen & Brennan, 2014). The common items can be either internal or external depending on whether they contribute to the total test score. The CINEG design does not assume random equivalence of examinee groups since the groups are from different populations. Thus, group and form differences need to be separated in the CINEG design, which requires strong statistical assumptions compared to the other designs.

After an equating design has been determined, an equating method should be chosen. Traditional equating methods such as the mean, linear, and equipercentile methods and item response theory (IRT) equating methods are commonly used. Deciding which method to use depends on equating designs, statistical assumptions, and other practical situations. This dissertation focuses only on IRT equating methods. In the next section, the characteristics of IRT and equating methods using IRT are outlined.

**Item Response Theory**

Item response theory (IRT) is broadly used in testing situations such as test development, computerized adaptive testing, item banking, differential item functioning, and test equating (Kolen & Brennan, 2014). An IRT model shows a mathematical relationship between an examinee’s latent trait ($\theta$) and the probability of answering an item correctly. IRT models exist for both dichotomous and polytomous items. Multiple-choice (MC) items typically are scored as 0 or 1, which are referred to here as
dichotomous items. Free response (FR) items that have more than two score categories are denoted as polytomous items.

An IRT model requires strong statistical assumptions about the dimensionality of the data to which the model applies. The assumption widely used in many IRT models is that a single ability is measured by item responses, which is called unidimensionality. The unidimensionality assumption is often violated in practice. For example, consider a math problem-solving item with a reading passage. This item violates the unidimensionality assumption because it requires both reading and math skills to answer the item correctly. Models assuming more than a single trait are referred to as multidimensional models, which will be discussed in the next section. An extension of the concept of unidimensionality is local independence (Lord, 1980). The local independence assumption means that one item’s responses are independent of responses to other items. The local independence assumption can also be violated when a set of items are associated with a common reading passage.

**Multidimensional Item Response Theory**

If a test measures more than one ability, a model that adequately reflects the multidimensionality of the test is needed; therefore, multidimensional IRT (MIRT) has been developed for modeling multidimensional data. The MIRT models can be either compensatory or non-compensatory (See Ackerman, 1994; McKinley & Reckase, 1983; Reckase, 1985). The compensatory models assume that high ability in one dimension can compensate for low abilities in other dimensions. For instance, suppose a situation in which a math item depends on two underlying constructs: arithmetic and algebra. If the
item involves only a small amount of the arithmetic ability but requires a considerable amount of the algebraic ability, the algebraic ability can compensate for the arithmetic ability. However, in the non-compensatory models, both dimensions need high abilities to have a high probability of a correct response.

In this dissertation, three compensatory MIRT models for dichotomous items are used: the simple structure (SS) model (Cai, 2010; Lee & Brossman, 2012; Thurstone, 1947), the bifactor (BF) model, (Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992), and the multidimensional two-parameter logistic (M2PL) model (Reckase, 1985), which is a direct extension of the unidimensional two-parameter logistic model. The M2PL model is considered as the full MIRT (FM) model in this dissertation. For polytomous items, the multidimensional graded response model (MGR: Muraki & Carlson, 1995) is used.

The M2PL model allows discrimination of more than one dimension; thus, it has more than one non-zero discrimination parameter. In the BF model, there is one general dimension and at least one specific dimension. The general factor denotes the principal latent ability a test is designed to measure, whereas the specific factors exhibit loadings beyond the general factor. The specific factors are orthogonal to each other and to the general factor. The specific factors can be due to two cases: (a) The specific factors may be due to any fixed clusters. Each cluster is associated with a separate specific factor in addition to the general factor on which all items load. For example, the specific factors can be content subdomains or item formats. Suppose a math test that has two dimensions: algebra and geometry. The general factor would be an overall math ability, while the
specific factors are the two sub-content areas (i.e., algebra and geometry). If the specific factors are due to item types, the specific factors designate each trait (i.e., MC and FR ability). That is, the specific factors reflect latent traits that are specific to the content subdomains or the item types, beyond the general factor. (b) The specific factors may be due to random nuisance. Compared to the first case, each specific trait is considered as somewhat secondary. For example, suppose a reading test with common reading passages. Responses to items within a passage are related to a secondary trait which may be skills specific to the passage. Typically, the specific traits here would be regarded as nuisance factors. In this dissertation, the BF model is applied to content subdomains and a mixed-format test; therefore, each specific trait is of interest after controlling for the general factor as well as the general factor. The SS model states that items discriminate only one dimension. For example, when the SS model is applied to a mixed-format test, MC items measure an MC-specific ability and FR items measure an FR-specific ability, while the two latent abilities can be correlated. Since each trait is of interest, the SS model can be applied to the first case in the BF model.

After item and ability parameters are estimated from any IRT models, these parameters should be placed on the same scale (i.e., scale linking) before equating. The next section describes linking methods for IRT models.

**Linking Methods**

Under the CINEG design, three linking methods are commonly used in the UIRT framework to place the new form (Form X) and old form (Form Y) parameter estimates on the same scale: separate calibration, multiple-group concurrent calibration, and fixed

In separate calibration, item parameters for the two forms are estimated separately (i.e., two runs of an estimation program). Since the new and old form examinee groups are not assumed to be equivalent, parameter estimates from the two separate runs are not on the same scale. Therefore, estimates from the common items that appear on both forms are used to find a scale transformation function. Scale linking then uses this transformation function to place all the item and ability estimates on the same metric.

In multiple-group concurrent calibration, the item parameters and ability distributions for both groups are estimated simultaneously (i.e., a single run of an estimation program). It combines data from both groups and treats items not taken by a particular group as not reached or missing (Lord, 1980). One group serves as the base scale. The new-form parameters are gradually placed on the old-form scale with common items. Estimating parameters for all items simultaneously assures that all parameter estimates are on the same scale. Hence, concurrent calibration does not need an additional scale linking process.

In contrast, fixed parameter calibration begins with calibration of the old form, followed by estimation of the non-common items on the new form. In this sense, the parameters of the common items on the new form retain their values from the old form calibration. The $\theta$ distribution of the new group is estimated based on the item responses for the common items on the new form and the common-item parameter estimates on the old form; therefore, this estimated distribution of $\theta$ are on the scale of the old form.
Because of this fixing process, the item parameter estimates for the new form non-common items are expected to be on the old form scale. Thus, fixed parameter calibration does not require an additional scale linking procedure. Kim (2006) recommended updating the prior ability distribution multiple times and the use of multiple EM cycles during fixed parameter calibration. However, a computer program that deals with full MIRT models for fixed parameter calibration has not been developed yet; therefore, fixed parameter calibration is not considered in this study.

The three linking methods mentioned above can be extended to the MIRT framework; however, linking methods for MIRT models are more complex than those for UIRT models. The MIRT scale linking methods generally estimate the scalar and the matrix coefficient to adjust for differences in (a) rotation, (b) correlation, (c) translation, which is similar to origin in UIRT, and (d) dilation, which corresponds to the unit of measurement in UIRT (Davey, Oshima, & Lee, 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2007; Oshima, Davey, & Lee, 2000). Under the RG design, Brossman and Lee (2013) stated that the MIRT scale linking procedures may not always be needed to conduct MIRT equating; however, it is a requirement under the CINEG design to have MIRT scale linking to adjust for rotational indeterminacy, correlational indeterminacy, translation indeterminacy, and dilation indeterminacy.

Various linking methods have been developed for MIRT models for separate calibration (Hirsch, 1989; Li & Lissitz, 2000; Min, 2007; Oshima, Davey, & Lee, 2000), while concurrent and fixed parameter calibration for MIRT models have gained little attention (Cai, 2010; Kim, 2017; Zhang, Lee, & Wang, 2016).
IRT Equating Methods

Both UIRT and MIRT equating involve multiple steps. The first step is to investigate model fit and any assumption violations. Then, item parameters and ability distributions should be estimated (i.e., calibration). If needed, scale linking should be conducted to place item and ability estimates on the same scale before equating.

After all the preliminary steps are performed, UIRT equating can be done in two ways: IRT true score equating and IRT observed score equating. IRT true score equating involves equating true scores through test characteristic curves (TCCs). General procedures for true score equating are as follows (Kolen & Brennan, 2014). First, choose a true score on a new form. Then, use an iterative method such as the Newton-Rapson method to find a latent ability (θ) that corresponds to that true score. Finally, use the old form to find the equivalent true score matching that ability. In UIRT observed score equating, conditional observed score distributions are estimated. Then, the conditional distributions are summed or integrated across all ability levels to obtain marginal observed score distributions. The new and old estimated observed score distributions are used to conduct equipercentile equating.

Both UIRT true score and observed score equating have advantages and disadvantages. UIRT true score equating has an advantage in that it uses a conversion that does not depend on the ability distribution; however, a drawback of UIRT true score equating is that true scores of examinees are not known. In practice, true scores are treated as converting number-correct observed scores on the new form to number-correct...
observed scores on the old form, where there is no justification for treating true scores in this way (Kolen & Brennan, 2014).

More importantly, UIRT true score equating has some complications when they are extended to MIRT. In the UIRT framework, true score equating is done by relating the forms to be equated through each test characteristic curve (TCC). The true score on the new form given $\theta$ is considered to be equivalent to the true score on the old form associated with that $\theta$; that is, it is a one-to-one relationship. However, true scores are related to ability levels through the test characteristic surface (TCS) in MIRT. The TCS maps several combinations of $\theta$’s to a single true score, which makes it challenging to extend UIRT true score equating to the MIRT framework. For this reason, this dissertation only considers IRT observed score equating. The next section describes MIRT equating methods when the test measures more than one ability.

**MIRT Equating Methods**

MIRT equating has the same purpose as UIRT equating, but the MIRT equating methods involve multiple abilities rather than a single ability. As explained previously, MIRT true score equating is less straightforward than MIRT observed score equating since it is difficult to find the corresponding true score and ability in a multidimensional surface. In MIRT observed score equating, observed score distributions are conditioned on a vector of abilities, and the marginal observed score distributions are obtained by a multivariate distribution (Brossman & Lee, 2013). The MIRT equating procedures have been developed for SS (Lee & Brossman, 2012), BF (Lee & Lee, 2016), and FM (Brossman & Lee, 2013) models, which will be discussed later in Chapter 2. These studies
suggest that the MIRT equating procedures can be more feasible than the UIRT equating methods for multidimensional data.

**Research Objectives**

MIRT can be widely applied in terms of its psychometric, cognitive, and test development perspectives. Multiple traits often are needed to answer an item correctly in many educational or psychological tests (Reckase, 2009). For example, scores on MC and FR items for a mixed-format test tend to be highly correlated but are likely to measure two distinct constructs for examinees (Cao, 2008). Moreover, tests are developed to align with content domains according to test specifications. For example, the ACT (2018) math test contains five content areas that are prerequisite for performance in college-level math.

For equating tests that measure several distinct proficiencies, procedures that reflect the multidimensional structure of the data are needed. Consequently, MIRT observed score equating methods have been developed using the SS, BF, and FM models (Lee & Brossman, 2012; Brossman & Lee, 2013; Lee & Lee, 2016). However, the aforementioned studies do not address multidimensional equating of tests under the CINEG design. As with UIRT equating, item parameter estimates should be placed on the same multidimensional scale when equating is conducted under the MIRT framework. While a MIRT scale linking process may not be required for the RG design (Brossman & Lee, 2013; Peterson & Lee, 2014), the CINEG design does require a scale linking process.
Although there have been studies regarding equating procedures under the MIRT framework, there is a need for further research in this area. Especially, few studies have presented equating procedures for SS, BF, and FM observed score equating under the CINEG design. There are very few studies that have investigated the BF (Zhang, Lee, & Wang, 2016), and SS (Kim, 2018) observed score equating method under the CINEG design. Furthermore, a comparison of SS, BF, and FM observed score equating methods under the CINEG design has never been attempted in previous studies. The purpose of this dissertation is to fill the gap in the literature.

Therefore, the primary objectives of this dissertation are to consolidate and expand MIRT observed score equating research with a specific focus on the CINEG design, which requires scale linking. Content areas and item types are two focal points of dimensionality. This dissertation uses two studies with different data types and comparison criteria to address the research objectives.

Study 1 describes MIRT equating procedures under the CINEG design. The study addresses assumptions regarding test forms and examinee groups in the MIRT framework, and what characteristics common items should have in MIRT equating. Moreover, dimensionality assessment procedures, scale linking, and equating procedures are explained with real datasets. Here, the source of multidimensionality is content subdomain.

The second study compares and evaluates performance of the MIRT observed score equating methods using pseudo-forms and pseudo-groups data. Mixed-format test datasets are manipulated to pseudo forms so that common items contain both MC items...
and FR items. Item types are considered to measure two distinct constructs. Therefore, the multidimensional structure for the second study is due to different item types.

More specifically, the purposes of this dissertation are:

1. To describe specific processes of MIRT equating under the CINEG design.
2. To compare the differences in equating results for UIRT, SS, BF, FM, and traditional equipercentile methods.
3. To compare scale linking methods (i.e., separate calibration and concurrent calibration) for MIRT observed score equating.
4. To compare the impact of group ability differences on MIRT observed score equating.

Study 1 addresses the first, second, and third research questions, while Study 2 answers the second, third, and fourth research questions.
CHAPTER 2
LITERATURE REVIEW

This chapter provides theoretical background for item response theory, scale linking, and equating. The chapter consists of sections discussing primarily: (a) UIRT/MIRT models and their characteristics; (b) UIRT and MIRT scale linking methods; and (c) UIRT and MIRT equating procedures.

Unidimensional Item Response Theory

As mentioned in the previous chapter, IRT is intended to model the probability of an examinee answering a given item correctly. One of the most widely used UIRT models is the three-parameter logistic (3PL) model, which includes an item discrimination parameter \(a\), an item difficulty parameter \(b\), and a lower asymptote parameter \(c\). The 3PL model (Birnbaum, 1968) for dichotomously scored items is denoted as

\[
P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_j(\theta_i - b_j)}},
\]

where \(\theta_i\) is the latent ability parameter for person \(i\), \(a_j\) is the item discrimination parameter of item \(j\), \(b_j\) is the difficulty parameter, \(c_j\) represents the pseudo-guessing (i.e., lower asymptote) parameter, and 1.7 is a scaling factor to make the logistic function comparable to the normal ogive function. An item characteristic curve (ICC) is shaped according to these item parameters. Higher \(a_j\) results in a steeper ICC, indicating that the item discriminates well among examinees within proficiency levels. The location of the ICC is determined by \(b_j\), which increases with increasing item difficulty. \(c_j\) signifies
the probability that an examinee with an extremely low ability level correctly answers the item (Lord, 1980).

The 3PL model can be constrained to the two-parameter logistic model (2PL) or the one-parameter logistic model (1PL) model. The 3PL model is reduced to the 2PL model when the pseudo-guessing parameter is assumed to be 0. The 1PL model is imbedded in both the 2PL and the 3PL in that the pseudo-guessing parameter is assumed to be 0 and all items have equal discrimination parameters (de Ayala, 2009).

There are IRT models that handle polytomous items (e.g., free-response items), in which item responses are scored with more than two categories. Examples of the polytomous items may include math items that require open-ended answers, writing prompts, etc. The graded response (GR) model (Samejima, 1969) and the generalized partial credit (GPC) model (Muraki, 1992) are frequently used polytomous IRT models, and this study used the GR model. The GR model assumes an examinee’s responses can be ordered; therefore, the response categories for an item are organized from low to high.

The probability of an examinee score for item $j$ in one of $k$ categories can be expressed as (Samejima, 1969):

$$P_{jk}^*(\theta_i) = \begin{cases} 1, & k = 1, \\ \frac{e^{D_{aj}(\theta_i-b_{jk})}}{1+e^{D_{aj}(\theta_i-b_{jk})}}, & k = 2, \ldots, m_j \end{cases}$$  \hspace{1cm} (2.2)

where all parameters are as previously defined, $b_{jk}$ is the difficulty for $k^{th}$ category, and $D$ is a scaling constant (i.e., 1.7). The category response function is computed by taking the difference between the cumulative category response functions (Samejima, 1969):
Multidimensional Item Response Theory

Cognitive processes in the context of psychology and education tend to be complex. MIRT was developed to more accurately represent items and persons in multidimensional space. Similar to ICC in UIRT, MIRT models describe an examinee’s latent abilities through the item characteristic surface (ICS); here, the examinee can have more than two abilities. Although there are compensatory and noncompensatory MIRT models, for ease of estimation this dissertation focuses on compensatory models.

The M3PL Model and its Characteristics

The multidimensional 3PL (M3PL) model is an extension of the UIRT 3PL model, which includes a pseudo-guessing parameter. The mathematical expression for the M3PL is given by:

\[
P_j(\theta_i) = c_j + (1 - c_j) \frac{e^{a_j^\prime \theta_i^\prime + d_j}}{1 + e^{a_j^\prime \theta_i^\prime + d_j}},
\]

where \( \theta_i \) is a \((1 \times m)\) vector of ability parameters, \( m \) designates the number of dimensions, \( a_j \) denotes a \((1 \times m)\) vector of discrimination parameters, \( d_j \) is the intercept, and \( c_j \) is the pseudo-guessing parameter.

The general idea of parameters related to item discrimination, difficulty, and lower asymptote is similar for both UIRT and MIRT. Nevertheless, the parameters have some differences stemming from the nature of MIRT. First of all, each element of \( a_j \)
coincides with a dimension and indicates the discrimination parameter (i.e., the slope of an ICS in the dimension). $d_j$ is related to the difficulty of the item; however, it is not a unique indicator of the difficulty of the item. Contrary to the difficulty parameter in UIRT, the term $\frac{-d_j}{a_{jk}}$ designates the relative difficulty of the item in the $k$-th dimension, where $a_{jk}$ is the $k$-th element of $a_j$. The pseudo-guessing parameter, $c_j$, has the same meaning for both UIRT and MIRT.

Discrimination and difficulty statistics for the compensatory MIRT model have been developed by Reckase and Mckinley (1991). The multidimensional discrimination index ($MDISC$) can be expressed as:

$$MDISC_j = \sqrt{\sum_{k=1}^{m} a_{jk}^2},$$  \hspace{1cm} (2.5)

which denotes the slope of an ICS at the steepest point in the direction specified by the multidimensional difficulty index ($MDIFF$) (Reckase & Mckinley, 1991). $MDIFF$ is denoted as

$$MDIFF_j = \frac{-d_j}{MDISC_j},$$  \hspace{1cm} (2.6)

which is the point at which the ICS steepest when the slope is decided in the direction from the origin of the space.

The direction where the item best measures is determined by the angles between each coordinate axis. The angles are computed by cosine corresponding to each dimension; that is,

$$\cos \gamma_{jk} = \frac{a_{jk}}{MDISC_j},$$  \hspace{1cm} (2.7)
where $\gamma_{jk}$ is the angle between the $k$-th axis and the line from the origin to the point in radians. $\gamma_{jk}$ is calculated as

$$\gamma_{jk} = \arccos \frac{a_{jk}}{MDISC_j}.$$  \hspace{1cm} (2.8)

If the angle between an item and a dimension is 0°, then the item measures only one dimension. If the angle is 45°, the item is interpreted as measuring a composite of two dimensions.

**MIRT Item Parameter Recovery**

MIRT item parameter recovery is not the primary focus of this dissertation; however, MIRT equating results may be affected by (a) how correlations between latent abilities are specified, and (b) how item parameters are recovered, and will further discussed in Chapter 3. Therefore, this section briefly summarizes MIRT item parameter recovery studies that addressed correlations.

Finch (2011) investigated difficulty and discrimination values for non-simple structure items using unweighted least squares (ULS) estimation with the normal-ogive harmonic analysis robust method (NOHARM). Along with ULS, two unidimensional approaches using BILOG were used: (a) the multidimensionality in the data was ignored and all the items were fit to a single ability, and (b) each item was separately grouped with each dimension. The data were generated from the compensatory MIRT model. Five study factors were used: (a) number of examinees (250, 500, 1000, or 2000), (b) intertrait correlation (0.0, 0.3, 0.5, or 0.8), (c) distribution of latent traits (normal or non-normal), (d) pseudo-guessing (M3PL or M2PL), and (e) type of latent structure (semi-complex or
complex). The results of the study indicated that item parameter estimates obtained using the MIRT model showed smaller bias in both discrimination and difficulty parameter estimates for non-complex items than did unidimensional estimation when two latent traits were present. As with discrimination parameter estimates, higher bias values were exhibited for the BILOG estimates as compared with ULS for the non-simple structure items. When the correlations between the latent abilities was higher, bias values were higher for all estimation methods. In terms of the difficulty parameter estimates, bias was greater when the data were not normally distributed, regardless of the estimation method. Both ULS and BILOG showed lower standard errors when the latent traits were normally distributed regardless of the correlation between the abilities; however, the ULS standard errors increased with increasing correlation values.

Svetina et al. (2017) examined MIRT item parameter recovery when tests showed complex structures or when latent traits were non-normal. Data were generated from the pool of estimated parameters for 62 dichotomous items from the reading assessment. Item responses of 2,000 examinees for 30 dichotomous items were generated to fit a two-dimensional compensatory model. Study factors included: two levels of model type (2PL or 3PL), three levels of correlation between dimensions (0.0, 0.4, or 0.7), three levels of distribution of the latent variables (normal for both $\theta$s, one $\theta$ was normal and the other skewed, or skewed for both $\theta$s), five levels of complexity, and two levels of discrimination balance. The results of the study indicated the impact of correlation between the latent abilities was consistent across three levels of correlations. Bias and RMSE values were larger for discrimination parameters when correlations became larger;
however, difficulty parameters were recovered most accurately at 0.7 correlation. Especially for complex items, their locations were best recovered at 0.4 correlation. Therefore, different levels of latent trait correlations did not consistently impact the MIRT item parameter recovery. Also, MIRT item parameter recovery was poorer when one or both latent traits were generated as skewed. When latent traits were assumed normal for both abilities, recovery of discrimination parameters was reasonable.

The previous studies had mixed findings about the relationship between item parameters and correlations; therefore, a simulation study was conducted to investigate how correlation between latent abilities impact item parameters, which is described in Chapter 3.

**Multidimensional Graded Response Model**

MIRT models have been developed for items with more than two response categories. This study used multidimensional graded response model (MGR) that was developed by Muraki and Carlson (1995).

The MGR model assumes that reaching step $k$ requires achievement of step $k - 1$. The lowest and highest scores for an item $j$ are zero and $m_j$, respectively. The probability of scoring in a specific category $k$ is the difference between the probability of successfully achieving $k$ or more steps and successfully achieving for $k + 1$ or more steps. Therefore, the probability that an examinee will have a score of $k$ is

$$P_{jk}(\theta_i) = P_{jk}^*(\theta_i) - P_{j(k+1)}^*(\theta_i),$$

(2.9)
where $P_{jk}^*(\theta_i) = 0$ is always 1 since all examinees must be in step 0 or more, and $P_{j(m_j+1)}^*(\theta_i) = 0$ since the examinees cannot reach more than $m_j$. The form of the normal ogive MGR models is presented as

$$P_{jk}(\theta_i) = \frac{1}{\sqrt{2\pi}} \int a_j^\prime \theta_i + d_{jk} e^{-\frac{t^2}{2}} dt,$$

where $k$ denotes the examinees’ score on item $j$, $a_j$ is the discrimination parameters that are in a vector form, and $d_{jk}$ is the parameter related to the easiness of achieving the $k^{th}$ step of the item. Relatively easy items result in high values for $d_{jk}$, whereas negative values occur when it is hard for examinees to complete a particular step of the item.

The Bifactor Model

The Bifactor (BF) model (Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992) is a constrained MIRT model in that items are allowed to load on one general dimension and at most one group-specific dimension. The BF model can be expressed as:

$$P_j(\theta_{iG}, \theta_{iS}) = c_j + \frac{1-c_j}{1+e^{-\lambda_j(\theta_{iG}+\theta_{iS})}},$$

Here, $\theta_{iG}$ indicates the general ability for person $i$, and $\theta_{iS}$ denotes specific abilities for person $i$, $a_{jG}$ and $a_{jS}$ are general and specific discrimination parameters for item $j$, and all other terms are as previously defined. An example of a four-item test measuring one general dimension and two specific dimensions is expressed in matrix form:

$$
\begin{pmatrix}
\lambda_{1G} & \lambda_{11} & 0 \\
\lambda_{2G} & \lambda_{21} & 0 \\
\lambda_{3G} & 0 & \lambda_{32} \\
\lambda_{4G} & 0 & \lambda_{42}
\end{pmatrix}
$$
As demonstrated by the matrix, all items measure the general dimension, the first two items have an additional loading on the first specific dimension, and the third and fourth item have a loading on the second dimension. No item should load on more than one specific dimension in the BF model. The BF model can be viewed as a restricted version of the two-tier item factor analysis (two-tier IFA) model (Cai, 2010) in that there is one general dimension in the BF model, whereas the two-tier IFA model allows general dimensions to be correlated among themselves.

The BF model can be used for modeling complex multidimensional situations with greater computational ease compared to other MIRT models. It allows for consideration of residual variance due to specific dimensions. For example, the general dimension could exhibit subject area (e.g., math) and the specific dimensions could be item formats, passages, or content domains. In this dissertation, content domains and item formats are considered as specific dimensions.

Cai et al. (2011) argue that the BF version of the GR model is comparable to its UIRT and MIRT counterparts. The BF models can be reduced to UIRT models when an item loads only on the general dimension. For the BF extension of the GR model, the cumulative response function is denoted as:

$$P_j(\theta_{\text{G}}, \theta_{\text{S}}) = \frac{1}{1 + \exp[-1.7(d_{j,m-1} + a_{jG}\theta_{\text{G}} + a_{jS}\theta_{\text{S}})]},$$

(2.12)

where $d_{j,m-1}$ is a category intercept, and the other parameters are the same as previously denoted. As with the GR and MGR models, the category response probability is the difference between two adjacent cumulative response probabilities.
The Simple Structure Model

The early work of Thurstone (1947) introduces the concept of simple structure (SS). In the SS model, each item loads on a single dimension, and the correlation between the dimensions are dependent on a person’s ability. Within each item, only one nonzero $\alpha$-parameter is needed. The SS model can also be thought of as a simplified version of the two-tier IFA model; that is, the SS model has only general dimensions that are correlated, and does not have specific dimensions. Take the example of a test with two primary dimensions. In the SS model, one set of items is assumed to load on the first dimension and the other set of items on the second dimension. Approximate simple structure (Roussos, Stout, & Marden, 1998) is often used for practical situations where a test has one large discrimination parameter and the others are near zero. Therefore, each item could have some amounts of discrimination on the other dimensions.

The SS model is appropriate for multidimensional situations such as content areas or mixed-format tests. Lee and Brossman (2012) stated that in the SS model, MC items only measure MC proficiency (i.e., $\theta_{MC}$) and FR items only measure FR proficiency (i.e., $\theta_{FR}$). The two abilities associated with the two item types are allowed to be correlated. Also, the same item type could be modeled by a UIRT model.

Dimensionality Assessment

As will be discussed later, dimensionality is an interaction between test items and examinee groups. The number of dimensions is dependent upon the examinee’s item responses. For example, one form is developed to measure two distinct abilities, but the examinee group turns out to be varied to one dimension. In this case, choosing a UIRT
model may be appropriate. However, if this form is found to have two dimensions to another examinee group, the unidimensionality assumption is violated. Therefore, questions may arise: (a) Is this form unidimensional or multidimensional? (b) How many dimensions can be found in this form? These questions can be answered through dimensionality assessments, although there is no correct answer because the dimensionality assessment procedures are sample dependent. The result of dimensionality assessments can be used to have an idea of which IRT model is appropriate before choosing an equating method.

