A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT

Author: Feng Tian

BOSTON COLLEGE
Lynch School of Education

Department of
Educational Research, Measurement, and Evaluation

A COMPARISON OF EQUATING/LINKING USING THE STOCKING-LORD METHOD AND CONCURRENT CALIBRATION WITH MIXED-FORMAT TESTS IN THE NON-EQUIVALENT GROUPS COMMON-ITEM DESIGN UNDER IRT

Dissertation
by

FENG TIAN

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

August 2011

Abstract

A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT

By Feng Tian

Larry Ludlow Ph.D., Chair

There has been a steady increase in the use of mixed-format tests, that is, tests consisting of both multiple-choice items and constructed-response items in both classroom and large-scale assessments. This calls for appropriate equating methods for such tests. As Item Response Theory (IRT) has rapidly become mainstream as the theoretical basis for measurement, different equating methods under IRT have also been developed. This study investigated the performances of two IRT equating methods using simulated data: linking following separate calibration (the Stocking-Lord method) and the concurrent calibration. The findings from this study show that the concurrent calibration method generally performs better in recovering the item parameters and more importantly, the concurrent calibration method produces more accurate estimated scores than linking following separate calibration. Limitations and directions for future research are discussed.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who have supported me during this doctoral work. First of all, I want to thank the Health and Disability Research Institute (HDRI) at Boston University and its associate director Dr. Steve Haley. It is through working with HDRI and Dr. Haley that the topic for this dissertation study originated. Unfortunately, Dr. Haley passed away before he could see this study completed. This dissertation is dedicated to the memory of Dr. Haley as a mentor and a friend.

I am grateful for the support I received throughout this long journey by my advisor and committee chair Dr. Larry Ludlow, who has given his time and continuous support from the beginning to the completion of this process. His guidance, patience, kindness and advice were instrumental in providing me with the focus needed to complete this work. I also would like to thank the committee members, Dr. Joseph Pedulla and Dr. Walt Haney for their insightful thoughts into the interpretation of the analyses results and the writing style, and for their time and commitment, which by all means made this dissertation a much better piece of work.

I owe a special note of gratitude to Dr. Pengsheng Ni for his recommendation of the software and help in carrying out the analysis.

Finally I am thankful to my family for their ongoing support, especially my parents who have been waiting for me to complete my degree with their patient love. I finally made their wish come true.

TABLE OF CONTENTS

List of Tables ..... **v**

List of Figures ..... **vi**

Chapter One: Introduction **1**
Background of the Study 1
Overview of Equating/Linking 2
Data Collection Designs for Equating 6
Overview of Equating Methods 8
Statement of the Problem 12
Purpose and Research Questions 14
Significance of the Study 15
Summary 16

Chapter Two: Review of Related Literature **17**
Traditional Equating methods 17
IRT Equating methods 19
A brief Introduction of Item Response Theory 19
The Transformation of IRT Scales 23
Four IRT Linking Methods Following Separate Calibration 27
Characteristic of the Anchor Test 31
Content Representativeness 33
Length of the Anchor Test 34
No DIF items in the Anchor Test 36
Construct/Trait Equivalence of MC and CR Items 36
Correlational Studies 38
Studies Using Factor Analysis 40
Research Studies Comparing Different Equating Methods 44
Overview of Research Methods in Comparative Studies of Equating methods 50
Summary 54

Chapter Three: Methodology 55
Instrument 55
A Brief Introduction of TIMSS 55
The Instrument in This Study 57
Factors Investigated 60
Two Types of Linking: Concurrent Calibration versus the Stocking-Lord Method 60
Three Different Lengths of common items: 5,10 and 15 Common Items 61
Three Types of Common-item Sets: Both MC & CR Items, MC Items Only and CR Items Only 61
Data Generation 64
Evaluation Criteria 65
Recovery of Item Parameters After Equating 65
The Difference Between the Estimated Person Parameters and the "True" Person Parameters 66
Data Analysis Procedure 67
Summary 68

Chapter Four: Results 70
Recovery of Item Parameters 70
Recovery of the Discrimination Parameter *a* 70
Recovery of the Difficulty Parameter *b* 73
Recovery of the Guessing Parameter *c* 76
The Difference Between the Estimated Person Parameters and the "True" Person Parameters 78
Summary 81

Chapter Five: Conclusions 83
Review of the Goal and the Methodology of the Study 83
Summary of the Findings 84
Recovery of Item Parameters 82
The Difference Between the Estimated Person Parameters and the "True" Person Parameters 85
Conclusions, Implications and Discussions 86

Limitations 91
Suggestions for Future Studies 93
Summary 94

References 95

Appendix A: ICL Codes 109

LIST OF TABLES

Table 3.1: Summary of Item Characteristics TIMSS 2003 Mathematics Grade 8 56
Table 3.2: Item Characteristics Used for Simulating Two Mixed-format Test forms 58
Table 3.3: Characteristics of Common Items 63
Table 4.1: MSEs of the *a* Parameter 71
Table 4.2: SDs of the *a* Parameter 71
Table 4.3: MSEs of the *b* Parameter 73
Table 4.4: SDs of the b Parameter 73
Table 4.5: MSEs of the *c* Parameter 76
Table 4.6: SDs of the *c* Parameter 76
Table 4.7: RMSDs between the Estimated Person Parameters and the "True" Person Parameters 79

LIST OF FIGURES

Figure 2.1: Equipercentile Equating on Two Hypothetic Tests.....19
Figure 2.2: Hypothetic TCCs for Two Test Forms.....29
Figure 4.1: MSEs of the *a* Parameter.....71
Figure 4.2: SDs of the *a* Parameter.....72
Figure 4.3: MSEs of the *b* Parameter.....74
Figure 4.4: SDs of the *b* Parameter.....74
Figure 4.5: MSEs of the *c* Parameter.....77
Figure 4.6: SDs of the *c* Parameter.....77
Figure 4.7: RMSDs between the Estimated Person Parameters and the "True" Person Parameters.....80

Chapter One: Introduction

Background of the Study

There are different formats of items that can be used in a test, each with its strengths and weaknesses. The formats usually fall into two categories: multiple-choice (MC) items and constructed-response (CR) items. The term *constructed-response* is generally used to refer to any question format that requires the examinees to produce a response in any way other than selecting from a list of alternative answers as they usually do for a MC item. Typically, MC items are dichotomously scored (DS) and CR items are polytomously scored (PS). MC items are economically practical and allow reliable and objective scoring; however, MC items are generally considered not to be able to measure students' higher-order thinking skills and limit the opportunity for demonstrating in-depth knowledge (Madaus, Haney & Kreitzer, 1992). CR items are considered to be able to measure traits that cannot be tapped by MC items; however, they are expensive to develop and difficult to score objectively and reliably (Wainer & Thissen, 1993). So a test composed of a mixture of different item formats is often used in assessments because combinations of different item formats allow for the measurement of a broader set of skills than the use of a single format (Kim & Lee, 2004). Such a test is referred to as a mixed-format test. There has been a steady increase in the use of mixed-format tests in both classroom and large-scale assessments. In 2008, 49 states used mixed-format tests in the subject of English language and 28 states used mixed-format tests in subjects other than English language in their state assessment programs (Quality Counts, 2009).

Overview of Test Equating/Linking

Most major testing programs, particularly large-scale or high-stakes testing programs, require the construction and administration of multiple forms of the same test. The main reason for this is that many testing programs administer tests several times in a year, and testing programs rarely administer the same test more than once to protect the security of the tests and meet the demands of an examinee population for flexible testing dates. Therefore, the construction and administration of alternate forms of the same test is a necessary requirement for operating these testing programs (Cook, 2007). The use of different forms of the same test raises the issue of the comparability of test scores. In order to use the scores from different forms of a test interchangeably, they must be put on a common scale. Different terms have been used to describe the procedure of transforming scores from different tests or test forms to make them comparable. Two frequently used terms are equating and linking.

Earlier efforts (Linn, 1993; Mislevy, 1992) were made to bring coherence to the definitions of linking and equating; however, the literature was not completely consistent in the use of the terminology (Feuer, Holland, Green, Bertenthal*, &* Hemphil, 1999). As more and more attention is paid to establishing linkage between tests, practitioners gradually realize that there are different types of linking and become aware of the importance of making distinctions among linkage types and linking scenarios, which on the surface appear to be essentially the same (Dorans, Pommerich, & Holland, 2007). Various frameworks (Feuer et al., 1999; Holland, 2007; Holland & Dorans, 2006; Kolen & Brennan, 2004; Pommerich & Dorans, 2004) have been presented to distinguish

among and develop terminology for different types of linking. The Holland (2007) and Holland and Dorans (2006) frameworks are the most up-to-date and most comprehensive ones and are followed in this study. According to Holland and Dorans (2006), the term *linking* refers to the general class of transformations between the scores from one test and those of another test, and there are three basic categories of linking: predicting, scale aligning and equating.

Predicting is the oldest form of linking (Holland, 2007). The goal of predicting is to predict an examinee's score on one test from other information sources, which are also called predictors by Dorans and Walker (2007) about that examinee such as a score from one other test, scores from several other tests or demographic information (Holland, 2007). Clearly, predicting is based on the linear regression method. It was recognized very early that linear regression was not a satisfactory way to find comparable scores (Thorndike, 1922, cited in Holland 2007; Otis, 1922, cited in Holland, 2007).

Scale aligning is the second oldest group of linking methods (Holland, 2007). The goal of scale aligning is to transform the scores from different tests onto a common scale (Dorans & Walker, 2007). Holland (2007) reported that scale aligning has several subcategories that can be used in different situations, including battery scaling (Kolen, 2004), anchor scaling (Holland & Dorans, 2006), vertical scaling (Kolen & Brennan, 2004), calibration (Holland & Dorans, 2006) and concordance (Pommerich & Dorans, 2004). Battery scaling is used when two or more tests that measure different constructs are administered to a common population; anchor scaling is used when two or more tests that measure different constructs are administered to different populations of examinees

and a common measure called an anchor measure is available for all the examinees in these different populations. Vertical scaling is used when tests that measure similar constructs are administered to different populations of examinees and a common measure is available for all examinees. Calibration is used in situations where the tests measure the same construct, have similar level of difficulty but differ in test length. Dorans (2007) pointed out that calibration also refers to the process of estimating the parameters of an item in the parlance of Item Response Theory.

Equating is the strongest kind of linking and is viewed widely as possessing the following desirable properties (Lord, 1980):

1. The same construct property. The two tests must measure the same construct.

2. Equity property. Once the two test forms have been equated, it should not matter to the examinees which form of the test is administered.

3. Symmetry property. The equating transformation should be symmetric. The equating of Form A to Form B should be the inverse of equating to Form B to Form A.

4. Group invariance property. The equating relationship should be the same regardless of the group of examinees used to conduct the equating.

In addition to these properties, Dorans and Holland (2000) added another property for test equating, that is,

5. Equal reliability property. The tests to be equated should have equal reliability.

In practice, these properties mean that different forms of a test need to be built to the same explicit content and statistical specifications and administered under the same conditions. These forms are referred to as *alternate forms of a test* or sometimes *parallel*

or *equivalent* forms, which according to (Kolen, 2007) have nearly identical content features and differ only in the particular items that appear on the alternate forms. Other cases of score linking are likely to violate at least one of the five properties for equating.

There are two main forms that equating can take: horizontal equating and vertical equating. For horizontal equating, the tests to be equated are designed to be as psychometrically identical as possible, that is, the tests are constructed to be parallel in both content and difficulty (Kolen, 2007). Most equating applications are of this type. Equating new versions to old versions of standardized tests such as SAT or ACT is an example of horizontal equating. For vertical equating, the tests to be equated are intentionally designed to be different in difficulty level but still measure the same construct. Most often vertical equating occurs in the context of an achievement test battery. For example, a fourth grader takes a mathematics achievement test at the fourth-grade level in a year and the following year the same student takes the same test battery at the fifth-grade level. This is a typical situation where vertical equating is required to compare the scores from the two tests. However, Kolen (2007, 1988) does not think that vertical equating should be included in test equating, because the content of the tests administered to students at various educational levels is different and scores on the tests at different levels cannot be used interchangeably. According to Kolen (2007), this process can be referred to as *vertical scaling*. In the terminology of the Standards for Educational and Psychological Testing (AERA, APA, NEME, 1995), this process is referred to as *scaling to achieve comparability*.

The major distinction between the terms *linking* and *equating* and between *vertical equating* and *vertical scaling* is conceptual in nature and there is not so much distinction in terms of the data collection design and statistical procedure to establish the relationship between the scores from different tests or test forms because the methodology remains essentially the same regardless of whether all of the properties for equating are met or not. For this reason the two terms *linking* and *equating* are used interchangeably in this study and they both generally refer to the procedure of establishing comparable scores from different tests or test forms.

Data Collection Designs for Equating

Test equating starts with data collection. Three factors may have an impact on the linking function, which are respectively differences in conditions of the test, differences in test content**,** and differences in examinees (Kolen & Brennan, 2004). The role of data collection is crucial to successful linking. It is the key to control for differences in test content and in examinees. Three commonly used data collection designs are (1) single-group design (2) equivalent groups design and (3) non-equivalent groups common-item design or non-equivalent groups anchor test (NEAT) design (von Davier, Holland, & Thayer, 2004).

In the single-group design, two test forms are administered to the same group of examinees. The advantage of this design is that it directly controls for the differences in both the test content and examinees. A disadvantage is that there could be an order effect. If fatigue is a factor in examinee performance, the form administered later may tend to be

more difficult than the form administered earlier; if familiarity with the test is related to examinee performance, then the form administered later may tend to be easier. One way to deal with the order effect is to counterbalance the order of administration of the forms. Because two forms must be administered to the examinees, it requires more testing time than is practically allowed. So this design is rarely used in practice.

In the equivalent-group design, two test forms are administered to two equivalent groups of examinees. The groups are randomly formed, so this design is also referred to as random groups design (Kolen & Brennan, 2004). The advantage of this design is that it avoids the issue of the order effect that can arise in the single-group design due to fatigue and familiarity with the test and it also requires less testing time compared to the single-group design. So the difference in the performance between the two groups can be taken as a direct indication of the difference in difficulty between the two test forms. However, equating is usually done to link a new test form to an old test form in practice. So this design is not appropriate for some high-stakes testing for purposes like admission, certification, and licensure because many of these testing programs usually use a new test at every administration to help maintain test security. Some other disadvantages are that this design requires a relatively large sample and random assignments of test forms to a large sample is not always practically possible.

Finally, in the NEAT design, two test forms are administered to two different groups of examinees. A set of common items or anchor test is included in both test forms. The role of the anchor test is to quantify the differences between the two groups. The anchor test may be internal if the score on t he common items contributes to the

examinee's score on the test, or external if the scores on the common items do not contribute to the examinee's score. The NEAT design is probably the most prevalent design in practice (Cook, 2007; Skaggs & Lissitz, 1986a). A major reason for its popularity is that this design requires administration of only one test form per test date, and the groups do not need to be from a common population. In contrast, the other two designs mentioned previously require administration of more than one test per test date. So the NEAT design has greater operational flexibility than the other two designs. However, this flexibility comes at a price. For this design to work best, certain conditions have to be met: similarity of the two groups taking the two forms, similarity of the two forms and a high correlation between the scores of the two forms and the anchor test (Cook, 2007). So this design is also the most difficult one to implement. Because of the prevalence, the importance and the complexities of the NEAT design, this study focuses on equating based on this design.

Overview of Equating Methods

Two theories dominate the field of measurement, Classical Test Theory (CTT) and Item Response Theory (IRT). Test equating methods, according to the testing theory on which they are based, can generally be classified into two categories: traditional or conventional equating methods and IRT equating methods. Traditional equating methods are based on CTT, which include (a) mean equating, (b) linear equating, and (c) equipercentile equating. There are two alternative procedures of using IRT in equating. One is to calibrate all the items in different forms together so that they are on a same

metric, which is known as concurrent calibration; another is to estimate the item parameters separately for different forms first and then put them on a same metric through a linking process.

The major advantages of CTT are its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). However, the analysis of item responses in CTT is heavily dependent on the sample that completed the test. This causes the major disadvantage of CTT models, that is, item parameter estimates (item difficulty) are dependent on examinees and person parameter estimates (ability) are dependent on the test. These dependencies limit the utility of the person and item statistics in practical test development work and complicate analyses (Hambleton & Jones, 1993). In the case of test equating/linking, each of the CTT methods is of limited utility because they do not satisfy conditions of equity, symmetry and invariance (Hambleton, Swaminathan & Rogers, 1991). Kolen (1981) clearly stated that '*...conventional equipercentile or linear equating can be strictly used only with parallel tests...'* (p.1-2). However, this is not the case for IRT. Compared to CTT, IRT has several advantages (Hambleton & Swaminathan, 1985). One of them is that item parameter estimates are independent of the ability level of examinees responding to the item and at the same time, ability estimates are also independent of the performance of other examinees and the items used in the test. So even if different samples are used in the calibration of the items, the item parameter estimates are not expected to change and if different sets of items are used, the person ability estimates are not expected to change. This is known as the "invariance" feature of IRT. Another advantage is that item

difficulty and person ability estimates are set on the same scale. So equity, symmetry and invariance are the basic properties of IRT models (Hambleton & Swaminathan, 1985; Lord, 1980). In theory, IRT models can be used to equate both parallel and nonparallel tests (Kolen, 1981).

In addition to the theoretical advantages of IRT equating over CTT equating, there are also several practical advantages through the use of IRT equating (Cook & Eignor, 1983). First, IRT equating offers better equating results than CTT methods at the upper ends of score scales where important decisions are often made; whereas for CTT methods, reasonable equating can only take place for only those scores actually obtained by the test takers. Second, IRT methods provide greater flexibility in choosing previous test forms for equating purposes. Because all previous test forms are calibrated on the same scale, the new form can be equated to any or all of the old forms. Third, re-equating becomes easier. If an item is dropped, the shortened form can easily be reconstructed based on the item information from the remaining test forms. Finally, IRT equating makes pre-equating possible. Pre-equating is an attempt to prepare a raw-to-scale score conversion table even before a test or form is administered. This is possible when item-level pretest data are available and can be calibrated. The parameters will be invariant when applied to new groups.

Studies have also been conducted to compare different equating methods including the comparison between the CTT and IRT equating methods (Marco, Petersen & Stewart, 1983; Kolen, 1981; Petersen, Cook & Stocking, 1983; Cook & Eignor, 1983). The results demonstrated some superiority of IRT equating methods over CTT equating

methods. For example, Marco et al. (1983) showed that when tests of unequal difficulty were equated, the IRT methods displayed the least amount of error. This study suggests the superiority of IRT methods when samples are not randomly chosen. Peterson et al. (1983) compared IRT methods and CTT methods for the SAT-Verbal tests and found that the IRT methods produced substantially smaller equating errors than the CTT methods. In another study, Cook and Eignor (1983) compared IRT methods and CTT methods for a variety of achievement tests of the College Board Admission Testing Program and Graduate Record Examination (GRE) and found that the equipercentile method was extremely inadequate in all cases. They felt that this method suffered from a lack of data at extreme scores.

IRT, however, also has some disadvantages. It has some restrictive assumptions such as unidimensionality and local independence to meet to be satisfactorily implemented. Whereas a single dimension is implicit in any test equating, IRT methods might be less robust with respect to violation of this assumption and failure to meet the unidimensionality assumption could potentially be a major source of problems for IRT equating methods (Skaggs & Lissitz, 1986a). Also, IRT models, especially the three-parameter model, usually require a l arge sample size for stable estimation of item parameters. Studies (Patience, 1981; Kolen & Whitney, 1982) attributed small sample size to be a contributing factor to the poor performance of IRT equating methods using three-parameter model because the estimation procedure either could not converge or produced extreme (unstable) item parameter estimates with small sample size.

