EVALUATING EQUATING RESULTS IN THE NON-EQUIVALENT GROUPS WITH ANCHOR TEST DESIGN USING EQUIPERCENTILE AND EQUITY CRITERIA

By

Minh Quang Duong

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2011

ABSTRACT

EVALUATING EQUATING RESULTS IN THE NON-EQUIVALENT GROUPS WITH ANCHOR TEST DESIGN USING EQUIPERCENTILE AND EQUITY CRITERIA

By

Minh Quang Duong

Testing programs often use multiple test forms of the same test to control item exposure and to ensure test security. Although test forms are constructed to be as similar as possible, they often differ. Test equating techniques are those statistical methods used to adjust scores obtained on different test forms of the same test so that they are comparable and can be used interchangeably.

In this study, the performance of four commonly used equating methods under the non-equivalent group with anchor test (NEAT) design - the frequency estimation equipercentile method (FE), the chain equipercentile method (CE), the item response theory (IRT) true score method (TS), and the IRT observed score method (OS) – were examined. In order to evaluate equating results, four evaluation criteria - the equipercentile criterion (EP), the full equity criterion (E), the first-order equity criterion (E_1), and the second-order equity criterion (E_2) – were used. Simulated data were used in various conditions of form and group differences.

Several major findings were obtained in this study. When the distributions used to simulate ability for the groups were equal, the four methods produced similar results, regardless of the criterion used.

When group difference existed in the distributions used to simulate the data, the results produced by different methods diverged significantly when the EP, E, and E_1 criteria were used. The difference was small when the E_2 criterion was used. In general, the OS method

outperformed the others in regarding to the EP and E criteria. The TS method performed the best in regarding to the E_1 criterion followed by the OS, CE, and FE methods. Between the two observed score methods (i.e., FE and CE), which were outperformed by the two IRT methods, the CE method produced much better results and they were close to those produced by the two IRT methods. The FE method produced the worst results, regardless of the criterion used.

It was also found that test form difference had clear effects on all methods, regardless of the criterion used. Larger difference between test forms led to worst equating results.

While the two IRT methods were not clearly affected by group differences in the generating distributions, the two observed score equating methods were. Larger group differences produced worse equating results obtained from the CE and the FE methods. In addition, the impacts of group differences were much stronger for the FE method than for the CE method.

Group and form interaction effects were not found for the IRT methods. They were, however, present for the FE and CE methods although those effects were small.

When evaluated with the E_2 criterion, the four equating methods produced results that were not better than those obtained from using directly raw scores from test forms without equating.

These results are discussed in more details and some recommendations are made for equating practice. Limitations of the study and suggestions for further research are also presented.

DEDICATIONS

To my parents,
who gave me my life, raised me with great love, taught me the value of education, and encouraged me to pursue further education in order to be a better person.

And to my beloved wife, LIEN KIM NGUYEN,
who sacrificed her teaching career to come with me to US and has always been by my side in all steps, sharing with me all ups and downs during our time at MSU and beyond.

ACKNOWLEGEMENTS

There are many people who have given me much more than I can ever possibly repay.

I would like to express my deep gratitude to my adviser and dissertation chair, Dr. Mark Reckase. My first meeting with him, which turned out to be one of the most important conversations of my life, inspired me to pursue advanced studies in psychometrics. Without his rich guidance, tremendous support, insightful comments, and great patience, this dissertation would not have been possible and my doctoral studies would never have been completed. I could not have asked for a better adviser.

My special thanks go to Dr. Tenko Raykov who, for the last six years, has become my mentor. I would like to thank him for his strong support, both academically and emotionally, during my time at MSU.

Since my first step at MSU, Dr. Richard Houang has become not only a good mentor but also a great friend. I am grateful to him for many things he has done for me. I will miss the conversations I had with him about educational measurement.

I thank Dr. Alexander Von Eye for his helpful comments and great support to my dissertation. I am privileged and proud to have him in my dissertation committee.

My sincere gratitude goes to Dr. Richard Prawat, Chair of the CEPSE department, and to Dr. Karen Klomparens, Dean of the MSU Graduate School, for their great support.

With all my heart, I would like to thank Dr. Christopher Wheeler for everything he has done for me. It was Dr. Wheeler who built a bridge connecting me to my success today.

Finally, I thank many other professors, graduate students, and friends who have enriched my life at MSU.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This introductory chapter presents the foundations of this study. Major points include context and nature of the problem, the approach used to address the problem, purpose of the study, specific research questions to be answered, research expectations, and the significance of this study to test equating research and practice.