There are mainly two procedures of assessing dimensionality. One is to investigate whether unidimensionality holds and the other is to specify the number of dimensions. Within this framework, exploratory and confirmatory approaches have been developed to answer the dimensionality questions. The exploratory approaches are conducted when there is no clear evidence regarding the dimensional structure of the data. In contrast, the confirmatory approaches are used to test whether a pre-specified dimensional structure fits the data.

One way to estimate dimensionality in an exploratory approach is Principal component analysis (PCA). PCA reduces the complexity of the interrelationships among observed variables to a small number of linear combinations of principal components. It uses the inter-item correlation matrix to investigate dimensionality. The first principal component has the largest variance, and the rest of the components have the largest variance possible given that they are orthogonal to the preceding components. Thus, each principal component is mutually uncorrelated. One of the criteria that can be used to
judge dimensionality with PCA is a scree plot. The scree plot can determine the amount of variance accounted for by each principal component. In the scree plot, eigenvalues of a correlation matrix are plotted according to their number of components. By looking for an elbow in the scree plot, the number of dimensions in the given data can be chosen; that is, choosing the point in the steep curve before the first point indicates the number of dimensions.

Another approach for assessing dimensionality is DIMTEST (Stout et al., 1992). DIMTEST is a nonparametric procedure that was developed to test the null hypothesis that two test forms are essentially unidimensional. DIMTEST can be used as both exploratory and confirmatory modes. Each test form is divided into two subtests, which are denoted the assessment subtest (AT) and the partitioning subtest (PT). The AT items should be dimensionally distinct from the remaining PT items. If $p < .05$, the AT items are dimensionally distinct from the PT items. An extension of DIMTEST is Poly-DIMTEST (Li & Stout, 1995) which handles both dichotomous and polytomous items.

DETECT (Zhang & Stout, 1999b) is used to answer how many dimensions are needed to represent the data. By using estimated conditional covariance, DETECT explores item clusters with maximum index where within-cluster conditional covariances are positive and between-cluster conditional covariances are negative. DETECT can be run as an exploratory analysis when it searches for the optimal partition, while a confirmatory analysis can be done by specifying partitions. Both DIMTEST and DETECT are based on conditional covariances or correlations but DETECT provides more specific answers than DIMTEST. Three indices are provided in DETECT to
evaluate the degree of multidimensionality: (a) the DETECT index, (2) the ASSI (approximate simple structure) index, and (c) the R (ratio) index (Zhang, 2007; Zhang & Stout, 1999b). The DETECT index values of 1 or greater indicate strong multidimensionality; values of 0.4 to 1 indicate moderate multidimensionality; values of 0.2 to 0.4 indicate weak multidimensionality; and values below 0.2 indicate essential unidimensionality. The ASSI index values close to 1 indicate approximate simple structure; the values greater than .25 indicate essential deviation from unidimensionality; and the values below .25 indicate essential unidimensionality. The R index values close to 1 indicate approximate simple structure; the values greater than .36 indicate essential deviation from unidimensionality; and the values below .36 indicate essential unidimensionality. Poly-DETECT (Yu & Nandakumar, 2001) is a modified version of DETECT that can handle both dichotomous and polytomous items.

Performing a confirmatory factor analysis (CFA) is another way to estimate dimensionality. CFA is used to test whether the given data fit a researcher’s hypothesized model, and statistical tests are used to evaluate how well the model fits to the data. Statistics that are commonly used are: (a) the chi-square test, (b) standardized root mean square error (SRMR), (c) root mean square error of approximation (RMSEA), (d) the Akaike Information Criteria (AIC; Akaike, 1974), and (e) Bayesian Information Criterion (BIC; Schwarz, 1978). The chi-square test investigates the null hypothesis that the sample covariance matrix is equal to the covariance matrix predicted by the CFA model. The SRMR index is also used to test the difference between sample and predicted covariance matrices; therefore, smaller values are considered to have a reasonable fit. A value of .1 or
smaller is considered as a good model fit (Kline, 2010). The RMSEA index indicates the discrepancy between the hypothesized model and the population covariance matrix. An RMSEA of .05 or less indicates a good model fit, while .1 or more indicates a poor model fit (Kline, 2010). AIC investigates the discrepancy between the true likelihood function and the estimated parameters in the model. BIC is closely related to AIC, but it operates under a Bayesian assumption. Smaller AIC and BIC values are closer to the true model. AIC and BIC introduce a penalty for the number of parameters in the model because increasing the number of parameters in the model improves the fit, and the penalty is larger in BIC than in AIC.

Another method for assessing dimensionality in a confirmatory approach is a disattenuated correlation. The disattenuated correlation is the correlation between true scores after considering unreliability of the cluster (e.g., subsection) scores; that is, it is calculated between item clusters such as content subdomain or item format. Disattenuated correlations close to 1 indicates multidimensionality, while disattenuated correlations close to 0 implies unidimensionality.

**UIRT Scale Linking**

Scale linking procedures must be used prior to equating in order to place Form X and Form Y parameter estimates on a common scale. For the CINEG design with separate calibration, the scale linking process is required to place item parameter estimates and ability estimates from the new form scale onto the old form scale. Let $I$ and $J$ represent the scales for the new and old forms, respectively, and $A$ and $B$ represent scaling
coefficients for transforming scale $I$ to scale $J$. The scales are linearly related; therefore, the ability, discrimination, difficulty, and pseudo-guessing parameters are transformed as:

$$\theta_{JI} = A\theta_{II} + B,$$

(2.13)

$$a_{IJ} = \frac{a_{II}}{A},$$

(2.14)

$$b_{JJ} = Ab_{IJ} + B,$$

(2.15)

$$c_{IJ} = c_{IJ}. $$

(2.16)

Once ability and item parameters are transformed, the probabilistic relationships for the new and the old form are the same.

**Scale Linking Methods**

In order to perform scale linking under the CINEG design, common items are often used as an anchor between the new and old forms. There are three types of scale transformation methods for dichotomous items: the mean/sigma method, the mean/mean method (Loyd & Hoover, 1980; Macro, 1977), and characteristic curve transformation methods (Haebara, 1980; Stocking & Lord, 1983). To estimate $A$ and $B$, the mean/sigma method (Marco, 1977) utilizes the means and standard deviations of the $b$-parameter estimates of common items:

$$A = \frac{\sigma(b_J)}{\sigma(b_I)} \text{ and } B = \mu(b_J) - A\mu(b_I).$$

(2.17)

The mean/mean method (Loyd & Hoover, 1980) uses the mean of the $a$ and $b$ parameter estimates of common items to compute $A$ and $B$:

$$A = \frac{\mu(a_J)}{\mu(a_I)} \text{ and } B = \mu(b_J) - A\mu(b_I).$$

(2.18)
Contrary to the mean/sigma and mean/mean methods, the characteristic curve transformation methods consider item parameter estimates simultaneously through the IRT characteristic curve. The Haebara (1980) method calculates the sum of the squared differences between the ICCs for each item for a particular ability so that the differences between the ICCs are minimized:

$$H_{\text{diff}}(\theta_i) = \sum_{j,V} \left[ P_{ij}(\theta_{ij}; \hat{a}_{jj}, \hat{b}_{jj}, \hat{c}_{jj}) - P_{ij} \left( \theta_{ij}; \frac{\hat{a}_{ij}}{A}, A\hat{b}_{ij} + B, \hat{c}_{ij} \right) \right]^2. \quad (2.19)$$

The $H_{\text{diff}}$ statistic is computed over the common items ($j:V$) and summed over all examinees. Then the estimation continues until finding $A$ and $B$ that minimize the criterion, $H_{\text{crit}}$:

$$H_{\text{crit}} = \sum_i H_{\text{diff}}(\theta_i). \quad (2.20)$$

The Stocking and Lord (1983) method takes the squared difference of sums over items:

$$S_{\text{diff}}(\theta_i) = \left[ \sum_{j:V} P_{ij}(\theta_{ij}; \hat{a}_{jj}, \hat{b}_{jj}, \hat{c}_{jj}) - \sum_{j:V} P_{ij} \left( \theta_{ij}; \frac{\hat{a}_{ij}}{A}, A\hat{b}_{ij} + B, \hat{c}_{ij} \right) \right]^2. \quad (2.21)$$

In other words, the squared difference is calculated between the test characteristic curves (TCCs) for a given $\theta_i$. Then $S_{\text{diff}}$ is summed over examinees, and the estimation proceeds to find $A$ and $B$ that minimize

$$S_{\text{crit}} = \sum_i S_{\text{diff}}(\theta_i). \quad (2.22)$$

The characteristic curve transformation methods use ICC or TCC to estimate linking coefficients. These methods could alleviate the weakness of the mean/sigma and mean/mean methods that consider item parameter estimates separately. Several studies
found that the characteristic curve methods produce more stable results than the mean/sigma and mean/mean methods (Lee & Ban, 2010; Li, Jiang, & von Davier, 2012).

The scale linking methods for dichotomous items could be extended for polytomous items. Cohen and Kim (1998) suggested the mean/sigma and mean/mean methods for polytomous items for the GR model. For the mean/sigma method, the mean and standard deviation of the $b$-parameters for the common items are computed separately for both forms. Then, they are replaced with Equation 2.17 to calculate the slope and intercept of the transformation equation. For the mean/mean method, both the mean of the $b$-parameters and the mean of the $a$-parameters for the common items are calculated. For the transformation equation, Equation 2.18 is used.

For the characteristic curve methods for polytomous items, it is necessary to set the criteria over items and categories within items. The Haebara method for the GR model can be expressed as

$$
H_{diff}(\theta_i) = \sum_{j \in V} \sum_{k} \left[ P_{ijk} \left( \theta_{ji}; \hat{a}_{ij}, \hat{b}_{ij2}, \ldots, \hat{b}_{ijk}, \ldots, \hat{b}_{ijm_j} \right) - P_{ijk} \left( \theta_{ji}; \frac{\hat{a}_{ij}}{A}, A\hat{b}_{ij2} + B, \ldots, A\hat{b}_{ijk} + B, \ldots, A\hat{b}_{ijm_j} + B \right)^2 \right].
$$

(2.23)

The first summation is the criteria over items ($j: V$) and the second summation is of the categories within the item ($k: j$). Therefore, the function is the sum of squared differences between the category responses for all categories and items. $H_{crit}$ is found by substituting Equation 2.23 into Equation 2.20.

The Stocking-Lord method for the GR model can be calculated as:
$$SL_{diff}(\theta_i) = \left[ \sum_{j:V} \sum_{k:j} W_{jk} P_{ijk} \left( \theta_{ji}; \hat{\alpha}_{jj}, \hat{b}_{jj2}, \cdots, \hat{b}_{jjk}, \cdots, \hat{b}_{jjm_j} \right) \right.$$ 

$$- \sum_{j:V} \sum_{k:j} W_{jk} P_{ijk} \left( \theta_{ji}; \frac{\hat{\alpha}_{jj}}{A}, A\hat{b}_{jj2} + B, \cdots, A\hat{b}_{jjk} + B, \cdots, A\hat{b}_{jjm_j} \right)$$

$$+ B \right]^2.$$

As shown earlier, this equation is the squared difference between TCCs. $W_{jk}$ denotes the scoring function related to the category $k$. $SL_{crit}$ is found by substituting Equation 2.24 into Equation 2.22.

**Concurrent Calibration**

Since each distribution of $\theta$ is assumed to follow a standard normal distribution, the item parameter estimates from separate calibration are not on the same scale; thus, separate calibration needs an additional scale linking procedure. On the contrary, concurrent calibration is a simultaneous process. Concurrent calibration estimates the distributions of $\theta$ for the two groups and the item parameters for the new form and the old form at the same time (Wingersky & Lord, 1984). Concurrent calibration combines data from multiple groups and treats items not taken by a specific group as missing (Lord, 1980), which allows concurrent calibration to handle this issue using a simultaneous computer run. IRT calibration programs such as BILOG-MG (Zimowski et al., 2003), PARSCALE (Muraki & Bock, 2003), flexMIRT (Cai, 2017), and MULTILOG (Thissen, Chen, & Bock, 2003) have the capacity to conduct multiple-group calibration (i.e., concurrent calibration). Several researchers have presented concurrent calibration (Baker
& Kim, 2004; Bock & Zimowski, 1997; Woodruff & Hanson, 1997), which uses an EM algorithm (Dempster, Laird, & Rubin, 1977) to acquire maximum likelihood estimates for the item parameters. Detailed mathematical formulas are provided in Woodruff and Hanson (1997).

The basic idea of the EM algorithm is as follows: an expectation (E) step evaluates the expectation of the log-likelihood using the current estimates for the parameters. A maximization (M) step calculates maximizing parameters for the expected log-likelihood found on the E step. Then, these estimates are used to find the distribution of the latent variables in the next E step. After updating the distributions of $\theta$ via the EM algorithm, the concurrent calibration procedure continues as follows: To fix the $\theta$ scale, the distributions for the old group and the new group are transformed to have a mean of 0 and a standard deviation of 1. The item parameter estimates are also linearly transformed. These item parameter estimates and distributions of $\theta$ are used at the next EM cycle. The iteration continues until the EM algorithm converges. Due to common items, the new-form parameters are gradually placed on the old-form scale.

**Comparison Studies**

There has been research regarding the relative performance of separate calibration and concurrent calibration. Kim and Cohen (1998) compared separate calibration with the Stocking-Lord method to concurrent calibration. A simulation study was conducted for this study under the CINEG design. The study used the computer program BILOG for separate calibration and MULTILOG for concurrent calibration based on the marginal maximum likelihood estimation. A 50-item test with four different numbers of common
items (5, 10, 25, and 50) were considered in the study. The first evaluation criterion was the root mean square difference (RMSD) between the item parameter estimates and the generating parameters. The other evaluation criterion was the mean Euclidean distance (MED) that shows the overall quality of recovery for the item parameter estimates and the generating item parameters. Their results showed that concurrent calibration produced larger RMSD and MED values than separate calibration when the number of common items was small. Furthermore, item parameter estimates were accurate when the number of common items was larger. However, the study used two different estimation programs, which might have caused biased results.

Hanson and Béguin (2002) compared the performance of separate and concurrent calibration under the CINEG design. The study pointed out the confounding effects of Kim and Cohen (1998). Therefore, this study used BILOG-MG and MULTILOG for both separate and concurrent calibration. The study used two 60-item ACT Mathematics forms, and the estimated item parameters were considered as population item parameters for simulation. The study had five study factors: (a) group equivalence, (b) estimation methods (mean/mean, mean/sigma, Stocking-Lord, Haebara method, and concurrent calibration), (c) estimation program, (d) sample size, and (e) the number of common items. The performance of linking was examined using IRT true score equating function and ICC criterion. The equating results for the two test forms were compared using the true score equating criterion, and the differences between the estimated and true ICCs were evaluated using the ICC criterion.
The results showed that there was less error with concurrent calibration than with separate calibration except when groups were nonequivalent. The mean squared error (MSE) based on both criteria for the nonequivalent groups was markedly larger than for equivalent groups. A larger MSE for nonequivalent groups was seen for both separate and concurrent calibration. A larger number of common items resulted in lower MSE, and the MSE was smaller when the sample size was larger. Among separate calibrations, the Stocking-Lord and Haebara methods had lower MSE than the mean/mean and mean/sigma methods. The authors also concluded that the two programs performed similarly for both separate and concurrent calibration with the exception of the MULTILOG N (1, 1) condition.

Later, Lee and Ban (2010) compared concurrent calibration, separate calibration with the Stocking-Lord and Haebara methods, and proficiency transformation procedures, which identifies a linking function based on $\theta$-values. The authors supposed a situation where forms $A_1$ and $A_2$ are each administered at Time 1 and Time 2. The authors also supposed that $A_2$ and another form $B_2$ are administered at the same time. Therefore, it can be assumed that examinees who took $A_2$ and $B_2$ are randomly equivalent groups. Using $A_2$ as an anchor, the authors applied item parameter estimates from Form $B_2$ to the scale of Form $A$.

Two 60-item ACT English forms were used to generate item parameters for simulation. BILOG-MG was used for item calibration, and the ST (Hanson, Zeng, & Chien, 2004) program was used for the scale transformation methods. $\theta$-values were sampled from an $N(0, 1)$ for $A_1$ and $\theta$-values for $A_2$ and $B_2$ were sampled from $N(0, 1)$,
$N(0.5, 1)$, and $N(1, 1)$ distributions. The study conditions that were investigated were:
(a) three sampling conditions, (b) four linking methods, (c) sample size ($n = 3,000$ and 500), and (d) the number of items per form ($k = 75$ and 25). To evaluate the four linking methods, the test characteristic curve (TCC) and expected observed score distribution (ESD) were used as criteria. The ESD criterion is determined by the item parameters of the IRT model.

The authors evaluated the performance of each linking procedure. The separate calibration procedures performed better than the concurrent calibration and proficiency transformation procedures. However, concurrent calibration performed better than separate calibration and proficiency transformation when all samples were from the same population. Between the two separate calibration methods, the Haebara method outperformed the Stocking-Lord method.

Kang and Petersen (2011) compared separate calibration with the Stocking-Lord method, concurrent calibration, and fixed parameter calibration under the CINEG design. Item parameters were chosen from two 60-item math achievement test forms and the data were manipulated to create 50-item forms with 10, 20, and 40 common items, respectively. The 3PL model is used and the ability distribution of the old form group is assumed to be $N(0, 1)$. Two levels of sample size (500 and 2,000) and three levels of ability distributions for the new form group (i.e., $N(0, 1)$, $N(0.25, 1.1^2)$, and $N(0.25, 1.2^2)$) distributions were used. BILOG-MG was implemented for separate, concurrent, and fixed parameter calibration and the PARSCALE program was also used for fixed parameter calibration. The results of the study found that separate calibration
and concurrent calibration were comparable for all conditions. Fixed parameter calibration with PARSACLE was similar to those obtained with the other two calibrations; however, fixed parameter calibration with BILOG-MG had poor performance in linking, especially when both groups had different ability distributions.

**MIRT Scale Linking**

As indicated in UIRT scale linking, scale linking procedures are also required for multidimensional data to place the ability and item parameter estimates on the same scale. In UIRT scale linking, the origin and unit of measurement are adjusted for differences; however, in MIRT scale linking, rotation and correlation should be considered in addition to origin and unit of measurement. That is, MIRT scale linking should adjust ability and item parameters on different scales due to indeterminacy in rotation, correlation, origin, and unit of measurement (Brossman & Lee, 2013). Origin and unit of measurement in UIRT are equivalent to translation and dilation in MIRT, respectively. Similar to UIRT, item and ability estimates are on the same scale with respect to translation and dilation under the RG design. Therefore, translation and dilation would not be required for MIRT scale linking under the RG design. However, the orientation of the coordinate axes might be different for the two forms; that is, there still exists rotational indeterminacy. Along with the rotational indeterminacy, item and ability parameters are dependent on correlational indeterminacy. The correlational indeterminacy is often resolved by designating multidimensional abilities to follow a multivariate normal distribution (i.e., $MVN(0, I)$) and to be uncorrelated to one another. To solve the rotational indeterminacy, Thompson, et al. (1997) illustrated a process to estimate an orthogonal rotation matrix ($T$)
that can be applied under the RG design. However, Brossman and Lee (2013) suggested that MIRT scale linking under the RG design is not always required to conduct MIRT observed score equating.

The multidimensional scale linking procedures have been illustrated under different assumptions and designs (Davey, Oshima, & Lee, 1996; Li & Lissitz, 2000; Min, 2007; Oshima, Davey, & Lee, 2000; Thompson, Nering, & Davey, 1997). Depending on each procedure, mathematical expressions vary for transforming ability and item parameters from the new scale to the old scale. However, general expressions are stated as:

\[
\mathbf{a}_j^T = \mathbf{a}_j^TR, \tag{2.25}
\]

\[
d_j = d_j - \mathbf{a}_j^TR\mathbf{\beta}, \tag{2.26}
\]

\[
c_j = c_j, \tag{2.27}
\]

\[
\theta_j = R^{-1}\theta_i + \mathbf{\beta}. \tag{2.28}
\]

where \( I \) and \( J \) are the new form and old form scale, respectively. The new form parameter estimates before scale linking are denoted as \( \hat{\mathbf{a}}_j, \hat{d}_j, \hat{c}_j, \) and \( \hat{\theta}_i. \) The old form parameter estimates are expressed as \( \mathbf{\hat{a}}_j, \mathbf{\hat{d}}_j, \mathbf{\hat{c}}_j, \) and \( \mathbf{\hat{\theta}}_i. \) \( R \) adjusts for differences in rotation and dilation and \( \mathbf{\beta} \) adjusts for difference in translation. The next three sections explain the indeterminacies of compensatory MIRT models.

**Translation of the Origin**

Translation of the origin refers to moving \( \theta \) to a new location. For a two-dimensional case, suppose the origin of the space were moved from \((0,0)\) to \((-1, -1)\).
While the location of the coordinate axes changed, the examinee’s location has not changed. The vector of coordinates of the examinee $i$ can be expressed as

$$\mathbf{v}_i = \theta_i - \delta,$$

where $\mathbf{v}_i$ is the vector of coordinates for examinee $i$ in the new coordinate system, $\theta_i$ denotes similarly with $\mathbf{v}_i$ in the old coordinate system, and $\delta$ represents the translation vector (i.e., (-1, -1)). The $d$-parameter vector is converted to the new axes as

$$\tilde{\mathbf{d}} = \mathbf{d} + a\delta'.$$

where $\tilde{\mathbf{d}}$ denotes the $d$-parameter vector in the new coordinate system. After the translation in origin, the $a$-parameters remain the same. However, the person parameters and the $d$-parameters should be changed to the new coordinate system.

**Rotation of the Coordinate Axes**

Another source of indeterminacy in the MIRT models is related to the orientation of the coordinate axes. Transformation of coordinates to a set of rotated coordinate axes can be done by multiplying an orthogonal rotation matrix to the original $\theta$-coordinate axes. In a two-dimensional space, the rotation matrix can be expressed as

$$R = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix},$$

where $\alpha$ is the degree of rotation in the clockwise direction. The vector of abilities and vector of item discriminations are then multiplied by the rotation matrix. The inverse of the orthogonal rotation matrix is its transpose. Therefore, $R' R = R R' = I$, the identity matrix. As a consequence, the exponent of the model has equal values after the rotation is applied (Reckase, 2009):
\[ aR(\theta R)' + 1d = aRR'\theta' + 1d = a\theta' + 1d. \] (2.32)

The \( d \) parameter remains unchanged because distances from the origin are not changed by an orthogonal rotation. To find the rotation matrix in MIRT linking, a procedure called the Procrustes method (Schönemann, 1966) is used. The Procrustes method is a general way to find an orthogonal matrix \( T \) that closely maps a matrix \( P \) to a matrix \( Q \), where the residual matrix \( E \) is defined as \( E = PT − Q \). Detailed procedures for the Procrustes method can be found in Schönemann (1966), Schönemann and Carroll (1970), and Gower and Dijksterhuis (2004).

**Change of the Unit of Measurement**

A third source of indeterminacy in the MIRT models is the unit of measurement used with each coordinate axis. The unit of measurement for each coordinate axis can be altered by multiplying a constant. Usually, the unit of measurement is set by specifying the standard deviation of the location coordinates to a particular value such as 1.0 in computer programs. To change the person coordinates from one unit to another, the \( \theta \)-matrix is multiplied by a scaling constant matrix \( (C) \),

\[ C = \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix}. \] (2.33)

The diagonal matrix is often called the dilation matrix. The transformed person coordinates \( (v) \) can be found by post-multiplying \n\[ v = \theta C, \] (2.34)

where \( \theta \) is the matrix of original coordinates.
Comparison Studies

This section reviews four studies that have presented MIRT scale linking methods for the MIRT models.

**Hirsch (1989).** Hirsch (1989) conducted a scale linking study under a common-examinee design. Both simulated and real test data were used. The simulation study used six pairs of simulated data with 2,000 responses and 40 items from the College Level Academic Skills Test (CLAST) Reading. Two data characteristics were considered: the number of common examinees (groups of 100, 200, and 300) and the difference in mean ability of the groups. For the real data analysis, six real datasets of 2,000 examinee responses were used. The data sets were the 36-item reading and the 35-item writing tests from the 1984 Florida CLAST. The linking procedure in Hirsch (1989) involved four steps: (a) estimate item and ability parameters using the M2PL model for the base and new forms, (b) orient the parameter estimates of the linked form to have the same position in the \( \theta \)-space as the base form, (c) conduct orthogonal Procrustes rotation, and (d) determine the means and standard deviations of \( \theta \)-estimates for the common examinees on the two forms. To evaluate the linking procedure, differences between the estimated true score for the base form and the estimated true score for the new form were calculated and the difference between the \( \theta \)-estimates for each form was computed.

Findings from this study are summarized as follows: (a) Using common-examinee design, approximately 100 common examinees showed adequate linking results when considering the true scores, (b) when the mean abilities of the populations of unique
Oshima, Davey, and Lee (2000). Oshima et al. (2000) extended the study by Davey et al. (1996) and presented four MIRT linking methods. The four methods are: (a) the direct method, (b) the equated function method, (c) the test characteristic function (TCF) method, and (d) the item characteristic function (ICF) method. The direct method is the multivariate extension of the unidimensional method by Divgi (1985). Linking parameters are estimated by the sum of the squared differences between $\hat{a}_j$ and $\hat{a}^*_j$, $\hat{d}_j$ and $\hat{d}^*_j$. That is,

$$f_1(R, \beta) = \sum_{j \in V} (\hat{a}_j - \hat{a}^*_j)^T (\hat{a}_j - \hat{a}^*_j) + \sum_{j \in V} (\hat{d}_j - \hat{d}^*_j)^2$$

(2.35)

where $V$ is the common-item set, $\hat{a}_j$ and $\hat{d}_j$ are item discrimination and intercept parameter estimates for item $j$ of the old form, and $\hat{a}^*_j$ and $\hat{d}^*_j$ are the parameter estimates that are transformed from the new form to that of the old form.

The equated function method minimizes the sum of the squared differences between defined functions $T_p$ and $T^*_p$ on separate sets of elements from $(\hat{a}_j, \hat{a}^*_j)$ and $(\hat{d}_j, \hat{d}^*_j)$:

$$f_2(R, \beta) = \frac{1}{p} \sum_{i=1}^{n} (T_p - T^*_p)^2,$$

(2.36)

where $p$ is the separate sets of elements of $\hat{a}_j$ and $\hat{d}_j$. The choice of the function type and the method of categorizing the items into $p$ groups are arbitrary.
The TCF method (the extension of the Stocking-Lord method) minimizes the sum of squared differences between test characteristic surfaces across the common items. The two linking coefficients \( (\mathbf{R}, \mathbf{\beta}) \) in Equations 2.25-2.28 are converted as \( (\mathbf{R}^{-1}, \mathbf{\beta}) \) in the TCF method. The TCF method can be expressed as:

\[
\begin{align*}
f_3(\mathbf{R}, \mathbf{\beta}) &= \frac{1}{q} \sum_{\theta} W_\theta \left[ \sum_{i=1}^{n} P_{1i}(\theta) - \sum_{i=1}^{n} P^*_{2i}(\theta) \right]^2,
\end{align*}
\]

(2.37)

where \( \mathbf{R} \) is an \( m \times m \) rotation and dilation matrix, \( \mathbf{\beta} \) is a \( m \times 1 \) translation vector, \( q \) is the number of \( \theta \) vectors, and \( W_\theta \) is for weighting differentially at each ability level, \( P_{1i}(\theta) \) denotes the TCC for the old form, and \( P^*_2(\theta) \) indicates the transformed TCC of the new form parameters. A modified Newton method is used to find \( \mathbf{R} \) and \( \mathbf{\beta} \) in the TCF method (Oshima, Davey, & Lee, 2000). The ICF method is the extension of the Haebara method to the case of multidimensional data. It minimizes the squared differences between the item characteristic curves over all of the common items:

\[
\begin{align*}
f_4(\mathbf{R}, \mathbf{\beta}) &= \frac{1}{nq} \sum_{\theta} W_\theta \left( \sum_{i=1}^{n} (P_{1i}(\theta) - P^*_{2i}(\theta)) \right)^2,
\end{align*}
\]

(2.38)

where terms are as defined for the TCF method.