Statement of the Problem

Although CTT has served test development well over several decades, IRT has rapidly become mainstream as the theoretical basis for measurement (Embretson & Reise, 2000). In practice, IRT models are applicable to various formats of items on a mixed-format test. As Baker & Kim (2004) pointed out, any combination of dichotomous models such as the 3PL model and a polytomous model such as Graded Response (GR) model or Generalized Partial Credit (GPC) model can be used to analyze data from mixed-format tests in situations where the choice is feasible. As the use of IRT in testing applications has grown considerably over the last few decades, different equating methods under IRT methods have also been developed. A key issue in IRT equating when dealing with test data from multiple test forms with multiple groups of examinees is how to calibrate the data and place all the item parameter and ability estimates on a common scale so that they are comparable and can be used interchangeably. This metric issue is usually dealt with using one of two types of calibrations: (a) separate calibration with linking and (b) concurrent calibration (Hanson & Béguin, 2002; Kim & Cohen, 2002; Kolen & Brennan, 2004; Vale, 1986). Separate calibration is performed by form, in which the item parameters from different forms are estimated using separate runs of the estimation software. When separate calibration of test forms is conducted in each examinee group, a process of linking the resulting scales should be followed to develop a common scale because each scale is group or form dependent. The linking process uses IRT linking methods to find a linear transformation between different scales usually via a

set of common items. There are different IRT linking methods such as the mean/mean method (Loyd & Hoover, 1980), mean/sigma method (Marco, 1977), Haebara method (Haebara, 1980), and Stocking-Lord method (Stocking & Lord, 1983). The former two are also called *moment methods* (Hanson & Béguin, 2002; Kim & Lee, 2004) and the latter two, *characteristic curve methods* (Hanson & Béguin; 2002; Stocking & Lord, 1983). These methods have been developed first under the dichotomous IRT model and then extended to polytomous IRT models (Baker, 1992, 1993, 1997; Cohen & Kim, 1998; Kim & Cohen, 1995; Kim & Hanson, 2002). In contrast, concurrent calibration does not involve the linking process and therefore, is more efficient and is easy to apply in practice. In concurrent calibration, parameters of all items from multiple forms are simultaneously estimated through a single calibration run with all response data from the multiple forms being combined together. As a result, all of the estimates are placed on a common scale.

Studies have been conducted to compare different IRT equating procedures. In regard to the different separate calibration procedures, two studies (Kim & Lee, 2004; Hanson & Béguin, 2002) used simulated data and drew a similar conclusion that characteristic curve methods are preferable to the others. Several studies also compared the performance of methods following separate calibration and concurrent calibration. Petersen et al. (1983) and Wingersky, Cook, & Eignor (1987) both concluded that concurrent calibration performed somewhat better than methods following separate calibration. Hanson & Béguin (2002) studied separate versus concurrent calibration with simulated data and concluded that concurrent calibration generally resulted in lower error than separate calibration. Kim & Cohen (1998) also studied separate versus concurrent

calibration with simulated data and concluded that separate and concurrent calibration provided similar results except when the number of common items was small, in which case separate calibration provided more accurate results. However, most of the studies were conducted for tests consisting only of MC items and little research has been done to compare the different IRT equating procedures for mixed-format tests. Kim & Lee (2004) compared the performance of different equating approaches following separate calibration with mixed-format tests and concluded that the moment methods seem ineffective at linking mixed-format tests. Yet, this study did not investigate the performance of the linking methods following separate calibration as compared to the performance of concurrent calibration and called for further investigation into this.

Purpose and Research Questions

The purpose of this study is to evaluate the performance of linking methods with mixed-format tests following separate calibration versus concurrent calibration. The research question addressed in this study is: with mixed format tests, which IRT equating method performs better under different conditions for the non-equivalent groups common item design: the characteristic curve method following separate calibration or the concurrent calibration method?

Significance of the Study

Test scores are often used as an important information source for important decision-makings, be it at individual level, institutional level or public policy level. Regardless of the type of decision that is to be made, it should be based on the most accurate information possible: the more accurate the information, the better the decision (Kolen & Brennan, 2004). However, due to security and other reasons, different forms of a test are often administered to different examinees and equating is often employed to adjust the scores on different forms to achieve fairness; therefore, test equating is critical in making important decisions and accurate test equating is highly desired.

More and more assessments contain a mixture of MC and CR items. If IRT is used to analyze the test data, the comparability of IRT item parameter estimates across different test forms is an important matter, since all decisions about examinees are derived from these estimates. Different IRT equating methods do not yield the same parameter estimates; therefore, it is essential to make comparisons between different IRT equating methods to determine what is the best equating procedure especially for mixed-format tests, since there has been a steady increase in the use of them. Unfortunately, little research has been done in this regard. Research has been done to compare different IRT equating methods following separate calibration, but no comparison has been made between the equating methods following separate calibration versus concurrent calibration for mixed-format tests. This study tries to fill the gap. It is expected that the findings of this study will provide useful guidelines on which IRT equating method is more appropriate under different conditions for mixed-formats equating in practice.

Summary

Multiple test forms are usually used in a test program for test security reasons and the scores from different forms have to be statistically transformed onto a common scale so that they are comparable, a procedure referred to as equating or linking. The increasing use of mixed-format tests, that is, tests that consist of both Multiple-choice items and Constructed-response items, calls for appropriate equating methods. This chapter provided a brief introduction to the concept of equating/linking, the different types of equating/linking approaches and the different data collection designs used in equating/linking. The purpose of this study is to compare two Item Response Theory (IRT) equating methods for mixed-format tests, namely the Stocking-Lord method and the concurrent calibration method for the non-equivalent groups common item design. The focus on IRT equating methods is because IRT has become mainstream as the theoretical basis for measurement and the focus on the non-equivalent groups common item design is because it is the most widely used equating design. A complete review of related literature is presented in Chapter Two.

Chapter Two: Review of related literature

Traditional Equating Methods

As mentioned previously, there are three methods based on CTT, namely, mean equating, linear equating and equipercentile equating. The three traditional equating methods are introduced briefly here and more discussions of these equating methods can be found in Angoff (1982, 1984), Braun and Holland (1982), Livingston (2004), and Kolen and Brennan (2004).

In mean equating, one test form is considered to differ in difficulty from another test form by a constant amount along the score scale (Kolen & Brennan, 2004). Scores on the two forms that are an equal distance away from their respective means are set equal:

$$x - \mu_x = y - \mu_y \quad (2.1)$$

where x and y are the raw scores and μ_x and μ_y are the means of the two test forms respectively. Mean equating involves the addition of a constant, which is the difference between the means of the two test forms. Mean equating assumes that differences in difficulty between the two forms are constant throughout the entire score range (Barnard, 1996). This assumption might be overly restrictive in many testing situations (Kolen & Brennan, 2004) and therefore, this method is seldom used in practice.

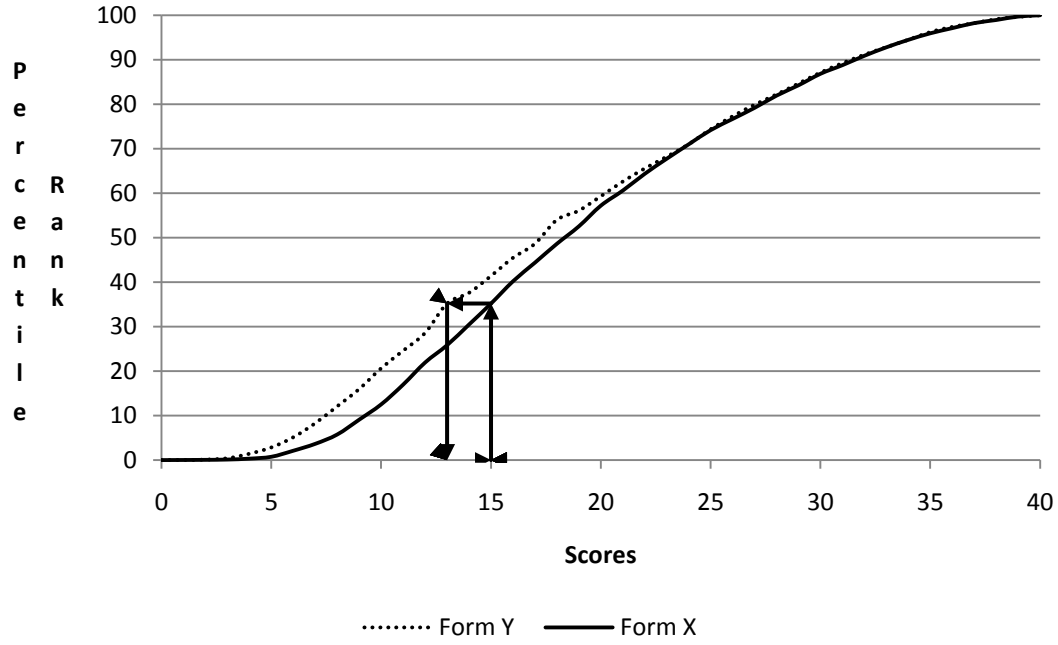
In linear equating, scores that are an equal distance from their mean in standard deviation units are set equal (Kolen & Brennan, 2004), or mathematically:

$$\frac{x - \mu_x}{\sigma_x} = \frac{y - \mu_y}{\sigma_y} \quad (2.2)$$

where x and y are the raw scores, μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of the two tests respectively. Actually, the above equation is also the equation for z-score transformation, and linear equating can conceptually be considered as establishing equivalent z-scores for two different tests. Linear equating assumes that, apart from differences in means and standard deviations, score distributions on two forms of a test are the same.

In equipercentile equating, scores that have the same percentile rank in the two forms are considered to be equal. The equipercentile equating procedure can be demonstrated using graphical methods. The first step of equipercentile equating is to determine the percentile ranks for the score distributions on each of the two tests to be equated. Percentile ranks are then plotted against the raw scores for each of the two tests and percentile rank-raw score curves can be constructed. Fig. 1 illustrates such a plot of two hypothetical tests, each with 40 items. As long as the percentile rank-raw score curves are constructed, it is fairly easy to convert equivalent scores from the plot. For example, a score of 15 in Test X is equivalent to a score of 13 in Test Y.

Fig 2.1 Equipercentile Equating on Two Hypothetical Tests



The main problem with equipercentile equating is that the score distributions on real tests taken by real test-takers are often irregular. The percentage of the test-takers with a given score does not change gradually as the scores increase; it fluctuates. Irregularities in the score distributions cause problems for equipercentile equating. They produce irregularities in the equipercentile equating adjustment, and those irregularities do not generalize to the population. So it usually requires a very large sample size for the score distribution and equipercentile relationship to be reasonably smooth (Kolen & Brennan, 2004). Or it requires a complicated technique called "smoothing" to produce estimates of the empirical distribution and the equipercentile relationship will have the smoothness property which is characteristic of the population. A detailed introduction of smoothing can be found in Kolen and Brennan (2004).

IRT Equating Methods

A Brief Introduction of Item Response theory

To better understand the different IRT equating methods, some basic knowledge about IRT is necessary. IRT, or *Item response theory*, as the name suggests, is a response-centered model. Briefly, IRT describes what happens when an examinee takes an item. The purpose of IRT is to estimate the probability of answering an item correctly. The probability is a function of the examinee's ability (considered a latent trait and Item Response Theory is also called latent trait theory) and the difficulty level of the item. The higher an examinee's ability, the more likely he/she will get an item right. The function, often referred to as item response function (IRF), can be graphically depicted as an s-shaped curve, called the Item Characteristic Curve (ICC). The summation of the ICCs of all the items in a test will form the Test Characteristic Curve (TCC), also an s-shaped curve. The TCC is a function of the examinee's ability and his/her expected score on the whole test: the higher the ability, the higher the score he/she is expected to achieve.

Several IRT models that vary in assumptions and in item parameters required to define the ICC are available. For dichotomous data, IRT models can be classified into one-, two- and three-parameter models depending on how many item parameters are incorporated in the model. In the one-parameter model, also called the Rasch model, named after its developer, only the *b* parameter or the item difficulty parameter is incorporated. Item difficulty is the point on the ability continuum at which individuals have a 50% chance of answering the item correctly. Statistically, the Rasch model can be expressed by the following equation which defines the ICC:

$$p_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} \qquad (2.3)$$

What is expressed in the equation is that the probability of a randomly selected examinee with ability θ succeeding on an item i is a function of the ability level of the person θ in relation to the difficulty level of the item b_i. In the model, exp = e is the natural logarithm base and has a value of roughly 2.718. In the Rasch model, it is assumed that the all items have the same discrimination and there is no guessing.

In the two-parameter model, a second item parameter, the item discrimination parameter, *a,* is incorporated. The addition of *a* allows the examination of item discrimination, which refers to how well the item can distinguish among individuals with different latent trait levels. Statistically, the two-parameter model can be expressed by the following equation:

$$p_i(\theta) = \frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]} \qquad (2.4)$$

where D is a scaling factor and is customarily set to equal to 1.7. It is also assumed that there is no guessing in this model.

In the three-parameter model, a third parameter *c*, the pseudo guessing parameter is incorporated, which estimates the likelihood of examinees getting an item correct by guessing. The three-parameter model can be expressed by the following equation:

$$p_i(\theta) = c_i + (1 + c_i)\frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]} \qquad (2.5)$$

For polytomous data, two popular models are available, which are both extended from the two-parameter model. One is *the graded response model* (GRM) developed by Samejima (1969, 1997). In the GRM, an item *i* has *m* ordered polytomous categories. The GRM specifies the cumulative probability of a response in category *k* or higher versus a response in categories lower than *k* and incorporates another parameter, the category parameter. The probability that an examinee obtains a certain score category is explained as follows,

$$p_{ik}(\theta) = p^{*}_{ik}(\theta) - p^{*}_{ik+1}(\theta) \qquad (2.6)$$

where $p^{*}_{ik}(\theta)$ and $p^{*}_{ik+1}(\theta)$ represent the cumulative probability of a response above category *k* and *k+1* for item *i* and

$$p^{*}_{ik}(\theta) = \frac{\exp[Da_i(\theta - b_{ik})]}{1 + \exp[Da_i(\theta - b_{ik})]} \qquad (2.7)$$

where b_{ik} is the so called item-category threshold parameter, a_i is the discrimination parameter for item *i.* Another model for polytomous data is the *generalized partial credit model* (GPC) developed by Muraki (1992). In the GPC, the probability of choosing a category *k* for an item *i* is expressed by the conditional probability of responding in category *k*, given the probability of responding in categories *k – 1*. The models are constructed by recursively applying a dichotomous model to the probability of choosing category *k* over another adjacent category *k -1* for each pair of binary categories. The probability function of scoring in category *k* on item i given the examinee's trait level, θ, for the partial credit model is defined as:

$$p_{ik}(\theta) = \frac{\exp\left[\sum_{v=0}^{k} a_i(\theta - b_{iv})\right]}{\sum_{c=0}^{m-1} \exp\left[\sum_{v=0}^{c} a_i(\theta - b_{iv})\right]} \qquad (2.8)$$

where b_{iv} is the so called item step parameter, a_i is the discrimination parameter for item *i*.

All IRT models share some common assumptions. The first assumption is unidimensionality. It assumes that there is only one underlying trait (latent trait) which decides the performance of examinees. However, this assumption is hardly tenable in practice since there are always cognitive, personality, and test taking factors that may affect test performance. So if test performance can be determined by a dominant factor, that is, the examinees' ability, this assumption is considered met. The assumption of unidimensionality also implies another assumption: local independence, which means that when examinees' ability is held constant, there is no other reason which could explain examinees' performance. Or the test performance of examinees at the same ability level is independent, so the response of an examinee to one item will not affect his/her response to the next one.

The Transformation of IRT Scales

Since invariance of item parameters is assumed in an IRT model, theoretically, there is no need for equating. The item parameters, once calibrated, can theoretically be used to estimate the abilities of any group of examinees. However, it is not completely true in real testing practice. The invariance feature will hold only if there is a single calibration in which a scale is set up. If two test forms are administered to different groups of examinees sampled from different populations and item parameters for the two

forms are estimated separately, then the item parameter estimates for the two forms will in general be on different IRT scales. This is due to the intrinsic indeterminacy of the latent variable scale in IRT, that is, there is not an absolute origin and unit of measurement for the latent variable scale. To resolve this problem, the scale, which could be either the persons or the items because in IRT they are on the same metric, is often arbitrarily set as having a mean of 0 and a standard deviation of 1 for the set of data being analyzed when conducting IRT analysis to fix the metric. Since the mean and the standard deviation are usually different for different groups, item parameter estimates are also usually not on a same scale. This indeterminant nature has an impact on how equating is conducted under different equating designs (Kolen & Brennan, 2004).

In the single group design, the examinees take both test forms and the parameters for the two forms are estimated together in a single calibration on the same examinees, so the parameters are assumed to be on the same scale. And there is no need for further action. In the equivalent/random groups design, if the same scaling convention of mean of 0 and standard deviation of 1 is used in separate calibration, then the parameter estimates for the two forms can be assumed to be on the same scale without further transformation. This is because the groups are randomly equivalent and the abilities are scaled to have the same mean and standard deviation in both groups. If different scaling conventions are used in the equivalent groups design, then a transformation is needed to adjust the differences in the scaling convention to make scores and item parameters on two IRT scales comparable. When conducting equating with nonequivalent groups as in the NEAT design, the parameter estimates that result from different calibrations are often

different on IRT scales. Thus, a transformation of IRT scales is also needed to convert one scale to the other. Actually as the NEAT design is the commonly used one (Kim & Hanson, 2002; Peterson, 2007) or even the most prevalently used one in practice (Cook, 2007), it is the focus of IRT equating methods and related IRT equating studies because most studies are based on this design and the present study takes no exception.

Kolen and Brennan (2004) show that the relationship between the two sets of item parameters from two separate calibrations is actually a linear one. So equating tests under IRT involves estimating two constants for a linear transformation that can be used to convert the item parameter estimates from one scale to the other one. If there are two scales, J and I and we use θ_J for scale J and θ_I for scale I, then the values on the two scales are related as follows:

$$\theta_J = \mathrm{A}\,\theta_I + \mathrm{B} \qquad (2.9)$$

where A and B are the constants of the linear function. A and B are often referred to as linking coefficients. Given the relation in equation (2.9), the item parameters on the two scales should also be related as follows (Baker 1992; Kolen & Brennan, 2004; Lord, 1980):

$$a_{Jj} = \frac{a_{Ij}}{\mathrm{A}} \qquad (2.10)$$

$$b_{Jj} = \mathrm{A}b_{Ij} + \mathrm{B} \qquad (2.11)$$

and

$$c_{Jj} = c_{Ij} \qquad (2.12)$$

where $a_{Jj,}$ $b_{Jj,}$ and c_{Jj} are the item parameters for item j on scale J and $a_{Ij,}$ $b_{Ij,}$ and c_{Ij} are the item parameters for item j on scale I. The pseudo guessing parameter c is independent of the scale transformation as is shown in equation (2.12). From equations (2.10) and (2.11), it is easy to get that

$$\mathrm{A} = \frac{a_{Ij}}{a_{Jj}} \tag{2.13}$$

$$\mathrm{B} = b_{Jj} - \mathrm{A}b_{Ij} \tag{2.14}$$

Equations (2.13) to (2.14) show the relationship between the two scales via the link by just one item administered to two different groups of examinees. In practice, a set of items is always used to link different test or test forms as in the NEAT design. Kolen and Brennan (2004) demonstrate that from equations (2.13) and (2.14), it follows that

$$A = \frac{\mu(a_{I)})}{\mu(a_J)}$$

$$= \frac{\sigma(b_J)}{\sigma(b_I)} \tag{2.15}$$

$$B = \mu(b_J) - A\mu(b_I) \tag{2.16}$$

Where $\mu(a_{I)})$ and $\mu(a_J)$ are the means of the a parameters of the common set of items from both scales, $\mu(b_J)$ and $\mu(b_I)$ are the means of the b parameters of the common items from both scales, and $\sigma(b_J)$ and $\sigma(b_I)$ are the standard deviations of the b parameters of the common items from both scales.