1.1. Test equating

In many testing programs, alternative forms of the same test are used in different administrations to maintain test security. For example, the SAT exam is given at several administrations each year with different forms. In developing various forms of the same test, test developers use test specifications to ensure that alternative forms are similar in contents and statistical characteristics. Despite test developers' efforts, it is almost inevitable that differences among test forms exist to some degree unless they are identical. As a result, one test form may be easier or more difficult than others. Therefore, some test takers might have advantages or disadvantages simply because they are administered a relative easy or difficult test form. In order to maintain test fairness, scores obtained from different test forms should not be used before some adjustment is made to ensure score comparability (i.e., being on the same scale). The adjustment process is called test equating or equating. Equating is often defined as a statistical process used to adjust scores on alternative test forms so that their scores can be used interchangeably (Kolen & Brennan, 2004). If equating is successfully performed, test fairness is maintained and it becomes possible to compare examinees or to measure their growth (Angoff, 1971; Petersen, Kolen, & Hoover, 1989).

In general, a test equating process consists of two important components: an equating design, and one or more equating methods. Equating design refers to a plan to collect equating data. For that reason, it is sometimes called data collection design. The most commonly used design is the non-equivalent groups with anchor test (NEAT) design. In this design two test forms, which share some common items, called anchor items, are administered to two samples from two, usually distinct, populations of test takers (von Davier, Holland, & Thayer, 2004a). If the total score includes the score on the anchor items, the anchor is called internal. If the score from the anchor items is not included in the total score, it is called an external anchor. This design is also called the common-item nonequivalent groups design (Kolen & Brennan, 2004). Other common designs are single group design and random groups design. In each equating design, different equating methods can be used. An equating method is a framework to derive the equating function which places scores of one test form on the scale of another test form. Equating methods can be generally classified into two different groups: the observed score equating (OSE) methods, and the item response theory (IRT) methods. The OSE methods are usually referred to as traditional methods. Another way to classify equating methods is based on the assumed relationship between scores on the two test forms being equated. Within this framework, an equating method can be classified as either linear or equipercentile depending on whether the relationship between scores on the two test forms is assumed linear or non-linear.

1.2. Evaluating equating results

Given the importance of equating in making scores comparable, which in turn has crucial impacts on decision making, it is critical that equating results be evaluated for accuracy. Evaluation results are also useful in helping psychometricians compare and select appropriate equating procedures in a specific situation.

Evaluating equating requires a criterion or criteria to which equating accuracy can be judged. A variety of criteria have been proposed and used in research and practice (for a detailed review, see Harris & Crouse, 1993).

Traditionally, equating results from a very large sample are often used as benchmarks to evaluate other equating procedures (e.g., see Holland, Sinharay, von Davier, & Han, 2008; Livingston & Kim, 2010; Puhan, Moses, Grant, & McHale, 2009; Sinharay & Holland, 2007). However, comparing a method to another assesses the similarity between them, but not necessarily the accuracy of the former. In addition, the selection of what method to use on a large sample to obtain the criterion is arbitrary. Any equating method can be used to produce the criterion. Because different methods likely yield different results, this approach does not seem reasonable. According to Harris and Crouse (1993), large sample equating procedures do not necessarily provide the true equating results to which other methods should be compared.

Another popular equating criterion is the standard error of equating (SEE) which is defined as the standard deviation of equated scores over many hypothetical replications of an equating process on samples from a target population of test takers (Kolen & Brennan, 2004). Assessing SEE often involves drawing random samples from the same population under the same set of conditions. Other statistical techniques, such as bootstrap methods, can be used to assess SEE for a single equating. Equating processes with smaller SEE are preferred. Several studies used SEE to evaluate equating results (e.g., see Cui & Kolen, 2008; Hanson, Zeng, & Kolen, 1993; Liu, Schulz, & Yu, 2008; Lord, 1982a, 1982b; Wang, Hanson, & Harris, 2000; Zeng, Hanson, & Kolen, 1994). However, the use of SEE as means of comparing different equating methods has been criticized because it only accounts for random errors due to sampling examinees from the population and ignores other sources of errors (Harris & Crouse, 1993).