Oshima et al. (2000) compared the four linking methods using a simulation study. For the simulation, a 40-item ACT mathematic test was used to obtain the generating item parameters for the M2PL model. The results of this study indicated that the TCF and ICF methods were more stable and had better recovery of true linking parameters than the direct and equated function methods; however, the difference between the minimized functions across the four methods were small.
Li and Lissitz (2000). Later, Li and Lissitz (2000) criticized the study of Oshima et al. (2000) and suggested the transformation of one matrix to another scale includes an orthogonal Procrustes rotation, a translation vector, and a dilation parameter. The Li and Lissitz (LL) method uses different parameterizations so that the linking coefficients \((R, \beta)\) in Equations 2.25-2.28 are converted into \((kR, -(1/k)m)\), where \(k\) is a dilation parameter and \(m\) is a translation vector. The LL method starts with the estimation of rotation matrix \(R\) through the minimization of differences between the old form and new form discrimination parameters for the common-item set:

\[
E_1 = A_I R - A_J,
\]

(2.39)

where \(E_1\) is the residual matrix, and \(A_I\) and \(A_J\) are the discrimination parameters for new and old forms, respectively. Then, the translation vector \(m\) is found by minimizing \(Q\):

\[
Q = \sum_{i=1}^{n} (d_{ij} - d_{ij}^*)^2,
\]

(2.40)

where \(d_{ij}^*\) transformed item difficulty parameter, and \(n\) is the number of common items. The dilation constant \(k\) is calculated to minimize the sum of squared errors of the residual matrix \(E_2\):

\[
E_2 = (kA_I R) - A_J,
\]

(2.41)

where \(R, A_I,\) and \(A_J\) are as defined previously. The rotation matrix \(R\) and the unit change coefficient \(k\) are derived using
where $A_{CI}$ and $A_{CJ}$ are the column-centered item discrimination parameter matrix for the new and old forms, respectively. After obtaining the rotation matrix $R$, the translation vector $m$, and the central dilation parameter $k$, the transformation can be done by substituting Equations 2.25-2.28 with LL method parameterizations.

Li and Lissitz (2000) conducted two simulation studies: Study 1 examined the effect of parameter estimates on the recovery of the transformation parameters. The study considered sample sizes, linking methods, test length, and ability distributions. Study 2 used the results of Study 1 to investigate the best MIRT linking method. ACT Form 24B (Reckase, 1985) item parameters were considered as the generating item parameters. The results of the study showed that biases from each of the methods were small, but RMSE values were different across the estimation methods. The least squares method produced the most exact scaling parameter and estimated the translation vector more precisely than the matching test response surface method.

**Min (2007).** More recently, Min (2007) developed a linking procedure from the LL method. A diagonal dilation matrix $K$ is added to replace the dilation constant $k$ in the LL method. Transformation can be conducted by substituting $(R, \beta)$ in Equations 2.25-2.28 into $(K^{-1}R, K^{-1}m)$, where $K$ is the diagonal dilation matrix.

The dilation matrix has $k_1$ (first dimension) and $k_2$ (second dimension) diagonal elements and off-diagonal elements are zero. The dilation vector is calculated as:
\[ K = \text{diag}[A'j, A_i R] \times (\text{diag}[R'A'_j, A_i R])^{-1}, \quad (2.43) \]

where terms are as previously defined.

Min (2007) compared the new linking method with two other linking methods: (a) the test characteristic method by Oshima et al. (2000), and (b) a composite transformation with three components by Li and Lissitz (2000). The author simulated the item parameters under approximate simple structure and mixed structure. For the simulation, an approximate simple structure was constructed using two sets of items, each containing ten items. Ten items loaded primarily on the first dimension and the other ten items loaded on the second dimension. For the mixed structure, two sets of five items loaded mostly on one of the two dimensions. Five bivariate normal distributions with several means and variances/covariances were considered. Moreover, two levels of examinees (1,000 and 2,000) were investigated. Real data were also implemented based on mean difference and difference variation of item parameter estimates.

The results of the study showed that modeling a unique dilation parameter for each dimension improved the orthogonal Procrustes transformation. The orthogonal Procrustes transformation was more accurate than the method by Li and Lissitz (2000) which had only one dilation parameter for all dimensions. Also, the oblique rotation by Oshima et al. (2000) was closer to the generating item parameters.

**IRT Equating Methods**

After item parameters and ability parameters are placed on the same scale through scale linking, IRT equating can be conducted. Two methods for IRT equating are IRT observed score equating and IRT true score equating. This dissertation focuses on IRT
observed score equating, which estimates observed score distributions on both forms, followed by traditional equipercentile equating using the observed score distributions. This section briefly explains IRT true score equating, and IRT observed score equating will be explained in more detail.

**IRT True Score Equating**

IRT true score equating relates true scores on the new form with true scores on the old form. The UIRT true score equating procedure is conducted in steps below:

First, a true score on the new form ($\tau_N$) is selected. The true score at a particular ability level ($\theta_i$) is defined as the sum of the probabilities of a correct response for each item; that is, $\tau_N(\theta_i) = \sum_{j:N} P_{ij}(\theta_i; a_j, b_j, c_j)$. In order to determine $\theta_i$ corresponding to the particular true score on the new form, the Newton-Raphson method is used. This method is an iterative procedure that minimizes the difference by taking the derivative with respect to $\theta_i$. After $\theta_i$ is computed, this value is substituted to the true score on the old form; that is, $\tau_O(\theta_i) = \sum_{j:O} P_{ij}(\theta_i; a_j, b_j, c_j)$.

However, some complications occur when UIRT true score equating is extended to the MIRT framework (Brossman & Lee, 2013). While a TCC relates a single ability in UIRT, true scores are related to ability levels through the TCS in the MIRT framework. Therefore, more than one combination of ability levels correspond to a single true score. When the TCS is computed for the old form, different combinations of ability levels corresponding to the new form true score may map to different true scores on the old form.
The problem can be detoured using the results by Zhang (1996), Zhang and Stout (1999a), and Zhang and Wang (1998). The authors demonstrated that any set of item responses that can be adequately modeled by a multidimensional compensatory model can be approximated by a UIRT model with estimated unidimensional ability and item parameters. Therefore, the authors defined the unidimensional ability as a standardized linear composite of multidimensional abilities. Later, Brossman and Lee (2013) developed a MIRT true score equating procedure using this approximation. The procedure for unidimensional approximation is described in greater detail in Brossman and Lee (2013). As explained in this subsection, more than one $\theta$ can map to the same true score, which makes it difficult to apply true score equating to the MIRT framework. Therefore, this study only focuses on IRT observed score equating methods.

**UIRT Observed Score Equating**

In order to conduct UIRT observed score equating, it is necessary to estimate the conditional distributions of observed number-correct scores on both forms. For dichotomous IRT models and polytomous IRT models, the conditional distributions for the number-correct scores are determined by a compound binomial distribution and compound multinomial distributions, respectively.

Using the Lord and Wingersky recursion formula (Lord & Wingersky, 1984), the conditional observed score distributions are estimated at each ability level ($\theta_i$). The probability of correctly answering the first item is $f_1(x = 1|\theta_i) = P_{i1}$ and the probability of incorrectly answering the first item is $f_1(x = 0|\theta_i) = 1 - P_{i1}$. Therefore,
\( f_r(x|\theta_i) \) is defined as the distribution of number-correct score distribution over the first \( r \) items with ability \( \theta_i \). For a test with \( r > 1 \), the recursion formula is denoted as:

\[
\begin{align*}
  f_r(x|\theta_i) &= f_{r-1}(x|\theta_i)(1 - P_{ir}), & x = 0, \\
  &= f_{r-1}(x|\theta_i)(1 - P_{ir}) + f_{r-1}(x - 1|\theta_i)P_{ir}, & 0 < x < r, \\
  &= f_{r-1}(x - 1|\theta_i)P_{ir}, & x = r.
\end{align*}
\]

(2.44)

After the recursion formula is used to find the conditional number-correct score distributions for each form, the conditional distributions are integrated over \( \theta \) levels to create a marginal distribution for each form:

\[
f(x) = \int_{\theta} f(x|\theta)\psi(\theta)\,d\theta \approx \sum_q f(x|\theta_q)\psi(\theta_q),
\]

(2.45)

where \( \psi(\theta) \) is the \( \theta \) distribution and \( q \) represents quadrature points for \( \theta \). The integration can be approximated with summation using finite quadrature points. Once marginal distributions of Form X and Form Y are obtained, equipercentile equating is conducted to equate scores on both forms.

For a mixed-format test, a compound multinomial distribution is used since the polytomous items have more than two response categories. Let \( f_1(x = W_{11}|\theta_i) = p_{i11}(\theta_i) \) be defined as the probability of earning a score in the first category of item 1, and \( f_1(x = W_{12}|\theta_i) = P_{i12}(\theta_i) \) as the probability of having a score in the second category of item 1. Then, for \( r > 1 \), the recursion formula of earning a score \( x \) after \( r \)th item is (Hanson, 1994):

\[
f_r(x|\theta_i) = \sum_{k=1}^{m} f_{r-1}(x - W_{jk})P_{ijk}(\theta_i).
\]

(2.46)
This function is defined as \( x \) values between \( Min_r \) and \( Max_r \) where each denote the minimum and maximum possible scores after adding the \( r^{th} \) item. Separate recursive formulas (i.e., dichotomous and polytomous) should be incorporated before calculating the marginal distributions. Once the conditional distributions are calculated, the marginal distributions are computed by applying Equation 2.46. Traditional equipercentile equating is conducted for the marginal distributions.

**Simple Structure MIRT Observed Score Equating**

The SS observed score equating process is similar to UIRT observed score equating, with some exceptions. Lee and Brossman (2012) suggested the SS observed score equating method for mixed-format tests. The SS observed score equating method has major differences with UIRT equating: (a) MC and FR items are calibrated separately, (b) a bivariate proficiency distribution for both item types is estimated, and (c) composite total score distributions are used. Even when dimensions are correlated, the SS procedure under RG design does not have rotational indeterminacy. Even though this dissertation illustrates the SS observed score equating method in a two-dimensional case, the SS observed score equating method can be generalized to more than two dimensions.

Let \( \theta_1 \) be MC proficiency and \( \theta_2 \) be FR proficiency. Then, \( f_1(x_1|\theta_1) \) and \( f_2(x_2|\theta_2) \) are the conditional observed score distribution for MC and FR sections, respectively. The conditional total score distribution is computed as:

\[
f(x|\theta_1, \theta_2) = \sum_{X=x_1+X_2} f_1(x_1|\theta_1) f_2(x_2|\theta_2),
\]

(2.47)
where summation occurs over all pairs of $w_1x_1$ and $w_2x_2$ scores to obtain the total score $x$. The conditional distributions can be calculated using the recursive algorithm by Lord and Wingersky (1984) for dichotomous items and Hanson (1994) for polytomous items. A marginal observed score distribution is found by integrating conditional total score distributions over a bivariate theta distribution:

$$f(x) = \int_{-\infty}^{\infty} f(x|\theta_1, \theta_2) g(\theta_1, \theta_2) d\theta_1 d\theta_2 \approx \sum_{\theta_1} \sum_{\theta_2} f(x|\theta_1, \theta_2) q(\theta_1, \theta_2),$$

(2.48)

where $q(\theta_1, \theta_2)$ is the bivariate discrete quadrature a specific pair of $\theta_1$ and $\theta_2$. The joint bivariate distribution, $g(\theta_1, \theta_2)$, is estimated using the Mislevy (1984) procedure. Traditional equipercentile equating is then conducted using the two fitted observed score distributions for both forms. Results from a real data analysis and a simulation study by Lee and Brossman (2012) indicated that the SS observed score equating produced stable equating results than the UIRT procedure. That is, the SS equating method produced the most similar results to the traditional equipercentile method. The SS equating procedure also had less equating error when used with multidimensional data.

**Bifactor Observed Score Equating**

BF observed score equating has a similar process to SS observed score equating. However, in BF observed score equating, a general factor is added to produce conditional distributions for both forms. Recently, Lee and Lee (2016) suggested a BF observed score equating procedure for mixed-format tests. Item format was considered as dimensionality in this study. The 2PL and GR Bifactor models were fit to three datasets: (a) matched samples, (b) pseudo-forms, and (c) simulated data. Let $\theta_G$, $\theta_1$, and $\theta_2$ denote a general
ability, MC-specific ability, and FR-specific ability, respectively. A marginal observed score distribution is calculated by integrating conditional observed score distributions over a trivariate theta distribution:

\[
f(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x|\theta_G, \theta_1, \theta_2) g(\theta_G, \theta_1, \theta_2) d\theta_G d\theta_1 d\theta_2
\]

\[
\approx \sum_{\theta_G} \sum_{\theta_1} \sum_{\theta_2} f(x|\theta_G, \theta_1, \theta_2) q(\theta_G, \theta_1, \theta_2)
\]

where \( g(\theta_G, \theta_1, \theta_2) \) denotes mutually uncorrelated trivariate distribution, \( q(\theta_G, \theta_1, \theta_2) \) represents the trivariate discrete quadrature.

From the results of the study by Lee and Lee (2016), the BF observed score equating method produced equating results that were similar to those for the UIRT equating method. The differences between the two methods were insignificant since almost all differences were smaller than the difference that matters (DTM) criterion. The UIRT and BF observed score equating methods provided somewhat different results when MC and FR abilities had very low correlation. The BF method produced equating results that were closer to the equipercentile method when a moderate degree of multidimensionality existed.

Zhang et al. (2016) conducted BF observed score equating with testlet-based data under the CINEG design. Separate calibration and concurrent calibration were used to estimate the item and examinee parameters. The Li and Lissitz (2000) method, equated function method (Oshima, Davey, & Lee, 2000), and test characteristic function method (Oshima, Davey, & Lee, 2000) were used for separate calibration. Simulated data were used to investigate the effect of local item independence (LID) and group difference. The test length was 50, and items were assumed to have five testlets with 10 items per testlet.
The results revealed that BF equating with Li and Lissitz (2000) method was preferred. The Li and Lissitz (2000) method led to small systematic and random equating errors even when the data showed high LID. When the LID level was high, the performance of the UIRT methods was negatively affected. In summary, the BF observed score equating methods may be promising for testlet-based exams.

Lim and Lee (2016) compared numerous equating methods for subscores using operational and simulated datasets. Four study factors were considered: test dimensionality, subtest length, form difference in difficulty, and sample size. The results showed that 3PL observed score equating with separate calibration performed the best. The worst performing method was BF true score equating. Also, BF observed score equating performed the best when a test (a) was multidimensional, (b) had small form difference, (c) had short subtest length, or (d) had small sample size.

**Full MIRT Observed Score Equating**

The FM equating procedure for MC items was developed by Brossman and Lee (2013). The conditional observed score distribution for MC items was found for each combination of $m$ abilities and is expressed as $f(x|\theta_i)$, where $\theta_i$ denotes a vector of $m$ abilities. For a test with $r > 1$, the conditional distributions are computed as (Brossman & Lee, 2013):

\[
\begin{align*}
  f_r(x|\theta_i) &= f_{r-1}(x|\theta_i)(1-P_{ir}), & x = 0, \\
  &= f_{r-1}(x|\theta_i)(1-P_{ir}) + f_{r-1}(x-1|\theta_i)P_{ir}, & 0 < x < r, \\
  &= f_{r-1}(x-1|\theta_i)P_{ir}, & x = r.
\end{align*}
\]
To obtain the marginal distributions for each form, the conditional distributions are integrated over the multivariate ability density $\psi(\theta)$ or summed over all combinations of the $m$ latent abilities:

$$f(x) = \int_{\theta_1} \int_{\theta_2} \cdots \int_{\theta_m} f(x|\theta)\psi(\theta) d(\theta) \approx \sum_{\theta_1} \sum_{\theta_2} \cdots \sum_{\theta_m} f(x|\theta)\psi(\theta),$$  \hspace{1cm} (2.51)

where terms are as previously described.

Brossman and Lee (2013) described two observed score equating methods (FM observed score equating and unidimensional approximation to MIRT observed score equating) and a true score method (unidimensional approximation to MIRT true score equating). UIRT observed score equating, UIRT true score equating, and equipercentile equating were evaluated for comparison purposes. Three sets of the Iowa Test of Educational Development (ITED) exams were selected to conduct the analyses. The exams differed in dimensionality and only contained MC items. The results showed that both unidimensional equating procedures performed similarly, and the three multidimensional equating methods performed similarly. The unidimensional procedures and the multidimensional procedures performed differently. Also, the multidimensional methods performed more similarly to the equipercentile method than did the unidimensional methods. Therefore, the results suggested that the multidimensional equating procedures could provide an acceptable alternative to UIRT equating when the data does not conform unidimensionality.

Peterson and Lee (2014) suggested the FM observed score equating method for mixed-format exams using Advanced Placement (AP) exams. For polytomous items, the recursion formulas by Hanson (1994) can be extended to multidimensional settings. Let
\( f_1(x = W_{11} | \theta_i) = P_{i11}(\theta_i) \) be defined as the probability of earning a score in the first category of item 1, and \( f_1(x = W_{12} | \theta_i) = P_{i12}(\theta_i) \) as the probability of having a score in the second category of item 1. Then, for \( r > 1 \), the recursion formula of earning a score \( x \) after the \( r^{th} \) item is:

\[
f_r(x | \theta_i) = \sum_{k=1}^{m_j} f_{r-1}(x - W_{jk}) P_{ijk}(\theta_i),
\]

(2.52)

where \( W_{jk} \) denotes the scoring function of the \( k^{th} \) category. The marginal distributions are found in the same manner as in Equation 2.52.

Using the proposed approach, Peterson and Lee (2014) compared the performance of FM equating with UIRT, BF, and equipercentile methods under the RG design. Matched samples datasets and single population datasets were used for the study. For the BF methods, group-specific dimensions were specified to item format or content domain. With the FM methods, two- and four-dimension models were evaluated. Identity and the equipercentile methods were used as equating criteria. The results of the study indicated that the MIRT equating methods worked better for exams that had more multidimensionality, and the UIRT equating methods performed better for unidimensional data. Single population datasets and AP grade agreement percentages dominantly showed this trend. Also, it was discovered that there were not large differences among UIRT and MIRT equating methods according to the AP grade scale.

Lee, Lee, and Brennan (2014) conducted various MIRT equating procedures using simulated datasets. The performance of the six equating procedures under the RG design was compared: (a) UIRT observed score equating, (b) UIRT true score equating, (c) full
MIRT observed score equating, (d) unidimensionalized MIRT observed score equating, (e) unidimensionalized MIRT observed score equating, (f) unidimensionalized MIRT true score equating, and (g) equipercentile equating. Four factors (test length, sample size, form differences, and correlations) were considered. The results confirmed that when the correlation between dimensions was low, the full MIRT procedure provided more accurate equating results than other equating procedures. Another interesting conclusion was that even when multidimensional data were used, the UIRT equating procedures yielded accurate equating results.

Summary

This chapter discussed UIRT and MIRT frameworks, scale linking procedures, and equating procedures with extensive review of the literature. First, UIRT and MIRT model characteristics were presented. Second, UIRT and MIRT scale linking methods and comparison studies were reviewed. Third, UIRT and MIRT equating methods were reviewed with recent research. Although MIRT studies have been receiving attention in recent years, little research exists on MIRT equating. This study aims to add to the current MIRT equating literature by evaluating the performance of MIRT scale linking and equating methods using various dimensionality sources.
CHAPTER 3

METHODOLOGY

This chapter consists of two sections that describe procedures for MIRT observed score equating under the CINEG design with real data analyses (Study 1), and pseudo-form and pseudo-group analyses (Study 2). The first section provides details on MIRT observed score equating under the CINEG design: assumptions, dimensionality assessment, item calibration, scale linking procedures, and equating procedures. The first section also provides real data examples for the MIRT equating procedures. The second section compares and evaluates the performance of the MIRT observed score equating methods by using pseudo-form and pseudo-group data. AP exams were manipulated to pseudo forms so that common items contain both MC and FR items. Pseudo groups were created by resampling original examinees to investigate the effect of group differences.

MIRT Observed Score Equating Under the CINEG Design

Although equating with MIRT is an extension of UIRT equating, equating under the MIRT framework needs stronger requirements than UIRT equating because the MIRT observed score equating procedure involves a vector of abilities rather than a single ability. Under the CINEG design, the two test forms are different in terms of origin, unit of measurement, and the orientation of the coordinate axes. Therefore, the item parameter estimates from the new form should be placed on the coordinate system of the old form. Thus, the goal of MIRT equating under the CINEG design is to adjust test scores on two forms from multidimensionally nonequivalent groups so that scores on the two forms can be used interchangeably. To simplify illustration, this section only considers a two-
dimensional case. The following assumptions are made in terms of examinee groups and test forms when MIRT observed score equating is conducted under the CINEG design:

**Test Forms**

Original and alternate forms for MIRT observed score equating must be carefully developed in the test development stage. The test forms should be constructed based on the same test specifications. For example, the common items should appear the same positions on the two forms. The text should be exactly the same in the original and alternate forms. In terms of content specifications, the forms are built to the same content areas including the same item types and the number of items. The content areas should cover all subdomains that both forms are intended to measure, and there should be a sufficient number of items in each content subdomain. Also, the two forms should be checked with respect to statistical specifications. Descriptive statistics, item difficulties, and discriminations for both examinee groups are investigated before equating. These requirements should be emphasized in test development processes, so that the same constructs are measured in precisely the same way across the two test forms, which will be explained in the next subsection.

**Dimensionality and Examinee Groups**

Before explaining dimensionality, a distinction should be made between dimensions and constructs in a MIRT model. Dimensions refer to the coordinate axes that represent the data, while constructs denote the cognitive aspects measured by a test. Generally, the number of coordinate axes is less than or equal to the number of constructs in a test. This distinction is needed because the coordinate axes of the original coordinate
system might not represent the constructs. The original coordinate system is often rotated so that the constructs are aligned with the coordinate axes.

Dimensionality is an interaction between examinee groups and test forms. Therefore, the number of dimensions (i.e., the number of coordinate axes) should be the same for the old and new forms that each examinee group takes. Also, the same number of constructs should be measured for both examinee groups and test forms.

If the two forms differ in their number of dimensions, item parameter estimates cannot be placed on the same coordinate system and consequently MIRT observed score equating is not feasible. Suppose the two forms (old and new) of a math exam are developed that require 3rd grade math skills (dimension 1) and 3rd grade reading skills (dimension 2). If the old form is administered to 3rd grade examinees, this form would have two dimensions; that is, both math and reading skills discriminate the old form group. Although it is not realistic, suppose that the new form is administered to 6th grade examinees. This new form might not exhibit two dimensions because the reading skills do not discriminate 6th grade examinees. In this case, the new form group does not have \( a_2 \) parameters, while the old form has both \( a_1 \) and \( a_2 \) parameters. Since one of forms does not have \( a_2 \) parameters, it is impossible to decide the location for the origin of the space and determine the unit of measurement on each of the coordinate axes.

The two multidimensionally nonequivalent groups are different in terms of: (a) origin and unit of measurement (mean and SD), and (b) correlation between dimensions. These differences are group-dependent, meaning that they can differ depending on which population takes the two forms. Using a common-item set, the means and SDs are
adjusted so that the item and ability parameters for the two forms are on the same scale. However, the correlations between dimensions cannot fully be adjusted through the common items. The next subsection describes this correlational issue.

**Correlations between Dimensions and Calibration Options**

Even though two test forms are thoroughly designed in terms of their dimensions, it is likely that the correlations between the latent abilities of two groups taking these two forms vary because examinee groups that were administered the two forms are different. The correlation can be rotated through common items; however, the rotation process does not alter the absolute magnitude of the correlation. This correlational issue arises when FM models are calibrated. If the correlation between abilities does not impact item parameters, then there is no need for concern regarding the adjustment for the correlation. The correlations between dimensions can be specified (e.g., freely estimated, orthogonal, or fixed at certain values) in the calibration process for FM models, thus how the specification of correlations impacts item parameters should be investigated.

The following questions about correlations should be answered before MIRT observed score equating is conducted: (a) if the correlations between dimensions of the two forms are different, do they impact on item parameters? For example, if the old form has a correlation of 0.7 between dimensions but the new form has a correlation of 0.9 between dimensions, does the difference matter to item parameters? (b) when calibrating FM models, how are correlations between dimensions specified?

A simulation study was conducted to answer these questions. A more detailed description about the simulation study is provided in the Preliminary Study section.
Examinees’ latent abilities are not necessarily orthogonal in the educational setting. Therefore, the correlation between dimensions was either (a) orthogonally estimated or (b) freely estimated for the simulation study. From the simulation analysis in Chapter 4, it was discovered that item parameters were well recovered when correlations between dimensions were freely estimated. Also, the correlation between the latent traits did not have a systematic impact on the recovery of item parameters. Thus, in the multidimensionally nonequivalent groups, correlations between the dimensions are not necessarily identical for both the old and new form groups. Furthermore, it is preferable that correlations between dimensions are freely estimated as opposed to using the orthogonal solution.

**Common Items**

Common items need the same assumptions with the test forms. The common items should measure the same constructs and have the same number of dimensions as the total test. From the math exam example explained earlier this section, the common items should have both $a_1$ and $a_2$ parameters. If one of the two parameters is missing from the common-item set, then the items cannot serve as common items in the MIRT framework because the common items should have the same dimensional structure as the test forms. Therefore, if the source of dimensionality is item format, the common-item set should contain both MC and FR items. If content subdomain is a source of dimensionality, the common-item set should contain each subdomain. In this way, common items with the same dimensional structure adjust for score differences that are solely due to form difficulty.
To summarize, MIRT observed score equating under the CINEG design is conducted under these assumptions:

- The number of dimensions and constructs should be the same for the old and new examinee groups that take the old and new forms, respectively;
- Origin and unit of measurement (mean and SD), and correlation between dimensions can be different for the old and new examinee groups that take the old and new forms, respectively;
- When calibrating FM models, it is suggested that correlations between latent traits be freely estimated.
- Common items should have the same dimensional structure and measure the same constructs as the test forms.

**Preliminary Study**

Latent traits and item parameters are invariant in parameters; however, when item parameters are estimated, they are group dependent so that the correlation between latent traits could impact item parameter estimates. Therefore, the purpose of Preliminary Study is to investigate how the correlation between dimensions of the two forms impact item parameter estimates. Also, the study seeks to answer how correlations between dimensions are specified when calibrating FM models.

A simulation study was conducted as follows:

1. Item parameters were generated for a 10-item test form using M2PL. The slope and difficulty parameters were generated such that $a \sim \ln(0,0.2)$, $b \sim N(0,1)$, where $ln$ designates a log-normal distribution. Although these
distributions were previously used for UIRT simulations in a study by Kim and Lee (2006), there is no empirical evidence available regarding application to a multidimensional setting.

2. Using the item parameters in Step 1, item responses of 5,000 examinees were simulated. The true correlation between two dimensions varied from extremely multidimensional (0.2), moderately multidimensional (0.75, 0.8), and to near unidimensional (0.95). That is, the form had four datasets and was replicated 100 times, resulting in a total of 400 datasets.