Equations (2.15) to (2.16) are just theoretical models. In real testing situations, the true item parameters are seldom known and item parameter estimates are only available.

As a result, the linking coefficients have to be properly estimated so as to minimize linking error due to sampling error. There are different approaches to estimate the appropriate A and B values for the transformation of different IRT scales when separate calibration is conducted for each test or test form; hence, there are different IRT linking methods. These methods are introduced in the following section.

Four IRT Linking Methods Following Separate Calibration

Four commonly used IRT linking methods are mean/mean, mean/sigma, Haebara, and Stocking-Lord methods. In mean/mean method as described by Loyd & Hoover (1980), the means of *a* parameter estimates for the common items of the two scales are used to replace the parameters in equation (2.15) to estimate the slope A and the means of *b* parameter estimates from the common items are used to replace the parameters in equation (2.16). This method can be summarized in the following equations:

$$\bar{A} = \frac{\mu(\bar{a}_I)}{\mu(\bar{a}_J)} \qquad (2.17)$$

$$\bar{B} = \mu(\bar{b}_J) - \bar{A}\,\mu(\bar{b}_I) \qquad (2.18)$$

where $\bar{A}$ and $\bar{B}$ are the linking coefficient estimates, $\mu(\bar{a}_I)$ and $\mu(\bar{a}_J)$ are the means of the *a* parameter estimates from the common items in scale I and J respectively, and $\mu(\bar{b}_J)$ and $\mu(\bar{b}_I)$) are the means of the *b* parameter estimates from the common items in both scales.

In the mean/sigma method as described by Marco (1977), the standard deviations and means of the *b* parameter estimates from the common items are used to replace the

parameters in equations (2.15) and (2.16). As a result, the equation for estimating the B constant in the mean/sigma method is the same as in the mean/mean method. The difference lies in how the A constant is estimated. This method can be summarized as in the following equations.

$$\bar{B} = \mu(\bar{b}_J) - \bar{A}\,\mu(\bar{b}_I) \tag{2.18}$$

$$\bar{A} = \frac{\sigma(\bar{b}_J)}{\sigma(\bar{b}_I)} \tag{2.19}$$

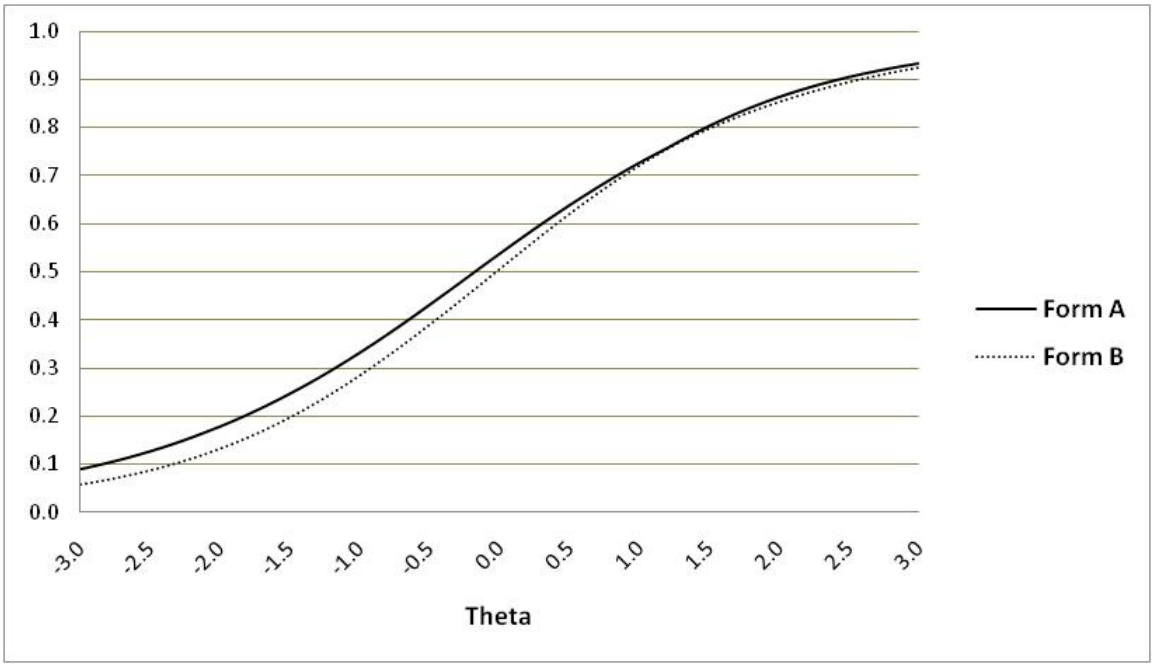
where $\sigma(\bar{b}_J)$ and $\sigma(\bar{b}_I)$ are the standard deviations of the b parameter estimates of the common items from both scales and the equation for estimating the B constant remains the same as in the mean/mean method.

Equations (2.17), (2.18) and (2.19) are appropriate for the dichotomous IRT model. Cohen and Kim (1998) extended these methods to the polytomous IRT model. For the mean/mean method, the means of the category or step parameter estimates of the common items from both scales are calculated then substituted for the parameters in equation (18). For the mean/sigma method, the means and the standard deviations of the category or step parameter estimates of the common items from both scales are calculated then substituted for the parameters in equation (2.18) and (2.19).

The moment methods follow a very straightforward way to transform the scales by substituting the means and standard deviations of the item parameter estimates of the common items. So they are conceptually simple and computationally easy. However, Kolen and Brennan (2004) pointed out one potential problem with the moment methods,

that is, the item parameters are not considered simultaneously. The characteristic curves of the common items from different groups are supposed to be the same or very similar after the parameter estimates are transformed onto a same scale. The characteristic curve of an item is determined simultaneously by the *a-, b-,* and *c-* parameters. In the moment methods, all the item parameters are treated separately. In this case, the moment methods, especially the mean/sigma method could be overly influenced by the differences between the *b-* parameter estimates. In contrast, the two characteristic curve methods developed by Haebara (1980) and Stocking and Lord (1983) consider all the item parameters simultaneously and therefore, are more complicated. The two characteristic curve methods attempt to find the linking coefficients that minimize the difference in test characteristic curves between the common items in different groups. The following figure gives a hypothetical example of test characteristic curves of two forms.

Fig 2.2 Hypothetical TCCs of Two Test Forms



As can be seen from the above graph, there are gaps between the two TCCs at different θ levels. The characteristic curve methods' strategy is to take the parameters of the items common to both measures and apply a transformation to them to find two equating coefficients A and B such that the TCC of one form, usually the new form, is as similar as possible to that of the other, usually the old form. To put it in another way, the characteristic curve methods find two equating coefficients that minimize the differences between the two TCCs.

The difference between the two characteristic curve methods lies in how they express the difference between the item characteristic curves.

Stocking and Lord (1983) expressed the difference between the characteristic curves as the square of the sum of the difference between the item characteristic curves for each item for examines of a particular ability θ. For a given θ_i , the sum of the squared difference over the common items (j:V) can be displayed in the following equation according to Kolen & Brennan (2004):

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}(\theta_{J_i}; \hat{a}_{J_j}, \hat{b}_{J_j}, \hat{c}_{J_j}) - \sum_{j:V} p_{ij}(\theta_{J_i}; \frac{\hat{a}_{I_j}}{A}, A\hat{b}_{I_j} + B, \hat{c}_{I_j}) \right]^2 \quad (2.20)$$

Then *SLdiff* is then cumulated to get $\sum_i SLdiff(\theta_i)$. The *SLdiff* can be cumulated in different ways. One is to sum over the examinees who are administered one form, usually the old form to get $\sum_i^N SLdiff(\theta_i)$, where N is the number of examinees who take the old form. Another is to sum over an arbitrary set of quadrature points along the ability scale

to get $\sum_{i}^{Q} SLdiff(\theta_i)$, where Q is the number of the quadrature points within the theta scale and θ_i is the theta value at the *i*-th quadrature point. The next and final step is to find the combination of A and B in equation (2.20) through a multivariate search that minimizes $\sum_{i} SLdiff(\theta_i)$.

In contrast to the Stocking-Lord method, Haebara (1980) expressed the difference between the characteristic curves as the sum of the squared difference between the item characteristic curves for each item for examinees of a particular ability (Kolen & Brennan, 2004). For a given ability θ_i, according to Kolen and Brennan (2004) the sum of the squared difference over the common items can be displayed :

$$Hdiff(\theta_i) = \sum_{j:V} \left[p_{ij}(\theta_{J_i}; \hat{a}_{J_j}, \hat{b}_{J_j}, \hat{c}_{J_j}) - p_{ij}(\theta_{J_i}; \frac{\hat{a}_{I_j}}{A}, A\hat{b}_{I_j} + B, \hat{c}_{I_j} \right]^2 \qquad (2.21)$$

The summation is over the common items (j:V). Then *Hdiff* is cumulated over all examinees to get $\sum_{i} HLdiff(\theta_i)$ in the same way as how $\sum_{i} SLdiff(\theta_i)$ is cumulated. The next and final step is to find the combination of A and B that minimizes $\sum_{i} HLdiff(\theta_i)$.

Equations 2.20 a nd 2.21 c onceptually describe how the characteristic curve methods work. As can be seen from the equations, the characteristic curve methods are more complicated than the moment methods. The mathematical procedure to find A and

B is also very complicated and computationally intensive. The application of the characteristic methods in practice often requires specially developed computer programs.

Characteristics of the Anchor Test

Equating designs have two key components: the design for data collection and the statistical model used to equate the score. As mentioned previously, the NEAT design provides great administrative flexibility and probably is the most prevalently used design in practice (Cook, 2007). However, this design is also the most difficult one to implement (Cook, 2007). In this design, the groups taking different tests or test forms are not considered to be equivalent. The central purpose of the anchor test is to separate group differences from form differences. However, studies have shown that different choices for anchor tests may have different equating results. Cook, Eignor, and Taft (1985) contrasted the equating results of using four different anchor tests to equate two forms of a biology test and the results of equating based on different sets of common items are not always the same. Cook and Petersen (1987) later discussed the results and pointed out that when the groups are similar in ability, the various anchor tests yield similar equating results and when the groups differ in level of ability, the different anchor tests yield very disparate equating results. So special care must be taken when selecting the set of common items constituting the anchor test (Cook & Petersen, 1987).

For this design to function well, a number of considerations have to be taken into account on how to choose the anchor test or common items, especially when the groups differ in level of ability. Some important considerations of the anchor test include content

representativeness, length of the anchor test and whether an item shows differential item functioning (DIF) or not. To reflect group differences accurately, the set of common items should be proportionally representative of the total test in content. That is, the common-item set should be a "mini-version" of the total test form (Kolen & Brennan, 2004).

Content Representativeness

Content representativeness means that anchor tests should be built to have the same specifications proportionally as the test itself. Klein and Jarjoura (1985) defined content representativeness as a match between anchor test and total test of the percentage of items in each of the several content areas. Whether the anchor items are representative of the overall items of tests being equated, in terms of content and statistical properties is especially important when groups vary in ability (Budescu, 1985; Cook & Pertersen, 1987). Budescu (1985) pointed out that the magnitude of the correlation between the anchor test and the unique components of each test form was the single most important determinant of the efficiency of the equating process. Brennan and Kolen (1987a) also pointed out that any substantial content change entailed a re-scaling of the test with a new "origin" form to which subsequent forms were equated. Klein and Jaroura (1985) evaluated the importance of the content representativeness of the anchor test and concluded that it was quite important to use content representative anchors with nonrandom groups and a failure to do so may lead to substantial equating errors. Cook and Petersen (1987) reported that inadequate content representation of the common-item

set creates especially serious problems when the examinee groups that take the alternate forms differ considerably in achievement.

Length of Anchor Test

Long tests are usually more reliable than short tests. So anchor tests have to be long enough to provide reliable measures so as to reflect group differences accurately. Too few common items could lead to equating problems (Petersen et al., 1983) while a larger number of common items results in less random equating error (Budescu, 1985; Wingersky et al., 1987). Klein and Kolen (1985) investigated the relationship between anchor-test length and accuracy of equating result and concluded that longer anchors did result in more accurate equating when the groups of examinees are dissimilar. However, long anchor tests also add to the burden of cost and time, especially when used as an external anchor.

There are no universal guidelines for selecting the length of the anchor. When considering the length of the anchor tests, Budescu (1985) suggested that each testing program needs to take into account the time, cost and context constraints as well as the particular index of efficiency when determining the length of the anchor test for its specific purposes; Kolen and Brennan (2004) pointed out that in constructing the common item sections, they should be long enough to represent test content adequately.

Although there is no absolute standard in regard to an appropriate length of an anchor test, a rule of thumb is given by Angoff (1984) that the appropriate number for anchor items should be at least 20 items or 20 percent of the total number of items in a test, whichever is larger. Kolen and Brennan (2004) recommended a similar rule of

thumb: the common item set should be at least 20 percent of a total test containing 40 or more items, unless the test is very long, in which case 30 common items might suffice.

Other numbers were also proposed as appropriate for the length of the anchor. Based on theoretical values of standard errors of item estimates, Wright (1997) considered 10 to 20 common items as sufficient for most equating situations and 10 common items may be adequate if the items are good. Mckinley and Reckase (1981) investigated effects of anchor test length on precision of the item parameter estimates and concluded that a 5-item anchor might be adequate but a 15-item anchor was suggested. Raju, Edwards, and Obsberg (1983) and Lord (1980) suggested that as few as five or six carefully chosen items could perform as satisfactory anchors in IRT equating when the items parameters of both tests were estimated in a single analysis using IRT concurrent methods. Hills, Subhiyah, and Hirsch (1988) studied the effects of anchor test length and found that five randomly chosen anchor items of a mathematics test were not sufficient to produce satisfactory equating results. An anchor of 10 items was found satisfactorily sufficient when an IRT method was adopted.

No DIF Items in the Anchor Test

Differential item functioning, or simply DIF, is present when individuals of the same ability but from different groups have different probabilities of success on a given item (Hambleton et al., 1991). Items showing DIF do not behave in the same manner across different groups. If test items operate in a different fashion, then the scores for different groups are not comparable. When the anchor test consists of DIF items, they tend to lower the reliability and validity of the anchor test and pose a serious threat in the

accurate detection of the group differences. So when an anchor test is used, differential item functioning analyses should be run to evaluate whether the items on the anchor test perform in the same way across different tests or test forms. In practice, the common items should be administered in approximately the same position across different test or test forms to avoid having the common items function differently across groups (Cook & Petersen, 1987).

Construct/Trait Equivalence of MC and CR Items

According to Frederiksen (1984), item format affects the meaning of the test scores by restricting the nature of the content and processes that can be measured. Tatsuoka (1991) thinks that multiple-choice items are suitable for measuring static knowledge. In the same line of research, researchers argue that multiple-choice assessments tend to encourage the teaching and learning of discrete facts and decontextualized procedures as well as rote memorization at the expense of deep conceptual understanding and the development of problem-solving skills (Resnick & Resnick, 1992; Shepard, 1991).

The limitation of MC items in measuring in-depth knowledge has prompted a search for alternatives to multiple-choice testing (Pollack, Rock, & Jenkins, 1992). CR items are thought to offer such an alternative. The primary motivation for the use of constructed-response formats thus stems from the idea that they can measure traits that cannot be tapped by multiple-choice items, for example, assessing dynamic cognitive processes (Bennett, Ward, Rock, & Lahart, 1990; Fiske, 1990; Fredericksen & Collins,

1989; Guthrie, 1984; Nickerson, 1989), identifying students' misconceptions in diagnostic testing (Birenbaum & Tatsuoka, 1987), and communicating to teachers and students the importance of practicing real-world tasks (Sebrechts, Bennett, & Rock, 1991).

A key assumption of IRT is that of unidimensionality, which assumes that there is only one underlying trait (latent trait) that decides the performance of examinees. To apply IRT models to mixed-format tests, MC items and CR items have to be calibrated together; a critical question to ask is: Do MC and CR items measure the same construct? Or to put it in a different way, is a mixed-format test unidimensional? The answer to the question affects the appropriateness of the use of unidimensional IRT models for mixed-format tests.

Messick (1993) regards the comparability of construct between multiple-choice and constructed-response items in a single test as an issue of trait equivalence and trait equivalence is an important part of evidence for construct validity. This psychometric issue has been raised in one form or another from almost the day that MC items were first used to test human subjects (Hogan, 1981; Traub & MacRury, 1990; cited in Traub, 1993). Many empirical studies were conducted to investigate this issue following different approaches. These studies can be categorized into two groups based on the methods used to examine trait equivalence. Some reported correlations between the two formats and some used factor analysis.

Correlational Studies

The correlational studies were based on the traditional psychometric framework, which examines whether or not different measures may be considered to be congeneric, that is, whether they have perfectly correlated true-scores (Joreskog, 1971; Lord, 1973 cited in Messick, 1993). In correlational studies, correlations between forms composed of MC items and forms composed of CR items are reported. Because measurement errors may affect the true-score correlations and different formats may differ in measurement precision, the reported correlations are usually corrected using the estimated reliability for attenuation due to measurement error as follows:

$$r_{xy.corrected} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \qquad (2.22)$$

Where r_{xy} is the estimated correlation between forms X and Y and r_{xx} and r_{yy} are the estimated reliability of forms X and Y respectively.

In his review of the related studies, Rodriguez (2003) pointed out three primary types of investigations regarding the construct equivalence of MC and CR items. The first type employs stem-equivalent items in both formats, where the CR items use the same stem as the MC items without the options (in some cases with minor editing to make the stem sensible). The second type employs content-equivalent items with independent stems, where these items in the two formats tap the same content and cognitive domains. The third type employs CR items that are qualitatively different than the MC items and they were explicitly written to tap a different aspect of the content domain or cognitive ability. Examples of such CR items include an essay item or an extended-response item.

The investigation of trait equivalence began with correlational studies. Rodriguez (2003) located a total number of twenty-nine correlational studies dated back to 1925. Of the twenty-nine studies, eleven used stem-equivalent forms. Six studies reported correlations at unity (Ackerman & Smith, 1988; Davis & Fifer, 1959; Frisbie & Cantor, 1995; Heim & Watts, 1967; Magill, 1934; Smith & Smith, 1984, cited in Rodriguez, 2003). The remaining five studies reported lower corrected correlations of between 0.78 and 0.95 (Hurd, 1932; Hurlbut, 1954; Ruch & Stoddard, 1925; Thiede, Klockars, & Hancock, 1991; Traub & Fisher, 1977, cited in Rodriguez, 2003).

Of the remaining eighteen studies, six used content-equivalent forms. These studies reported corrected correlations ranging between 0.68 and 0.98, mostly near 0.92 (DeMars, 1998; Hancock, 1992; Harke, Herron, & Leffler, 1972; Paterson,1926; Vernon, 1962; Ward, 1982, cited in Rodriguez, 2003).

The other twelve studies used essay items or CR items that were neither stem-equivalent nor content-equivalent. These studies reported corrected correlations ranging from 0.48 to 0.90 and mostly near 0.80 (Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990; Bracht & Hopkins, 1970; Breland & Gaynor, 1979; Coffman, 1966; Fisher, 1996; Godshalk, Swineford, & Coffman, 1966; Hogan & Mishler, 1980; Horn, 1966; Loyd & Steele, 1986; Manhart, 1996; Moss, Cole, & Khampalikit, 1982, cited in Rodriguez, 2003).