Cross-validation and replication are also frequently used to evaluate equating results. Cross-validation applies equating transformation obtained on one sample to another independent sample. Replication requires recalculation of the equating transformation on another sample. Both methods use results from two different applications to check the stability of equating results. Examples of research of this kind are those conducted by Holmes (1982), and Kolen (1981). Circular equating, which equates a test form to itself through a chain of equatings, is another commonly used equating criterion. Traditionally, the circular equating criterion is intended to assess systematic error. Ideally, the final result must be an identity (i.e., a score is transformed to an identical score). Many studies used this criterion to evaluate equating results (e.g., see Gafni & Melamed, 1990; Han, Kolen, & Pohlmann, 1997; Klein & Jarjoura, 1985; Lord & Wingersky, 1984; Philips, 1985; Puhan, 2010; Skaggs, 2005; Wang, Hanson, & Harris, 2000). Cross-validation, replication, and circular equating have the same setback as SEE. Stability obtained in those methods may not be an appropriate criterion for choosing an equating method because an incorrect equating procedure may produce more stable equating relationships than correct procedures (Lord & Wingersky, 1984). Kolen and Brennan (1987) recommended that circular equating should be used with considerable caution. Wang, Hanson, and Harris (2000) showed in their simulation study that the accuracy of equating methods cannot be determined by circular equating because this equating criterion does not take into account systematic error (bias) embedded in the equating.

1.3. Concerns regarding equating criteria

As presented previously, most widely used equating criteria have shortcomings. Although many equating criteria have been proposed and used, no one criterion is unambiguously preferable to others. For many years, researchers have recognized that there is a

problem in evaluating equating because no definitive criterion exists (Harris & Crouse, 1993). Using different criteria may lead to different conclusions about equating adequacy in a given context (Skaggs, 1990). Kolen (1990) indicated that there is no universally agreed upon equating criterion. This does not mean that all criteria are equally problematic. Some criteria can be better than others in a specific situation. However, the lack of a common equating criterion makes it difficult to compare results across equating studies. Even the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), which states that technical information should be provided on the accuracy of the equating, provides no guideline on how the adequacy of equating should be assessed.

1.4. The approach taken: equating definition and equating criterion

When implementing an equating evaluation process, it is crucial to consider the adopted definition of equating. The goal of equating evaluation must be to assess the extent to which the definition of equating holds. In other words, equating criteria should be closely linked to what it means for the two test forms to be equated.

Theoretically, equating is often defined as a statistical process to adjust scores of multiple test forms so that their scores are comparable and interchangeable (Kolen & Brennan, 2004; Petersen, Kolen, & Hooever, 1989). Definition of equating presented this way does not carry much usefulness to the process of selecting an appropriate criterion. What does it mean to say scores are comparable and interchangeable? How does one determine if scores are actually comparable after being equated? In order to be able to select a correct and fair criterion, one needs a definition that can be operationally applied. In other words, an operational definition of equating is necessary for criterion selection. In order to be useful, any operational definition must be able to specify what comparability and interchangeability mean.

This study focused on two operational definitions of equating that have been proposed in the literature: the equipercentile definition proposed by Angoff (1971), and equity definition proposed by Lord (1980).

1.4.1. Equipercentile definition

According to Angoff (1971), "two scores, one on form *X* and the other on form *Y* (where *X* and *Y* measure the same function with the same degree of reliability), may be considered equivalent if their corresponding percentile ranks in any given group are equal" (p. 563). This statement is commonly regarded as Angoff's equipercentile definition of equating (Harris & Crouse, 1993). The equipercentile definition implies that the distributions of scores on two test forms in a population should be identical after equating (Kolen & Brennan, 2004, p.12). The equipercentile definition is also labeled the definition of observed-score equating.

1.4.2. Equity definition

The equity definition of equating, also called the definition of true-score equating, was proposed by Lord (1980) as "if an equating of test *X* and *Y* is to be equitable to each applicant, it must be a matter of indifference to applicants at every given ability level θ whether they are to take test *X* or test *Y*" (p.195). Equity requires that for every θ, the conditional distributions of scores on the two test forms to be equated must be identical after equating.

The two definitions are not unrelated. If one applies the equipercentile definition to a group of examinees with the same ability θ level, it is equivalent to the equity definition (Divgi, 1981). In addition, both definitions are based on score distributions. The difference is that the equipercentile definition is focused on marginal score distributions while the equity definition is defined on conditional score distributions.

Lord (1980) proved that equity is never satisfied unless the test forms being equated are perfectly reliable or strictly parallel, in which case equating is unnecessary. In practice test forms are never perfectly reliable nor strictly parallel. In other words, equity is unlikely to be fully satisfied in practice. Nevertheless, equity can be considered as a gold standard for evaluating equating in the sense that it represents an ideal equating.

Since full equity is unlikely to be satisfied in practice, some weaker versions of equity have been proposed. Two popular weakened versions are the first-order equity (Divgi, 1981) and the second-order equity (Morris, 1982). Those equity definitions require only the first-order moment (i.e., expected value) or the second-order moment (i.e., variance or standard deviation) of the two conditional score distributions be the same, respectively. Kolen, Hanson, and Brennan (1992) argued that the second-order equity should be nearly satisfied in order for the two test forms being equated to be used interchangeably.