3. Each dataset was calibrated three times using the computer program flexMIRT (Cai, 2017) with the correlation between dimensions: (a) freely estimated, (b) fixed at zero (orthogonal), and (c) fixed at 0.2, 0.75, 0.8, or 0.95 (i.e., true correlation). For example, a dataset with true correlation of 0.2 was calibrated with the correlation between dimensions (a) freely estimated, (b) orthogonal to each other, and (c) fixed at 0.2.

4. The generating item parameters and estimated item parameters were compared in terms of Bias, Standard Errors (SE), and Root Mean Squared Error (RMSE):

\[
Bias_i = \frac{\sum_{r=1}^{100}(t_{ir} - \hat{t}_i)}{100}, \quad (3.1)
\]

\[
SE_i = \sqrt{\frac{\sum_{r=1}^{100}(t_{ir} - \hat{t}_i)^2}{100}}, \quad (3.2)
\]

\[
RMSE_i = \sqrt{\frac{\sum_{r=1}^{100}(t_{ir} - \hat{t}_i)^2}{100}} = \sqrt{Bias_i^2 + SE_i^2}, \quad (3.3)
\]
where $\hat{t}_{ir}$ is an estimate for item $i$ at replication $r$; $\bar{t}_i$ represents the average, over 100 replications, of the estimates for item $i$; and $t_i$ is either the discrimination parameter or the intercept parameter for item $i$. Bias, SE, and RMSE were computed for each item and then the averages were calculated over all the items at the test level. Bias indicates that how the expected value of an estimator is systematically different from the true parameter value and SE indicates sampling error. RMSE reflects the overall error by considering both estimation bias and sampling error.

**Study 1: Real Data Analyses**

The purpose of Study 1 is to illustrate assumptions, dimensionality assessment, item calibration and scale linking procedures, and equating process for MIRT observed score equating under the CINEG design. The MIRT observed score equating methods (SS, BF, FM) are applied to real datasets to investigate whether they are reasonably similar to UIRT and equipercentile equating methods. This study also compares scale linking methods (i.e., separate and concurrent calibration) for MIRT observed score equating. Sub-content domains in AP exams are considered as the source of multidimensionality using only MC items.

**Data Preparation**

This study used test forms for two administrations of the AP Spanish Language (Spanish) and French Language (French) exams. Although the exams included both MC and FR items, only the MC items were used in Study 1. All data were collected under the CINEG design with internal anchor items; therefore, data modification was not required.
In this dissertation, the form administered in the previous year was considered as the old form, and the form administered in the following year as the new form. The two Spanish test forms, administered in 2011 and 2012, each contained 70 MC items, and the two French test forms, administered in 2012 and 2013, contained 65 MC items. The MC section of both Spanish and French had five response options addressing the content areas of Listening and Reading. The dataset for the Spanish old form included 118,176 examinees, and the dataset for the Spanish new form included 20,000 examinees. The French old and new forms were administered to a total of 17,006 and 20,000 examinees, respectively. To reduce calibration time, item responses for 5,000 examinees were randomly selected from the Spanish and French forms.

Table 3-1 presents the characteristics of the Spanish and French exams, and Table 3-2 presents the descriptive statistics of the Spanish and French exams based on the sampled data. Table 3-2 shows that the new form is easier than the old form for Spanish, and the old form is slightly easier than the new form for French. The effect sizes between the two Spanish forms and two French forms were .0427 and 0.009, respectively.

This study used a nonlinearly transformed scale score ranging from 10 to 40. The 10-40 score scale was developed to be normally distributed with a mean of 25 and standard deviation of 5 for the old form. Equating results are obtained for raw and scale scores. The raw-to-scale score conversion tables for Spanish and French are provided in Tables 3-3 and 3-4, respectively. Note that even though this dissertation used AP exams, results may not be generalizable to the operational AP exams since Study 1 used only the MC section.
**Dimensionality Assessment**

Dimensionality assessment is a critical tool for deciding which IRT equating method to use. If the results of dimensionality assessment indicate the unidimensionality assumption has been violated, MIRT equating methods would be more suitable than UIRT equating methods. Note that in Study 1, content subdomain is considered a source of multidimensionality. Several assessment methods were used to investigate the dimensionality of the Spanish and French exams including (a) item vector plots, (b) principal component analysis (PCA), (c) DETECT, (d) confirmatory factor analysis (CFA), and (e) disattenuated correlation. The first three methods were used as exploratory approaches, while the last two methods were used as confirmatory approaches. In addition to the dimensionality assessment procedures, model fit was investigated by comparing the fitted marginal observed-score distributions of (M)IRT models to the actual observed-score distributions to allow for visual assessment of model fit.

A simple way to detect dimensional structure is to draw item vector plots of two test forms to determine which items tend to be more discriminating or more difficult, whether different items from different content areas measure different ability composites, and how similar the vector profiles are between the two forms. The dimensional structure of the Spanish Language and French Language forms was examined using item vector plots. When an item is represented as a vector in a two-dimensional space, the overall discrimination of the item is denoted by the value of $MDISC$ (see Equation 2.5). The amount of discrimination is specified by the length of the vector, with the tail of the vector lying along the $P = .5$ equiprobability contour, for which $P$ is the probability of a
correct response. The discrimination parameters are constrained to be positive, so that vectors are located in either the first or the third quadrant. Therefore, easy items lie in the third quadrant and difficult items lie in the first quadrant. The item vector plots were drawn using the computer program *mathematica* (Wolfram Research, Inc., 2018).

Second, PCA with tetrachoric and polychoric correlations was conducted using the R package *psych* (Revelle, 2018) to investigate the dimensionality on the new form and old form datasets of Spanish and French. Scree plots of eigenvalues were plotted based on the PCA analyses to evaluate the amount of variance accounted for by each dimension. Third, the R package *sirt* (Robitzsch, 2019) was used to conduct DETECT to investigate the number of dimensions needed to adequately represent the data. A DETECT analysis was chosen because it provides more information than DIMTEST. The DETECT analysis results will be provided in Chapter 4 with respect to (a) the DETECT index, (b) the ASSI index, and (c) the R index.

While exploratory approaches do not assume any dimensional structure, dimensional structure is pre-defined in confirmatory approaches. A confirmatory factor analysis (CFA) was conducted based on the test specifications and was performed using the R package *lavaan* (Rosseel, 2012). Five statistics were investigated: the chi-square test, standardized root mean square error (SRMR), root mean square error of approximation (RMSEA), the Akaike Information Criteria (AIC; Akaike, 1974), and Bayesian Information Criterion (BIC; Schwarz, 1978). The AIC and BIC are calculated as,
\[ AIC = 2k - 2lnL, \]  
(3.4)

\[ BIC = \ln(N)k - 2lnL, \]  
(3.5)

where \( k \) is the number of estimated parameters and \( N \) is the number of examinees.

Since the minimization criterion is used to select a model, a model with the smallest statistic is selected. When the number of parameters is increased, the AIC gives less penalty; therefore, the AIC tends to select a more complex model. However, the BIC gives more penalty as the model complicates more.

The last dimensionality assessment procedure is disattenuate correlation. Content subdomains are defined here as the source of dimensionality. Let \( X_1 \) represent the listening section and \( X_2 \) represent the reading section of the Spanish and French forms. The disattenuated correlation is the correlation between true scores, which is calculated as

\[
\rho_{T_1T_2} = \frac{\rho_{X_1X_2}}{\sqrt{(\rho_{X_1X_1}')(\rho_{X_2X_2}')}},
\]  
(3.6)

where \( \rho_{X_1X_2} \) represents the observed correlation between scores on \( X_1 \) and \( X_2 \), \( \rho_{X_1X_1}' \) is the coefficient alpha of \( X_1 \) scores, and \( \rho_{X_2X_2}' \) is the coefficient alpha of \( X_2 \) scores. The coefficient alpha was used assuming that items within each subsections are essentially tau-equivalent. Disattenuated correlations close to 1 indicate the two subsections measure the same construct, meaning the unidimensionality assumption holds for the test form.
**Item Calibration and Scale Linking**

The computer program *flexMIRT* (Cai, 2017) was used for calibrating MC items for the Spanish and French forms. The 2PL versions of UIRT, SS, BF, and FM models were fitted to the data. The 2PL model was chosen to minimize possible errors when compared to the 3PL model, which contained the pseudo-guessing parameters. For the UIRT calibrations, scale indeterminacy was handled by letting ability distributions take the values of $\theta \sim N(0,1)$. Ability distributions for each form were specified to follow $\theta \sim MVN(\mathbf{0}, \mathbf{\Sigma})$ for the SS calibration, $\theta \sim MVN(\mathbf{0}, \mathbf{I})$ for the BF calibration, and both for the FM calibration, respectively. The off-diagonal elements of $\mathbf{\Sigma}$ are positive values between 0 and 1.

The SS and BF models are determined as confirmatory models with fixed content subdomains. In the SS model, items were assigned to load on one of the dimensions (i.e., listening, or reading) as specified by the test specifications. The BF model was estimated with one general dimension and two specific dimensions, with each item in the test belonging to either listening or reading. In contrast, the FM models were calibrated, resulting in two dimensions with correlations between dimensions freely estimated (abbreviated as FMC) or fixed at zero (abbreviated as FMO). In the educational testing context, it is not common to consider examinees’ latent traits as orthogonal. Therefore, the effect of treating latent abilities as freely estimated or orthogonal was investigated in this study.

For all item calibration, 21 evenly spaced quadrature points from -4 to 4 were used. Prior distributions were specified to estimate the slope parameters to avoid
convergence issues. The $a$-parameter prior was lognormal (0, 0.5) for all IRT models. This prior is used in the BILOG-MG (Zimowski et al., 2003) program so that item parameter estimates are within reasonable ranges for the UIRT model. The lognormal priors also let specific factors be non-negative estimates in the BF model (DeMars, 2013). The \textit{flexMIRT} program calibrates item parameters in the logistic metric. Therefore, in order to convert item parameters to the normal metric, “NormalMetric3PL=Yes” option should be used for UIRT models. However, this command only works for the UIRT models. In MIRT models, a slope parameter estimate ($a$) should be divided by 1.702 and an intercept parameter estimate ($c$) should be converted to a difficulty parameter estimate ($b$) by taking $b = -c/a$. This transformation was needed to be able to use other computer programs that employ different parameterization from \textit{flexmirt}.

After item and ability parameters were calibrated, scale linking procedures were conducted for UIRT and MIRT models. Detailed scale linking procedures for each model are described in Chapter 2. The next paragraph explains how to link the two test forms using the separate calibration method.

For the UIRT model, scale transformation was conducted for the new form using the Stocking-Lord method and the computer program \textit{STUIRT} (Kim & Kolen, 2004) was used. For the SS, BF, and FM models, scale transformation was conducted using the Min method (2007). The Min method was chosen for computational ease; however, the Min method was modified by substituting the rotation matrix with an identity matrix for the SS and BF models (Kim, 2017). The idea of substituting the rotation matrix was adapted from Kim (2017), although Kim (2017) only considered the BF model. The modification
is needed because rotating the coordinate axes may alter the constraint imposed on the pre-specified dimensions. For SS, all the items are loaded on one of the dimensions specified by test specifications. When the coordinate axes are rotated, some items will be outside the two-dimensional surfaces, which is not a simple structure anymore. Likewise, in the BF model, all the items lie on one of the two-dimensional surfaces, which are composed of the general dimension on one axis and the specific dimension on the other axis; the BF structure is lost when the coordinate axes are rotated. If only the scale indeterminacy is considered in the linking process, the problem of SS and BF structure can be solved. Therefore, the rotation matrix in the Min method was substituted by an identity matrix for both SS and BF models (i.e., rotation was not conducted). The R package plink (Weeks, 2010) and R (R Core Team, 2018) code written by the author were used for the MIRT scale linking procedures.

In multiple-group concurrent calibration, the old and new form item parameters were estimated simultaneously for both groups. The common item parameters were constrained to be equal for both groups. The latent distribution for the old form group was specified as a reference group where the means and variances are fixed. The means, variances, and covariances for the new form group were freely estimated. The concept of concurrent calibration is the same for both UIRT and MIRT frameworks. The end result of the concurrent calibration is that all of the item parameter and ability estimates are on the same scale. Therefore, an additional UIRT and MIRT scale linking process is not required in this case. Concurrent calibration was implemented by the flexMIRT program. Sample syntaxes of flexMIRT are provided in Appendix.
Equating Procedures

The UIRT, SS, BF, and FM observed score equating procedures were conducted as described in Chapter 2 for dichotomous items. For the UIRT equating method, number-correct score distributions were estimated for the old and new forms, and traditional equipercentile equating was applied. To compute the conditional observed score distributions, Lord and Wingersky (1984)’s recursive algorithm was used. Then, the conditional distributions were used to find the marginal distribution. The UIRT observed score equating methods were conducted using the computer program MULTeq (Lee, 2012).

Even though two populations are involved under the CINEG design, the equating function is defined for a single population. Therefore, the two populations should be combined to obtain a single population. To construct a synthetic population, a synthetic weight of 1.0 was given to the examinee group taking the new form. The chained equipercentile (CE) method without smoothing was used for comparison purposes because it does not assume the dimensionality of the data. The sample size for both Spanish and French was large enough (N=5,000), so smoothing was not considered.

The procedures involved in the SS observed score equating method are similar to those in the UIRT equating method; however, number-correct scores were conditioned on each content area. An estimated correlation between the listening and reading dimensions for the new form was used to create the quadrature distribution for the new form (i.e., synthetic population). Item parameter estimates for the old form and transformed item parameter estimates for the new form with the transformed new form quadrature
distribution were used to find the marginal ability distribution for the two forms. Using Equation (2.48), marginal ability distributions were found by summing conditional distributions over the new-form bivariate theta quadrature distribution. Finally, the estimated marginal distributions were used to equate the new form scores to the old form scale, using the traditional equipercentile method. The program MULTeq (Lee, 2012) was used for SS observed score equating.

In the BF observed score equating method, number-correct scores were conditioned on the general dimension ($\theta_G$) and the specific dimensions ($\theta_s$). Here, the number of specific dimensions corresponds to the number of content areas. A correlation between the listening and reading dimensions for the new form (i.e., zero) was used to create the quadrature distribution for the new form (i.e., synthetic population). Item parameter estimates for the old form and transformed item parameter estimates for the new form with the transformed new form quadrature distribution were used to find the marginal distribution for the two forms. Using Equation (2.49), marginal ability distributions were calculated by summing conditional distributions over an orthogonal trivariate ability distribution. After the marginal ability distributions were computed for the two examinee groups, traditional equipercentile equating was conducted using the program MULTeq (Lee, 2012).

For the FM observed score equating method, the latent ability distributions were specified to follow (a) a multivariate normal with correlations equal to the estimated values (i.e., FMC), or (b) a multivariate normal with fixed zero correlations (i.e., FMO). An estimated correlation between the listening and reading dimensions for the new form
(i.e., estimated values or zero, respectively) was used to create the quadrature distribution for the new form (i.e., synthetic population). Item parameter estimates for the old form and transformed item parameter estimates for the new form with the transformed new form quadrature distribution were used to find the marginal ability distribution for the two forms. Marginal ability distributions were found by summing conditional distributions over the new-form multivariate theta quadrature distribution using Equation (2.51). The FM equating methods were implemented using R code written by the author.

**Evaluation Criteria**

The Difference That Matters (DTM) index (Dorans et al., 2003) was used to determine an acceptable level of equating error. Differences between equating results greater than DTM mean that they are practically different. The DTM value of 0.5, which is a half of a score unit, was used in this study. Differences between (M)IRT equating methods and the CE equating method at each new form score were assessed for both conditional and overall statistics using the 0.5 criterion. The DTM index can be an alternative method to evaluate equating error among equating methods because the true equating relationship is not known.

**Study 2. Pseudo-Form and Pseudo-Group Analyses**

The purpose of Study 2 is to compare and evaluate the performance of the MIRT observed score equating methods using pseudo-form and pseudo-group data. One benefit of creating pseudo forms is that it is possible to include both MC and FR items in the common-item set. An AP exam was manipulated to pseudo forms so that common items contain both MC items and FR items, allowing item types to serve as the source of
dimensionality in this study. Criterion equating relationships could be established under the SG design because all examinees took the original test.

**Data Preparation**

**Intact forms.** The Spanish and French new forms from Study 1 were selected for use in Study 2. In addition to the MC items – Listening and Reading skills – included in Study 1, Study 2 also includes the FR items – Writing and Speaking skills. Writing and speaking skills were measured as FR items for both Spanish and French forms and all responses were scored from 0 to 5. The number of items and score points of the intact forms are provided in Table 3-5, and descriptive statistics of the intact forms are provided in Table 3-6.

**Pseudo forms.** The same procedures were conducted for both Spanish and French forms. Pseudo forms were created by dividing the items from one test form into two similar half test forms. The two pseudo forms were constructed to be as similar as possible to the original test based on their p-values and content areas. For example, Figure 3-1 illustrates pseudo-form construction for Spanish. Originally, the Spanish dataset contained MC score points of 70 and FR score points of 20, respectively. Five MC items were dropped to adjust score points evenly, and one FR item was dropped since each form needed one FR items and one common FR item. Therefore, the dataset was divided into two forms with MC score points of 25 and FR score points of 10 and with internal common items having MC score points of 10 and FR score points of 5. That is, each form has 25 MC and 2 FR items in total, and the common-item set has 10 MC and 1 FR items, so that common items occupied 30% of the total test score to expect stable
equating results. Kolen and Brennan (2014) suggested the common items should be at least 20% of the length of a total test containing 40 or more items. The common items were selected to be representative of the pseudo forms in consideration of the $p$-values and content subdomains. Descriptive statistics of the pseudo forms are presented in Table 3-7.

All examinees took the original test, which was divided into two forms; therefore, criterion equating relationships could be established as traditional equipercentile equating under the SG design. The SG equipercentile equating can serve as a criterion in the following sense. First, since all examinees took both pseudo forms, direct equating can be conducted using the entire pseudo forms instead of equating through the common-item set. Second, with the SG data, there is no group difference, which is a potential source of equating error. Finally, a large sample size could be used based on the original dataset with $N = 20,000$.

After the pseudo forms were created, the effect of group differences was examined. Kolen and Brennan (2014) noted that an effect size 0.3 or more could lead to considerable differences among equating methods, while an effect size of 0.1 or less may not cause dissimilarities in equating results. Therefore, this study used three levels of the common-item effect sizes of (a) .05, (b) .1, and (c) .3 to investigate the effect of group differences on equating results. Sampling 3,000 examinees for each form with external variables (i.e., parental education and ethnicity) was used to create pseudo groups. The sample size of 3,000 was chosen to guarantee stable equating results as well as the ease of computation time. The sampling process was as follows:
1. Categorize data with external variables: The original test had categories of 0-9 for parental education and 0-8 for ethnicity. Considering the number of examinees within each category, the external variables were recoded so that each category had categories of 0 to 4.

2. Sample 3,000 examinees for each of the old and the new forms based on the categorized external variables without replacement. To obtain group differences, for example, more examinees were sampled from the lower categories of parental education and ethnicity on the new form and fewer examinees were sampled from the higher categories of parental education and ethnicity on the old form.

3. Calculate effect size until 0.05, 0.1 or 0.3 is achieved, using Equation 3.7 below. Here, the effect size is a modification of the Glass’s $\Delta$ statistic (Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985), which is defined as the mean difference between the experimental and control group divided by the standard deviation of the control group. The experimental group and the control group are analogous to the new form and the old form groups, respectively. Glass (1981) argued that if several treatments were compared to the control group, it would be better to use the standard deviation from the control group. Therefore, this study adopted the Glass’s statistic for calculating the effect size:

$$Effect\ Size = \frac{x_n - \bar{x}_o}{s_o},$$

(3.7)
where \( n \) and \( o \) represents the new form and old form, respectively; \( \bar{x} \) is the average common-item scores; and \( s_o \) is the standard deviation of common-item scores of the old form.

4. Steps 2-3 are replicated until 50 pseudo-datasets are obtained. The 50 replications were employed to reduce the estimation time.

**Dimensionality Assessment**

As with the process used in Study 1, dimensionality assessment procedures were used to investigate the dimensionality of the Spanish and French pseudo forms including (a) PCA, (b) Poly-DETECT, (c) CFA, and (d) disattenuated correlation. Note that the dimensionality assessment was conducted based on the original data rather than the pseudo data due to the large number of pseudo forms (i.e., 50 pairs). Since the dimensionality assessment for each of the 50 pairs of the pseudo forms could not be conducted, it was assumed that the dimensional structure of the original data would not be substantially different from the dimensional structure of the pseudo forms. Recall that Study 2 defines multidimensionality source as item format. Therefore, the DETECT analysis was substituted with Poly-DETECT, which can handle both dichotomous and polytomous items. For the confirmatory approaches (i.e., CFA and disattenuated correlation), subsections were specified as being either MC or FR sections. In addition to the dimensionality assessment procedures, fitted marginal distributions of (M)IRT models were compared to the actual distributions.
Item Calibration and Scale Linking

The item calibration procedures follow the same process as Study 1; however, Study 2 was conducted using mixed-format tests, and thereby involved polytomous items. The computer program flexMIRT (Cai, 2017) was used for calibrating MC and FR items on each form. The 2PL versions of UIRT, SS, BF, and FM models and the graded response model (GR) were fitted for the polytomous items. The SS and BF models were estimated with factors defined by item format. The two-dimensional FM model was calibrated with correlation between dimensions freely estimated or fixed at zero. Then, scale linking for UIRT and MIRT models was conducted. The scale linking process is the same as Study 1 except it involved polytomous items. For the separate calibration method, common items that contain both MC and FR items were used to estimate the scale transformation by the Stocking-Lord method for the UIRT model, and by the Min method for the MIRT models. Concurrent calibration was also implemented as described in Study 1 with the mixed common items.

Equating Procedures

UIRT, SS, BF, and FM equating procedures were conducted as described in Chapter 2. Equating procedures are illustrated in this section that are specific to Study 2. For the UIRT equating method, a polytomous version of the Lord-Wingersky recursion formula (Hanson, 1994) was used to obtain the conditional distribution of number-correct scores. For SS equating, number-correct scores were conditioned on each item type and then conditional total-score distribution over the two item types for each form were calculated. For BF equating, number-correct scores were conditioned on one general
dimension and two group-specific dimensions that correspond to the MC and FR item types. For the FM equating method, conditional observed score distributions for MC items and FR items were found for two dimensions. Once conditional distributions were computed for UIRT and MIRT models, marginal distributions were specified using the same process and traditional equipercentile equating procedures were conducted as described in Study 1.

As an equating criterion, traditional equipercentile equating without smoothing was conducted under the SG design by using RAGE-RGEQUATE (Kolen, 2005). Each dataset had a large sample size of \( N = 20,000 \), and therefore, smoothing was not conducted. Total study conditions for Study 2 are presented in Table 3.8. Note that the MIRT equating procedures (i.e., SS, BF, FMC, and FMO) have the same number of conditions. Study 2 uses a total of 36 conditions (2 linking methods \( \times \) 3 effect sizes \( \times \) 6 equating methods). Equating was done with two types of score scales (i.e., raw scores and unrounded scale scores). The raw-to-scale score conversion tables for Spanish and French are provided in Tables 3.9 and 3.10, respectively.

**Evaluation Criteria**

To evaluate the observed score equating results, evaluation criteria were used in three ways: (a) conditional statistics, (b) overall statistics, and (c) difference that Matters. Bias, Standard Errors (SE), and Root Mean Squared Error (RMSE) were calculated at each raw and scale score point as:

$$\text{Bias}_i = \frac{\sum_{r=1}^{50} (dY(x_i) - eY(x_i))}{50},$$  
(3.8)
\[ SE_i = \sqrt{\frac{\sum_{i=1}^{50}(\hat{e}_Y(x_i) - \bar{e}_Y(x_i))^2}{50}}, \]  \tag{3.9} \\
\[ RMSE_i = \sqrt{\frac{\sum_{i=1}^{50}(\hat{e}_Y(x_i) - e_Y(x_i))^2}{50}} = \sqrt{Bias_i^2 + SE_i^2}, \]  \tag{3.10}

where \( i \) is a score point; \( x_i \) is raw or scale score at point \( i \); \( \hat{e}_Y(x_i) \) represents the old form equivalent of new form score \( x_i \); \( e_Y(x_i) \) is the criterion equivalent; and \( \bar{e}_Y(x_i) \) represents the average, over 50 replications, of the equivalents of new form score \( x_i \). The overall statistics were calculated by averaging across all score points. Here, the unweighted and weighted averages were used. The weighted averages were calculated using the proportion of examinees scoring at particular raw or scale score point for the new form in the criterion equating relationship.