Although according to the definition of trait equivalence described by Traub (1993) that the true-score correlations of unity suggest trait equivalence, coefficients of correlation less than one often arise as a consequence of measurement error, so higher

correlations are often considered as evidence for trait equivalence. The review of correlational studies showed mixed results. The reported correlation coefficients range from as low as 0.4 to as high as unity. After a close examination of the methods used in different studies, Rodriguez (2003) found that the magnitude of the correlations is related to the item design characteristics. Where the items are more similar (particularly stem-equivalent), the resulting corrected correlations tend to be higher. Where the items in the two formats are more dissimilar (particularly with the use of essay-type items), the corrected correlations are lower. One limitation with the correlational studies is that MC items and CR items had to be treated as two separate tests by design in these studies for the calculation of correlation coefficients and therefore correlational studies could not examine whether MC items and CR items measure the same construct in a whole single test, that is, they could not provide direct evidence whether the mixed-format test is unidimensional.

Studies Using Factor Analysis

One way to examine whether MC and CR items measure the same construct simultaneously is to conduct factor analysis and the dimensionality of mixed-format tests has been examined by several researchers using factor analysis. In a study of the College Board's Advanced Placement Computer Science (APCS) examination, Bennett, Rock, and Wang (1991) posed a two-factor confirmatory factor analysis (CFA) model composed of multiple-choice and free-response (CR) factors to test the relationship of the skills measured by the multiple-choice and free-response items. They found that the MC items have high loadings on the MC factor and CR items have high loadings on the CR

factor and a two-factor model fit the data reasonably well. However, the two factors were highly correlated (r =.93 and .97 for two samples). So next they fit a one-factor model restricting all items to load on the same factor and found that little or no loss in fit occurred in moving from the two-factor model to the one-factor model. They concluded that one-factor model provided the most parsimonious model fit.

In their study of the Analytical scale of the Graduate Record Examination (GRE) General Test, Bridgeman and Rock (1993) took a similar approach to what Bennett et al. (1991) did by posing a two-factor CFA model to the data. The findings are also similar. MC items and CR items have higher loadings on two different factors respectively but the two factors have very high correlations (r = .93). They concluded that the open-ended version (CR items) does not appear to tap any significantly different new dimension and despite the apparent differences in the surface features of the tasks in the different formats, the underlying skills may be nearly identical.

In another study, Thissen, Wainer, and Wang (1994) examined the relationship between MC and CR sections of the computer science and chemistry tests of the College Board's Advanced Placement Program using an alternative approach. Instead of fitting a two-factor model to the data, they posed what they called a “general-plus-specific” model to the data. This model restricted all items to load on one general factor and the CR items to load on a separate factor. One advantage of this approach is that it makes it possible to compare the factor loadings of the CR items on different factors. After comparing the result of their “general-plus-specific” model to those of the two-factor model and one factor model by Bennett et al. (1991) for the APCS, Thissen et al. (1994) concluded that

their model fit the data considerably better and there are clearly free-response factors. However, they also found that although these free-response factors have loadings that are significantly different from zero, they account for relatively little of the observed covariance among the items. The loadings of the free-response items on the free-response factors are uniformly smaller than the loadings of the free-response items on the general factor. So their conclusion is that the free-response items measure the same proficiency as the multiple-choice items for the most part.

Ercikan and Schwarz (1995 cited in Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998) examined the relationship between MC and CR items in reading, mathematics and science tests in a state assessment program. These researchers found that the two-factor model, with separate factors for each item type, fit the data consistently better than the one-factor model. Although the two factors are highly correlated with correlation coefficients between 0.71 and 0.8 for the three tests, they concluded that CR items in this study measure a different construct than their MC counterparts in all three content areas

Pollock and Rock (1997) also examined the relationship between MC items and CR items in the mathematics and science tests in the High School Effectiveness Study (HSES) of the National Education Longitudinal Study of 1988 (NELS:88). They first conducted an exploratory factor analysis to examine the factor structure of the two tests. Two distinct factors were identified in each of the two tests, and were associated with the two different test formats, that is, all of the CR questions had high factor loadings on one factor, and all of the MC questions loaded on the other. Next, they tried to fit a two-factor

CFA model to the data to "statistically reproduce the results of the exploratory solutions" (p.37). They did not report how the model fit the data but reported that MC items had high loadings on the MC factor and CR items on the CR factor and the correlation between the two factors was .86. They concluded that although the correlation between the two factors is relatively high, it still is low enough to suggest that while they share much in common, the two formats still have some unique variance.

Tian (2009) examined the assumption of unidimensionality of the 12 booklets of the science and mathematics test of the Trends in International Mathematics and Science Study (TIMSS), which consists of both MC and CR items by fitting a one-factor CFA model to the data using samples from four different countries. The results are mixed. Some booklets in both content areas show a good model fit and some do not. The twelve booklets in both content areas are of different lengths. All the booklets where the unidimensionality assumption does not hold are those with fewer items (less than 30) in that subject. The assumption holds pretty well for booklets with more items (more than 50) in that subject.

Like the correlational studies reviewed previously, studies using factor analysis also showed mixed findings. It is especially interesting to see studies using the same method draw different conclusions. Bennett et al. (1991) and Ercikan & Schwarz (1995 cited in Ercikan et al., 1998) both fit a one-factor model and a two-factor CFA model to the data, but after comparing the two models Bennet et al. (1991) found that the one-factor model provides the most parsimonious fit and concluded that MC items and CR items measure the same construct while Ercikan & Shwarz (1995, cited in Ercikan et al.

1998) found that the two-factor model fit better than the one-factor model and concluded that CR items measure a different construct than their MC counterparts. Bridgeman & Rock (1993) and Pollock & Rock (1997) both fit a two-factor model to the data. In spite of the similar results that the two factors were highly correlated, they drew somewhat different conclusions.

Research Studies Comparing Different Equating Methods

Since the advances in computer technology permitted the application of IRT around late 1970s and early 1980s, interest in equating research has intensified (Skaggs & Lissitz (1986a). IRT equating has been researched and compared empirically in many studies. Skaggs and Lissitz (1986a) provided a comprehensive review of the early literature concerning IRT equating methods. Their review shows that the focus of many early equating studies was on t he comparison of effectiveness between IRT equating methods and conventional equating methods such as linear equating and equipercentile equating and on the development of new equating techniques.

One comparative study was conducted by Marco et al. (1983). In this study, the anchor test design was used. Linear equating, equipercentile equating and IRT characteristic curve equating using the one-parameter model and the three-parameter model were examined under a variety of conditions including random and dissimilar samples, the relative difficulty levels of the anchor test to the total tests and unequal difficulty of the two tests. When tests of equal difficulty were equated, the following conclusions were drawn from the study. When the anchor test was equal, the type of

sample mattered very little and all the methods performed well except one of the variations of equipercentile method. When the anchor test was easier or more difficult than the total tests, equating with random samples showed very little error. On the other hand, in the case of dissimilar samples, the IRT methods were greatly superior to the conventional methods regardless of which IRT model was used. When tests of unequal difficulty were equated, the IRT method using the three-parameter model was superior while the linear method and IRT method using the one-parameter model displayed larger total error.

Kolen (1981) also conducted a study comparing linear equating, equipercentile equating and the IRT mean and sigma equating method using the one-parameter model and the three-parameter model for tests of equal difficulty level and of different difficulty level using the random group design. The results showed that in general the IRT equating using the three-parameter model and the equipercentile equating performed well whether the tests to be equated were of similar or dissimilar difficulty. Linear equating performed well when the equating tests were of similar difficulty. IRT equating using the one-parameter model did not produce satisfactory results. Kolen (1981) attributed the inadequacy of the IRT equating using the one-parameter model to failure to account for guessing in the model. This conclusion is consistent with the findings from Skaggs and Lissitz (1986b) that the one-parameter model equating worked well when the same degree of chance scoring (guessing) was present in both tests and that when test discriminations or the degree of chance scoring were unequal between the two tests, the one-parameter model was inadequate.

Another comparative study was conducted by Petersen et al. (1983), in which different IRT equating methods were compared against the linear equating method and the equipercentile equating method for the SAT-Verbal and SAT-Mathematical tests. In this study, one form of the SAT was equated to itself through several intervening forms. Each equating in this chain was conducted using the anchor test design. Three IRT equating methods involved in this study were the concurrent calibration method, the Stocking- Lord method and what is referred to as the "fixed b's method". In the fixed b's method, the anchor tests were calibrated together with the first test. Then, the item parameter estimates for the anchor test were fixed to estimate the item parameters for other tests so that all the item parameters from different tests were automatically placed on the same scale. All the IRT equating methods were based on t he three-parameter model. Their results showed that all the three IRT equating methods produced substantially smaller errors than any of the conventional methods for the Verbal test and they performed similarly to the linear equating method for the Mathematics. The equipercentile equating method produced the worst results. This was mainly because no smoothing was used for this method in their study. Of the three IRT equating methods, the concurrent calibration method produced the most stable results overall.

Other studies that compared IRT equating methods and conventional equating methods include Cook, Dunbar, and Eignor (1981), Cowell (1981), Kolen and Whitney (1982), and Skaggs and Lissitz (1986b).

One general finding from many comparative studies is that IRT equating methods, especially those using the three-parameter model, perform as well as, if not better than,

the conventional methods. However, the superiority of IRT equating over conventional equating is not consistent. Violation of the unidimensionality assumption is a potentially major source of problems for IRT equating (Skaggs & Listz, 1986a). Also small sample size may lead to poor performances of IRT equating. This is because the application of IRT models often requires a l arge sample size to get stable parameter estimates. For example, Patience (1981) found that IRT equating was outperformed by equipercentile equating when equating tests of different difficulty with a sample size of 1,000 for each test. He suggested that small sample size might be a co ntributing factor because the estimation procedure had difficulty converging. Another study by Kolen and Whitney (1982) found that with small sample sizes (170-198), a number of extreme item parameter estimates were produced with the IRT model. This suggests a problem with the estimation procedure that contributed significantly to equating errors, a finding consistent with Patience's results. So Skaggs and Listz (1986a) do not recommend the use of the three-parameter model IRT equating with small sample sizes (less than 1,000 responses per item) because the equating results will not be accurate.

Studies were also conducted to compare different IRT equating methods. The focus of the studies was first on t he comparison between different methods for the dichotomous IRT models. One study (Ogasawara, 2000) compared the two moment methods using simulated data and real data and concluded that the mean/mean method produced linking coefficients with smaller standard errors and therefore was superior to the mean/sigma method. More studies focused on comparing characteristic curve methods to the moment methods. Baker and Al-Karni (1991) compared the Stocking-

Lord method and the mean/mean method using both simulated data and real data and found that the two methods produced similar linking coefficients in real data settings and the Stocking-Lord method recovered the true parameters with less error than the mean/mean method using simulated data. Ogasawara (2001a, 2001b) also compared the characteristic curve methods, which are referred to as the test response functions (TRFs) method in the studies and the mean/mean method using both simulated and real data and concluded that the former produced more stable results than the latter. Other comparative studies include Hanson and Beguin (2002), Kim and Cohen (1992) and Way and Tang (1991). A common finding from all these studies is that the characteristic curve methods (the Stocking-Lord method and the Haebara method) are more accurate than, and should be preferred over, the moment methods. Studies were also conducted to examine the performance of the Stocking-Lord method to that of the Haebara method and found that both methods produced very similar results (Way & Tang, 1991; Hanson & Beguin, 2002).

Concurrent calibration was also compared to the four equating methods following separate calibration. Petersen et al. (1983) compared the results using different IRT equating methods and concluded that concurrent calibration performed somewhat better than the Stocking-Lord method. Using simulation procedures in which the data fit the 3PL model, Kim and Cohen (1998) compared scale linking using the Stocking-Lord method to concurrent calibration but drew a somewhat different conclusion. For a small number of common items, Kim and Cohen (1998) found that the Stocking-Lord method produced more accurate results than concurrent calibration. Both methods yielded similar

results with a large number of common items. However, Hanson and Buguin (2002) pointed out one limitation of this study, that is, different IRT software programs were used for separate and concurrent estimation. Thus, the differences between separate calibration and concurrent calibration were confounded with the difference between computer programs. In order to provide further information concerning the relative performance of concurrent calibration versus equating/linking following separate calibration, Hanson and Buguin (2002) conducted another study. Also using simulation procedures, they compared the two moment methods, the two characteristic methods and concurrent calibration. In that study, the concurrent calibration procedures were found to produce more accurate results than the characteristic curve methods and as was found in many studies, the two moment methods did not perform as well as the characteristic methods and concurrent calibration.

The comparative studies were also extended to polytomous IRT models. Cohen and Kim (1998) conducted a study using simulation procedures to compare the two moment methods and the Stocking-Lord method under the graded response model. They concluded that the methods produced similar results. Kim and Cohen (2002) compared linking using the Stocking-Lord method and concurrent calibration for data that were simulated to fit the graded response model. They found that concurrent calibration was slightly more accurate.

Little research has been conducted that compares different equating methods for mixed-format tests. Kim and Lee (2004) conducted a simulation study to examine the performance of the characteristic methods and the moment methods when linking mixed-

format tests that consists of MC and CR items. In that study, the MC items were calibrated using the 3PL model and the CR items, the GPC model. They concluded that the characteristic curve methods are superior to the moment methods, which is consistent with the findings from previous studies that deal with single-format tests. No research has been done to investigate how the performance of the linking methods following separate calibration compares to the performance of concurrent calibration. The present study tries to fill the gap.

Overview of Research Methods in Comparative Studies of Equating Methods

One difficulty in comparing and evaluating different equating methods is that no definitive criterion exists, which makes it difficult to judge the relative merits of equating methods. Researchers have come up with different ways to overcome such difficulty. Three commonly used approaches are cross-validation or replication, circular equating and simulation.

In cross-validation or replication, researchers attempt to replicate or cross validate the results found in one sample in another sample. This approach was used in a number of studies conducted to compare different equating methods. For example, Kolen and Whitney (1982) used the cross-validation method in their comparison of four equating methods; Loyd and Hoover (1980) tried to equate several subtests from an achievement test battery using the Rasch model. To examine their results, they equated the same tests again using examinee groups of comparable abilities

When no t rue equating relationship is known, the cross-validation/replication method is quite useful because it provides some external criteria for comparing the results of the different equating methods. However, the qualities measured by cross-validation or replication are not sufficient though they are desirable in equating (Harris & Crouse, 1993). This is because this method mainly provides a measure of the stability of the results by different methods rather than a measure of the accuracy of the equating. Moreover, Lord and Wingersky (1984, cited in Harris & Crouse, 1993) have suggested that it is possible for inaccurate procedures to yield more stable results than accurate procedures. Thus, the results of studies using these criteria fail to address the question of which equating is better in a particular situation in terms of which equating is closer to the true equating relationship (Harris & Crouse, 1993).

In circular equating, the test under consideration is equated to itself, usually through a chain of intervening test forms. An example of circular equating would be equating Form A to Form B, Form B to Form C, and Form C back to Form A. In circular equating there could be a different number of intervening forms. The procedure always starts with the test under study, which serves as the initial scale. The focus is how closely the transformed scale from the circular chain of equating agrees with the initial scale. This approach has been used in multiple equating studies. This is the approach Petersen et al. (1983) took to compare the performance of a variety of IRT and conventional equating methods for a standardized verbal test. Cook and Eignor (1985) equated tests in a chain and compared the final and initial conversions in an investigation of the feasibility of using IRT methodology to equate achievement tests. Kingston and Holland

(1986) used equating in a circle to examine alternate ways of equating the Graduate Record Exam. Klein and Kolen (1985) equated a test to itself in an examination of the effect of the number of common items in common-item equating with nonrandom groups.

Circular equating has the advantage of having a known criterion against which to compare the accuracy of different equating methods. However, there is some doubt as to whether equating a test to itself can really accomplish this and the use of equating in a circle as a paradigm for comparing equating methods has been challenged (Harris & Crouse, 1993). Harris and Crouse (1993) reviewed related literature and found several problems with this approach. One problem is that equating methods involving the estimation of only one or two moments will frequently appear more accurate than methods involving the estimation of more moments (such as mean or linear equating compared to equipercentile equating). Brennan and Kolen (1987a) cautioned that the use of equating in a circle to compare equating methods be limited to those methods that estimate the same number of moments. However, in their study of three linear methods, Gafni and Melamed (1990) discovered an interaction between the type of linear method used and the paradigm used to evaluate them. Based on their findings, Gafni and Melamed (1990) questioned the use of the circular paradigm even for comparing methods with the same number of estimated moments. Another problem with this method discussed by Brennan and Kolen (1987b) is that the results obtained may be dependent on the form one chooses to start and end on. In their article they provide an example of this phenomenon, in which different results are obtained when the equating begins and ends on one form rather than on another.

Simulation or using simulated/generated data is a technique often used in equating studies. This approach has a great advantage over the other two approaches, that is, when using simulated data, one knows the true equating relationship he/she is trying to recover and therefore, there exists a definitive criterion against which to compare different equating methods. Simulation has been frequently used in equating studies. For examples, Baker and Al-Karni (1991) used both real and simulated data for computing IRT-equating coefficients to compare two IRT equating methods; Stocking, Eignor, and Cook (1988) examined factors affecting the invariant properties of four linear and curvilinear equating procedures using simulations; also Way and Tang (1991) compared four different equating methods using both real and simulated data. Other studies using this approach include Kim and Cohen (1998, 2002), Ogasawara (2000, 2001a, 2001b), and Hanson and Beguin (2002). However, the use of simulation is not without problems. First, there is also a question of bias accruing to the data-- that is, if the data are generated from a specific IRT model, will that bring bias to the results if the same model is also used in one of the equating methods being compared (Hwang & Cleary, 1986); also, the generated data may not closely resemble the real data to which the study results will be applied. Harris and Crouse (1993) suggested that using simulated data is most useful when the data closely resembles the real data so that results of the simulation studies could be generalized to real data equating situations and when it is accompanied by analyses involving the real data.

Summary

This chapter first presented an introduction of the traditional equating methods, introduction of the Item Response Theory (IRT), its assumptions and IRT based equating methods and an introduction of the characteristics of the anchor test. Then it reviewed related literature from three perspectives.

First, do MC items and CR items measure the same construct or does the unidimensionality assumption of IRT hold for mixed-format tests? The review showed mixed findings in regard to the construct equivalence of the MC and CR items. Some concluded that MC and CR items measure the same construct and some concluded that they do not measure the same construct.

Second, how do different equating methods compare? The review showed that in general the IRT equating methods perform as well as, if not better than, the conventional methods when the IRT assumptions hold and the sample size is sufficient and that in general the two characteristic curve methods perform similarly and better than the two moment methods and the concurrent calibration method performs slightly better than the characteristic curve methods.

Lastly, what are the methods used in comparative studies of equating methods? The review showed that one widely used technique in those comparative studies is simulation. The full research design for this study is presented in Chapter Three.

Chapter Three: Methodology

As reviewed in the previous chapter, there are three commonly used approaches in equating studies: cross-validation or replication, circular equating and simulation, each with its shortcomings. However, simulation is the technique most frequently used in the IRT equating studies that compared linking following separate calibration versus concurrent calibration and that compared different linking methods for mixed formats. For this reason, this study also used simulated data so that the true equating relationship is known.

Instrument

To mimic a real data situation as Harris and Crouse (1993) suggested, the simulated examinee responses were based on actual item parameter estimates obtained from the TIMSS 2003 8th grade Mathematics assessment, which are available from the TIMSS 2003 Technical Report (Martin, Mullis, Gonzalez, & Chrostowski, 2004).

A Brief Introduction of TIMSS

TIMSS, or Trends in International Mathematics and Science Study, a project of the International Association for the Evaluation of Educational Achievement (IEA), is the largest and longest running international comparative education study in mathematics and science. It is conducted on a 4-year cycle, with the first data collection in 1995. TIMSS 2003, the third data collection in the TIMSS cycle of studies, was administered at the fourth and eighth grade in 49 countries, of which 26 participated at the fourth grade and 46 at the eighth grade.