**Summary**

Chapter 3 describes MIRT equating assumptions and procedures regarding assumptions, dimensionality assessments, scale linking methods, and equating procedures. Table 3.11 illustrates major differences between the two studies. While Study 1 was conducted using only MC items, the focus of Study 2 was mixed-format tests that contain both MC and FR items. The two studies have different sources of dimensionality: content subdomains for Study 1 and item types for Study 2, respectively. Since the true equating relationship is unknown, DTM was used as an equating criterion for Study 1, whereas Study 2 has an equipercentile equating criterion under the SG design. Study 1 was conducted one time, whereas Study 2 was replicated 50 times with pseudo groups and pseudo forms.
Table 3-1. Study 1: Characteristics of the Spanish and French exams

<table>
<thead>
<tr>
<th>Test</th>
<th>Subdomains</th>
<th>Test Length</th>
<th>Number of Common Items</th>
<th>Sample Size (New / Old)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>New</td>
<td>Old</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>Total</td>
<td>70</td>
<td>70</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Listening</td>
<td>34</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>36</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td>French</td>
<td>Total</td>
<td>65</td>
<td>65</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Listening</td>
<td>35</td>
<td>35</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>30</td>
<td>30</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 3-2. Study 1: Descriptive statistics of samples for Spanish and French

<table>
<thead>
<tr>
<th>Test</th>
<th>Form</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>New</td>
<td>47.4134</td>
<td>11.2351</td>
<td>-0.4827</td>
<td>3.0139</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>45.4898</td>
<td>11.3850</td>
<td>-0.4363</td>
<td>2.9357</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>French</td>
<td>New</td>
<td>43.4698</td>
<td>11.5669</td>
<td>-0.4412</td>
<td>2.4083</td>
<td>7</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>43.6844</td>
<td>11.5268</td>
<td>-0.3602</td>
<td>2.4022</td>
<td>7</td>
<td>65</td>
</tr>
</tbody>
</table>
### Table 3-3. *Study 1: Raw-to-scale score conversion table for Spanish*

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Unrounded scale score</th>
<th>Raw Score</th>
<th>Unrounded scale score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9.5213</td>
<td>36</td>
<td>20.8711</td>
</tr>
<tr>
<td>1</td>
<td>10.6268</td>
<td>37</td>
<td>21.2685</td>
</tr>
<tr>
<td>2</td>
<td>10.6980</td>
<td>38</td>
<td>21.6659</td>
</tr>
<tr>
<td>3</td>
<td>10.7511</td>
<td>39</td>
<td>22.0631</td>
</tr>
<tr>
<td>4</td>
<td>10.7936</td>
<td>40</td>
<td>22.4805</td>
</tr>
<tr>
<td>5</td>
<td>10.8381</td>
<td>41</td>
<td>22.8833</td>
</tr>
<tr>
<td>6</td>
<td>10.8928</td>
<td>42</td>
<td>23.2529</td>
</tr>
<tr>
<td>7</td>
<td>10.9294</td>
<td>43</td>
<td>23.6562</td>
</tr>
<tr>
<td>8</td>
<td>10.9976</td>
<td>44</td>
<td>24.0640</td>
</tr>
<tr>
<td>9</td>
<td>11.1091</td>
<td>45</td>
<td>24.4765</td>
</tr>
<tr>
<td>10</td>
<td>11.2046</td>
<td>46</td>
<td>24.8899</td>
</tr>
<tr>
<td>11</td>
<td>11.3357</td>
<td>47</td>
<td>25.3124</td>
</tr>
<tr>
<td>12</td>
<td>11.4910</td>
<td>48</td>
<td>25.7349</td>
</tr>
<tr>
<td>13</td>
<td>11.6624</td>
<td>49</td>
<td>26.1461</td>
</tr>
<tr>
<td>14</td>
<td>11.8750</td>
<td>50</td>
<td>26.5623</td>
</tr>
<tr>
<td>15</td>
<td>12.1300</td>
<td>51</td>
<td>26.9956</td>
</tr>
<tr>
<td>16</td>
<td>12.4342</td>
<td>52</td>
<td>27.4367</td>
</tr>
<tr>
<td>17</td>
<td>12.7909</td>
<td>53</td>
<td>27.8928</td>
</tr>
<tr>
<td>18</td>
<td>13.1934</td>
<td>54</td>
<td>28.3609</td>
</tr>
<tr>
<td>19</td>
<td>13.6038</td>
<td>55</td>
<td>28.8442</td>
</tr>
<tr>
<td>20</td>
<td>14.0243</td>
<td>56</td>
<td>29.3493</td>
</tr>
<tr>
<td>21</td>
<td>14.4762</td>
<td>57</td>
<td>29.8730</td>
</tr>
<tr>
<td>22</td>
<td>14.9341</td>
<td>58</td>
<td>30.4146</td>
</tr>
<tr>
<td>23</td>
<td>15.3842</td>
<td>59</td>
<td>30.9875</td>
</tr>
<tr>
<td>24</td>
<td>15.8317</td>
<td>60</td>
<td>31.5949</td>
</tr>
<tr>
<td>25</td>
<td>16.2754</td>
<td>61</td>
<td>32.2306</td>
</tr>
<tr>
<td>26</td>
<td>16.7203</td>
<td>62</td>
<td>32.9073</td>
</tr>
<tr>
<td>27</td>
<td>17.1604</td>
<td>63</td>
<td>33.6377</td>
</tr>
<tr>
<td>28</td>
<td>17.5886</td>
<td>64</td>
<td>34.4598</td>
</tr>
<tr>
<td>29</td>
<td>18.0058</td>
<td>65</td>
<td>35.3719</td>
</tr>
<tr>
<td>30</td>
<td>18.4204</td>
<td>66</td>
<td>36.3714</td>
</tr>
<tr>
<td>31</td>
<td>18.8388</td>
<td>67</td>
<td>37.5646</td>
</tr>
<tr>
<td>32</td>
<td>19.2573</td>
<td>68</td>
<td>39.0874</td>
</tr>
<tr>
<td>33</td>
<td>19.6694</td>
<td>69</td>
<td>40.5</td>
</tr>
<tr>
<td>34</td>
<td>20.0765</td>
<td>70</td>
<td>40.5</td>
</tr>
<tr>
<td>35</td>
<td>20.4765</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: Conversion table is for research purposes only.*
Table 3-4. *Study 1: Raw-to-scale score conversion table for French*

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Unrounded scale score</th>
<th>Raw Score</th>
<th>Unrounded scale score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6.6108</td>
<td>36</td>
<td>22.1401</td>
</tr>
<tr>
<td>1</td>
<td>6.6108</td>
<td>37</td>
<td>22.5129</td>
</tr>
<tr>
<td>2</td>
<td>6.6108</td>
<td>38</td>
<td>22.8697</td>
</tr>
<tr>
<td>3</td>
<td>6.6108</td>
<td>39</td>
<td>23.2076</td>
</tr>
<tr>
<td>4</td>
<td>6.6108</td>
<td>40</td>
<td>23.5571</td>
</tr>
<tr>
<td>5</td>
<td>7.6724</td>
<td>41</td>
<td>23.9099</td>
</tr>
<tr>
<td>6</td>
<td>7.6724</td>
<td>42</td>
<td>24.2588</td>
</tr>
<tr>
<td>7</td>
<td>7.9441</td>
<td>43</td>
<td>24.6219</td>
</tr>
<tr>
<td>8</td>
<td>7.9441</td>
<td>44</td>
<td>24.9959</td>
</tr>
<tr>
<td>9</td>
<td>7.9441</td>
<td>45</td>
<td>25.3895</td>
</tr>
<tr>
<td>10</td>
<td>8.6274</td>
<td>46</td>
<td>25.7788</td>
</tr>
<tr>
<td>11</td>
<td>9.4117</td>
<td>47</td>
<td>26.1615</td>
</tr>
<tr>
<td>12</td>
<td>10.1608</td>
<td>48</td>
<td>26.5597</td>
</tr>
<tr>
<td>13</td>
<td>11.0566</td>
<td>49</td>
<td>26.9764</td>
</tr>
<tr>
<td>14</td>
<td>11.7514</td>
<td>50</td>
<td>27.4139</td>
</tr>
<tr>
<td>15</td>
<td>12.2482</td>
<td>51</td>
<td>27.8673</td>
</tr>
<tr>
<td>16</td>
<td>12.8516</td>
<td>52</td>
<td>28.3208</td>
</tr>
<tr>
<td>17</td>
<td>13.4382</td>
<td>53</td>
<td>28.8069</td>
</tr>
<tr>
<td>18</td>
<td>14.0088</td>
<td>54</td>
<td>29.3572</td>
</tr>
<tr>
<td>19</td>
<td>14.5931</td>
<td>55</td>
<td>29.9331</td>
</tr>
<tr>
<td>20</td>
<td>15.1098</td>
<td>56</td>
<td>30.5381</td>
</tr>
<tr>
<td>21</td>
<td>15.5974</td>
<td>57</td>
<td>31.1824</td>
</tr>
<tr>
<td>22</td>
<td>16.0861</td>
<td>58</td>
<td>31.8981</td>
</tr>
<tr>
<td>23</td>
<td>16.5454</td>
<td>59</td>
<td>32.744</td>
</tr>
<tr>
<td>24</td>
<td>16.9803</td>
<td>60</td>
<td>33.7372</td>
</tr>
<tr>
<td>25</td>
<td>17.4088</td>
<td>61</td>
<td>34.9349</td>
</tr>
<tr>
<td>26</td>
<td>17.8365</td>
<td>62</td>
<td>36.4273</td>
</tr>
<tr>
<td>27</td>
<td>18.2698</td>
<td>63</td>
<td>38.3369</td>
</tr>
<tr>
<td>28</td>
<td>18.6741</td>
<td>64</td>
<td>40.5</td>
</tr>
<tr>
<td>29</td>
<td>19.0648</td>
<td>65</td>
<td>40.5</td>
</tr>
<tr>
<td>30</td>
<td>19.4653</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>19.8481</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>20.246</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>20.6616</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>21.0487</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>21.4197</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Conversion table is for research purposes only.
Table 3-5. *Study 2: Characteristics of Spanish and French*

<table>
<thead>
<tr>
<th>Test</th>
<th>Section</th>
<th>Intact Forms</th>
<th>Pseudo Forms</th>
<th>Score Points</th>
<th>Score Points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N items</td>
<td>Score</td>
<td>N items</td>
<td>Points</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Points</td>
<td></td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>New</td>
<td>Old</td>
<td>CI</td>
</tr>
<tr>
<td>Spanish</td>
<td>MC</td>
<td>70</td>
<td>(0-1)</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>4</td>
<td>(0-5)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>French</td>
<td>MC</td>
<td>65</td>
<td>(0-1)</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>4</td>
<td>(0-5)</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3-6. *Study 2: Descriptive statistics of intact forms for Spanish and French*

<table>
<thead>
<tr>
<th>Test</th>
<th>Section</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>Total</td>
<td>60.855</td>
<td>13.561</td>
<td>-0.5527</td>
<td>3.1119</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>2012</td>
<td>MC</td>
<td>47.662</td>
<td>11.275</td>
<td>-0.5461</td>
<td>3.1873</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>13.217</td>
<td>3.147</td>
<td>-0.6888</td>
<td>3.7711</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>French</td>
<td>Total</td>
<td>55.650</td>
<td>14.754</td>
<td>-0.3757</td>
<td>2.4648</td>
<td>1</td>
<td>85</td>
</tr>
<tr>
<td>2012</td>
<td>MC</td>
<td>43.864</td>
<td>11.450</td>
<td>-0.3826</td>
<td>2.4225</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>11.786</td>
<td>4.015</td>
<td>-0.2890</td>
<td>2.6826</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>
Table 3-7. Study 2: Descriptive statistics of pseudo forms for Spanish and French

<table>
<thead>
<tr>
<th>Test</th>
<th>Section</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New Form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composite</td>
<td></td>
<td>29.2331</td>
<td>6.8681</td>
<td>-0.4243</td>
<td>3.0171</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td>23.1819</td>
<td>5.8457</td>
<td>-0.4565</td>
<td>3.0663</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>6.0513</td>
<td>1.5766</td>
<td>-0.1253</td>
<td>3.3602</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Old Form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>Composite</td>
<td>31.5427</td>
<td>7.0891</td>
<td>-0.5779</td>
<td>3.0270</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td>24.3581</td>
<td>5.8886</td>
<td>-0.5555</td>
<td>3.1135</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>7.1846</td>
<td>1.8312</td>
<td>-0.6967</td>
<td>3.3083</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Common Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td>6.3152</td>
<td>2.3493</td>
<td>-0.4640</td>
<td>2.6418</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>3.2910</td>
<td>0.9358</td>
<td>-0.0674</td>
<td>3.0769</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>New Form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composite</td>
<td></td>
<td>29.1204</td>
<td>7.8717</td>
<td>-0.3519</td>
<td>2.5149</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td>22.9772</td>
<td>6.3630</td>
<td>-0.3447</td>
<td>2.4479</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>6.1432</td>
<td>1.9729</td>
<td>-0.2415</td>
<td>2.8293</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Old Form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>Composite</td>
<td>30.4690</td>
<td>8.1401</td>
<td>-0.4457</td>
<td>2.4672</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td>24.3433</td>
<td>6.5115</td>
<td>-0.4874</td>
<td>2.4793</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>6.1256</td>
<td>2.0921</td>
<td>-0.2348</td>
<td>2.5302</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Common Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td>6.9729</td>
<td>2.2152</td>
<td>-0.5757</td>
<td>2.6144</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>3.0970</td>
<td>1.0026</td>
<td>0.0295</td>
<td>2.7670</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
### Table 3-8. Study 2: Study conditions

<table>
<thead>
<tr>
<th>Test</th>
<th>Model</th>
<th>Linking Method</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UIRT</td>
<td>Stocking-Lord</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>Concurrent</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>SS, BF,</td>
<td>Min</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>FMC, FMO</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Concurrent</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chained equipercentile</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>UIRT</td>
<td>Stocking-Lord</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>Concurrent</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>SS, BF,</td>
<td>Min</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>FMC, FMO</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Concurrent</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chained equipercentile</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
</tbody>
</table>
Table 3-9. *Study 2: Raw-to-scale score conversion table for Spanish*

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Unrounded scale score</th>
<th>Raw Score</th>
<th>Unrounded scale score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4.7219</td>
<td>36</td>
<td>27.6281</td>
</tr>
<tr>
<td>1</td>
<td>4.7219</td>
<td>37</td>
<td>28.4607</td>
</tr>
<tr>
<td>2</td>
<td>4.7219</td>
<td>38</td>
<td>29.3345</td>
</tr>
<tr>
<td>3</td>
<td>4.7219</td>
<td>39</td>
<td>30.2567</td>
</tr>
<tr>
<td>4</td>
<td>6.9235</td>
<td>40</td>
<td>31.2548</td>
</tr>
<tr>
<td>5</td>
<td>8.2360</td>
<td>41</td>
<td>32.3567</td>
</tr>
<tr>
<td>6</td>
<td>8.9201</td>
<td>42</td>
<td>33.6023</td>
</tr>
<tr>
<td>7</td>
<td>9.7576</td>
<td>43</td>
<td>35.1234</td>
</tr>
<tr>
<td>8</td>
<td>10.7236</td>
<td>44</td>
<td>37.0791</td>
</tr>
<tr>
<td>9</td>
<td>11.4918</td>
<td>45</td>
<td>39.6225</td>
</tr>
<tr>
<td>10</td>
<td>12.0594</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>12.5396</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>13.0228</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13.5568</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>14.1230</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>14.7131</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>15.2552</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>15.8045</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>16.3922</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>16.9774</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>17.5732</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>18.1702</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>18.7500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>19.3205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>19.9129</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>20.4932</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>21.0562</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>21.6425</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>22.2316</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>22.8335</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>23.4555</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>24.0806</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>24.7288</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>25.4062</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>26.1017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>26.8388</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Conversion table is for research purposes only.
Table 3-10. Study 2: Raw-to-scale score conversion table for French

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Unrounded scale score</th>
<th>Raw Score</th>
<th>Unrounded scale score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4.9122</td>
<td>36</td>
<td>27.6753</td>
</tr>
<tr>
<td>1</td>
<td>4.9122</td>
<td>37</td>
<td>28.3609</td>
</tr>
<tr>
<td>2</td>
<td>6.6108</td>
<td>38</td>
<td>29.1099</td>
</tr>
<tr>
<td>3</td>
<td>6.6108</td>
<td>39</td>
<td>29.9403</td>
</tr>
<tr>
<td>4</td>
<td>7.3374</td>
<td>40</td>
<td>30.8826</td>
</tr>
<tr>
<td>5</td>
<td>8.1732</td>
<td>41</td>
<td>31.9326</td>
</tr>
<tr>
<td>6</td>
<td>9.1543</td>
<td>42</td>
<td>33.1449</td>
</tr>
<tr>
<td>7</td>
<td>10.1608</td>
<td>43</td>
<td>34.7360</td>
</tr>
<tr>
<td>8</td>
<td>11.0566</td>
<td>44</td>
<td>36.8348</td>
</tr>
<tr>
<td>9</td>
<td>11.8824</td>
<td>45</td>
<td>39.6674</td>
</tr>
<tr>
<td>10</td>
<td>12.6979</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>13.4328</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>14.1450</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>14.8491</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>15.5626</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>16.2594</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16.9238</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>17.5459</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>18.1286</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>18.6938</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>19.2415</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>19.7653</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>20.2806</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>20.7843</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>21.2666</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>21.7542</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>22.2642</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>22.7675</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>23.2566</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>23.7482</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>24.2566</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>24.7707</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>25.2924</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>25.8318</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>26.3926</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>27.0135</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Conversion table is for research purposes only.
Table 3-11. Differences between Study 1 and Study 2

<table>
<thead>
<tr>
<th></th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis</td>
<td>Intact form</td>
<td>Pseudo group and Pseudo form</td>
</tr>
<tr>
<td>Test Composition</td>
<td>MC</td>
<td>MC+FR</td>
</tr>
<tr>
<td>Source of Dimensionality</td>
<td>Content subdomain</td>
<td>Item types</td>
</tr>
<tr>
<td>Equating Criteria</td>
<td>DTM</td>
<td>Single group equipercentile</td>
</tr>
<tr>
<td>Replication</td>
<td>1</td>
<td>50</td>
</tr>
</tbody>
</table>

*Figure 3-1. Pseudo-form construction for the Spanish exam*
CHAPTER 4

RESULTS

Chapter 4 is composed of four sections. The first section provides preliminary study results. The preliminary study examined the relationship between correlation and item parameters and the possible impact on equating results. The second section presents Study 1 results and the third section provides Study 2 results, including equating results for the UIRT, SS, BF, and full MIRT methods (i.e., freely estimated and orthogonal) for Spanish and French with dimensionality assessments. The final section summarizes findings from the studies.

Preliminary Study

The main purpose of the preliminary study was to examine how the correlation between dimensions of the two forms impacts item parameter estimates using a small-scale simulation study. The study also explored how correlations between dimensions are set when calibrating FM models. Thus, the study provided a preliminary evidence for how correlations between dimensions impact MIRT equating results under the CINEG design.

First, Table 4-1 compares true item parameters with mean values of estimated item parameters with correlations between dimensions (a) freely estimated, (b) fix at zero, and (c) fix at 0.2. As can be seen from the table, the $a_1$ and $a_2$ parameters were well recovered when correlations were freely estimated. On the contrary, the $d$ parameters were recovered most accurately when the correlation between dimension was zero. A similar tendency was found in a previous study (Svetina et al., 2017), in which the
discrimination and difficulty parameters were well-recovered across different conditions. The same table is presented for the correlation of 0.75 in Table 4-2. Compared to Table 4-1, the \(a_1\) and \(a_2\) parameters were best recovered for correlations freely estimated or fixed at 0.75. The \(d\) parameters were well-recovered for the “Fix at 0” or “Fix at 0.75” conditions. Tables 4-3 and 4-4 represent the item parameter recovery when correlations were 0.8 and 0.95, respectively. As the correlations became larger, the discrimination parameters (i.e., \(a_1\) and \(a_2\)) were recovered most accurately when the correlation was fixed at the true value. On the other hand, the \(d\) parameters were best recovered for the “Fix at zero \((r=0.8)\)” and “Fix at true value \((r =0.95)\)” conditions, respectively.

Table 4-5 represents the values of ABAIS, ASE, and ARMSE for all three calibration conditions (i.e., freely estimated, fix at zero, and fix at true value) with respect to MIRT item parameters. One notable observation was that the ABIAS, ASE, and ARMSE values were smaller when the correlation between dimensions was specified at true values. This phenomenon was predictable because the true item parameters were generated from its true correlation. In practice, however, the true correlations are never known. Therefore, the “Freely estimated” option is the second best in recovering MIRT item parameters and is the most realistic way to specify correlations. Some of the discrimination parameters with the “Freely estimated” option were recovered better than the “Fix at true value” option when the correlations were relatively lower (i.e., \(r = 0.2\) and 0.75). For most of the conditions, the recovery of the MIRT item discrimination parameters was poorer when the correlations were fixed at zero. This is an expected finding because an examinee’s latent traits are not strictly orthogonal. On the contrary,
the MIRT item difficulty parameters were recovered inconsistently across the four correlation conditions.

When the correlations became larger, ABIAS and ARMSE tended to be larger for the discrimination parameters. The ASE values for the discrimination parameters were larger, but slightly lower at the largest correlation (i.e., \( r = 0.95 \)). However, no clear trends were apparent in ABAIS, ASE, and ARMSE for the difficulty parameters. Overall, it seems that the correlation between dimensions may not impact heavily on recovering MIRT item parameters.

Several conclusions can be made from the Preliminary Analysis. First, correlation between dimensions does not have substantial effects on the recovery of MIRT item parameters. Therefore, even though the correlation between the dimensions for the two test forms differ, it may not greatly impact the MIRT equating results under the CINEG design. Second, when FM models are calibrated, correlation between dimensions should be freely estimated to better reflect the underlying structure of examinees’ latent abilities.
Study 1: Real Data Analyses

Dimensionality Assessment

In this section, dimensionality assessment procedures are presented for each subject area, followed by equating results.

In Figure 4-1, item vector plots for the Spanish form administered in 2011 and its common items are displayed. Also, in Figure 4-2, item vector plots are plotted for the Spanish form administered in 2012 and its common items. Both figures suggest that the common items had vector patterns similar to the total test, supporting the assumption that the common items should have the same dimensional structure as the total test. When comparing Figures 4-1 and 4-2, the Spanish 2011 and 2012 forms demonstrate similar vector patterns. The arrows point in multiple directions, suggesting that both forms measure more than one dimension. It can be observed that the Spanish 2011 form was more difficult and more discriminating than the Spanish 2012 form because more arrows were spotted in the first quadrant. Also, the arrows in the Spanish 2011 form were steeper than those of the Spanish 2011 form.

In Figure 4-3, item vector plots for the French form administered in 2012 and its common items are displayed. Also, in Figure 4-4, item vector plots are plotted for the French form administered in 2013 and its common items. Similar to Spanish, the common items had vector patterns similar to the total test, and the French 2012 and 2013 forms displayed similar vector patterns. It can be observed that the French 2012 form was more difficult and more discriminating than the French 2013 form because more arrows
were spotted in the first quadrant. Also, the arrows in the French 2012 form were little steeper than those of the French 2013 form.

Second, principal components analysis (PCA) was conducted to test the unidimensionality assumption. Scree plots were drawn in Figure 4-5 for Spanish and French, with the upper plot describing the Spanish forms and the lower for the French forms. The scree plots suggested a multi-dimensional solution may provide better fit for Spanish. For Spanish, the line flattened from the third factor for the new form, whereas the line started to flatten from the second factor for the old form. On the contrary, French showed more evidence of unidimensionality.

Dimensionality was next examined by running DETECT in an exploratory mode. Note that DETECT already assumes multidimensionality, so the results are presented from a two-factor solution, rather than a 1-factor solution. In this study, DETECT was conducted to explore two- to four-dimensional solutions. The DETECT index that maximized the partition was chosen. Table 4-6 presents the DETECT results for Spanish and French. As expected, higher dimensional solutions were favored for both the Spanish and French forms. For the Spanish new and old forms, the DETECT index showed weak multidimensionality (i.e., $0.2 < \text{DETECT index} < 0.4$) when a four-dimensional solution was used. The ASSI and R indices also demonstrated essential deviation from unidimensionality (i.e., $\text{ASSI} > .25$ and $R > .36$, respectively) with the four-dimensional solution. For French with a four-factor solution, the DETECT index showed weak multidimensionality for the new form, but essential unidimensionality for the old
form. In contrast, the ASSI and R indices demonstrated essential deviation from unidimensionality for both the new and old forms.

Both PCA and DETECT results indicated that the Spanish and French forms had potential for multidimensionality. Note that, in this study, the number of dimensions for each of the exams was determined by the table of specifications, the results for the exploratory dimensionality assessment are used only for supporting evidence.

A confirmatory factor analysis (CFA) was next conducted for Spanish and French to evaluate the fit of a unidimensional structure versus a two-dimensional structure. The first CFA model assumed that all items in each of the Spanish and French forms load on to a single factor, whereas the second CFA model assumed a two-factor (listening and reading) structure. Results for several goodness-of-fit measures are summarized in Table 4-7. Based on the chi-square $p$-values in Table 4-7, the predicted covariance matrices for all CFA models were statistically different from the sample covariance matrices for both Spanish and French. Other fit indices were the standardized root mean square error (SRMR) and root mean square error approximation (RMSEA). According to the SRMR and RMSEA indices, a two-factor CFA model had a good model fit for Spanish and French. The two-factor CFA model had slightly smaller SRMR and RMSEA values than the one-factor CFA model. Finally, in terms of AIC and BIC values, the two-factor CFA model favored overall for both forms of Spanish and French. The last dimensionality procedure was disattenuated correlation. In Table 4-8, boldface numbers are disattenuated correlations, italic numbers are observed correlations, and numbers in the diagonal are Cronbach’s alpha. Disattenuated correlations indicated that two sections within the same
form measure somewhat different constructs. For Spanish, the disattenuated correlation between listening and reading was 0.8731 for the new form and 0.8579 for the old form. For French, the disattenuated correlation between listening and reading was 0.9308 for the new form and 0.9212 for the old form. Overall, the old forms demonstrated slightly lower disattenuated correlation than the new forms for both Spanish and French.

Finally, model fit was investigated by inspecting the fitted distributions of (M)IRT models. The fitted distributions of (M)IRT models are presented in Figure 4-6 for Spanish and Figure 4-7 for French. The fitted distributions for all models appeared to fit the actual distributions reasonably well for Spanish, with the exception of the FMO method. FMO slightly underfitted the distributions for the both forms of Spanish compared to the other equating methods. For French, the fitted distributions closely approximated the actual distributions, although some discrepancies were found for the new form using the UIRT method.

**Equating Results**

Figures 4-8 and 4-9 display six equating results for Spanish and French raw scores, respectively. The upper plot presents equating results using separate calibration and the lower plot presents concurrent calibration. Equating results are only displayed for a raw (or scale) score of 10 to the maximum possible score because very few frequencies were found for the raw (or scale) scores below 10. The graphs were drawn with respect to differences in the results between each equating method and identity equating (i.e., new form raw score). As seen in Figure 4-8, results for (M)IRT equating methods for both separate and concurrent calibration showed a similar curvilinear equating relationship,
although differences between equating methods were somewhat attenuated in concurrent calibration. Figure 4-9 demonstrates that the plot of separate calibration resembled that of concurrent calibration.

Figures 4-10 and 4-11 present differences between the (M)IRT equating methods and the chained equipercentile equating method for Spanish and French, respectively. The upper two plots depict separate calibration for raw score and unrounded scale score and the lower two plots represent concurrent calibration for raw score and unrounded scale score. The two straight lines represent the DTM criterion (i.e., ± 0.5) which can be used to determine equating results from a practical standpoint.

**Spanish.** In Figure 4-10, generally similar patterns were found between equating results using separate and concurrent calibration. For raw score equating using separate calibration, FM methods were closest to the CE method and were within DTM across the entire score scale. The UIRT and BF methods deviated more from the DTM boundaries than the other MIRT methods, falling outside DTM at a score range of 10-40 and 10-55, respectively. For raw score equating using concurrent calibration, FMC, FMO, and SS were for the most part within the DTM standard. The UIRT and BF methods fell outside the DTM criterion at a score range of 10-40, but were within the DTM boundaries at the score range of 40-70. For scale score equating using both separate and concurrent calibration, there were not large discrepancies among the equating methods. Most patterns were within DTM except for the maximum score of 70. This is possibly due to there being very few frequencies at the maximum scores.
**French.** Differences between the (M)IRT equating methods and the CE method for French are presented in Figure 4-11. For raw score equating using separate calibration, the UIRT and BF methods were outside of the DTM standard at a score range of 10-20. The equivalents of FMC, FMO, and SS generally maintained similar patterns and were within the DTM criterion across the score scale. For raw score equating with concurrent calibration, most of the UIRT and BF methods fell outside of the DTM boundaries, except for within a score range of 50-65. Several of the FMO equivalents deviated from the DTM criterion, while FMC and SS remained consistently inside DTM. For unrounded scale score equating using both separate and concurrent calibration, discrepancy patterns were relatively flat across the score scale. Equivalents for the equating methods were for the most part within the DTM boundaries for both separate and concurrent calibration, although some equivalents were outside of the criteria at the lower and upper score ranges.

**Study 2: Pseudo-Form and Pseudo-Group Analyses**

**Dimensionality Assessment**

Similar to Study 1, Study 2 investigates four dimensionality assessment approaches. To determine the amount of variance accounted for by each dimension, scree plots were drawn in Figure 4-12, with the upper plot describing the Spanish forms and the lower for the French forms. For Spanish, the line flattened from the second factor for the new form, whereas the line started to flatten from the third factor for the old form. Therefore, the data for Spanish was better explained by more than one dimension. French showed evidence of unidimensionality in that the scree plots were relatively flat.
Considering that the data used in Study 2 were created by splitting the data used in Study 1, the patterns of the scree plots were expectedly similar to Study 1.

Table 4-9 presents the poly-DETECT results for Spanish and French. As expected, higher dimensional solutions were favored for both the Spanish and French forms. For the Spanish new and old forms, the DETECT index indicated moderate multidimensionality when a four-dimensional solution was used, and the ASSI and R indices indicated essential deviation from unidimensionality. French showed weak multidimensionality with the four-dimensional solution in terms of the DETECT index. The ASSI and R indices indicated essential deviation from unidimensionality, except for the ASSI index for the old form.

In addition to the two exploratory approaches, CFA results are provided in Table 4-10. In contrast to Study 1, the second CFA model assumed a two-factor (MC and FR) structure. Based on the chi-square p-values, the predicted covariance matrices of all CFA models were statistically different from the sample covariance matrices. According to the SRMR and RMSEA indices, both of CFA models were a good fit for Spanish and French; therefore, it was hard to decide which model fitted better with SRMR and RMSEA indices. Instead, smaller values of AIC and BIC indicated the two-factor CFA model was favored for both forms of Spanish and French.