The TIMSS 2003 Framework (Mullis, Martin, Smith, Garden, Gregory, Gonzalez, Chrostowski, & O'Connor, 2003) at the eighth grade categorized mathematics items into five content domains, which are respectively, number, algebra, measurement, geometry, and data. There were 194 items in mathematics. Of the 194 mathematics items, 66 were CR items, requiring students to generate and write their own answers, and the rest were MC items with 4 to 5 options. There were two types of CR items: short-answer questions and extended-response questions. In scoring the items, correct answers to the MC and short-answer questions were worth one point. Responses to extended response questions were evaluated for partial credit, with a fully correct answer being awarded two points and a partially correct answer being awarded one point. A summary of the TIMSS 2003 mathematics items by item format and content domain is shown in Table 3.1.

Table 3.1 Summary of Item Characteristics TIMSS 2003 Mathematics, Grade 8

		Number of Items
Item Format	Multiple Choice	128
	Constructed-response	66
Content domain	Algebra	47
	Data	28
	Geometry	31
	Measurement	31
	Number	57

In TIMSS 2003, a family of IRT models was used in the scaling procedure. For MC items, the three-parameter model was used. For CR items, two parameter model was used. For extended response items, the GPC model was used; but for short answer items, the two-parameter model was used instead of a typically used polytomous IRT model for

constructed-response items, because the short answer items had only two response options and were scored as correct or incorrect in TIMSS 2003.

In addition to the 194 mathematics items, there is also a total of 189 science items in TIMSS 2003 at 8th grade. It is impossible for every student to complete all the items, because it would require much more testing time than could be allotted for individual students. Therefore, TIMSS 2003 used a matrix-sampling technique that involved dividing the entire assessment item pool into a set of unique item blocks, distributing these blocks across a set of booklets, and rotating the booklets among the students. Each student took just one booklet. The TIMSS design for 2003 divided the 194 mathematics items at eighth grade into 14 item blocks. Each block contained 12-15 content-balanced mathematics items.

The Instrument in This Study

For this study, a total of 70 items were selected from the 194 items including 55 MC items and 15 extended response items to create under different simulated conditions two mixed-format test forms: Form A and Form B. At the 8th grade, the usual number of items administered in a mathematics test is around 50. For example, in the Florida Comprehensive Assessment Test (FCAT), each student answered 50 mathematics questions; in the Massachusetts Comprehensive Assessment System (MCAS), each student answered 49 mathematics questions; in the New Jersey Assessment of Skills and Knowledge (NJASK), each student answered 46 mathematics questions; and in the New York State Testing Program (NYSTP), each student answered 45 mathematics questions. So in this study, each form consisted of 50 items. In the design of a mixed-format test, the

balance of item formats is important. One way to balance the item formats is to divide the testing time evenly between the MC items and CR items as is done in the National Assessment of Educational Progress (NAEP). However, as CR items usually require students to consider a situation that requires more than a numerical response or a short verbal communication and the student is expected to take more time to complete them than the MC items, the number of CR items is often less than that of MC items. CR items usually account for about 20 to 40 percent of the total number of the items. For example, about 24% (11 out of 46) of the items are CR items in the NJASK, about 27% (13 out of 49) in the MCAS, 40% (20 out of 50) in the FCAT, and 40% in the NYSTP. In the present study, 30 percent (15 out of 50) of the total items were CR items. Special care was taken in the formation of the two forms to balance the content in each form such that the content and statistical characteristics of the three sets were as similar as possible. Both forms have the same item difficulty range between -1.3 and 2.207 and both forms are of the same difficulty level, with a same mean difficulty of 0.26. Obviously the two forms are actually parallel forms and this study is an equating study. Since there is not so much distinction between equating and linking in terms of the data collection design and statistical procedure, it is believed that the results from this study may also apply to other forms of linking using the same data collection design and the same methods.

A summary of items characteristics for both forms is presented in Table 3.2.

Table 3.2 Item Characteristics Used for Simulating Two Mixed-format Test Forms

Item #	domain	format	slope	location	guessing	step 1	step 2	Forms	
								A	B
1	algebra	CR	0.465	-0.402		-2.368	2.368	√	√
2	algebra	CR	0.413	-0.107		-2.75	2.75	√	√
3	algebra	CR	0.530	0.021		-0.861	0.861	√	√

4	algebra	CR	0.772	0.554		-1.484	1.484	√	√
5	algebra	CR	0.706	1.001		-2.048	2.048	√	√
6	algebra	CR	1.297	1.131		-0.134	0.134	√	√
7	data	CR	1.066	0.648		-0.25	0.25	√	√
8	data	CR	1.089	0.820		-0.016	0.016	√	√
9	data	CR	0.842	1.460		-0.313	0.313	√	√
10	data	CR	0.839	1.590		-0.694	0.694	√	√
11	geometry	CR	0.687	0.022		-1.006	1.006	√	√
12	geometry	CR	0.499	2.207		-1.288	1.288	√	√
13	measurement	CR	0.742	0.612		-0.632	0.632	√	√
14	measurement	CR	0.521	1.498		-2.52	2.52	√	√
15	number	CR	0.815	0.925		-1.549	1.549	√	√
16	algebra	MC	0.585	-1.304	0.001			√	√
17	algebra	MC	0.706	-0.611	0.106			√	√
18	algebra	MC	1.116	0.037	0.132			√	√
19	data	MC	0.583	-0.998	0.073			√	√
20	data	MC	0.562	0.600	0.135			√	√
21	data	MC	0.701	0.702	0.044			√	√
22	geometry	MC	0.662	-0.315	0.171			√	√
23	geometry	MC	1.108	0.260	0.167			√	√
24	geometry	MC	1.244	0.370	0.185			√	√
25	measurement	MC	0.983	-0.299	0.355			√	√
26	measurement	MC	0.911	-0.197	0.073			√	√
27	measurement	MC	1.275	0.491	0.137			√	√
28	number	MC	1.013	-0.498	0.154			√	√
29	number	MC	0.819	0.181	0.173			√	√
30	number	MC	1.236	0.945	0.157			√	√
31	algebra	MC	0.858	0.050	0.233			√	
32	algebra	MC	0.841	0.831	0.102			√	
33	data	MC	1.064	-0.838	0.278			√	
34	data	MC	0.822	-0.768	0.104			√	
35	data	MC	0.721	-0.549	0.183			√	
36	data	MC	1.334	0.366	0.226			√	
37	geometry	MC	0.558	-0.121	0.288			√	
38	geometry	MC	1.089	0.041	0.129			√	
39	geometry	MC	0.555	0.540	0.086			√	
40	geometry	MC	0.978	0.552	0.010			√	
41	measurement	MC	0.979	-0.382	0.086			√	
42	number	MC	0.615	-0.839	0.086			√	
43	number	MC	0.833	-0.674	0.086			√	
44	number	MC	0.405	-0.538	0.184			√	

45	number	MC	1.330	-0.234	0.053			√	
46	number	MC	1.046	-0.112	0.122			√	
47	number	MC	0.407	0.018	0.000			√	
48	number	MC	1.466	0.030	0.154			√	
49	number	MC	0.902	0.046	0.139			√	
50	number	MC	1.434	0.088	0.334			√	
51	algebra	MC	1.217	-0.357	0.082				√
52	algebra	MC	0.733	0.077	0.127				√
53	algebra	MC	1.627	1.121	0.142				√
54	algebra	MC	0.826	1.272	0.158				√
55	data	MC	0.864	-0.777	0.205				√
56	data	MC	1.163	-0.020	0.285				√
57	geometry	MC	0.868	-0.545	0.058				√
58	geometry	MC	0.830	0.224	0.182				√
59	geometry	MC	0.986	1.078	0.204				√
60	measurement	MC	1.231	0.136	0.193				√
61	measurement	MC	1.579	0.387	0.141				√
62	measurement	MC	0.755	0.830	0.241				√
63	measurement	MC	1.060	1.029	0.275				√
64	number	MC	0.790	-0.596	0.118				√
65	number	MC	0.801	-0.272	0.202				√
66	number	MC	1.133	0.064	0.170				√
67	number	MC	1.235	0.231	0.243				√
68	number	MC	1.440	0.462	0.289				√
69	number	MC	1.352	0.510	0.211				√
70	number	MC	1.330	1.157	0.182				√

Factors investigated

Three factors are investigated in this study: (a) two types of linking, (b) three types of length for the common items, and (c) three types of common-item set. There are a total of 18 conditions studied (2 × 3 ×3).

Two Types of Linking: Concurrent Calibration versus the Stocking-Lord Method

This study focuses on the comparison between the concurrent calibration method and the Stocking-Lord method following separate calibration. This is for two reasons.

First, a general finding from related comparative studies shows that among the four different IRT linking methods following separate calibration, the two characteristic curve methods, namely the Stocking-Lord method and the Haebara method perform better than the two moment methods, namely the mean/mean and the mean/sigma methods; second, the two characteristic methods perform very similarly but the Stocking-Lord method is used more widely. So this study focuses on the comparison between the concurrent calibration method and the Stocking-Lord method for the sake of simplicity and efficiency.

Three Different Lengths of the Common Items: 5, 10 and 15 Common Items

Long tests are usually more reliable than short tests and anchor tests have to be long enough to provide a reliable measure and reflect group differences accurately. However, long anchor tests also add to the burden of cost and time, especially when used as external anchor. Although there is no absolute standard in regard to an appropriate length of an anchor test, a rule of thumb is given by Angoff (1984) and also recommended by Kolen and Brennan (2004) that the appropriate number for anchor items should be at least 20 items or 20 percent of the total number of items in a test. This study compares the performance of the concurrent calibration method and the Stocking-Lord method when the common items are of three different lengths: (1) less than 20% (5 out of 50) of the total items, (2) 20 % (10 out of 50) of the total items, and (3) more than 20% (15 out of 50) of the total items. Hopefully, the findings of this study can provide some evidence for an appropriate length of an anchor test.

Three Types of Common-item Set: Both MC & CR Items, MC Items Only and CR Items Only

Since mixed-format consists of only two types of items, anchor items can be selected in three possible ways: (1) both MC and CR items, (2) MC items only, and (3) CR items only. This study compares the performance of the concurrent calibration method and the Stocking-Lord method under the three different types of common items. The inclusion of both types of items is because when two mixed-format test forms have both MC and CR items in common, it is considered ideal to have both types of items included in the common items because they better represent the total test in content and characteristics (Kim & Lee, 2004). However, linking through MC items only is often chosen for practical reasons such as concerns about CR items in terms of reliability, security, and rater drift (Kim & Lee, 2004). The use of only CR items in linking is rare in practice, and it is included in this study only for comparative purposes. It is hoped that the findings from this study can provide some guidance in the selection of the appropriate types of common items. A summary of the characteristics of the common items is presented in Table 3.3.

Table 3.3 Characteristic of Common Items.

Item #	domain	format	slope	location	guessing	Number and type of common items								
						5 common items			10 common items			15 common items		
						MC & CR	MC only	CR only	MC & CR	MC only	CR only	MC & CR	MC only	CR only
1	algebra	CR	0.465	-0.402				√	√		√			√
2	algebra	CR	0.413	-0.107										√
3	algebra	CR	0.530	0.021							√	√		√
4	algebra	CR	0.772	0.554							√			√
5	algebra	CR	0.706	1.001										√
6	algebra	CR	1.297	1.131							√	√		√
7	data	CR	1.066	0.648				√	√		√			√
8	data	CR	1.089	0.820							√			√
9	data	CR	0.842	1.460								√		√
10	data	CR	0.839	1.590							√			√
11	geometry	CR	0.687	0.022										√
12	geometry	CR	0.499	2.207		√		√	√		√	√		√
13	measurement	CR	0.742	0.612										√
14	measurement	CR	0.521	1.498				√			√			√
15	number	CR	0.815	0.925				√			√	√		√
16	algebra	MC	0.585	-1.304	0.001	√	√		√	√			√	
17	algebra	MC	0.706	-0.611	0.106				√			√	√	
18	algebra	MC	1.116	0.037	0.132								√	
19	data	MC	0.583	-0.998	0.073					√		√	√	
20	data	MC	0.562	0.600	0.135							√	√	
21	data	MC	0.701	0.702	0.044	√	√		√	√		√	√	
22	geometry	MC	0.662	-0.315	0.171								√	
23	geometry	MC	1.108	0.260	0.167		√			√		√	√	
24	geometry	MC	1.244	0.370	0.185				√			√	√	
25	measurement	MC	0.983	-0.299	0.355	√	√		√	√			√	
26	measurement	MC	0.911	-0.197	0.073					√		√	√	
27	measurement	MC	1.275	0.491	0.137					√		√	√	
28	number	MC	1.013	-0.498	0.154					√			√	
29	number	MC	0.819	0.181	0.173	√			√	√		√	√	
30	number	MC	1.236	0.945	0.157		√		√	√		√	√	

Data Generation

Harris and Crouse (1993) suggested that using simulated data is most useful when it is accompanied by analyses involving the real data. However, in TIMSS 2003, all items were calibrated using a combined sample from participating countries and the calibration data are not available. This makes analyses involving the real data impossible, so all analyses in this study are based on the simulated data. The data generation procedure is described in detail in the following section.

The item parameters of the selected MC items and CR items listed in Table 3.3 were treated as the true item parameters for simulating data. Two sets of person parameters were also generated from a normal distribution as the true person parameters: for Form A with a mean of 0 and standard deviation of 1 and for Form B with a mean of 1 and a standard deviation of 1 so that they were non-equivalent groups. A dichotomous MC item response (U_{ij}) of an individual on one item was generated by comparing a random number (R) from the uniform distribution in the range between 0 and 1 to the probability of getting the item (P_{ij}) right computed based on the unidimensional three-parameter model. If $R \leq P_{ij}$, then $U_{ij}=1$; otherwise, $U_{ij}=0$. A polytomous CR item response (U_{ijk}, where k =1, 2, 3, is the number of category of the item responses) of an individual on one item was generated in a similar way. In this case, a random number (R) from the uniform distribution in the range between 0 and 1 was compared to the probability of getting each response category (P_{ijk}) computed based on the unidimensional GPC model. If $P_{ij(k-1)} < R \leq P_{ijk}$, then $U_{ijk}=k-1$, where $P_{ij0}=0$.

For each of 18 conditions investigated in this study, 50 s ets of samples with sample size of 10,000 will be generated using Wingen 2 (Han, 2007),

Evaluation Criteria

The results are evaluated from two aspects: the recovery of the item parameters after equating and how close the estimated person parameters are to the true person parameters.

Recovery of the Item Parameters after Equating

In each condition to be studied, there are 50 sets of item parameter estimates for Form A and corresponding item parameter estimates for Form B. The Form B item parameter estimates should be on the same metric as the known true parameters after a transformation in each of the replications. The closer the Form B parameter estimates are to the true parameters, the more accurate the transformation. The results were compared in terms of the equating errors and the stability of item parameter recovery. The magnitude of equating errors was evaluated by calculating the commonly used mean squared error (MSE) (Kim & Lee, 2004; Kim & Cohen, 1998; Hanson & Béguin, 2002; Yao & Boughton, 2009). Let f_{true} be the true parameter and let f_{ij} be one estimated parameter of item i from sample j,

$$MSE = \frac{1}{n}\frac{1}{m}\Sigma_{j=1}^{n}\Sigma_{i=1}^{m}(f_{ij} - f_{true})^2 \qquad (3.1)$$

where m is the number of items and n is the number of replications.

The standard deviation (SD) of the recovered item parameters across replications is a measure of the stability of the item recovery. It is calculated as

$$SD = \sqrt{\frac{1}{nm-1}\Sigma_{j=1}^{n}\Sigma_{i=1}^{m}(f_{ij} - \bar{f}_i)^2} \quad (3.2)$$

Where $\bar{f}$ is the average parameter estimates from all replications, and

$$\bar{f}_i = \frac{1}{n}\sum_{j=1}^{n} f_{ij} \quad (3.3)$$

The MSE and SD were computed separately for the discrimination parameter *a*, difficulty parameter *b* and guessing parameter *c*. The smaller the MSEs, the better the method in recovering the parameters; the smaller the SDs, the more stable the method in recovering the parameters.

The Difference between the Estimated Person Parameter sand the "True" Person Parameters

Since it is possible that a method may perform better in recovering one type of item parameter than another, it is more helpful to use a single index in evaluating the overall quality of the two equating methods. Also because the purpose of equating is to put scores from different metrics onto one metric, it is of greater importance to evaluate the quality of the different equating methods by examining how close the estimated person or ability parameters, which often appear in some form of test scores, are to the true person parameters. The root mean squared difference (RMSD) is a measure of the average difference between the estimated person parameters and the true person parameters for each examinee across all replications. It is calculated as:

$$RMSD = \sqrt{\frac{1}{n}\frac{1}{N}\Sigma_{j=1}^{n}\Sigma_{i=1}^{N}(\theta_{ij} - \theta_{trueij})^2} \quad (3.4)$$

Where N is the number of examinees, θ_{ij} is the estimated ability parameter of person *i* on Form B in the *n*th replication based on the equated item parameter estimates, and θ_{trueij} is the true ability parameter of person *i* on Form B in the *n*th replication based on the true parameters used to generate the data. The latter is treated as the true person parameter for the examinees because it is based on the true parameters. As equations 3.4 shows, the difference between the estimated person parameters and true person parameters are first summed up across all cases in the sample and all replications and then averaged out. So its magnitude is not related to the sample size and large sample size makes the calculation of RMSD more stable. One property of the RMSD is that it is measured at the same unit as the data, so it is a criterion not only to evaluate the relative performances of the two equating methods but also to evaluate the accuracy of the two equating methods independently.

The MSE, SD and RMSD were computed for each of the 18 conditions as the criteria to answer the question: which IRT equating method performs better under different conditions for the non-equivalent common item design: the characteristic curve method following separate calibration or the concurrent calibration method.

Data Analysis Procedure

Step 1. Concurrent calibration. Data from Form A and Form B are merged together for each replication with different sets of common items as listed in Table 3.3 to get 50 different sets of combined samples. Then the item parameters for the items on the two forms were estimated for each replication;

Step 2. Separate calibration. The item parameters for the items on the two forms were estimated separately for Form A and Form B for each replication;

Step 3. Linking Form B scale to Form A scale. The item parameter estimates for Form B were put on the scale of the item parameters for Form A following the Stocking-Lord method using the computer program IRTEQ (Han, 2009) for each replication; and

Step 4. Generation of true person parameters and estimates person parameters for Form B. The true person parameters were generated based on the parameters used to generate the data and the estimated person parameters were generated based on parameter estimates derived from the two equating methods.

All item calibration and score generation were conducted using the computer program IRT Command Language (ICL) (Hanson, 2002) since ICL can handle multiple group analysis. Sample ICL syntaxes are presented in Appendix A.

Summary

This Chapter outlines the complete analysis plan to address the research question posed in Chapter One. To compare the performance of the Stocking-Lord method and the concurrent calibration method with the mixed-format tests, the commonly used simulation technique was used. 70 MC and CR items from TIMSS 2003 mathematics assessment were used to form two 50-item mixed-format tests with 35 MC items and 15 CR items. The simulation study was conducted using simulated unidimensional data based on the parameters of these items, which were treated as the true parameters. The two test forms were equated using both the concurrent calibration method and the

Stocking-Lord method. Their performances were evaluated in two ways: the accuracy and stability of the item parameter recovery and how close the estimated person parameters were to the "true" person parameters based on the true parameters. Results of the above outlined analyses are presented in Chapter Four.

Chapter Four: results

This chapter presents the results from the simulation study described in Chapter Three. The performances of two equating approaches, the concurrent calibration method and the Stocking-Lord method, were investigated under different conditions. Their performances were evaluated based on the recovery of the item parameters and the difference between the estimated person parameters and the true person parameters. The present chapter is divided into two main sections: (a) the recovery of item parameters after equating, and (b) the differences between the estimated person parameters and the true person parameters.