The disattenuated correlations in Table 4-11 indicate that MC and FR sections in the Spanish and French forms measure distinct constructs. For Spanish, the disattenuated correlation between MC and FR sections were 0.8652 for the new form and 0.8211 for the old form. For French, the disattenuated correlation between MC and FR
sections for the new and old forms were 0.8856 and 0.9310, respectively. For Spanish, the old form was slightly more multidimensional than the new form, while the new form was more multidimensional than the old form for French. The underlying constructs were more distinct in Study 2 than those in Study 1 because the disattenuated correlations of MC and FR sections for both subjects were smaller.

Finally, model fit was investigated by the fitted distributions of (M)IRT models that are presented in Figure 4-13 for Spanish and Figure 4-14 for French, respectively. The fitted distributions for all (M)IRT models appeared to fit the actual distributions well for Spanish. The full MIRT models slightly deviated from the actual distribution more than the other equating methods for the both forms of Spanish. For French, the MIRT models closely approximated the actual distributions, although the UIRT model appeared to deviate from the actual distribution for the both forms.

Spanish: Conditional Statistics

Figures 4-15 to 4-20 provide conditional bias, SE, and RMSE for Spanish raw scores and unrounded scale scores, respectively. The upper three plots represent conditional statistics using separate calibration and the lower three plots represent those using concurrent calibration. The plots in each column represent the effect size for group differences (i.e., 0.05, 0.1, and 0.3). For illustrative purposes, the effect sizes of 0.05, 0.1, and 0.3 are denoted here as small, medium, and large group difference, respectively. Similar to Study 1, equating results were presented from a raw (or scale) score of 10 to the maximum possible score because very few frequencies were found for raw (or scale) scores below 10 for each subject. In each subsection, conditional bias is described first,
followed by conditional SE and conditional RMSE. The equating results for each statistic are presented first as raw scores, followed by scale scores.

**Conditional bias.** Figure 4-15 presents conditional bias for Spanish raw scores. For separate calibration, UIRT introduced a large amount of negative at the lower score range and positive bias at the upper score ranges when group difference was large (i.e., effect size = 0.3). The UIRT method yielded substantial negative bias that exceeded -0.5, particularly for scores of approximately 20 or below, regardless of group difference. Also, when comparing three plots for all three effect sizes, BF and UIRT were more affected by group difference than the other methods because the graph patterns drastically changed. The full MIRT methods (i.e., FMC and FMO) maintained similar values of bias across the score scale, except at the upper score range for the effect size of 0.3. The patterns of bias for the CE and SS methods were moderately flat across the entire score range, except at the lower score range for the effect size 0.3 condition. For concurrent calibration, the patterns of the conditional bias were less varied than separate calibration across the equating methods. The UIRT and BF methods had similar patterns and yielded larger bias values along the score scale than the other methods. The amount of bias for the CE and SS methods remained relatively constant throughout the score range for all effect size conditions, except negative bias at the lower end for CE when there was large group difference. The two full MIRT methods showed a similar equating relationship and the bias values were relatively small compared to UIRT.

Conditional bias for Spanish unrounded scale scores can be found in Figure 4-16. The overall patterns of the equating relationships were similar to those for the raw score
equating relationships. For separate calibration, the SS method produced a flat line across the entire score range. Larger bias values were associated with BF, full MIRT, and UIRT methods in the middle to upper region of the score scale, especially when there was a large group difference. For concurrent calibration, the amount of bias for the SS and CE methods remained constant and relatively small along the score scale. It was obvious that most of the bias values produced by all equating methods were less varied than those of separate calibration.

**Conditional SE.** Figure 4-17 shows conditional SE for Spanish raw scores. For separate calibration, the SS method produced smaller SEs than the other equating methods across the score range regardless of group difference. The UIRT and MIRT methods yielded larger SEs at the lower and upper ranges of the score scale, whereas the CE method had a bumpy pattern across the score range. Relatively smaller SE values were produced for all equating methods and all effect sizes in the score of 25-30 range, where only a few frequencies were found. The effect of group difference was not distinct for all equating methods, except for the UIRT method, which had relatively larger values at the lower end of the score scale for the effect size of 0.3. For concurrent calibration, SE values from all equating methods except CE tended to be smaller at the upper score range. Within the lower score range (i.e., at scores of 10-25), the SE values were largest for UIRT and BF and smallest for SS, with full MIRT in the middle. The CE method had its own pattern so that the largest values were observed for score ranges of 30-45. Overall, random errors were small along the score scale for all equating methods using both separate and concurrent calibration.
Conditional SEs for Spanish unrounded scale scores are presented in Figure 4-18. The patterns for the differences in scale scores were similar to those for the raw scores. As expected, the UIRT and BF methods produced relatively large SEs across the score scale, while the SS method remained constant along the entire score range regardless of group difference for both separate and concurrent calibrations. For the full MIRT methods, SE values tended to be smaller near the score range of 35-40, regardless of group difference or calibration method. The CE method had its own bumpy pattern across the score scale. The impact of group differences was not clear for both calibration methods, as similar patterns were associated with each effect size condition.

**Conditional RMSE.** Figure 4-19 presents conditional RMSEs for Spanish raw scores. In general, smaller overall errors were associated with either the full MIRT methods or the SS method across the score scale. For separate calibration, as group difference increased, larger RMSE values were found along the entire score range for all equating methods. The effect of group difference was noticeable: for the effect size 0.3 condition, the UIRT and full MIRT methods yielded larger RMSEs at the upper and lower score ranges, and BF produced larger RMSEs at lower and middle score ranges. The pattern for the CE method was less variable than the other equating methods but produced larger values at the lower part of the score scale when group difference was large. For concurrent calibration, equating methods were less affected by group difference, except UIRT and CE. RMSE values of all equating methods had relatively flat patterns and were smaller than those of separate calibration across the entire score range, although relatively larger values were found at a score range of 10-15 for the effect size of 0.3.
Conditional RMSEs for Spanish unrounded scale scores can be found in Figure 4-20. In general, scale score relationships were very similar to raw score equating relationships. Large group difference impact RMSE values, especially for BF and full MIRT methods. Values for the CE and SS methods remained relatively constant across the score scale even with the large group difference. The UIRT method produced larger RMSEs than the other equating methods at the upper score range for small and medium group differences, and the upper and lower score ranges for large group difference.

Similar to the values for raw scores, the RMSE values for concurrent calibration were smaller than separate calibration and were only minimally affected by group difference. The RMSE values of the SS, BF, full MIRT, and CE methods were fairly flat along the score scale.

**Spanish: Overall Statistics**

Overall equating results for Spanish raw scores and unrounded scale scores are provided in Tables 4-12 to 4-16. Within each score type, unweighted statistics are presented, followed by weighted statistics. Weighted statistics were calculated using the observed relative frequency corresponding to each new form score point of 10-45. For each effect size condition, the first column covers average absolute bias (ABIAS), the second column is for average standard error (ASE), and the last column contains average root mean square error (ARMSE).

**Raw Scores.** Unweighted statistics for Spanish raw scores can be found in Table 4-12. The smallest ABIAS values were associated with the SS method using concurrent calibration across all effect size conditions. The CE method had the second smallest
ABIAS values except for the large group difference condition. UIRT produced the largest ABIAS values when the effect size was 0.3, followed by BF. The UIRT and BF methods were more heavily affected by group difference, while the SS and CE methods were relatively robust. Concurrent calibration always resulted in smaller ABIAS than separate calibration across all equating methods and effect size conditions. When comparing three columns (i.e., effect sizes of 0.05, 0.1, and 0.3) of ABIAS values, the ABIAS values using separate calibration were more greatly affected by group difference than those using concurrent calibration, because the amount of increase across the effect sizes for separate calibration was larger than that of concurrent calibration for all equating methods. Moreover, when group difference was large, the difference between separate and concurrent calibration increased: when the effect size was 0.3, the ABIAS values using separate calibration were nearly one and a half times larger than those of concurrent calibration for all equating methods. In terms of ASE, smaller group difference reduced the values of ASE in equating. FMO and SS generally produced smaller values, whereas UIRT showed relatively larger ASEs. The ASE values were not appreciably affected by group difference or calibration methods compared to the ABIAS values. In general, ARMSE values indicated that SS showed the smallest overall error, followed by FMO and FMC. The CE method outperformed the BF and UIRT methods. The ARMSE values for UIRT were relatively larger than the other equating methods when there were medium or large group differences. In addition, it was clear that as the effect size became larger, the ARMSE values became larger. Group difference made a major impact on ARSME values between separate and concurrent calibration, except for SS, because the ARSME
values using separate calibration were twice as large as those using concurrent calibration for effect size of 0.3. Similar with ABIAS, the ARMSE values using separate calibration were more heavily affected by group difference than those using concurrent calibration.

Table 4-13 presents weighted statistics for Spanish raw scores. Overall patterns among equating procedures, group differences, and calibration methods remained similar to the unweighted statistics. The magnitudes of ABIAS, ASE, and ARMSE decreased after applying weights. In addition, weights slightly alleviated the difference between separate and concurrent calibration.

**Unrounded Scale Scores.** Table 4-14 presents unweighted statistics for Spanish unrounded scale scores. The largest ABIAS value was found with BF equating using separate calibration for effect size 0.3. UIRT method with separate calibration had the second largest ABIAS values when group difference was large. CE and SS using concurrent calibration had the smallest ABIAS values. The full MIRT methods (i.e., FMC and FMO) behaved similarly, and FMC had slightly larger values than FMO across the effect size conditions. ABIAS values for group difference did not appreciably increase relative to values for raw scores for all equating methods. Also, the difference between ABIAS values using separate and concurrent calibration for all equating methods was smaller than that of raw scores, regardless of group difference. With respect to ASE, values were not substantially different across equating and calibration methods for effect sizes of 0.05 and 0.1. BF and UIRT resulted in slightly larger ASE values compared to the other equating methods for effect size of 0.3. The magnitudes of ASE were generally smaller than those of raw scores, although the difference was not considerable. Based on
ARMSEs, the overall equating errors were the smallest in SS, followed by CE. BF equating using separate calibration and UIRT using both separate and concurrent calibration produced larger ARMSEs than the other equating methods when group difference was large. Similar to the ABIAS values, the ARMSE values were not much affected by group difference and the difference in the ARMSE values between separate and concurrent calibration was not considerable.

Table 4-15 presents weighted statistics for Spanish unrounded scale scores. The general patterns of the weighted statistics were similar to those found for the unweighted statistics. Overall, applying weights to unweighted statistics led to smaller values of ABIAS, ASE, and ARMSE.

**French: Conditional Statistics**

Figures 4-21 to 4-26 provide conditional bias, SE, and RMSE for French raw scores and unrounded scale scores. Results are presented similarly to the results for Spanish in the previous section.

**Conditional bias.** Figure 4-21 presents conditional bias for French raw scores. In terms of separate calibration, the UIRT method produced a substantial amount of negative bias across the entire score scale, with the largest values found near the score of 30-35 for all effect sizes. The BF method had the second largest of positive bias along the score scale at scores ranges of 10 to 25. The pattern of the CE method was relatively flat, although it had a little bumpy pattern for effect size of 0.3. The performances of SS, FMC, and FMO were similar and preferable to those of UIRT and BF. Large group difference had an impact on bias values: UIRT consistently produced larger bias values
across the score scale than the other equating methods for the effect size of 0.3. Even SS, FMC, FMO produced larger bias values when the group difference was large. For concurrent calibration, the patterns of the bias values were relatively even and smaller than those of separate calibration across the entire score range. When group difference was small (i.e., effect size = 0.05), the patterns of bias for all equating methods were flat, especially the SS method. UIRT and BF had similar bias patterns, while FMC and FMO showed a similar equating relationship for all effect sizes.

Conditional bias for French unrounded scale scores can be found in Figure 4-22. The general patterns of the bias values were very similar to those for raw score across all equating methods. With respect to separate calibration, the BF method yielded larger bias at the lower part of the score scale and its pattern differed from the other equating methods, especially for small and medium group differences. As group difference became larger, more score points of bias values from the UIRT method produced larger bias values than the other equating methods. For concurrent calibration, more bias values were smaller compared to separate calibration. Only middle score ranges of UIRT and BF yielded large bias values. The bias patterns for FMC, FMO, SS, and CE methods were reasonably constant, and the bias values were smaller than those of separate calibration along the entire score range, regardless of group difference.

**Conditional SE.** Figure 4-23 displays conditional SEs for French raw scores. In terms of separate calibration, the CE method displayed a unique pattern and resulted in the largest SE values among all equating methods for effect sizes of 0.1 and 0.3. When group difference was small, the difference among equating methods was not apparent
except for UIRT and CE. In contrast, large group difference led to discrepancy among the equating results. Most of SE values from all equating methods were remained constant along the score scale, with the exception of UIRT at the lower score range. For concurrent calibration, UIRT, BF, and full MIRT methods yielded larger SE values at the lower to middle score points, while the SS method remained relatively flat across the score scale. The CE method was heavily affected by group difference, with the SE values becoming noticeably larger when the effect size was 0.3. The SE pattern of the CE method was unique, while UIRT, SS, BF, and full MIRT methods had similar patterns.

Figure 4-24 presents conditional SEs for French unrounded scale scores. The general pattern for the differences in scale scores was similar to the differences in raw scores. For both calibration methods, the UIRT and CE produced the largest SE values across the entire score range for small and medium group differences. The SE values of BF, FMC, and FMO were smaller than those of UIRT and CE, but larger than SS along the score scale. The impact of group difference was apparent because the SE values of all equating methods tended to increase as the group difference increased.

**Conditional RMSE.** Figure 4-25 shows conditional RMSEs for French raw scores. In terms of separate calibration, the UIRT method resulted in the largest RMSE values across the score scale among all equating methods. The BF method had a unique pattern: its values substantially decreased at a score range of 25-30 when the effect size was 0.1 and 0.3. When group difference was large, most of the RMSE values tended to be larger along the score scale. For concurrent calibration, differences among the equating methods were less noticeable compared to those of separate calibration. When the effect
size was 0.05, only the UIRT method had noticeably larger RMSEs than the other
equating methods; however, when the effect size was 0.3, the discrepancy among the
equating methods were distinguishable. The SS, full MIRT, and CE methods performed
reasonably well compared to UIRT and BF, regardless of group difference or calibration
method.

Conditional RMSEs for French unrounded scale scores can be found in Figure 4-26. The general patterns of the scale score equating relationships were similar to raw
score equating relationships, although the RMSE values were relatively small. For
separate calibration, the UIRT method was heavily affected by group difference: when the
effect size was the largest, the UIRT method produced noticeably larger RMSEs than the
other equating methods. On the contrary, the SS, FMC, and FMO methods resulted in the
smallest RMSE values over the entire score range even with the large group difference.
For concurrent calibration, RMSE values were smaller than those of separate calibration
except for a few score points. The amount of bias for SS were relatively constant
compared to the other equating methods. When the group difference was small, all RMSE
values had similar patterns to each other, while large group difference made most RMSE
values larger.

French: Overall Statistics

Tables 4-16 to 4-19 provide overall equating results for French raw scores and
unrounded scale scores. Results are presented similarly to the results for Spanish in the
previous section.
**Raw Scores.** Unweighted statistics for French raw scores can be found in Table 4-16. The ABIAS value of the UIRT method was the second largest for effect size 0.05, and the largest for effect sizes 0.1 and 0.3. The SS and CE methods outperformed UIRT and BF using both calibration methods and full MIRT methods using separate calibration, regardless of group difference. For large group difference, the full MIRT and SS methods using concurrent calibration produced relatively smaller ABIAS values than UIRT, BF, and CE. In addition, it was observed that the gap between separate and concurrent calibration increased with large group difference: when the effect size was 0.5, the ABIAS values using separate and concurrent calibration were similar for all equating methods. On the other hand, when the effect size was 0.3, the ABIAS values using separate calibration were one and a half to two times larger than those using concurrent calibration, depending on equating methods. With respect to ASE values, the SS and full MIRT methods yielded smaller ASE values, while UIRT, BF, and CE produced larger ASEs. The magnitudes of ASE did not increase drastically compared to ABIAS values when the group difference increased. In terms of ARMSEs, it was noticeable that the BF and UIRT methods constantly produced larger RMSE values than the other equating methods for all effect sizes. Even SS and CE, which yielded relatively smaller values than the other equating methods, produced large values for effect size 0.3.

In Table 4-17, weighted statistics are reported for French raw scores. The general patterns of equating relationships were similar to those found for the unweighted statistics. After applying weights, the values of ABIAS, ASE, and ARSME were smaller than those of unweighted statistics.
**Unrounded Scale Scores.** Table 4-18 presents unweighted statistics for French unrounded scale scores. The UIRT method produced the largest ABIAS values across the effect size conditions. The SS and CE methods yielded the smallest ABIAS values. The ABIAS values of the full MIRT methods (i.e., FMC and FMO) were similar to each other. When comparing separate and concurrent calibration, the amount of difference was not substantial compared to raw score equating for all equating methods and effect sizes. In terms of ASE, most of the values were similar except for the UIRT method, which produced slightly larger values than the other equating methods. The ASE values were not appreciably affected by group difference because the values were not varied across the effect size conditions. For ARMSE, the overall equating errors were the largest in UIRT, followed by BF. It was clear that the ARMSE values of (a) the full MIRT methods using separate calibration, (b) BF and UIRT using both calibration methods had larger ARMSE values than the other equating methods. The SS and CE methods outperformed the other equating methods regardless of group difference. As with ABIAS, the difference in ARMSE values between separate and concurrent calibration was not substantial for all equating methods and effect sizes.

Table 4-19 presents weighted statistics for French unrounded scale scores. The overall patterns of the weighted statistics were very similar to those for the unweighted statistics. Weights resulted in smaller values of ABIAS, ASE, and ARMSE compared to the unweighted statistics.
Summary

Chapter 4 presented Preliminary Study results and equating results from two studies for the UIRT, SS, BF, FMC, FMO, and CE methods for Spanish and French.

In the Preliminary Study, it was found that item parameters were well recovered when correlations between dimensions were freely estimated. In addition, the correlation between the dimensions did not have a systematic impact on the recovery of item parameters, so that correlation between dimensions may not impact on MIRT equating results. Before conducting analyses for Study 1 and Study 2, dimensionality assessments were conducted, which provided evidence of multidimensionality of Spanish and French. In Study 1, real data analysis was conducted, and the results were examined in terms of raw and scale score equating relationships using separate and concurrent calibration with the DTM standard employing the CE method as a criterion. In Study 2, pseudo forms were created by splitting one test form into two test forms, so that single-group equipercentile equating served as a criterion equating relationship. Results were investigated in terms of raw and unrounded scale score equating relationships using the two calibration methods. The impact of group difference was investigated by manipulating the effect size conditions (i.e., 0.05, 0.1, and 0.3). Generally, a comparison between UIRT and MIRT methods suggested better performance of the MIRT methods over UIRT. The SS method and the full MIRT methods (i.e., FMC and FMO) showed more accurate equating results compared to UIRT. Furthermore, larger group difference led to imprecise equating results overall, and the SS method was less affected by group
difference. In terms of calibration methods, concurrent calibration outperformed separate calibration for all equating methods.
Table 4-1. Preliminary study: Comparison of true item parameters with three types of estimated item parameters ($r=0.2$)

<table>
<thead>
<tr>
<th>Item</th>
<th>True</th>
<th>Freely estimated</th>
<th>Fix at 0</th>
<th>Fix at 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$d$</td>
<td>$a_1$</td>
</tr>
<tr>
<td>1</td>
<td>1.3681</td>
<td>0.9031</td>
<td>-0.2061</td>
<td>1.2999</td>
</tr>
<tr>
<td>2</td>
<td>0.7322</td>
<td>0.9387</td>
<td>1.4620</td>
<td>0.8099</td>
</tr>
<tr>
<td>3</td>
<td>0.9618</td>
<td>0.5832</td>
<td>0.5091</td>
<td>0.9161</td>
</tr>
<tr>
<td>4</td>
<td>0.7922</td>
<td>0.9987</td>
<td>-0.1328</td>
<td>0.8628</td>
</tr>
<tr>
<td>5</td>
<td>1.0490</td>
<td>1.9431</td>
<td>0.0794</td>
<td>1.1139</td>
</tr>
<tr>
<td>6</td>
<td>1.0811</td>
<td>1.2124</td>
<td>-0.4258</td>
<td>1.1042</td>
</tr>
<tr>
<td>7</td>
<td>0.7205</td>
<td>0.7473</td>
<td>-1.5079</td>
<td>0.7856</td>
</tr>
<tr>
<td>8</td>
<td>0.8432</td>
<td>0.8526</td>
<td>0.4082</td>
<td>0.8753</td>
</tr>
<tr>
<td>9</td>
<td>0.9091</td>
<td>1.4491</td>
<td>1.1923</td>
<td>0.9780</td>
</tr>
<tr>
<td>10</td>
<td>1.4369</td>
<td>1.3853</td>
<td>1.1442</td>
<td>1.3962</td>
</tr>
</tbody>
</table>

Table 4-2. Preliminary study: Comparison of true item parameters with three types of estimated item parameters ($r=0.75$)

<table>
<thead>
<tr>
<th>Item</th>
<th>True</th>
<th>Freely estimated</th>
<th>Fix at 0</th>
<th>Fix at 0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$d$</td>
<td>$a_1$</td>
</tr>
<tr>
<td>1</td>
<td>1.3681</td>
<td>0.9031</td>
<td>-0.2061</td>
<td>1.1881</td>
</tr>
<tr>
<td>2</td>
<td>0.7322</td>
<td>0.9387</td>
<td>1.4620</td>
<td>0.8969</td>
</tr>
<tr>
<td>3</td>
<td>0.9618</td>
<td>0.5832</td>
<td>0.5091</td>
<td>0.8720</td>
</tr>
<tr>
<td>4</td>
<td>0.7922</td>
<td>0.9987</td>
<td>-0.1328</td>
<td>0.9407</td>
</tr>
<tr>
<td>5</td>
<td>1.0490</td>
<td>1.9431</td>
<td>0.0794</td>
<td>1.2709</td>
</tr>
<tr>
<td>6</td>
<td>1.0811</td>
<td>1.2124</td>
<td>-0.4258</td>
<td>1.0960</td>
</tr>
<tr>
<td>7</td>
<td>0.7205</td>
<td>0.7473</td>
<td>-1.5079</td>
<td>0.8118</td>
</tr>
<tr>
<td>8</td>
<td>0.8432</td>
<td>0.8526</td>
<td>0.4082</td>
<td>0.9143</td>
</tr>
<tr>
<td>9</td>
<td>0.9091</td>
<td>1.4491</td>
<td>1.1923</td>
<td>1.0828</td>
</tr>
<tr>
<td>10</td>
<td>1.4369</td>
<td>1.3853</td>
<td>1.1442</td>
<td>1.3531</td>
</tr>
</tbody>
</table>
### Table 4-3. Preliminary study: Comparison of true item parameters with three types of estimated item parameters ($r=0.8$)

<table>
<thead>
<tr>
<th>Item</th>
<th>True</th>
<th>Freely estimated</th>
<th>Fix at 0</th>
<th>Fix at 0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$d$</td>
<td>$a_1$</td>
</tr>
<tr>
<td>1</td>
<td>1.3681</td>
<td>0.9031</td>
<td>-0.2061</td>
<td>1.1564</td>
</tr>
<tr>
<td>2</td>
<td>0.7322</td>
<td>0.9387</td>
<td>1.4620</td>
<td>0.9204</td>
</tr>
<tr>
<td>3</td>
<td>0.9618</td>
<td>0.5832</td>
<td>0.5091</td>
<td>0.8469</td>
</tr>
<tr>
<td>4</td>
<td>0.7922</td>
<td>0.9987</td>
<td>-0.1328</td>
<td>0.9721</td>
</tr>
<tr>
<td>5</td>
<td>1.0490</td>
<td>1.9431</td>
<td>0.0794</td>
<td>1.2989</td>
</tr>
<tr>
<td>6</td>
<td>1.0811</td>
<td>1.2124</td>
<td>-0.4258</td>
<td>1.1259</td>
</tr>
<tr>
<td>7</td>
<td>0.7205</td>
<td>0.7473</td>
<td>-1.5079</td>
<td>0.8058</td>
</tr>
<tr>
<td>8</td>
<td>0.8432</td>
<td>0.8526</td>
<td>0.4082</td>
<td>0.9266</td>
</tr>
<tr>
<td>9</td>
<td>0.9091</td>
<td>1.4491</td>
<td>1.1923</td>
<td>1.1371</td>
</tr>
<tr>
<td>10</td>
<td>1.4369</td>
<td>1.3853</td>
<td>1.1442</td>
<td>1.3508</td>
</tr>
</tbody>
</table>

### Table 4-4. Preliminary study: Comparison of true item parameters with three types of estimated item parameters ($r=0.95$)

<table>
<thead>
<tr>
<th>Item</th>
<th>True</th>
<th>Freely estimated</th>
<th>Fix at 0</th>
<th>Fix at 0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$d$</td>
<td>$a_1$</td>
</tr>
<tr>
<td>1</td>
<td>1.3681</td>
<td>0.9031</td>
<td>-0.2061</td>
<td>1.1490</td>
</tr>
<tr>
<td>2</td>
<td>0.7322</td>
<td>0.9387</td>
<td>1.4620</td>
<td>0.9166</td>
</tr>
<tr>
<td>3</td>
<td>0.9618</td>
<td>0.5832</td>
<td>0.5091</td>
<td>0.8641</td>
</tr>
<tr>
<td>4</td>
<td>0.7922</td>
<td>0.9987</td>
<td>-0.1328</td>
<td>0.9775</td>
</tr>
<tr>
<td>5</td>
<td>1.0490</td>
<td>1.9431</td>
<td>0.0794</td>
<td>1.3397</td>
</tr>
<tr>
<td>6</td>
<td>1.0811</td>
<td>1.2124</td>
<td>-0.4258</td>
<td>1.1305</td>
</tr>
<tr>
<td>7</td>
<td>0.7205</td>
<td>0.7473</td>
<td>-1.5079</td>
<td>0.8261</td>
</tr>
<tr>
<td>8</td>
<td>0.8432</td>
<td>0.8526</td>
<td>0.4082</td>
<td>0.9332</td>
</tr>
<tr>
<td>10</td>
<td>1.4369</td>
<td>1.3853</td>
<td>1.1442</td>
<td>1.3298</td>
</tr>
</tbody>
</table>
Table 4-5. Preliminary study: Values of ABIAS, ASE, and ARMSE for all three conditions

<table>
<thead>
<tr>
<th></th>
<th>Freely estimated</th>
<th>Fix at 0</th>
<th>Fix at true value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2</td>
<td>0.75</td>
<td>0.8</td>
</tr>
<tr>
<td>$r$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_1$</td>
<td>ABIAS</td>
<td>0.0585</td>
<td>0.1378</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>0.1310</td>
<td>0.1532</td>
</tr>
<tr>
<td></td>
<td>ARMSE</td>
<td>0.1435</td>
<td>0.2060</td>
</tr>
<tr>
<td>$a_2$</td>
<td>ABIAS</td>
<td>0.0456</td>
<td>0.1273</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>0.1332</td>
<td>0.1706</td>
</tr>
<tr>
<td></td>
<td>ARMSE</td>
<td>0.1407</td>
<td>0.2129</td>
</tr>
<tr>
<td>$d$</td>
<td>ABIAS</td>
<td>0.0491</td>
<td>0.0438</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>0.0410</td>
<td>0.0425</td>
</tr>
<tr>
<td></td>
<td>ARMSE</td>
<td>0.0640</td>
<td>0.0610</td>
</tr>
</tbody>
</table>
Table 4-6. *Study 1: DETECT results for Spanish and French*

<table>
<thead>
<tr>
<th>Subject</th>
<th>Form</th>
<th># of factors</th>
<th>DETECT</th>
<th>ASSI</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>New</td>
<td>0.1712</td>
<td>0.1669</td>
<td>0.3335</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 factors</td>
<td>0.3100</td>
<td>0.3607</td>
<td>0.6038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 factors</td>
<td>0.3144</td>
<td>0.3665</td>
<td>0.6124</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>0.1572</td>
<td>0.0534</td>
<td>0.2688</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 factors</td>
<td>0.2300</td>
<td>0.1768</td>
<td>0.3933</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 factors</td>
<td>0.3548</td>
<td>0.3988</td>
<td>0.6068</td>
</tr>
<tr>
<td>French</td>
<td>New</td>
<td>0.1368</td>
<td>0.2260</td>
<td>0.3554</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 factors</td>
<td>0.1559</td>
<td>0.2740</td>
<td>0.4051</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 factors</td>
<td>0.1908</td>
<td>0.3587</td>
<td>0.4957</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>-0.0743</td>
<td>-0.2106</td>
<td>-0.1780</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 factors</td>
<td>0.1461</td>
<td>0.2413</td>
<td>0.3500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 factors</td>
<td>0.2377</td>
<td>0.4058</td>
<td>0.5695</td>
</tr>
</tbody>
</table>

Table 4-7. *Study 1: CFA results for Spanish and French*

<table>
<thead>
<tr>
<th>Subject</th>
<th>Form</th>
<th>CFA</th>
<th>Chi-square</th>
<th>SRMR</th>
<th>RMSEA</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>New</td>
<td>1 factor</td>
<td>0.000</td>
<td>0.035</td>
<td>0.031</td>
<td>338081.728</td>
<td>338549.263</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td>0.000</td>
<td>0.034</td>
<td>0.030</td>
<td>336993.047</td>
<td>337463.922</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>1 factor</td>
<td>0.000</td>
<td>0.037</td>
<td>0.031</td>
<td>379080.887</td>
<td>379548.423</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td>0.000</td>
<td>0.036</td>
<td>0.030</td>
<td>377779.510</td>
<td>378250.385</td>
</tr>
<tr>
<td>French</td>
<td>New</td>
<td>1 factor</td>
<td>0.000</td>
<td>0.022</td>
<td>0.022</td>
<td>344409.224</td>
<td>345256.459</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td>0.000</td>
<td>0.021</td>
<td>0.020</td>
<td>343929.036</td>
<td>344782.788</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>1 factor</td>
<td>0.000</td>
<td>0.026</td>
<td>0.026</td>
<td>348359.999</td>
<td>349207.234</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td>0.000</td>
<td>0.026</td>
<td>0.025</td>
<td>347642.776</td>
<td>348496.528</td>
</tr>
</tbody>
</table>
Table 4-8. Study 1: Disattenuated correlation, observed correlation, and Cronbach’s alpha for Spanish and French

<table>
<thead>
<tr>
<th>Test</th>
<th>Form</th>
<th>Subdomains</th>
<th>Listening</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>Listening</td>
<td>0.8184</td>
<td><strong>0.8731</strong></td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td>Reading</td>
<td>0.7338</td>
<td>0.8631</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>Listening</td>
<td>0.8132</td>
<td><strong>0.8579</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reading</td>
<td>0.7167</td>
<td>0.8582</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>Listening</td>
<td>0.8390</td>
<td><strong>0.9308</strong></td>
</tr>
<tr>
<td>French</td>
<td></td>
<td>Reading</td>
<td>0.7915</td>
<td>0.8617</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>Listening</td>
<td>0.8343</td>
<td><strong>0.9212</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reading</td>
<td>0.7824</td>
<td>0.8645</td>
</tr>
</tbody>
</table>

Note: Boldface numbers are disattenuated correlations; italic numbers are observed correlations; numbers in the diagonal are Cronbach’s alpha.