Recovery of Item Parameters

Two criteria were used in the evaluation of the recovery of item parameters: the mean squared error (MSE) between the estimated and true item parameters as calculated in equation 3.1 measures the magnitude of the equating errors and the standard deviation (SD) of the recovered item parameters as calculated in equation 3.2 measures the stability of the item recovery. The recovery of item parameters was evaluated separately for the discrimination parameter *a*, the difficulty parameter *b* and the guessing parameter *c*.

Recovery of the Discrimination Parameter a

The MSEs and SDs for the recovery of the *a* parameters after equating for all 18 conditions including two equating methods (the concurrent calibration and the Stocking-Lord method), three different lengths of common items (5, 10 and 15) and three types of

common items (MC items only, both MC & CR items and CR items only) are summarized in Tables 4.1 and 4.2 and are graphically depicted in Figures 4.1 and 4.2.

Table 4.1 MSEs of the *a* Parameter

	5 common items			10 common items			15 common items		
	MC	Both	CR	MC	Both	CR	MC	Both	CR
CC	0.1267	0.1249	0.1212	0.0991	0.0871	0.0847	0.0912	0.0852	0.0836
SL	0.1275	0.1268	0.1268	0.1079	0.0981	0.0973	0.1027	0.0868	0.0852

Note: 1. CC - Concurrent calibration, SL - the Stocking-Lord method
2. MC - multiple-choice items, CR - constructed response item, Both - multiple-choice and constructed response items

Table 4.2 SDs of the *a* parameter

	5 common items			10 common items			15 common items		
	MC	Both	CR	MC	Both	CR	MC	Both	CR
CC	0.0102	0.0101	0.0102	0.0098	0.0096	0.0097	0.0095	0.0095	0.0092
SL	0.0126	0.0125	0.0102	0.0118	0.0106	0.0100	0.0111	0.0101	0.0099

Note: 1. CC - Concurrent calibration, SL - the Stocking-Lord method
2. MC - multiple-choice items, CR - constructed response item, Both - multiple-choice and constructed response items

Figure 4.1 MSEs of the *a* Parameter

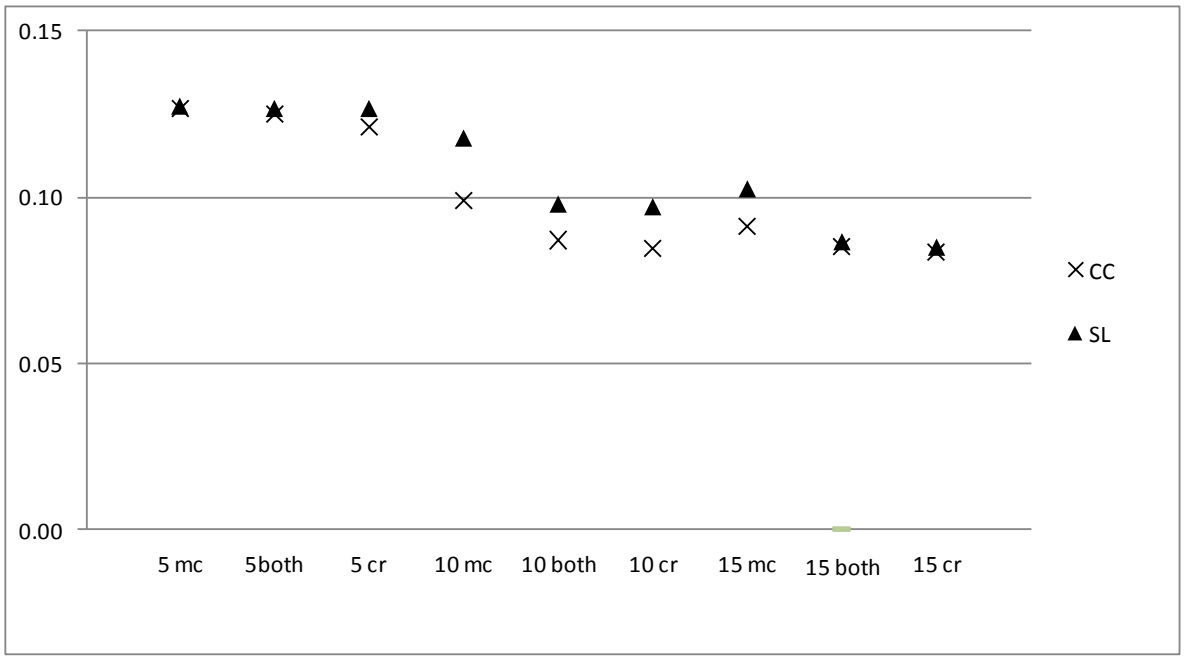
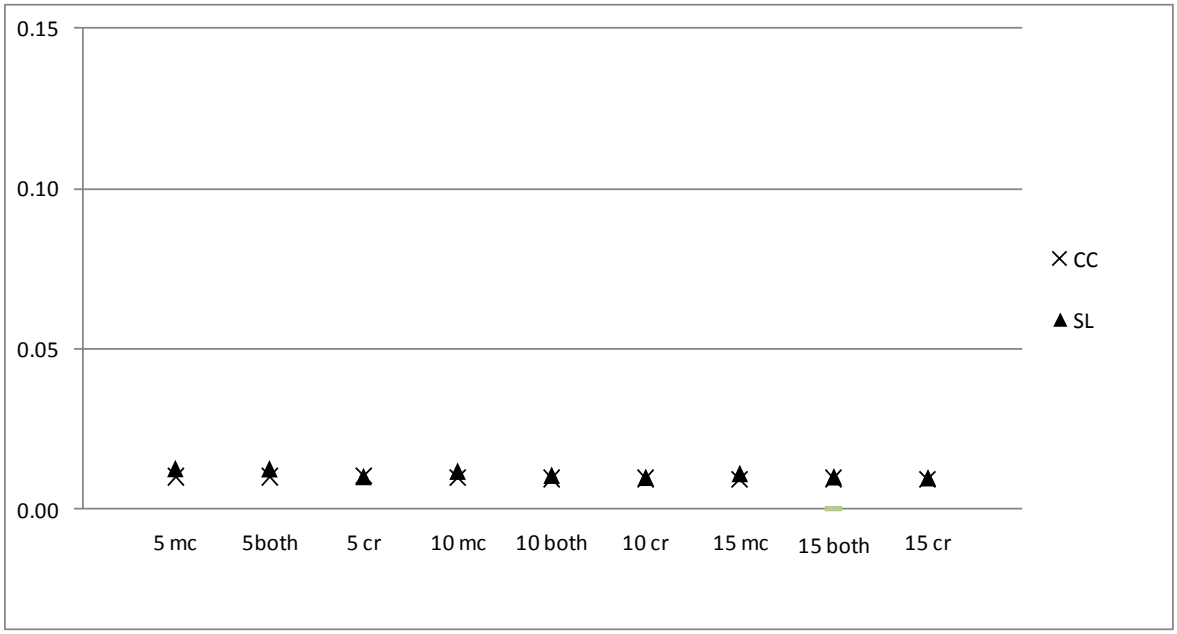


Figure 4.2 SDs of the *a* Parameter



From Tables 4.1 & 4.2 and Figures 4.1 & 4.2, the major findings are summarized as follows:

1). In general, the concurrent calibration method has smaller MSEs than the Stocking-Lord method in recovering the *a* parameter. The difference of the MSEs between the two methods is largest when the common items are all MC items; it is smaller when the common items consists both types of items; and it is smallest and remains almost the same when the common items are all CR items.

2). In general, the length of the common-item set impacts the recovery of the *a* parameter in a consistent way for both methods. The MSE tends to decrease as the length of the common-item set increases for the same type of common items.

3). The type of the common-item set impacts the recovery of the *a* parameter in a consistent way. For both the concurrent calibration method and the Stocking-Lord

method, when the common items are all CR items, the recovery of *a* parameter has the smallest MSEs for each of the three different lengths of the common-item sets and when the common items are all MC items, the recovery has the largest MSEs for each of the three different lengths of the common-item set. So the MSE tends to decrease as the number of CR items increases in the common-items set when the number of common items is the same.

4). The SDs of the recovered *a* parameter across all replications are very low, ranging from .0092 to 0.0126. The differences between the two methods are at the third or even the fourth decimal place, so the SDs remain very close for all 18 conditions as shown graphically in Figure 4.2.

Recovery of the Difficulty Parameter b

The MSEs and SDs for the recovery of the *b* parameters after equating for all 18 conditions are summarized in Tables 4.3 and 4.4 and are graphically depicted in Figures 4.3 and 4.4.

Table 4.3 MSEs of the *b* Parameter

	5 common items			10 common items			15 common items		
	MC	Both	CR	MC	Both	CR	MC	Both	CR
CC	0.0843	0.0668	0.0646	0.0837	0.0651	0.0606	0.0811	0.0656	0.0545
SL	0.1025	0.0994	0.0837	0.0981	0.0885	0.0832	0.0972	0.0867	0.0833

Note: 1. CC - Concurrent calibration, SL - the Stocking-Lord method
2. MC - multiple-choice items, CR - constructed response item, Both - multiple-choice and constructed response items

Table 4.4 SDs of the *b* Parameter

	5 common items			10 common items			15 common items		
	MC	Both	CR	MC	Both	CR	MC	Both	CR
CC	0.0482	0.0449	0.0446	0.0480	0.0414	0.0427	0.0475	0.0433	0.0386
SL	0.0538	0.0521	0.0484	0.0531	0.0492	0.0481	0.0520	0.0492	0.0482

Note: 1. CC - Concurrent calibration, SL - the Stocking-Lord method

2. MC - multiple-choice items, CR - constructed response item, Both - multiple-choice and constructed response items

Figure 4.3 MSEs of the *b* Parameter

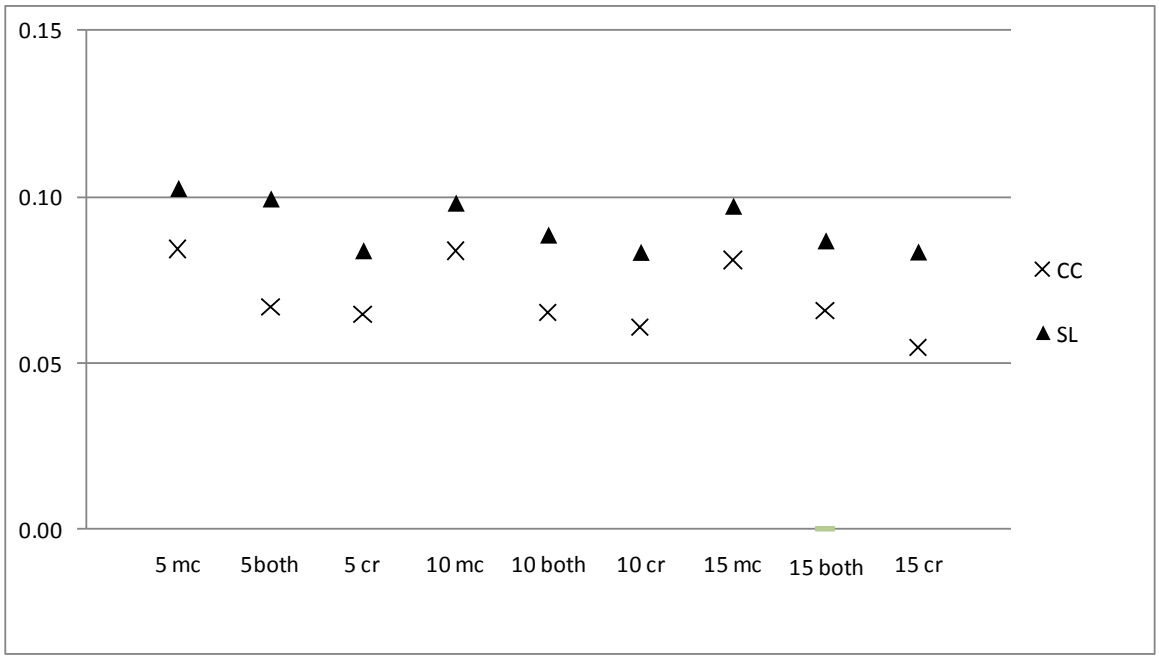
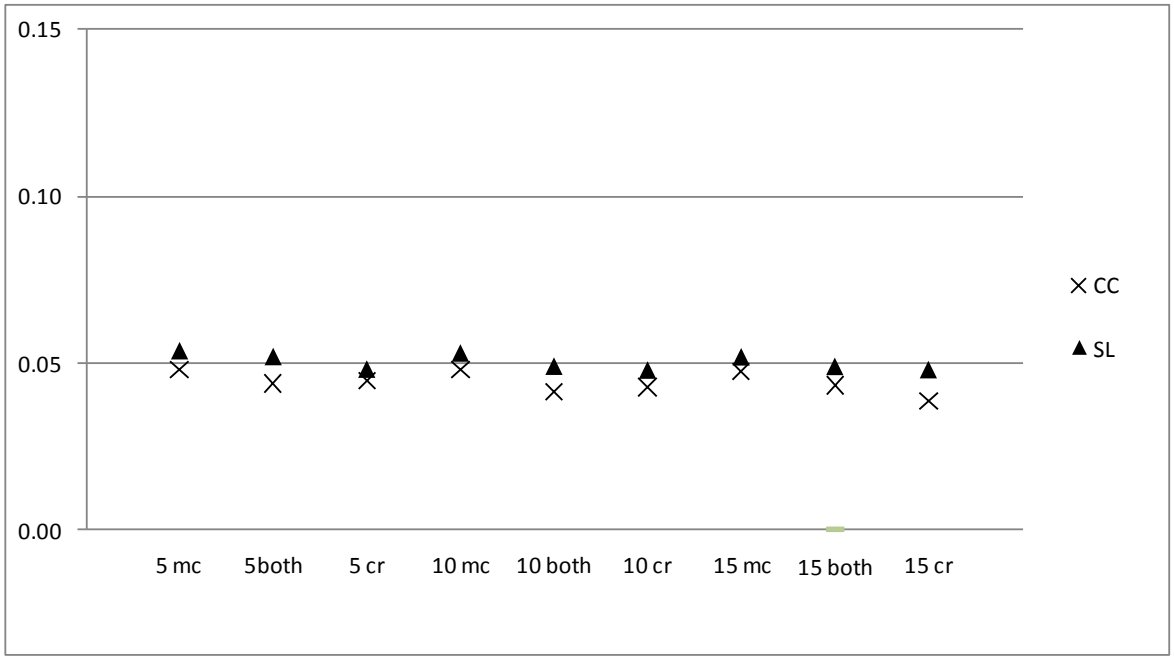


Figure 4.4 SDs of the *b* Parameter



From Tables 4.3 & 4.4 and Figures 4.3 & 4.4, the major findings are summarized as follows:

1). When the number and the type of the common items are the same, the concurrent calibration method consistently has smaller MSEs than the Stocking-Lord method in recovering the *b* parameter.

2). In general, the length of the common-item set impacts the recovery of the *b* parameter in a consistent way for both methods. The MSE tends to decrease as the length of the common-item set increases. For the Stocking-Lord method, however, when the common items are all CR items, the MSEs remain almost the same when the number of common items is 10 (20%) or more.

3). The type of the common items impacts the recovery of the *b* parameter also in a consistent way. For both the concurrent calibration method and the Stocking-Lord method, when the common items are all CR items, the recovery of the *b* parameter has the smallest MSEs for each of the three different lengths of the common-item set and when the common items are all MC items, the recovery of the *b* parameter has the largest MSEs for each of the three different lengths of the common-item set. So the MSE tends to decrease as the number of CR items increases in the common-item set when the number of the common items is the same.

4). For both the concurrent calibration method and the Stocking-Lord method, the SD tends to decrease as the number of common items increases and as the number of CR items increases in the common items. The concurrent calibration method has smaller SDs than the Stocking-Lord method under all conditions. The SDs of the recovered *b*

parameter across all replications are also very low, ranging from 0.0386 to 0.0538. The differences between the two methods are at the third decimal place and are very close to each other for all 18 conditions as shown graphically in Figure 4.4.

Recovery of the Guessing Parameter c

The MSEs and SDs for the recovery of the *c* parameters after equating for all 18 conditions are summarized in Tables 4.5 and 4.6 and are graphically depicted in Figures 4.5 and 4.6.

Table 4.5 MSEs of the *c* Parameter

	5 common items			10 common items			15 common items		
	MC	Both	CR	MC	Both	CR	MC	Both	CR
CC	0.0335	0.0259	0.0258	0.0340	0.0254	0.0239	0.0337	0.0250	0.0219
SL	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333

Note: 1. CC - Concurrent calibration, SL - the Stocking-Lord method
2. MC - multiple-choice items, CR - constructed response item, Both - multiple-choice and constructed response items

Table 4.6 SDs of the *c* Parameter

	5 common items			10 common items			15 common items		
	MC	Both	CR	MC	Both	CR	MC	Both	CR
CC	0.01927	0.01736	0.01772	0.01929	0.01667	0.01731	0.01929	0.01775	0.01608
SL	0.01926	0.01926	0.01926	0.01926	0.01926	0.01926	0.01926	0.01926	0.01926

Note: 1. CC - Concurrent calibration, SL - the Stocking-Lord method
2. MC - multiple-choice items, CR - constructed response item, Both - multiple-choice and constructed response items

Figure 4.5 MSEs of the *c* Parameter

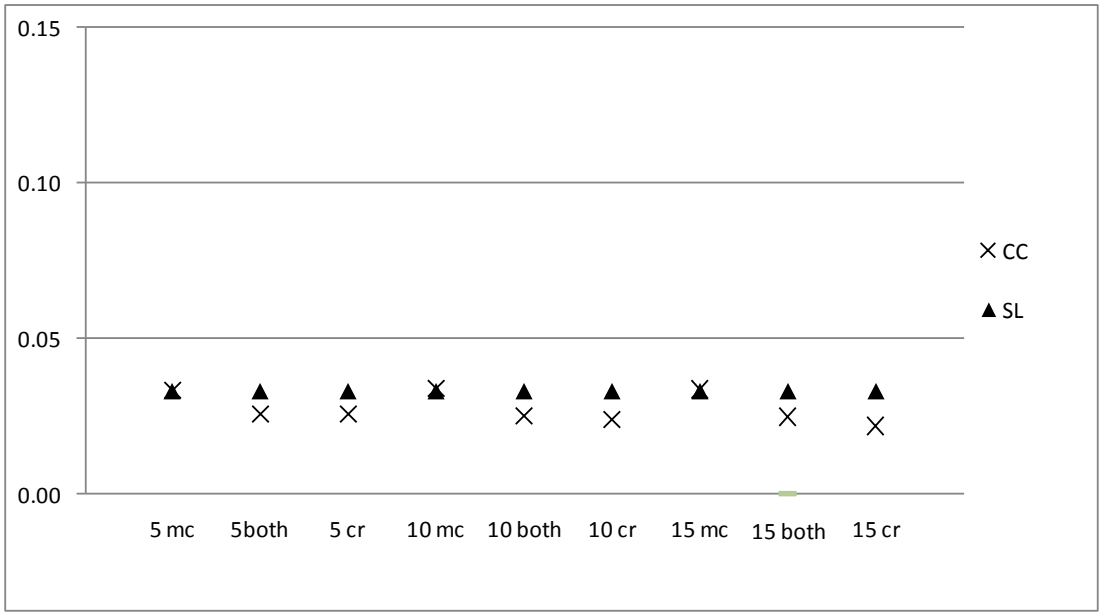
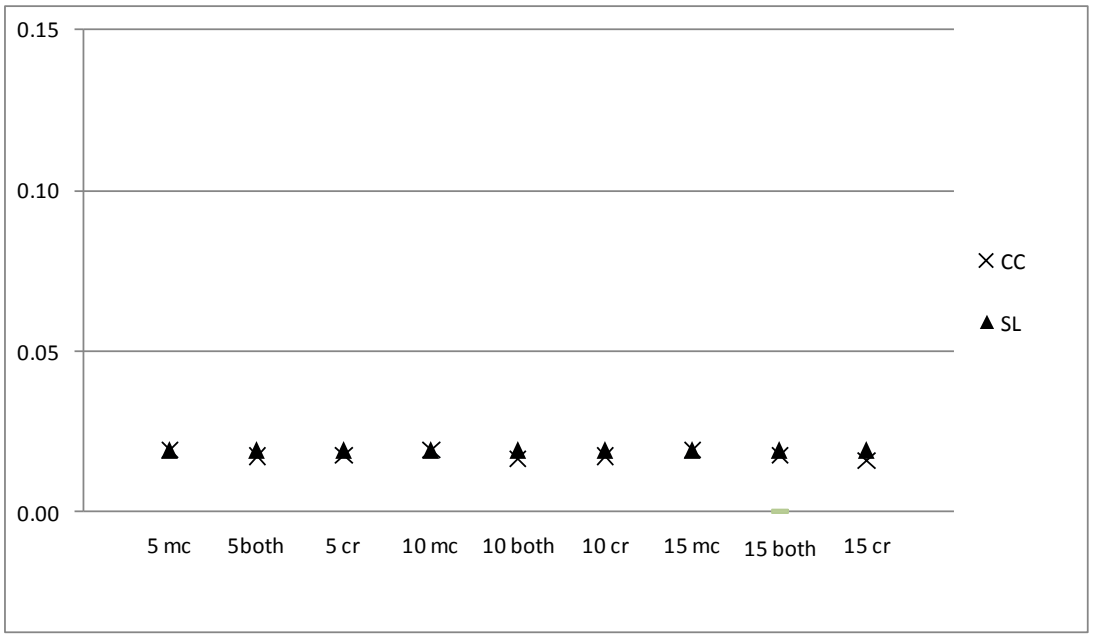


Figure 4.6 SDs of the *c* Parameter



From Tables 4.5 & 4.6 and Figures 4.5 & 4.6, the major findings are summarized as follows:

1). When the number and type of the common items are the same, the concurrent calibration method tends to have smaller MSEs than the Stocking-Lord method in recovering the *c* parameter.

2). For the concurrent calibration method, when the number of common items remains the same, the MSEs are smallest when the common items are all CR items and are biggest when the common items are all MC items. So the MSE tends to decrease as the number of the CR items increases in the common-item set. When the type of common items is the same, generally the MSEs tend to decrease as the number of common items increases.

3). For the Stocking-Lord method, the MSE remains the same regardless of the number and the type of the common items. This is because the guessing parameter *c* is independent of the scale transformation when different forms are calibrated separately as shown in equation 2.12 in Chapter 2.

4). The SDs of the recovered *c* parameter across all replications are very low, ranging from .0161 to 0.0193. The differences between the two methods are at the third or even the fifth decimal place, so the SDs remain very close for all 18 conditions as shown graphically in Figure 4.6.

The Differences between the Estimated Person Parameters and the "True" Person Parameters

In the previous section, the recovery of item parameters by the concurrent calibration method and the Stocking-Lord method was evaluated separately for the discrimination, difficulty and the guessing parameters. The results show that the

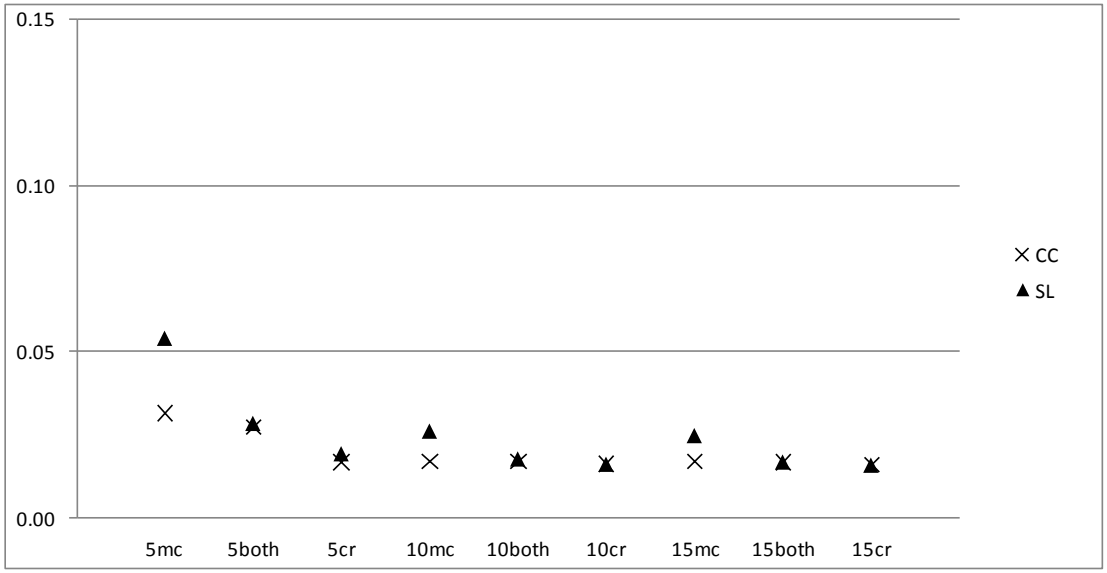
concurrent calibration method performs better in recovering all three parameters. The differences between the estimated person parameters generated based on the item parameters after equating and the true person parameters generated based on the item parameters to simulate the response data were also examined to evaluate the overall quality of the two equating methods. The criterion is the root mean squared difference (RMSD) as calculated in equation 3.4. The RMSD provides a single index in evaluating the overall performance of the two equating methods. The RMSDs between the estimated person parameters and the true person parameters for all 18 conditions are summarized in Tables 4.7 and are graphically depicted in Figures 4.7.

Table 4.7 RMSDs between the Estimated Person parameters and the "True" Person parameters

	5 common items			10 common items			15 common items		
	MC	Both	CR	MC	Both	CR	MC	Both	CR
CC	0.0314	0.0271	0.0167	0.0170	0.0169	0.0162	0.0170	0.0166	0.0159
SL	0.0540	0.0285	0.0194	0.0262	0.0178	0.0162	0.0248	0.0169	0.0159

Note: 1. CC - Concurrent calibration, SL - the Stocking-Lord method
2. MC - multiple-choice items, CR - constructed response item, Both - multiple-choice and constructed response items

Figure 4.7 RMSDs between the Estimated Person parameters and the "True" Person parameters



From Table 4.7 and Figure 4.7, the major findings are summarized as follows:

1). For the concurrent calibration method, the RMSD ranges from 0.0159 to 0.0314 and for the Stocking-Lord method, 0.0159 to 0.054. The true person parameter was simulated to have a standard deviation of 1, so on average the difference between the true person parameters and the estimated person parameters for each examine ranges from about 0.0159 standard deviation to about 0.0314 standard deviation when the concurrent calibration method is used and from about from about 0.0159 standard deviation to about 0.054 of a standard deviation when the Stocking-Lord method is used. When the number and type of the common items are the same, the concurrent calibration method has smaller RMSDs than the Stocking-Lord method. The gap between the two methods is widest with a difference ranging from 0.0078 to 0.0226 in the RMSD when the common items are all MC items, and it tends to get narrower or even disappear as the number of CR items increases in the common-item set.

2). In general, the RMSD tends to decrease as the length of the common-item set increases, when the type of the common items is the same. However, for the concurrent calibration method, the RMSDs remain almost the same when the number of the common items is 10 (20%) or more and for the Stocking-Lord method, the RMSDs remain almost the same when the number of the common items is 10 (20%) or more and the common-item set contains CR items.

3). The type of the common items impacts the differences between the true person parameters and the estimated person parameters in a consistent way. For both the concurrent calibration method and the Stocking-Lord method, when the common items are all CR items, the RMSDs are lowest and remain very close for each of the three different lengths of the common-item set. When the common items are all MC items, the RMSDs are highest for each of the three different lengths of the common-item set. So the RMSD tends to decrease as the number of CR items increases in the common-item set when the number of the common items is the same.

Summary

This Chapter presents the results of the analyses outlined in Chapter Three. In general, the analyses showed that both methods provide very stable results in recovering item parameters and the concurrent calibration method performs slightly better than the Stocking-Lord method in recovering all three item parameters. The estimated person parameters based on the concurrent calibration method were closer to the "true" person parameters than those based on the Stocking-Lord method. The analyses also showed that

the longer the common-item sets, the better the performance for both methods and the more CR items in the common-item sets, the better the performance for both methods.

Chapter Five: Conclusions

In this chapter, the goal and the methodology of this study are reviewed; the main findings are summarized, and the implications of these findings are discussed. Finally, the limitations of the study and suggestions for future research are presented.

Review of the Goal and Methodology of the Study

The steady increase in the use of mixed-format tests in large-scale assessments calls for appropriate equating methods. As IRT has become the mainstream as the theoretical basis for measurement, different IRT equating methods were also developed. The primary goal of this study is to compare the performance of two IRT equating methods, namely, the concurrent calibration method and linking following separate calibration using the Stocking-Lord procedure. Their performance was examined under different conditions with the non-equivalent groups common-item design. These conditions included three different numbers of common items (5, 10 and 15) and three types of common items (MC items only, both MC and CR items, and CR items only).

In this study, the simulation technique was used because the true equating relationship is known when simulated data are used and because it is the technique most frequently used in the IRT equating studies. For this study, a total of 70 items were selected from the TIMSS 2003 8th grade Mathematics assessment to form two mixed-format test forms, Form A and Form B, each consisting of 35 MC items and 15 CR items. The person parameters were generated from a normal distribution as the true person parameters with a mean of 0 and a standard deviation of 1 for Form A and a mean of 1

and a standard deviation of 1 for Form B. Item response data were then generated based on the item parameters of the selected items and the chosen unidimensional IRT models (three-parameter model for MC items and the GPC model for CR items). Fifty Form A and Form B samples with a sample size of 10,000 were generated.

For the concurrent calibration method, the item parameters were estimated in a single run using both Form A and Form B samples. For the separate calibration, the item parameters for Form A and B were estimated separately first; then the item parameter estimates for Form B were put on the scale of the item parameters for Form A following the Stocking-Lord method.

The results were evaluated from two aspects: the recovery of the item parameters after equating and how close the estimated person parameters are to the true person parameters. The criteria used were the MSE for the accuracy and the SD for the stability of the recovery of item parameters and the RMSD for the difference between the estimated person parameters and the true parameters. For all the criteria, the smaller the number, the better the equating method performs.

Summary of the Findings

Recovery of Item Parameters

From the results of the analysis, it can be concluded that the two methods perform differently in recovering the item parameters and that the length and the type of the common items have an impact on the performance of the two equating methods in different ways under different circumstances. The findings can be summarized as follows:

1). The concurrent calibration method yields smaller MSEs in all simulated conditions and therefore, performs better in recovering all three parameters than the Stocking-Lord method.

2). The length of the common-item set impacts the recovery of all three parameters in the same way for both methods. The longer the common-item set, the smaller the MSEs for both methods and the better the performance.

3). The type of the common items impacts the recovery of the three parameters for both the concurrent calibration method and the Stocking-Lord method. Both methods yield the smallest MSEs in recovering all three item parameters when the common items are all CR items and the biggest MSEs when the common items are all MC items. The MSE tends to decrease as the number of CR items increases in the common items. So both methods perform best when the common items are all CR items and their performance declines as the number of CR items decreases. It is also noticed that for the Stocking-Lord method, when the common items are all CR items, the length of the common-item set does not make any difference in recovering the difficulty parameter.

4). The SDs of the recovered item parameters across all replications are very low and very close to each other for all 18 conditions. This shows that both methods provide very stable results regardless of the number and the type of common items.

The Differences between the Estimated Person Parameters and the "True" Person Parameters

When the number and type of the common items are the same, the concurrent calibration method has smaller RMSDs than the Stocking-Lord method. So the estimated person parameters based on the concurrent calibration method are closer to the true

person parameters than those based on the Stocking-Lord method and therefore, the concurrent calibration method performs better than the Stocking-Lord method. However, a closer examination of the RMSDs shows that the difference between the two methods is very small. The largest difference in the RMSD between the two methods occurs when the common items consists of only 5 MC items with a value of 0.026, that is, on average the estimated person parameters for each examine based on the concurrent calibration method is at best just 0.026 standard deviation closer to the true person parameters than the estimated person parameters based on the Stocking-Lord method.

The RMSDs tend to decrease as the number of common items increases and as the number of CR items increases in the common items, so an increase in the number of common items or the number of CR items in the common items will make the estimated person parameters closer to the true person parameters. However, when the anchor test consists of 10 (20%) items or more of the total test, the RMSDs remain almost the same, especially when the common-item set contains CR items. So an increase in the number of common items when the number of common items is 20% or more of the total test may not result in a great improvement in the performance of the two equating methods.

Conclusions, Implications and Discussions

The findings from this study show that the concurrent calibration method generally performs better in recovering the item parameters and, more importantly, the concurrent calibration method produces more accurate estimated person parameters than the Stocking-Lord method. Therefore, overall the concurrent calibration method performs

better than the Stocking-Lord method with mixed format tests for the non-equivalent groups common-item design. But the difference in the performance between the two methods is very small. This result is consistent with what has been found in other studies (Petersen et al., 1983; Wingersky et al., 1987; Hanson & Beguin, 2002; Kim & Cohen, 2002) that dealt with only a single type of item. Hanson & Beguin (2002) pointed out that the advantage of the concurrent calibration method is due to the lower error on the common items, which is expected because the common item parameter estimates are based on larger samples from different groups.

Although the concurrent calibration method has been found to perform better than linking after separate calibration, researchers (Hanson & Buguin, 2002; Kolen & Brennan, 2004) do not recommend completely avoiding separate calibration in favor of concurrent calibration. This is because previous research has not shown consistent findings in favor of concurrent calibration. For example, Kim and Cohen (1998) concluded that the performance of separate calibration was equal to or better than concurrent calibration for MC items. Because not many studies have been done to compare the performance of the concurrent calibration method and linking after separate calibration for mixed-format tests, the evidence in this study is not sufficient to recommend completely avoiding separate calibration for mixed-format tests. Linking after separate calibration also has one potential benefit to identify possible individual item problems for the common items (Hanson & Beguin, 2002; Kolen & Brennan, 2004). This is of great importance because problematic items such as items showing DIF tend to lower the reliability and validity of the anchor test. Having two sets of item parameter estimates from separate calibration

may facilitate examining item parameter estimates for the common items. One simple way to do so as Kolen & Brennan (2004) suggested is to plot the parameter estimates for the common items from different calibrations to look for outliers. Items with estimates that do not appear to lie on a straight line may function differently in different groups and might need to be eliminated from the anchor test. Such examination is not possible for the concurrent calibration method because only one item parameter estimate for each common item is produced.

Another factor that may limit the use of concurrent calibration in practice is the availability of commercial computer programs. In his review of commercial computer software, Meng (2007) found that PARSCALE (Muraki, & Bock, 2003) and MULTILOG (Thissen, 1991) can be used to calibrate mixed-format test data. However, after trials on both programs, he found that neither is appropriate for non-equivalent groups common-item design, especially when more than two forms are involved. In his trials with PARSCALE, the program stopped running at PHASE II during the item parameter estimation iteration processes without providing any error messages and this problem could not be solved even by the PARSCALE technical support personnel (Meng, 2007). PARSCALE was also originally chosen as the IRT computer program for this study. However, the same problem occurred in the item parameter estimation iteration processes even though there were only two forms. Eventually, this study had to turn to the computer program ICL (Hanson, 2002) for the item and person parameter estimation. Hanson and Benguin (2002) also found that the concurrent calibration method puts more of a burden on the computer programs than separate calibration and results in some

performance problems, especially with more than two forms being equated simultaneously. In their analysis using MULTILOG to calibrate two test forms simultaneously, some runs failed to converge but there were no convergence problems when the two forms were calibrated separately. Although the concurrent calibration method may produce more accurate results, the difference between the two methods is not great. Given the potential benefit of separate calibration and the limit in the use of the concurrent calibration method, as Kolen and Brennan (2004) pointed out, separate calibration using the characteristic curve method seems to be safest and the concurrent calibration method could be used as an adjunct to the separate calibration method.

The findings from this study also show that an increase in the number of CR items when the number of common items is fixed or an increase in the number of common items may improve the performance of both equating methods. Kim (2004) provides some theoretical explanations for the relationship between the equating results and the type and the length of the common items, which has do w ith the number of response categories: the more response categories involved in the linking process, the more accurate and stable the results. Hence, the result here makes sense because when the number of common items is fixed in the tests, an anchor test consisting of only CR items usually contains the most response categories, an anchor test consisting of both types of items contains the second most response categories and an anchor test consisting of only MC items contains the least response categories. Also adding more common items in the tests will increase the number of response categories, and therefore improve the accuracy of the linking results. Similar results were also found in other studies. Meng (2007)

examined the performance of different equating methods with mixed-format tests in a vertical scaling situation and concluded that when the type of common-item set is changed from dichotomous-only to mixed-format, errors and biases are likely reduced. He also found that doubling the number of common items typically lowered errors and biases in the linking results. Kim and Lee (2004) also found that using both types of items, which they termed *simultaneous linking*, yielded more accurate results than linking through a single item type only.

Although an anchor test consisting of only CR items may produce better results, there are several limitations in using such an anchor test. First, CR items tend to be easy to memorize (Muraki, Hombo, & Lee, 2000). It may be difficult to find CR items that can be reused across forms considering that CR items are usually more expensive to develop. Second, even if the same CR items are used, the standards of the raters scoring the CR items almost always differ across the administrations (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998), making their item parameter estimates unstable and therefore, influencing the accuracy of equating. So the best approach in practice would be to include both types of the items in the anchor test. This is also considered ideal by Kim and Lee (2004) because they better represent the total test in both content and characteristics.

As for the appropriate length of the anchor test, both Angoff (1984) and Kolen and Brennan (2004) recommend that the common items should be at least 20 percent of the total test. The findings in regard to the difference between the estimated person parameters and the true person parameters from this study provide further support for their recommendations. When the anchor test consists of 10 (20%) items or more of the

total test, a further increase in the number of common items may not result in a great improvement in the performance of the two equating methods.

Limitations

As with any other simulation study, the findings in this study have some limitations because of the unique design using simulation. The following limitations are acknowledged.

First, Harris and Crouse (1993) suggested that using simulated data is most useful when the data closely resemble the real data and when it is accompanied by analyses involving the real data so that results of the simulation studies could be generalized to real data equating situations. In this study, although the data were simulated to resemble the real TIMSS 2003 data, the analyses did not involve the real data because the calibration data in TIMSS 2003 are not available. This may put some limitations on the generalizability of the findings from this study to real data equating situations.

Second, mixed-format tests may vary greatly in length and composition. So may the anchor test accordingly. However, the findings pertain to only the two specific forms used in this study, each consisting of 50 items (35 MC items and 15 CR items). In addition, only three levels of the number of common items and three different types of common items were considered. Therefore, considerable caution needs to be taken in generalizing the results to broader situations because of the small number of conditions investigated.