Table 4-9. Study 2: Poly-DETECT results for Spanish and French

<table>
<thead>
<tr>
<th>Subject</th>
<th>Form</th>
<th># of factors</th>
<th>DETECT</th>
<th>ASSI</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>2 factors</td>
<td>0.1394</td>
<td>-0.0480</td>
<td>0.2226</td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td>3 factors</td>
<td>0.4278</td>
<td>0.4625</td>
<td>0.6831</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 factors</td>
<td>0.4836</td>
<td>0.5465</td>
<td>0.7723</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>2 factors</td>
<td>0.0595</td>
<td>-0.3363</td>
<td>0.0762</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 factors</td>
<td>0.2549</td>
<td>-0.1291</td>
<td>0.3268</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 factors</td>
<td>0.5142</td>
<td>0.3153</td>
<td>0.6593</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>2 factors</td>
<td>-0.1644</td>
<td>-0.5676</td>
<td>-0.2617</td>
</tr>
<tr>
<td>French</td>
<td></td>
<td>3 factors</td>
<td>0.2716</td>
<td>0.1171</td>
<td>0.4325</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 factors</td>
<td>0.3441</td>
<td>0.2913</td>
<td>0.5479</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>2 factors</td>
<td>-0.2493</td>
<td>-0.4264</td>
<td>-0.4230</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 factors</td>
<td>0.0670</td>
<td>-0.2583</td>
<td>0.1137</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 factors</td>
<td>0.3116</td>
<td>0.2162</td>
<td>0.5288</td>
</tr>
</tbody>
</table>
Table 4-10. Study 2: CFA results for Spanish and French

<table>
<thead>
<tr>
<th>Subject</th>
<th>Form</th>
<th>CFA</th>
<th>Chi-square</th>
<th>SRMR</th>
<th>RMSEA</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>New</td>
<td>1 factor</td>
<td>0.000</td>
<td>0.028</td>
<td>0.030</td>
<td>808975.115</td>
<td>809559.973</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td>0.000</td>
<td>0.028</td>
<td>0.030</td>
<td>808703.936</td>
<td>809296.697</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>1 factor</td>
<td>0.000</td>
<td>0.036</td>
<td>0.037</td>
<td>782573.837</td>
<td>783158.696</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td>0.000</td>
<td>0.035</td>
<td>0.037</td>
<td>782088.297</td>
<td>782681.059</td>
</tr>
<tr>
<td>French</td>
<td>New</td>
<td>1 factor</td>
<td>0.000</td>
<td>0.022</td>
<td>0.026</td>
<td>763010.987</td>
<td>763583.845</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td>0.000</td>
<td>0.021</td>
<td>0.024</td>
<td>762123.504</td>
<td>762704.103</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>1 factor</td>
<td>0.000</td>
<td>0.021</td>
<td>0.026</td>
<td>713652.463</td>
<td>714225.321</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 factors</td>
<td>0.000</td>
<td>0.020</td>
<td>0.025</td>
<td>713268.102</td>
<td>713848.701</td>
</tr>
</tbody>
</table>

Table 4-11. Study 2: Disattenuated correlation, observed correlation, and Cronbach’s alpha for Spanish and French

<table>
<thead>
<tr>
<th>Test</th>
<th>Form</th>
<th>Sections</th>
<th>MC</th>
<th>FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>New</td>
<td>MC</td>
<td>0.8292</td>
<td><strong>0.8652</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FR</td>
<td>0.5704</td>
<td>0.5241</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>MC</td>
<td>0.8362</td>
<td><strong>0.8211</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FR</td>
<td>0.5669</td>
<td>0.5700</td>
</tr>
<tr>
<td>French</td>
<td>New</td>
<td>MC</td>
<td>0.8445</td>
<td><strong>0.8856</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FR</td>
<td>0.7004</td>
<td>0.7407</td>
</tr>
<tr>
<td></td>
<td>Old</td>
<td>MC</td>
<td>0.8642</td>
<td><strong>0.9310</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FR</td>
<td>0.7151</td>
<td>0.6827</td>
</tr>
</tbody>
</table>

Note: Boldface numbers are disattenuated correlations; italic numbers are observed correlations; numbers in the diagonal are Cronbach’s alpha.
Table 4-12. Study 2: Unweighted ABIAS, ASE, and ARSME for Spanish raw score

<table>
<thead>
<tr>
<th></th>
<th>Effect size 0.05</th>
<th>Effect size 0.1</th>
<th>Effect size 0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
</tr>
<tr>
<td>FMC</td>
<td>SEP</td>
<td>0.3027</td>
<td>0.1177</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.2203</td>
<td>0.1039</td>
</tr>
<tr>
<td>FMO</td>
<td>SEP</td>
<td>0.2917</td>
<td>0.1266</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.1981</td>
<td>0.0844</td>
</tr>
<tr>
<td>BF</td>
<td>SEP</td>
<td>0.2531</td>
<td>0.1297</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.2404</td>
<td>0.1278</td>
</tr>
<tr>
<td>SS</td>
<td>SEP</td>
<td>0.1331</td>
<td>0.0744</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.1223</td>
<td>0.0706</td>
</tr>
<tr>
<td>UIRT</td>
<td>SEP</td>
<td>0.4690</td>
<td>0.1392</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.4120</td>
<td>0.1273</td>
</tr>
<tr>
<td>CE</td>
<td></td>
<td>0.1780</td>
<td>0.1280</td>
</tr>
</tbody>
</table>

Note: ABIAS= Average absolute bias; ASE= Average standard error; ARMSE= Average root mean square error.
Table 4-13. Study 2: Weighted ABIAS, ASE, and ARSME for Spanish raw score

<table>
<thead>
<tr>
<th></th>
<th>ABIAS</th>
<th>ASE</th>
<th>ARMSE</th>
<th>ABIAS</th>
<th>ASE</th>
<th>ARMSE</th>
<th>ABIAS</th>
<th>ASE</th>
<th>ARMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMC</td>
<td>0.2411</td>
<td>0.0945</td>
<td>0.2590</td>
<td>0.3729</td>
<td>0.1180</td>
<td>0.3912</td>
<td>0.4804</td>
<td>0.1247</td>
<td>0.4963</td>
</tr>
<tr>
<td></td>
<td>0.1925</td>
<td>0.0884</td>
<td>0.2118</td>
<td>0.2785</td>
<td>0.1069</td>
<td>0.2983</td>
<td>0.3247</td>
<td>0.1211</td>
<td>0.3465</td>
</tr>
<tr>
<td>FMO</td>
<td>0.2147</td>
<td>0.0927</td>
<td>0.2338</td>
<td>0.3526</td>
<td>0.1147</td>
<td>0.3708</td>
<td>0.4838</td>
<td>0.1345</td>
<td>0.5022</td>
</tr>
<tr>
<td></td>
<td>0.1847</td>
<td>0.0790</td>
<td>0.2008</td>
<td>0.2469</td>
<td>0.0847</td>
<td>0.2610</td>
<td>0.2945</td>
<td>0.1014</td>
<td>0.3114</td>
</tr>
<tr>
<td>BF</td>
<td>0.2244</td>
<td>0.1147</td>
<td>0.2519</td>
<td>0.3017</td>
<td>0.1247</td>
<td>0.3264</td>
<td>0.4924</td>
<td>0.1645</td>
<td>0.5191</td>
</tr>
<tr>
<td></td>
<td>0.2180</td>
<td>0.1013</td>
<td>0.2404</td>
<td>0.2847</td>
<td>0.1147</td>
<td>0.3069</td>
<td>0.3758</td>
<td>0.1645</td>
<td>0.4103</td>
</tr>
<tr>
<td>SS</td>
<td>0.1246</td>
<td>0.0698</td>
<td>0.1428</td>
<td>0.1436</td>
<td>0.0685</td>
<td>0.1591</td>
<td>0.2843</td>
<td>0.1158</td>
<td>0.3070</td>
</tr>
<tr>
<td></td>
<td>0.1185</td>
<td>0.0683</td>
<td>0.1367</td>
<td>0.1135</td>
<td>0.0590</td>
<td>0.1279</td>
<td>0.1847</td>
<td>0.0945</td>
<td>0.2074</td>
</tr>
<tr>
<td>UIRT</td>
<td>0.4136</td>
<td>0.1214</td>
<td>0.4310</td>
<td>0.4973</td>
<td>0.1547</td>
<td>0.5208</td>
<td>0.6972</td>
<td>0.2017</td>
<td>0.7257</td>
</tr>
<tr>
<td></td>
<td>0.3685</td>
<td>0.1199</td>
<td>0.3875</td>
<td>0.3986</td>
<td>0.1488</td>
<td>0.4254</td>
<td>0.4653</td>
<td>0.2147</td>
<td>0.5124</td>
</tr>
<tr>
<td>CE</td>
<td>0.1546</td>
<td>0.1158</td>
<td>0.1932</td>
<td>0.1649</td>
<td>0.1433</td>
<td>0.2184</td>
<td>0.3017</td>
<td>0.1644</td>
<td>0.3435</td>
</tr>
</tbody>
</table>

Note: ABIAS= Average absolute bias; ASE= Average standard error; ARMSE= Average root mean square error.
Table 4.14. Study 2: Unweighted ABIAS, ASE, and ARSME for Spanish unrounded scale score

<table>
<thead>
<tr>
<th></th>
<th>Effect size 0.05</th>
<th>Effect size 0.1</th>
<th>Effect size 0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
</tr>
<tr>
<td>FMC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.2279</td>
<td>0.1101</td>
<td>0.2531</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2152</td>
<td>0.1034</td>
<td>0.2388</td>
</tr>
<tr>
<td>FMO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.2199</td>
<td>0.1157</td>
<td>0.2484</td>
</tr>
<tr>
<td>CONC</td>
<td>0.1926</td>
<td>0.0792</td>
<td>0.2083</td>
</tr>
<tr>
<td>BF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.2437</td>
<td>0.1143</td>
<td>0.2692</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2350</td>
<td>0.1011</td>
<td>0.2558</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.1109</td>
<td>0.0544</td>
<td>0.1235</td>
</tr>
<tr>
<td>CONC</td>
<td>0.1022</td>
<td>0.0548</td>
<td>0.1160</td>
</tr>
<tr>
<td>UIRT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.3907</td>
<td>0.1258</td>
<td>0.4105</td>
</tr>
<tr>
<td>CONC</td>
<td>0.3271</td>
<td>0.1086</td>
<td>0.3447</td>
</tr>
<tr>
<td>CE</td>
<td>0.1758</td>
<td>0.0914</td>
<td>0.1982</td>
</tr>
</tbody>
</table>

Note: ABIAS= Average absolute bias; ASE= Average standard error; ARMSE= Average root mean square error.
Table 4-15. *Study 2: Weighted ABIAS, ASE, and ARSME for Spanish unrounded scale score*

<table>
<thead>
<tr>
<th></th>
<th>Effect size 0.05</th>
<th></th>
<th></th>
<th>Effect size 0.1</th>
<th></th>
<th></th>
<th>Effect size 0.3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
<td>ABIAS</td>
<td>ASE</td>
</tr>
<tr>
<td>FMC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.2014</td>
<td>0.0990</td>
<td>0.2244</td>
<td>0.2549</td>
<td>0.1047</td>
<td>0.2756</td>
<td>0.3650</td>
<td>0.1247</td>
</tr>
<tr>
<td>CONC</td>
<td>0.1985</td>
<td>0.0895</td>
<td>0.2177</td>
<td>0.2135</td>
<td>0.0982</td>
<td>0.2350</td>
<td>0.3125</td>
<td>0.1135</td>
</tr>
<tr>
<td>FMO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.1985</td>
<td>0.1057</td>
<td>0.2248</td>
<td>0.2347</td>
<td>0.1135</td>
<td>0.2607</td>
<td>0.3547</td>
<td>0.1247</td>
</tr>
<tr>
<td>CONC</td>
<td>0.1747</td>
<td>0.0644</td>
<td>0.1861</td>
<td>0.1985</td>
<td>0.0985</td>
<td>0.2215</td>
<td>0.2847</td>
<td>0.1047</td>
</tr>
<tr>
<td>BF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.2135</td>
<td>0.1041</td>
<td>0.2376</td>
<td>0.2798</td>
<td>0.1358</td>
<td>0.3111</td>
<td>0.5132</td>
<td>0.1614</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2014</td>
<td>0.0947</td>
<td>0.2225</td>
<td>0.2543</td>
<td>0.1135</td>
<td>0.2785</td>
<td>0.3135</td>
<td>0.1543</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.1047</td>
<td>0.0514</td>
<td>0.1166</td>
<td>0.1147</td>
<td>0.0650</td>
<td>0.1318</td>
<td>0.2547</td>
<td>0.1047</td>
</tr>
<tr>
<td>CONC</td>
<td>0.0946</td>
<td>0.0599</td>
<td>0.1119</td>
<td>0.1044</td>
<td>0.0572</td>
<td>0.1190</td>
<td>0.1644</td>
<td>0.0846</td>
</tr>
<tr>
<td>UIRT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.3155</td>
<td>0.1135</td>
<td>0.3353</td>
<td>0.4325</td>
<td>0.1677</td>
<td>0.4638</td>
<td>0.4860</td>
<td>0.1984</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2846</td>
<td>0.0984</td>
<td>0.3011</td>
<td>0.3947</td>
<td>0.1325</td>
<td>0.4163</td>
<td>0.4135</td>
<td>0.1847</td>
</tr>
<tr>
<td>CE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1544</td>
<td>0.0850</td>
<td>0.1762</td>
<td>0.1733</td>
<td>0.1135</td>
<td>0.2071</td>
<td>0.3347</td>
<td>0.1474</td>
</tr>
</tbody>
</table>

*Note: ABIAS= Average absolute bias; ASE= Average standard error; ARMSE= Average root mean square error.*
Table 4-16. *Study 2: Unweighted ABIAS, ASE, and ARSME for French raw score*

<table>
<thead>
<tr>
<th></th>
<th>Effect size 0.05</th>
<th></th>
<th>Effect size 0.1</th>
<th></th>
<th>Effect size 0.3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
</tr>
<tr>
<td>FMC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.4178</td>
<td>0.0451</td>
<td>0.4202</td>
<td>0.4966</td>
<td>0.1301</td>
<td>0.5134</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2813</td>
<td>0.0433</td>
<td>0.2846</td>
<td>0.4054</td>
<td>0.1342</td>
<td>0.4270</td>
</tr>
<tr>
<td>FMO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.3962</td>
<td>0.0404</td>
<td>0.3982</td>
<td>0.4765</td>
<td>0.1411</td>
<td>0.4970</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2732</td>
<td>0.0392</td>
<td>0.2760</td>
<td>0.3741</td>
<td>0.1443</td>
<td>0.4010</td>
</tr>
<tr>
<td>BF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.5246</td>
<td>0.0829</td>
<td>0.5312</td>
<td>0.8092</td>
<td>0.1720</td>
<td>0.8273</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2971</td>
<td>0.0712</td>
<td>0.3055</td>
<td>0.5534</td>
<td>0.1373</td>
<td>0.5702</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.3664</td>
<td>0.0344</td>
<td>0.3680</td>
<td>0.3800</td>
<td>0.0793</td>
<td>0.3882</td>
</tr>
<tr>
<td>CONC</td>
<td>0.1879</td>
<td>0.0309</td>
<td>0.1904</td>
<td>0.2015</td>
<td>0.0724</td>
<td>0.2141</td>
</tr>
<tr>
<td>UIRT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.4912</td>
<td>0.1231</td>
<td>0.5064</td>
<td>1.3175</td>
<td>0.2143</td>
<td>1.3348</td>
</tr>
<tr>
<td>CONC</td>
<td>0.4875</td>
<td>0.1025</td>
<td>0.4982</td>
<td>0.7621</td>
<td>0.1505</td>
<td>0.7768</td>
</tr>
<tr>
<td>CE</td>
<td>0.2189</td>
<td>0.0752</td>
<td>0.2315</td>
<td>0.4527</td>
<td>0.2062</td>
<td>0.4974</td>
</tr>
</tbody>
</table>

*Note:* ABIAS = Average absolute bias; ASE = Average standard error; ARMSE = Average root mean square error.
<table>
<thead>
<tr>
<th></th>
<th>Effect size 0.05</th>
<th></th>
<th>Effect size 0.1</th>
<th></th>
<th>Effect size 0.3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
</tr>
<tr>
<td>FMC</td>
<td>SEP</td>
<td>0.3822</td>
<td>0.0411</td>
<td>0.3844</td>
<td>0.4528</td>
<td>0.1211</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.2152</td>
<td>0.0338</td>
<td>0.2178</td>
<td>0.3858</td>
<td>0.1285</td>
</tr>
<tr>
<td>FMO</td>
<td>SEP</td>
<td>0.3527</td>
<td>0.0392</td>
<td>0.3549</td>
<td>0.4352</td>
<td>0.1309</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.2546</td>
<td>0.0325</td>
<td>0.2567</td>
<td>0.3525</td>
<td>0.1335</td>
</tr>
<tr>
<td>BF</td>
<td>SEP</td>
<td>0.5138</td>
<td>0.0751</td>
<td>0.5193</td>
<td>0.7333</td>
<td>0.1525</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.2744</td>
<td>0.0692</td>
<td>0.2830</td>
<td>0.4532</td>
<td>0.1279</td>
</tr>
<tr>
<td>SS</td>
<td>SEP</td>
<td>0.2558</td>
<td>0.0346</td>
<td>0.2581</td>
<td>0.2979</td>
<td>0.0635</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.1537</td>
<td>0.0298</td>
<td>0.1566</td>
<td>0.1852</td>
<td>0.0611</td>
</tr>
<tr>
<td>UIRT</td>
<td>SEP</td>
<td>0.4535</td>
<td>0.1136</td>
<td>0.4675</td>
<td>0.9946</td>
<td>0.1925</td>
</tr>
<tr>
<td></td>
<td>CONC</td>
<td>0.4185</td>
<td>0.0994</td>
<td>0.4301</td>
<td>0.7152</td>
<td>0.1345</td>
</tr>
<tr>
<td>CE</td>
<td></td>
<td>0.1947</td>
<td>0.0613</td>
<td>0.2041</td>
<td>0.4015</td>
<td>0.1839</td>
</tr>
</tbody>
</table>

*Note: ABIAS = Average absolute bias; ASE = Average standard error; ARMSE = Average root mean square error.*
Table 4-18. Study 2: Unweighted ABIAS, ASE, and ARSME for French unrounded scale score

<table>
<thead>
<tr>
<th></th>
<th>Effect size 0.05</th>
<th></th>
<th></th>
<th>Effect size 0.1</th>
<th></th>
<th></th>
<th>Effect size 0.3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
</tr>
<tr>
<td>FMC</td>
<td>0.2262</td>
<td>0.0412</td>
<td>0.2299</td>
<td>0.3585</td>
<td>0.1286</td>
<td>0.3809</td>
<td>0.5729</td>
<td>0.1712</td>
<td>0.5979</td>
</tr>
<tr>
<td></td>
<td>0.1829</td>
<td>0.0397</td>
<td>0.1872</td>
<td>0.3098</td>
<td>0.1311</td>
<td>0.3364</td>
<td>0.4425</td>
<td>0.1419</td>
<td>0.4647</td>
</tr>
<tr>
<td>FMO</td>
<td>0.2016</td>
<td>0.0398</td>
<td>0.2055</td>
<td>0.3316</td>
<td>0.1292</td>
<td>0.3559</td>
<td>0.5928</td>
<td>0.1788</td>
<td>0.6192</td>
</tr>
<tr>
<td></td>
<td>0.1725</td>
<td>0.0363</td>
<td>0.1763</td>
<td>0.3511</td>
<td>0.1308</td>
<td>0.3747</td>
<td>0.3985</td>
<td>0.1501</td>
<td>0.4258</td>
</tr>
<tr>
<td>BF</td>
<td>0.2608</td>
<td>0.0744</td>
<td>0.2712</td>
<td>0.4953</td>
<td>0.1682</td>
<td>0.5231</td>
<td>0.6585</td>
<td>0.1983</td>
<td>0.6877</td>
</tr>
<tr>
<td></td>
<td>0.2435</td>
<td>0.0551</td>
<td>0.2497</td>
<td>0.4222</td>
<td>0.1332</td>
<td>0.4427</td>
<td>0.5497</td>
<td>0.1566</td>
<td>0.5716</td>
</tr>
<tr>
<td>SS</td>
<td>0.1185</td>
<td>0.0366</td>
<td>0.1240</td>
<td>0.2925</td>
<td>0.0718</td>
<td>0.3012</td>
<td>0.4133</td>
<td>0.1363</td>
<td>0.4352</td>
</tr>
<tr>
<td></td>
<td>0.1057</td>
<td>0.0269</td>
<td>0.1091</td>
<td>0.1252</td>
<td>0.0725</td>
<td>0.1447</td>
<td>0.2254</td>
<td>0.0856</td>
<td>0.2411</td>
</tr>
<tr>
<td>UIRT</td>
<td>0.4338</td>
<td>0.1206</td>
<td>0.4503</td>
<td>0.6525</td>
<td>0.1809</td>
<td>0.6771</td>
<td>0.9809</td>
<td>0.2223</td>
<td>1.0058</td>
</tr>
<tr>
<td></td>
<td>0.2856</td>
<td>0.0812</td>
<td>0.2969</td>
<td>0.5167</td>
<td>0.1492</td>
<td>0.5378</td>
<td>0.7295</td>
<td>0.1864</td>
<td>0.7529</td>
</tr>
<tr>
<td>CE</td>
<td>0.1050</td>
<td>0.0585</td>
<td>0.1202</td>
<td>0.2468</td>
<td>0.1497</td>
<td>0.2887</td>
<td>0.3552</td>
<td>0.2487</td>
<td>0.4336</td>
</tr>
</tbody>
</table>

Note: ABIAS= Average absolute bias; ASE= Average standard error; ARSME= Average root mean square error.
Table 4-19. Study 2: Weighted ABIAS, ASE, and ARSME for French unrounded scale score

<table>
<thead>
<tr>
<th></th>
<th>Effect size 0.05</th>
<th></th>
<th>Effect size 0.1</th>
<th></th>
<th>Effect size 0.3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
<td>ABIAS</td>
<td>ASE</td>
<td>ARMSE</td>
</tr>
<tr>
<td>FMC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.1958</td>
<td>0.0319</td>
<td>0.1984</td>
<td>0.3152</td>
<td>0.1165</td>
<td>0.3360</td>
</tr>
<tr>
<td>CONC</td>
<td>0.1547</td>
<td>0.0285</td>
<td>0.1573</td>
<td>0.2945</td>
<td>0.1196</td>
<td>0.3179</td>
</tr>
<tr>
<td>FMO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.1847</td>
<td>0.0305</td>
<td>0.1872</td>
<td>0.3032</td>
<td>0.1178</td>
<td>0.3253</td>
</tr>
<tr>
<td>CONC</td>
<td>0.1535</td>
<td>0.0272</td>
<td>0.1559</td>
<td>0.3433</td>
<td>0.1109</td>
<td>0.3608</td>
</tr>
<tr>
<td>BF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.2433</td>
<td>0.0625</td>
<td>0.2512</td>
<td>0.4256</td>
<td>0.1446</td>
<td>0.4495</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2131</td>
<td>0.0413</td>
<td>0.2171</td>
<td>0.3848</td>
<td>0.1165</td>
<td>0.4020</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.0928</td>
<td>0.0293</td>
<td>0.0973</td>
<td>0.2523</td>
<td>0.0595</td>
<td>0.2592</td>
</tr>
<tr>
<td>CONC</td>
<td>0.0847</td>
<td>0.0213</td>
<td>0.0873</td>
<td>0.1089</td>
<td>0.0435</td>
<td>0.1173</td>
</tr>
<tr>
<td>UIRT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.3928</td>
<td>0.1133</td>
<td>0.4088</td>
<td>0.5825</td>
<td>0.1653</td>
<td>0.6055</td>
</tr>
<tr>
<td>CONC</td>
<td>0.2462</td>
<td>0.0735</td>
<td>0.2569</td>
<td>0.4537</td>
<td>0.1313</td>
<td>0.4723</td>
</tr>
<tr>
<td>CE</td>
<td>0.0910</td>
<td>0.0517</td>
<td>0.1047</td>
<td>0.2074</td>
<td>0.1249</td>
<td>0.2421</td>
</tr>
</tbody>
</table>

Note: ABIAS= Average absolute bias; ASE= Average standard error; ARSME= Average root mean square error.
Figure 4-1. Item vectors for the Spanish language 2011 form and common items
Figure 4-2. Item vectors for the Spanish language 2012 form and common items
Figure 4-3. Item vectors for the French language 2012 form and common items
Figure 4-4. Item vectors for the French language 2013 form and common items
Figure 4-5. Study 1: Scree plots for Spanish and French
Figure 4-6. Study 1: Spanish observed and fitted distributions
Figure 4-7. Study 1: French observed and fitted distributions
Figure 4-8. Study 1: Spanish raw-to-raw score equivalents for all equating methods using separate and concurrent calibration
Figure 4-9. Study 1: French raw-to-raw score equivalents for all equating methods using separate and concurrent calibration
Figure 4-10. Study 1: Differences between (M)IRT and CE equated raw and unrounded scale scores for Spanish.
Figure 4-11. Study 1: Differences between (M)IRT and CE equated raw and unrounded scale scores for French
Figure 4-12. Study 2: Scree plots for Spanish and French
Figure 4-13. Study 2: Spanish observed and fitted distributions
Figure 4.14. Study 2: French observed and fitted distributions
Figure 4-15. Study 2: Conditional bias for Spanish raw scores
Figure 4-16. Study 2: Conditional bias for Spanish unrounded scale scores
Figure 4-17. Study 2: Conditional SE for Spanish raw scores
Figure 4-18. Study 2: Conditional SE for Spanish unrounded scale scores
Figure 4-19. Study 2: Conditional RMSE for Spanish raw scores
Figure 4-20. Study 2: Conditional RMSE for Spanish unrounded scale scores
Figure 4-21. Study 2: Conditional bias for French raw scores
Figure 4-22. Study 2: Conditional bias for French unrounded scale scores
Figure 4.23. Study 2: Conditional SE for French raw scores
Figure 4-24. Study 2: Conditional SE for French unrounded scale scores
Figure 4-25. Study 2: Conditional RMSE for French raw scores
Figure 4-26. Study 2: Conditional RMSE for French unrounded scale scores
CHAPTER 5
DISCUSSION

The main objectives of this dissertation were to consolidate and expand MIRT observed score equating research under the CINEG design. First, this dissertation presented specific processes of MIRT equating under the CINEG design. The subsequent chapters next compared the differences in (M)IRT equating results and scale linking methods for MIRT observed score equating, and finally investigated the impact of group differences on MIRT observed score equating. This dissertation used two studies with different data types and comparison criteria to answer research questions. Study 1 used real data analysis on two subjects (i.e., Spanish and French) to compare equating methods in terms of raw scores and unrounded scale scores. The DTM criterion (i.e., ± 0.5) was used to determine how the MIRT equivalents differed from the CE equivalents across the score scale. In Study 2, pseudo forms were created using the Spanish and French data from Study 1, and the equating methods were compared in terms of raw scores and unrounded scale scores. The equating criterion was established based on single-group equipercentile equating. Equating performance was evaluated using conditional bias, SE, and RMSE statistics, followed by overall statistics (i.e., ABIAS, ASE, and ARMSE).