Third, the findings from this study were based on an ideal situation where the data were simulated from the same model used for item parameter estimation and where the sample size was ample. So the assumption of unidimensionality for IRT held and the item parameter estimates were stable. A partial justification for such an ideal situation is that ideally simulated situations should be preferred over real situations to show that the theoretical extension has been made properly (Kim & Lee, 2004). Because of the ideal situation simulated in this study, the findings and conclusions drawn from this study are limited only to unidimensional data with a large sample size. However, the model fit may not hold well with real data. Although IRT equating is fairly robust to violations of the unidimensionality assumption when equating alternate forms of a test as long as the violation is not too severe (Bolt, 1999; Camilli, Wang, and Fesq, 1995; Cook, Dorans, Eignor, and Petersen, 1985; De Champlain, 1996; Dorans and Kingston, 1985; Yen, 1984; cited in Kolen and Brennan, 2004), such violation may affect the performance of the equating methods. One study (Beguin, Hanson and Glas, 2000) compared the Stocking-Lord method to the concurrent calibration method for MC items using simulated data that purposefully did not fit the IRT model due to multidimensionality. They found that when groups were nonequivalent, the Stocking-Lord method produced more accurate equating than the concurrent calibration method. This finding is different from what was found in this study and by others like Kim and Cohen (1998) and Hanson and Benguin (2002) in which the data were simulated to fit the model and provides more evidence that separate calibration using the characteristic curve method seems to be safest in practice before more studies are conducted to examine the performance of the two equating methods for

mixed-format tests when the data do not fit the IRT model well due to multidimensionality.

Suggestions for Future Research

One limitation of this study is that the data were simulated to be in an ideal situation where the assumption of the unidimensionality fits. However, the simple unidimensional model would probably be misspecified to some extent with real data. This is more likely to be the case with mixed-format tests as studies show mixed findings in regard to whether MC items and CR items measure the same construct. Such model misspecification could affect the relative performance of the linking method following separate calibration versus the concurrent calibration method. So the performance of the two equating methods needs to be further examined under different simulation situations and real situations. One important research question for future studies is how the linking following separate calibration and the concurrent calibration method perform in situations where the assumptions of the IRT models involved in a mixed-format test do not hold well.

There is another method to transform item parameters from different scales onto one. This procedure is called the fixed parameter or item anchoring method (Hanson & Benguin, 2002) or referred to as the "fixed b's method" (Petersen et al., 1983). In this method, item parameters are estimated first for one form and then the item parameters in the other form are estimated with the common item parameters fixed at their estimated values using the first form. So this method does not involve a linking procedure as in the

concurrent calibration method and the two forms are calibrated separately. Obviously, the item anchoring method combines features of the concurrent and separate calibration. However, this method has been rarely studied. Only one study (Petersen et al., 1983) was found to compare it to the concurrent calibration method, the Stocking-Lord method, the linear equating method and the equipercentile equating method for MC items and the results showed that when compared to the conventional equating methods, the item anchoring method performed similarly to the concurrent calibration method and the Stocking-Lord method and better than the conventional methods. How this method performs with mixed-format tests as compared to the concurrent calibration and linking after separate calibration requires further investigation.

Summary

This chapter provides a review of the goal of the study, the general research design and a summary of the findings from the analysis results. A conclusion was drawn based on the findings. The conclusion and its implications were discussed in detail. Three limitations of this study were acknowledged and directions for future research in this area were suggested.

References:

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement,12*(2), 117-128.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test Equating*. New York: Academic Press.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for Educational and Psychological Testing*. Washington ,DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*, 239-251.

Baker, F. B. (1997). Empirical sampling distributions of equating coefficients for graded and nominal response instruments. *Applied Psychological Measurement, 21*, 157-172.

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

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker, Inc.

Barnard, J. J. (1996). *In search of equity in educational measurement: traditional versus modern equating methods*. Paper presented at the ASEESA's national conference at the HSRC Conference Center, Pretoria, South Africa.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). *Effect of unidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. Available from http://www.ba-h.com/papers/paper0002.html

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system score constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14*, 151-162.

Bennett, R. E., Rock, D. A., & Wang, M. (1991).Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*(1), 77-92.

Bennett, R. E., Ward, W. C., Rock, D. A., & Lahart, C. (1990). *Toward a framework for constructed-response items* (ETS Research Report No. 90-7). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472).Reading, MA: Addison-Wesley.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats-It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 385-395.

Bolt, D. M. *(* 1999*).* Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education,* 12*,* 383-407.

Bracht, G. H., & Hopkins, K. D. (1970). The communality of essay and objective tests of academic achievement. *Educational and Psychological Measurement, 30,* 359-364.

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

Breland, H. M., & Gaynor, G. L. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement, 16(2),* 119-128.

Brennan, R. L., & Kolen, M. J. (1987a). Some practical issues in equating. *Applied Psychological Measurement, 11***,** 279-290.

Brennan, R. L., & Kolen, M. J. (1987b). A reply to Angoff. *Applied Psychological Measurement, 11,* 301-306.

Bridgeman, B., & Rock, D. A. (1993). Relationship among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement, 30*, 313-329.

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

Camilli, G., Wang, M-m., & Fesq, J. (1995). The effect of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement*, 32, 79-96.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3*, 151-156.

Cook, L. L. (2007). Practical problems in equating test scores: a practitioner's perspective. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (2nd ed., pp.73-87). New York: Springer.

Cook, L. L., Dorans, N.J., Eignor, D.R., & Petersen, N.S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (ETS Research Rep. No. RR-85-30). Princeton, NJ : Educational Testing Service.

Cook, L. L., Dunbar, S. B., & Eignor, D. R. (1981, April). *IRT equating: A flexible alternative to conventional methods for solving practical testing problems.* Paper presented at the Annual Meeting- of the American Educational Research Association, Los Angeles.

Cook, L. L., & Eignor, D. R. (1983, April). *An investigation of the feasibility of applying item response theory to equate achievement tests*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1985). *A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates* (RR-85-38). Princeton NJ: Educational Testing Service.

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

Cohen, A. S., & Kim, S. H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*(2), 116-130.

Cowell, W. R. (1981, April). *Applicability of a simplified three-parameter logistic model for equating tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Davis, F. B., & Fifer, G. (1959). The effect on test reliability and validity of scoring

aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement, 14*(2), 159-170.

De Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement, 33*(2), 181-201.

DeMars, C. (1998). *The impact of test consequences and response format on performance.* Unpublished doctoral dissertation, Michigan State University.

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research, 16*(s1), 85–94.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violation of unidimensionality on the estimation of item and ability parameters and on item response theory equating of GRE verbal scale. *Journal of Educational Measurement, 22*(4), 249-262.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkage. In Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales* (pp. 179-199). New York: Springer.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281–306.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Ercikan, K., & Schwarz, R. (1995, April). *Dimensionality of multiple-choice and constructed-response tests for different ability groups.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35,* 137-154.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press.

Fisher, G. L. (1996). *The validity of pre-calculus multiple-choice and performance-based testing as a predictor of undergraduate mathematics and chemistry achievement*. Unpublished master's thesis, University of California, Santa Barbara.

Fiske, E. B. (1990, January 31). But is the child learning? Schools trying new tests. *The New York Times*, pp. 1, B6.

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

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Frederiksen, J. R., & Collins, A. (1989). A system approach to educational testing. *Educational Researcher*, *18(9),* 27-32.

Frisbie, D. A., & Cantor, N. K. (1995). The validity of scores from alternative methods of assessing spelling achievement. *Journal of Educational Measurement, 32(1),* 55-78.

Gafni, N., & Melamed, E. (1990). Using the circular equating paradigm for comparison of linear equating models. *Applied Psychological Measurement, 14*(3)*,* 243-256.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability.* New York: College Entrance Examination Board.

Guthrie, J. T. (1984). Testing higher level skills. *Journal of Reading, 28*, 188-190.

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

Hambleton, R.K., & Jones, R.W. (1993). An NCME Instructional Module on Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Norwell, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item

responses. *Applied Psychological Measurement, 31*(5), 457-459.

Han, K. T. (2009). IRTEQ: Windows application that implements IRT scaling and equating [computer program]. *Applied Psychological Measurement, 33(6),* 491-493.

Hancock, G. R. (1992). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education, 62*(2), 143-157.

Hanson, B. A. (2002). *IRT Command Language (ICL)*. Computer software. [Available at http://www.b-a-h.com/software/irt/icl/index.html]

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Harke, D. J., Herron, J. D., & Leffler, R. W. (1972). Comparison of a randomized multiple-choice format with a written on-hour physics problem test. *Science Education, 56,* 563-565.

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

Heim, A. W., & Watts, K. P. (1967). An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology, 37*, 339-346.

Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum competency tests: Comparison of methods. *Journal of Educational Measurement, 25*, 221-231.

Hogan, T. P. (1981) *Relationship between free-response and choice-type tests of achievement: A review of the literature*. Green Bay, WI: University of Wisconsin. (Eric Document NO. ED 224 81).

Hogan, T. P., & Mishler, C. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. *Journal of Educational Measurement, 17*(3), 219-227.

Holland, P. W. (2007). A framework and history for score linking. In Dorans, N. J.,Pommerich, M. & Holland, P. W. (2007). *Linking and aligning scores and scales* (pp. 5-30). New York: Springer.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.),

Educational Measurement (4th ed., pp. 187-220). Westport, CT: Praeger.

Horn,J . L. (1966). Some characteristics of classroom examinations. *Journal of Educational Measurement, 3*, 292-295.

Hurd, A. W. (1932). Comparison of short answer and multiple-choice tests covering identical subject content. *Journal of Educational Research, 26,* 28-30.

Hurlbut, D. (1954). The relative value of recall and recognition techniques for measuring precise knowledge of word meaning-nouns, verbs, adjectives. *Journal of Educational Research, 47*, 561-576.

Hwang, C., & Cleary, T. A. (1986, April). *Comparing IRT pre-equating and section pre-equating: A simulation study.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika 36*, 109-133.

Kim, S., & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26*, 255-270.

Kim, S., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Applied Psychological Measurement, 29*(1), 51-56.

Kim, S., & Cohen, A. S. (1995). A minimum χ^2 method for equating tests under the graded response model. *Applied Psychological Measurement, 19*(2), 167-176.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131-143.

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25-41.

Kim, S., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report 2004–5). Iowa City, IA: ACT, Inc.

Kim, S. (2004). *Unidimensional IRT scale linking procedures for mixed-format tests and their robustness to multidimensionality*. Unpublished Ph.D. Dissertation, The University of Iowa, Iowa City.

Kingston, N. M., & Holland, P. W. (1986). *Alternative methods of equating the GRE*

general test (GRE Board Professional Report GREB No. 81-16P, ETS Research Rep. No. 86-16). Princeton, NJ: Educational Testing Service.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22*, 197-206.

Klein, L. W., & Kolen, M. J. (1985, March). *Effect of number of common items in common-item equating with nonrandom groups.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (2nd ed., pp.31-55). New York: Springer.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement, 28*, 219–226.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-36.

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

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

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of General Educational Development. *Journal of Educational Measurement, 19*(4), 279-293.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton: Educational Testing Service.

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

Lord, F. M. (*1973*). *Testing if two measuring* procedures measure the same dimension. *Psychological Bulletin, 79*, 71–72.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and

equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453-461.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

Loyd, B. H., & Steele, J. (1986). Assessment of reading comprehension: A comparison of constructs. *Reading Psychology, 7*, 1-10.

Madaus, G. F., Haney, W., & Kreitzer, A. (1992). *Testing and evaluation: Learning from the projects we fund*. New York: Council for Aid to Education.

Magill, W. H. (1934). The influence of the form of item on the validity of achievement tests. *Journal of Educational Psychology, 25*, 21-28.

Manhart, J. J. (1996, April). *Factor analytic methods for determining whether multiple-choice and constructed-response tests measure the same construct*. Paper presented at the annual meeting of the American Educational Research Association, New York.

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

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. White (Ed.), *New Horizons in* Testing (pp. 147-177). New York: Academic.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 technical report.* Chestnut Hill, MA: Boston College.

McKinley, R. L., & Reckase, M. D. (1981). *A comparison of procedures for constructing large item pools* (RR 81-3). Columbia, MO: University of Missouri-Columbia, Tailored Testing Research Laboratory.

Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling* (Doctoral dissertation). Retrieved from http://ir.uiowa.edu/etd/338.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 61-74). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and*

prospects. Princeton, NJ: Policy Information Center.

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at Grades 4, 7, and 10. *Journal of Educational Measurement, 19*(1), 37-47.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., Chrostowski, S. J., & O'Connor, K. M. (2003). *TIMSS Assessment Frameworks and Specifications 2003* (2nd Edition). Chestnut Hill, MA: Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., & Bock, R. D. (2003). *PARSCALE (version 4.1): IRT item analysis and test scoring for rating-scale data.* Chicago, IL: Scientific Software, Inc.

Muraki, E., Hombo, C. M., & Lee, Y. W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*, 325-337.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher, 18*(9), 3-7.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review, Otaru University of Commerce, 51*(1), 1-23.

Ogasawara, H. (2001a). Item response theory true score equating and their standard errors. *Journal of Educational Behavioral Statistics, 26*(1), 31-50.

Ogasawara, H. (2001b). Least square estimations of item response theory linking coefficients. *Applied Psychological Measurement, 25*(4), 3-21.

Otis, A. S. (1922). The method for finding the correspondence between scores in two tests. *Journal of Educational Psychology, 9*, 239-260.

Paterson, D. G. (1926). Do new and old type examinations measure different mental functions? *School and Society, 24,* 246-248.

Patience, W. (1981, April). *A comparison of latent trait and equipercentile methods of vertically equating tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Petersen, N. S. (2007). Equating: best practices and challenges to best practices. In N. J. Dorans M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales* (2nd ed., pp.59-71). New York: Springer.

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

Pommerich, M., & Dorans, N. J. (Eds.). (2004). Concordance [Special issue]. *Applied Psychological Measurement*, *28*(4). No Child Left Behind Act of 2001, Pub. L. No. 107–110 (2002).

Quality Counts 2009. (2009, January 8). Education Week, 28 (17).

Pollock, J. M., & Rock, D. A. (1997). *Constructed-response tests in the NELS:88 high school effectiveness study.* Washington, DC: National Center for Education Statistics, U.S. Department of Education, Office of Educational Research and Improvement.

Pollock, J. M., Rock, D. A., & Jenkins, F. (1992, April). *Advantages and disadvantages of constructed-response item formats.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983, April). *The effect of anchor test size in vertical equating with the Rasch and three-parameter models.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. O'Connor (Eds.), *Changing Assessments: Alternative views of Aptitude, Achievement, and Instruction* (pp.37-75). Norwell, MA: Kluwer Academic Publishers.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed response items: a random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Ruch, G. M., & Stoddard, G. D. (1925). Comparative reliabilities of five types of objective examinations. *Journal of Educational Psychology, 16,* 89-103.

Samejima, F. (1969). *Estimation of a latent ability using a response pattern of graded scores.* Psychometrika Monograph No. 17. Richmond, VA: Psychometrics Society.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). *Machine-scorable complex constructed-response quantitative items: Agreement between expert system and human raters' scores* (ETS Research Report No. 91-11). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20 (26),* 2-16.

Skaggs, G., & Lissitz, R. W. (1986a). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56,* 495-529.

Skaggs, G., & Lissitz, R. W. (1986b). An exploration of the robustness of four test equating models. *Applied Psychological Measurement, 10,* 303-317.

Smith, J. K., & Smith, M. R. (1984, April). *The influence of item format on measures of reading comprehension.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 207-210.

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988, April). *Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Tatsuoka, K. K. (1991). *Item construction and psychometric models appropriate for constructed-responses* (ETS Research Report No. 91-31). Princeton, NJ: Educational Testing Service.

Tian, F. (2009, April). *Examining the construct equivalence between Chinese, Russian, Arabic versions and English version of TIMSS 2003 using factor analysis, DIF analysis and DTF analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Thiede, K. W., Klockars, A. J., & Hancock, G. R. (1991, April). *Recognition versus recall test formats: A correlational analysis.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Thissen, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory* [Computer software]. Chicago, IL: Scientific Software International, Inc.

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-

choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*(2), 113-123.

Thorndike, E. L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology*, *6*, 29-33.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1,* 355-369.

Traub, R. E., & MacRury, K. (1990). Multiple-choice vs. Free-response in the testing of scholastic achievement. In K. Ingenkamp & R. S. Jäger (Eds.), *Tests und trends 8:Jahrbuch der padagogischen diagnostik* (pp. 128-159). Weinheim: Beltz Verlag.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement,10,* 333–344.

Vemon, P. E. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement, 9*, 430-449.

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

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103- 118.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6*, 1-11.

Way, W. D., & Tang, K. L. (1991). *A comparison of four logistic model equating methods.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration*[Computer program]. ETS Research Report 87-24. Princeton, NJ: Educational Testing Service.

Wright, B. D. (1977) Solving measurement problems with the Rasch model. *Journal of*

Educational Measurement, 14(2), 97-116.

Yao, L., & Boughton, K. A. (2009). Multidimensional linking for test with mixed item type. *Journal of Educational Measurement, 46 ,177-197*

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145

APPENDIX A: ICL CODES

A 1. Sample ICL codes used to separately calibrate Forms A and B

```
set data formA1.dat
output -log_file formA1.log
set model [concat [rep 3 15] [rep 1 35]]
allocate_items_dist 50 -models $model
read_examinees formA1.dat 50i1
starting_values_dichotomous
EM_steps
print -item_param
set fileID [open formA1.par w]
write_item_param_channel $fileID -format %+.4f -no_item_numbers
close $fileID
release_items_dist
```

A 2. Sample ICL codes used to calibrate Form A and Form B simultaneously

```
set data form5both1.dat
output -log_file form5both1.log
set model [concat [rep 3 15] [rep 1 35] [rep 3 14] [rep 1 31]]
allocate_items_dist 95 -num_groups 2 -models $model
read_examinees form5both1.dat {@2 95i1} {i1}
starting_values_dichotomous
EM_steps -estim_dist -scale_points -max_iter 200
print -item_param
set fileID [open form5both1.par w]
write_item_param_channel $fileID -format %+.4f -no_item_numbers
close $fileID
release_items_dist
```

A3. Sample ICL codes used to generate True scores for Form B

```
set data formB1.dat
output -log_file form5both1.log
set model [concat [rep 3 15] [rep 1 35]]
allocate_items_dist 50  -models $model
read_examinees formB1.dat  50i1
starting_values_dichotomous
EM_steps -estim_dist -scale_points -max_iter 200
print -item_param -latent_dist -latent_dist_moments
```

```
read_item_param true.par -no_item_numbers
set estep [new_estep]
estep_compute $estep 1 1
delete_estep $estep
set eapfile [open form5both1true.theta w]
for {set i 1} {$i <= [num_examinees]} {incr i} {
set resp [examinee_responses $i]
set numcorrect 0
foreach r $resp {
if {$r > 0} then {incr numcorrect}}
set eap [examinee_posterior_mean $i]
set mle [examinee_theta_MLE $i -6.0 6.0]
puts $eapfile [format "%+.5f\t%+.5f\t%d" $eap $mle $numcorrect]}
close $eapfile
release_items_dist
```

A3. Sample ICL codes used to generate estimated scores for Form B

```
set data formB1.dat
output -log_file form5both1score.log
set model [concat [rep 3 15] [rep 1 35]]
allocate_items_dist 50  -models $model
read_examinees formB1.dat  50i1
starting_values_dichotomous
EM_steps -estim_dist -scale_points -max_iter 200
print -item_param -latent_dist -latent_dist_moments
read_item_param form5both1.par -no_item_numbers
set estep [new_estep]
estep_compute $estep 1 1
delete_estep $estep
set eapfile [open form5both1.theta w]
for {set i 1} {$i <= [num_examinees]} {incr i} {
set resp [examinee_responses $i]
set numcorrect 0
foreach r $resp {
if {$r > 0} then {incr numcorrect}}
set eap [examinee_posterior_mean $i]
set mle [examinee_theta_MLE $i -6.0 6.0]
puts $eapfile [format "%+.5f\t%+.5f\t%d" $eap $mle $numcorrect]}
close $eapfile
release_items_dist
```