This chapter summarizes results from each research question. Limitations and suggestions for future research are discussed, then conclusions and implications are presented. Note that for simplicity, equating results for scale scores that were already discussed in Chapter 4 are not presented again here because they followed the same patterns as equating results for raw scores.
Summary of Results

Research Objective 1

To describe specific processes of MIRT equating under the CINEG design

The main purpose of this study was to declare assumptions for MIRT equating under the CINEG design. The goal of MIRT equating under the CINEG design is to adjust test scores on two forms from multidimensionally-nonequivalent groups so that scores on the two forms can be used interchangeably. In Chapter 3, assumptions were made in terms of (a) test forms, (b) dimensionality and examinee groups, (c) correlation between dimensions and calibration options, and (d) common items when MIRT observed score equating is conducted under the CINEG design. Each assumption is summarized below:

(a) Test forms: The test forms to be equated should be constructed based on the same test specifications. The two forms should be checked with respect to the table of content and statistical specifications.

(b) Dimensionality and examinee groups: The number of dimensions and constructs should be the same for the old and new examinee groups that take the old and new forms, respectively, although the origin and unit of measurement (mean and SD) can differ.

(c) Correlation between dimensions and calibration options: Correlation between dimensions can be different for the old and new examinee groups that take the old and new forms, respectively. When calibrating FM
models, it is suggested that correlations between latent traits be freely estimated.

(d) Common items: Common items should have the same dimensional structure and measure the same constructs as the test forms.

Preliminary Study was conducted to provide evidence toward assumption (c): correlation between dimensions and calibration options. From the preliminary study, it was found that correlation between dimensions did not impact the recovery of MIRT item parameters; therefore, the correlations between dimensions are allowed to be different when conducting MIRT equating under the CINEG design. It was also discovered that when calibrating FM models, correlation between dimensions should be freely estimated to reflect the underlying structure of examinees’ latent abilities.

**Research Objective 2**

*To compare the differences in equating results for UIRT, SS, BF, FM, and traditional equipercentile methods*

Another purpose of this dissertation was to compare the differences in equating results for UIRT, MIRT (i.e., SS, BF, FMC, and FMO), and traditional equipercentile equating (i.e., CE). Since the proposed equating method was the MIRT procedure in this dissertation, comparison of the performance of MIRT equating with UIRT and CE was needed to demonstrate the superiority of the MIRT procedure over UIRT or CE. This subsection answers the research question based on only Study 1 results. Study 2 results will be summarized with research objective #4 because research objectives #2 and #4 are closely related.
For Spanish, the UIRT and BF methods produced more different equivalents than SS, FMC, FMO, and CE methods. When differences between (M)IRT and CE equated raw scores were compared, the FM methods (i.e., FMC and FMO) were closest to the CE method. The UIRT and BF methods deviated more from the DTM boundaries than the other equating methods. For French, the CE method was the closest to the new form raw score, followed by SS. When differences between (M)IRT and CE equated raw scores were compared, the full MIRT and SS methods remained similar to each other and within DTM along the score scale.

The findings from Study 1 supported the results of previous studies, although study designs and conditions were different from this dissertation: UIRT and BF performed similarly for conditions with higher correlation values (Lee & Lee, 2016; Peterson & Lee, 2014); Full MIRT equating methods provided more accurate equating results than UIRT methods (Lee, Lee, & Brennan, 2014); Full MIRT equating procedures can be an alternative to UIRT when the data are not strictly unidimensional (Brossman & Lee, 2013); SS outperformed UIRT under the moderate degree of multidimensionality (Lee & Brossman, 2012; Kim, 2018).

Research Objective 3

To compare scale linking methods (i.e., separate calibration and concurrent calibration) for MIRT observed score equating

Spanish. For Spanish equating from Study 1, it was found that results for (M)IRT equating methods for both separate and concurrent calibration showed a similar curvilinear equating relationship, although differences among equating methods reduced
for concurrent calibration. In Study 2, conditional bias values indicated that the UIRT method yielded substantial negative bias using separate calibration. On the other hand, the patterns of the conditional bias produced by concurrent calibration were less varied than those of separate calibration across the equating methods. Conditional SE values of Spanish scores indicated that the SS method produced smaller SE than the other equating methods along the score scale for separate calibration. On the contrary, SE values from all equating methods tended to be smaller at the upper score range for concurrent calibration. SE values from all equating methods had similar values along the score scale using concurrent calibration. In terms of RMSE values, all equating methods yielded larger SEs at the lower and upper part of score range for separate calibration, while most RMSE values produced by concurrent calibration remained constant across the score range.

When investigating overall statistics, concurrent calibration always resulted in smaller ABIAS, ASE, and ARMSE values than separate calibration across all equating methods. This finding was consistent with Kim (2017) that concurrent calibration provided equating results with smallest equating error, while the separate calibration method provided equating results with the largest equating error, although Kim (2017)’s study was conducted to investigate the performance of BF linking only. The ABIAS and ARMSE values using separate calibration were much affected by group difference than those using concurrent calibration, while the ASE values were similar for both calibration methods.

**French.** Raw-to-raw score equivalents from Study 1 showed that the plot of separate calibration resembled that of concurrent calibration. SS was closest to the new
form score, while UIRT and BF were far from the new form score for both separate and concurrent calibration. When comparing differences between the (M)IRT equating methods, FMC and SS remained inside of DTM, while UIRT and BF fell outside of DTM for both separate and concurrent calibration. In Study 2, the largest bias values were found near the score range of 30-35 for separate calibration. In contrast, the patterns of the bias values were relatively flat and smaller than those of separate calibration across the score scale. Most of SE values from all equating methods for separate calibration had larger values than those for concurrent calibration. For RMSE values of concurrent calibration, differences among the equating methods were less noticeable compared to separate calibration.

When comparing overall statistics, concurrent calibration outperformed separate calibration because the former produced smaller ABIAS, ASE, and ARMSE values across all equating methods. The difference between separate and concurrent calibration was affected by group difference for the ABIAS and ARMSE values, while there was not much differences between separate and concurrent calibration found for the ASE values.

**Research Objective 4**

*To compare the impact of group ability differences on MIRT observed score equating*

The impact of group difference is a crucial factor for evaluating the performance of MIRT observed score equating under the CINEG design. In general, when group difference was larger, systematic and total errors of equating increased, while random error of equating was not much affected. This finding was similar to those noted in previous studies (Lee, Lee, & Brennan, 2014; Lim, 2016).
For Spanish conditional statistics, the impact of group difference was obvious: when group difference was large; UIRT and BF were much affected by group difference than the other equating methods because the two graph patterns changed radically according to effect size. The bias values of FMC, FMO, SS, and CE were relatively flat across the score range when group difference was either small or medium, but large group difference made the equating methods produce larger values at the upper or lower score ranges. The SE values were not affected by group difference. On the contrary, the effect of group difference was noticeable for the RMSE values. The UIRT, full MIRT, and BF methods produced larger RMSEs across the score scale when the effect size was 0.3. The difference between separate and concurrent calibration was also affected by group difference, because RMSE values of all equating methods using concurrent calibration had relatively flat patterns than those of separate calibration across the score range when group difference was large.

For Spanish overall statistics, it was clear that systematic (i.e., ABIAS) and total errors (i.e., ARMSE) became larger as the group difference was larger. The ABIAS and AMRSE values of the UIRT and BF methods were affected by group difference, whereas SS and CE were relatively similar across the effect size conditions. In addition, the ABIAS and ARMSE values using separate calibration were much affected by group difference than those using concurrent calibration. Also, the difference between separate and concurrent calibration increase when group difference was large. All ASE values remained similar across the effect size conditions compared to the other two statistics.
For French conditional statistics, group difference was significant in evaluating the equating results because most bias values of all equating methods had larger values with a large group difference. On the contrary, when group difference was small, most bias values across the score range had relatively flat pattern. In terms of SE, large group difference resulted in discrepancy among the equating results, while the difference among the equating methods was not clear when the group difference was small. For RMSE values, when group difference was large, most of the RMSE values were larger across the entire score range.

For French overall statistics, it was obvious that larger group difference caused the ABIAS and ARMSE values to increase. Most of the ABIAS and ARMSE values were relatively large when there was large group difference, except for SS and FMO using concurrent calibration. The amount of increase in SE values according to group difference was smaller than that of ABIAS or ARMSE.

**Limitations and Future Research**

There are several limitations regarding the generalization of the results of this dissertation. This section provides the limitations and suggestions for future research.

This dissertation used only two sets of real data, which limits the generalization of the results. In Study 2, pseudo forms were created using SG equipercentile as the equating criterion. Even though the SG equipercentile equating method often is considered as a reasonable criterion, the generalization about performances of all equating methods should be exercised with caution because the pseudo forms might be quite different from the original test forms in many aspects. Pseudo forms were created
from an intact form by splitting the items, so they were naturally shorter than the original
test. Study 2 also contained MC+FR common items and conducted MIRT equating,
which is not consistent with the operational settings.

This dissertation used data from two subject areas: Spanish and French. Since
these subjects involved reading items, there were passages within each reading section.
The passages (i.e., testlets) can be another source of dimensionality because items within
each passage tend to be highly correlated; however, for the purpose of simplicity this
study did not take the testlet effect into consideration.

Lastly, it is important to note that full MIRT models used in this dissertation are
exploratory in nature. As was discussed in Chapter 3, the full MIRT models can be used
to roughly investigate the dimensionality of the data. The full MIRT models are
exploratory in their nature, which means the latent dimensions are not pre-specified
unlike the confirmatory models such as SS and BF. In order to compare their relative
performance to the confirmatory SS and BF models, two-dimensional solutions were
specified in an exploratory manner for the full MIRT models. The dimensions of SS and
BF that already are specified from test blueprints can be directly interpretable, while the
full MIRT procedure have limitations in interpreting the dimensions.

Future studies may consider further expanding upon the work presented in this
dissertation by including a simulation study. Since the data used in this dissertation were
not strongly multidimensional (e.g., disattenuated correlations ranged from 0.82 to 0.93),
the degree of multidimensionality can be widely varied with the simulation study so that
various (M)IRT equating procedures can be studied under a wide range of correlations.
Also, the simulation study can include different levels of the proportion of common items such as 20% to 40% to the total test. The composition of MC and FR on common items can also be considered. For example, the common items can consist of MC items of 10% + FR items of 5%, MC items of 20% + FR items of 5%, etc.

Various linking methods could be employed to investigate the impact of MIRT linking methods on MIRT equating under the CINEG design. Kim (2017) presented concurrent and fixed calibration methods for the BF model, and compared the relative appropriateness of separate, concurrent, and fixed parameter calibration methods under the BF model. The study found that concurrent calibration provided better performance than separate calibration, while the fixed parameter calibration method depended on the proportion of common items. These three linking methods in conjunction with MIIRT equating based on various MIRT models would need to be further examined.

To maintain the SS and BF structure when using separate calibration, the rotation matrix was set to an identity matrix in this dissertation as suggested by Kim (2017). Even though the results from this approach provided evidence to its application for MIRT linking and equating in this dissertation, more theoretical and practical evidence is needed. With respect to scale linking of the SS model, another possible way would be to do UIRT linking for each dimension. That is, all items are calibrated using the SS model first; then, scale linking is conducted by each dimension, because items that are calibrated under the SS model are loaded on only one dimension. More research is necessary in this area.
Conclusions and Implications

The primary purpose of this dissertation was to introduce assumptions, as well as to present specific process of MIRT equating under the CINEG design and evaluate the relative performance of various equating methods. This dissertation was the first attempt to expand previous studies for MIRT equating under the CINEG design in terms of three aspects: (a) introduced MIRT equating assumptions for the CINEG design, (b) applied SS scale linking when separate calibration was used, and (c) compared the SS, BF, and FM equating methods. Based on the results of this dissertation, the following conclusions can be emphasized:

First, it was declared that the number of dimensions and constructs should be the same for the old and new forms, and common items also should have the same dimensional structure as the total test. If one dimension that is found on the new form is not found on the old form, item parameter estimates cannot be placed on the same coordinate system, so that MIRT observed score equating is not achievable. Likewise, if one dimension is missing from the common items, then the common items cannot serve as common items in the MIRT context. Descriptive statistics such as mean and standard deviations of the old and new forms, and correlations between the latent dimensions for the two test forms can be different, because the MIRT observed score equating is conducted under the two multidimensionally-nonequivalent groups.

Second, it was found that the MIRT methods provided more accurate equating results than the UIRT equating methods. Specifically, the SS and full MIRT methods outperformed UIRT, and this trend was found with both Spanish and French datasets. SS
always resulted in more accurate equating results than UIRT across the two studies regardless of group difference or calibration methods. Even for French, which was found to have only slight evidence of multidimensionality, the UIRT equating method did not work better than the MIRT equating methods.

Third, concurrent calibration always yielded better performance in equating than separate calibration. Even though both separate and concurrent calibration showed similar equating relationships, the difference among equating methods for concurrent calibration was generally smaller than that of separate calibration. For example, bias values of the UIRT method were generally larger than the other equating methods across all conditions for separate calibration, but discrepancy between the UIRT method and the other equating methods for concurrent calibration was smaller.

Finally, group ability difference had an impact on the performance of the MIRT equating methods. In general, larger group difference introduced more equating error for all equating methods. However, the UIRT and BF methods were more easily affected by group difference than the SS and full MIRT methods.

Although UIRT is an easy model for analyzing data, the unidimensionality assumption is not realistic in practical situations. The assumption that an examinee needs one ability to answer an item correctly is very ideal, but an examinee’s cognitive process to reach an answer is not that simple. Therefore, a MIRT approach is eventually needed in educational or psychological testing situations. For example, many testing programs such as ACT (ACT, 2018) and SAT (College Board, 2015) report subsection scores for each subject, which proves that an examinee’s ability is composed of multiple traits. In such
cases, equating with MIRT models could be considered as an alternative to UIRT equating. It is suggested that conducting equating should be preceded by dimensionality assessment. If there is any evidence of multidimensionality, MIRT equating would be preferred as long as the chosen MIRT model reflects appropriately the dimensionality structure of the data.
REFERENCES


APPENDIX. SAMPLE FLEXMIRT SYNTAX

A1. Sample flexMIRT syntax for separate calibration using the 2PL IRT model

<Project>
Title = "Separate calibration";
Description = "Spanish UIRT separate calibration";

<Options>
Mode = Calibration;
MaxE = 3000;
Etol = 0.005;
SavePRM = Yes;
Quadrature = 21, 4.0;
NormalMetric3PL = Yes;
SlopeThreshold = Yes;
Processors = 4;

<Groups>
%group%
File = "spa_new.txt";
Varnames = V1-V70;
N = 5000;
Ncats(V1-V70) = 2;
Model(V1-V70) = Graded(2);

<Constraints>
Prior (V1-V70), Slope: logNormal(0, 0.5);
A2. Sample flexMIRT syntax for concurrent calibration using the 2PL IRT model

<Project>
   Title = "Concurrent calibration";
   Description = "Spanish UIRT concurrent";

<Options>
   Mode = Calibration;
   MaxE = 3000;
   Etol = 0.005;
   SavePRM = Yes;
   Quadrature = 21, 4.0;
   NormalMetric3PL = Yes;
   SlopeThreshold = Yes;
   Processors = 4;

<Groups>
   %New%
   File = "spa_new.txt";
   Varnames = V1-V70;
   N = 5000;
   Ncats(V1-V70) = 2;
   Model(V1-V70) = Graded(2);

   %Old%
   File = "spa_old.txt";
   Varnames = V1-V70;
   N = 5000;
   Ncats(V1-V70) = 2;
   Model(V1-V70) = Graded(2);

<Constraints>
   Prior Old, (V1-V70), Slope: logNormal(0, 0.5);
   Prior New, (V1-V70), Slope: logNormal(0, 0.5);
   Free New, Mean(1);
   Free New, Cov(1, 1);
A3. *Sample flexMIRT syntax for separate calibration using the 2PL SS model*

```xml
<Project>
    Title = "Separate calibration";
    Description = "Spanish SS separate calibration";
</Project>

<Options>
    Mode = Calibration;
    MaxE = 3000;
    Etol = 0.005;
    SavePRM = Yes;
    Quadrature = 21, 4.0;
    Processors = 4;
</Options>

<Groups>
    %group%
    File = "spa_new.txt";
    Varnames = V1-V70;
    N = 5000;
    Dimensions = 2;
    Ncats(V1-V70) = 2;
    Model(V1-V70) = Graded(2);
</Groups>

<Constraints>
    Fix (V1-V70), Slope;
    Free (V1-V34), Slope(1);
    Free (V35-V70), Slope(2);
    Free Cov(2,1);
    Prior (V1-V34), Slope(1): logNormal(0, 0.5);
    Prior (V35-V70), Slope(2): logNormal(0, 0.5);
```
A4. Sample flexMIRT syntax for concurrent calibration using the 2PL SS model

<Project>
  Title = "Multiple group calibration";
  Description = "Spanish SS concurrent calibration";

<Options>
  Mode = Calibration;
  MaxE = 3000;
  Etol = 0.005;
  SavePRM = Yes;
  Quadrature= 21, 4.0;
  Processors=4;

<Groups>
  %old%
  File = "spa_old.txt";
  Varnames = V1-V70;
  N = 5000;
  Dimensions = 2;
  Ncats(V1-V70) = 2;
  Model(V1-V70) = Graded(2);

  %New%
  File = "spa_new.txt";
  Varnames = V1-V70;
  N = 5000;
  Dimensions = 2;
  Ncats(V1-V70) = 2;
  Model(V1-V70) = Graded(2);

<Constraints>
  Free New, Mean(1);
  Free New, Mean(2);
  Free New, Cov(2,1);

  Fix Old, (V1-V70), Slope;
  Free Old, (V1-V34), Slope(1);
  Free Old, (V35-V70), Slope(2);

  Fix New, (V1-V70), Slope;
  Free New, (V1-V34), Slope(1);
  Free New, (V35-V70), Slope(2);

  Equal Old, (V1-V4, V19-V26), Slope(1) : New, (V1-V4, V19-V26), Slope(1);
  Equal Old, (V35-V43), Slope(2) : New, (V35-V43), Slope(2);
  Equal Old, (V1-V4, V19-V26), Intercept : New, (V1-V4, V19-V26), Intercept;
  Equal Old, (V35-V43), Intercept : New, (V35-V43), Intercept;
Prior Old, (V1-V34), Slope(1): logNormal(0, 0.5);
Prior Old, (V35-V70), Slope(2): logNormal(0, 0.5);
Prior New, (V1-V34), Slope(1): logNormal(0, 0.5);
Prior New, (V35-V70), Slope(2): logNormal(0, 0.5);
A5. Sample flexMIRT syntax for separate calibration using the 2PL BF model

<Project>
  Title = "Separate calibration";
  Description = "Spanish BF new form separate calibration";

<Options>
  Mode = Calibration;
  MaxE = 3000;
  Etol = 0.005;
  SavePRM = Yes;
  Quadrature = 21, 4.0;
  Processors=4;

<Groups>
  %group%
  File = "spa_new.txt";
  Varnames = V1-V70;
  N = 5000;
  Dimensions = 3;
  Primary=1;
  Ncats(V1-V70) = 2;
  Model(V1-V70) = Graded(2);

<Constraints>
  Fix (V1-V70), Slope;
  Free (V1-V70), Slope(1);
  Free (V1-V34), Slope(2);
  Free (V35-V70), Slope(3);

  Prior (V1-V70), Slope(1): logNormal(0, 0.5);
  Prior (V1-V34), Slope(2): logNormal(0, 0.5);
  Prior (V35-V70), Slope(3): logNormal(0, 0.5);
A6. Sample flexMIRT syntax for concurrent calibration using the 2PL BF model

<Project>
Title = "Multiple group calibration";
Description = "Spanish BF concurrent calibration";

<Options>
Mode = Calibration;
MaxE = 3000;
Etol = 0.005;
SavePRM = Yes;
Quadrature= 21, 4.0;
Processors=4;

<Groups>
%Old%
File = "spa_old.txt";
Varnames = V1-V70;
N = 5000;
Dimensions = 3;
Primary= 1;
Ncats(V1-V70) = 2;
Model(V1-V70) = Graded(2);

%New%
File = "spa_new.txt";
Varnames = V1-V70;
N = 5000;
Dimensions = 3;
Primary= 1;
Ncats(V1-V70) = 2;
Model(V1-V70) = Graded(2);

<Constraints>
Free New, Mean(1);
Free New, Mean(2);
Free New, Mean(3);
Fix Old, (V1-V70), Slope;
Free Old, (V1-V70), Slope(1);
Free Old, (V1-V34), Slope(2);
Free Old, (V35-V70), Slope(3);
Fix New, (V1-V70), Slope;
Free New, (V1-V70), Slope(1);
Free New, (V1-V34), Slope(2);
Free New, (V35-V70), Slope(3);
Equal Old, (V1-V4,V19-V26), Slope(1) : New, (V1-V4,V19-V26), Slope(1);
Equal Old, (V1-V4,V19-V26), Slope(2) : New, (V1-V4,V19-V26), Slope(2);
Equal Old, (V1-V4,V19-V26), Intercept : New, (V1-V4,V19-V26), Intercept;

Equal Old, (V35-V43), Slope(1) : New, (V35-V43), Slope(1);
Equal Old, (V35-V43), Slope(3) : New, (V35-V43), Slope(3);
Equal Old, (V35-V43), Intercept : New, (V35-V43), Intercept;

Prior Old, (V1-V70), Slope(1): logNormal(0,0.5);
Prior Old, (V1-V34), Slope(2): logNormal(0, 0.5);
Prior Old, (V35-V70), Slope(3): logNormal(0, 0.5);

Prior New, (V1-V70), Slope(1): logNormal(0,0.5);
Prior New, (V1-V34), Slope(2): logNormal(0, 0.5);
Prior New, (V35-V70), Slope(3): logNormal(0, 0.5);
A7. Sample flexMIRT syntax for separate calibration using the 2PL FM model

<Project>
  Title = "Separate calibration"
  Description = "Spanish new form FM separate calibration"

<Options>
  Mode = Calibration;
  MaxE = 3000;
  Etol = 1e-4;
  Mtol = 1e-4;
  SavePRM = Yes;
  Quadrature = 21, 4.0;
  Processors = 4;

<Groups>
%group%
  File = "spa_new.txt"
  Varnames = V1-V70;
  N = 5000;
  Dimensions = 2;
  Ncats(V1-V70) = 2;
  Model(V1-V70) = Graded(2);

<Constraints>
  Free Cov(2,1);
  Prior (V1-V70), Slope(1): logNormal(0, 0.5);
  Prior (V1-V70), Slope(2): logNormal(0, 0.5);
A8. Sample flexMIRT syntax for concurrent calibration using the 2PL FM model

<Project>
Title = "Multiple group calibration";
Description = "Spanish FM concurrent calibration";

<Options>
Mode = Calibration;
MaxE = 3000;
Etol = 0.005;
SavePRM = Yes;
Quadrature= 21, 4.0;
Processors=4;

<Groups>
%old%
File = "spa_old.txt";
Varnames = V1-V70;
N = 5000;
Dimensions = 2;
Ncats(V1-V70) = 2;
Model(V1-V70) = Graded(2);

%New%
File = "spa_new.txt";
Varnames = V1-V70;
N = 5000;
Dimensions = 2;
Ncats(V1-V70) = 2;
Model(V1-V70) = Graded(2);

<Constraints>
Free New, Mean(1);
Free New, Mean(2);
Free New, Cov(2,1);

Equal Old, (V1-V4,V19-V26,V35-V43), Slope(1) : New, (V1-V4,V19-V26,V35-V43), Slope(1);
Equal Old, (V1-V4,V19-V26,V35-V43), Slope(2) : New, (V1-V4,V19-V26,V35-V43), Slope(2);

Prior Old, (V1-V70), Slope(1): logNormal(0, 0.5);
Prior Old, (V1-V70), Slope(2): logNormal(0, 0.5);

Prior New, (V1-V70), Slope(1): logNormal(0, 0.5);
Prior New, (V1-V70), Slope(2): logNormal(0, 0.5);