1.4.3. Equipercentile criterion and equity criteria

Four equating criteria can be formulated from the two basic operational definitions of equating. Let X, Y, and Y_e represent score on Form X (old form), score on Form Y (new form), and equated Form Y score, respectively. Also, let x, y, and y_e represent particular values of X, Y, and Y_e, respectively. The equipercentile criterion compares two marginal score distributions: one for Y_e, and one for X. The full equity criterion compare two conditional score distributions for Y_e and X at a specific value of latent ability θ. The first-order and second-order equity criteria compare the means and the variances (or standard deviations) of those two conditional distributions.

Note that equity-based criteria must be evaluated at all levels of θ in the range of

interest. Different criteria are appropriate for different equating definitions (or purposes). For example, if the equating purpose is to obtain the same marginal distributions of scores on two test forms after equating, the equipercentile criterion would be appropriate. If getting the same conditional distributions on two test forms after equating is the goal, the full equity criterion would be more suitable.

1.5. Motivation

Test equating is an important task in many testing programs to make scores from alternative test forms comparable. It is very crucial that equating results be evaluated based on appropriate criteria in accordance with predetermined equating purposes. There are several factors that motivated this study as follow:

- There are urgent needs for evaluating equating results properly using appropriate and fair criteria which are directly linked to the adopted definition of equating. In order to ensure test fairness, the evaluation process for equating must be correct and fair.
- Although equity is the most important aspect of equating (Lord, 1980), equity-based criteria have rarely been used in equating research and practice. In fact, no research using full equity criterion to evaluate equating results has been reported, at least up to the time this dissertation was written.
- Many testing program currently use equipercentile OSE methods and IRT equating methods. Therefore, it is necessary to compare their performances. Although many research studies comparing them have been conducted, most focused only between two methods of the same kind (i.e., either OSE or IRT equating), and they led to different conclusions. In addition, research comparing equipercentile OSE and IRT equating methods is sparse.

- The NEAT design is the most popular equating design used in practice. However, not much research has been conducted to compare equipercentile OSE and IRT equating methods in this design, especially using equity criteria.

1.6. Purpose of the study and research questions

1.6.1. Purpose

The primary purpose of this study was to use equipercentile and equity-based criteria to evaluate performance of four commonly used equating methods under the NEAT design. Specifically, those equating methods are (see Chapter 2 for more details):

- Presmoothed frequency estimation equipercentile method (FE)
- Presmoothed chain equipercentile method (CE)
- IRT true score equating method (TS)
- IRT observed score equating method (OS)

In addition, the identity equating (i.e., no equating) method was also used to examine possible conditions when no equating is preferred. The performance of those equating methods was investigated in various conditions of differences between test forms and differences between groups of test takers.

1.6.2. Research questions

Particularly, this study aimed to address the following research questions:

Question 1: Overall, how do those equating methods compare to one another in terms of equipercentile and equity criteria?

Question 2: How do test form differences affect equating results for each method?

Question 3: How do group differences affect equating results for each method?

Question 4: Are there interaction effects between test form differences and group differences for each method?

Question 5: In what conditions is the identity equating preferred to the others?

1.7. Research expectations

Kolen and Brennan (2004) stated that each equating method tends to function optimally under certain situations. Therefore, it was expected that the investigated methods would perform differently relative to different criteria. Specifically, it was expected that

- FE, CE, and OS perform relatively well under the equipercentile criterion.
- TS produces the most accurate results under the first-order equity criterion.
- OS performs better than TS when the equipercentile criterion was used.
- All equating methods perform similarly when test form differences and group differences are small.
- When form differences and group differences are large, equating results, regardless of methods used, would be worse than when these differences are negligible.
- Identity equating is preferred when form differences and/or group differences are very large.

1.8. Significance of the study

Given the lack of research on using equipercentile and equity criteria to evaluate equating results, and the scarcity of studies comparing equipercentile OSE and IRT equating methods in the NEAT design, this study was initiated to fill the gap. It was hoped that this study would make significant contributions to the research literature by providing an alternative perspective on how to evaluate equating results in such a way that is well aligned with equating purposes in specific

contexts. In addition, results from this study would provide more comprehensive guidance for practitioners to select appropriate methods based on their adopted purposes. It was also expected that this study will inform equating practice by suggesting the size of form difference and group difference that can cause a specific equating method to perform well or poorly relative to various criteria.

1.9. Additional notes

- In this study, the equating direction was from Form Y to Form X. In other words, Form Y was the new form and Form X was the base (old) form.
- The anchor used in this study was internal which means the anchor score was included in the total score.
- Although the words 'definition', 'purpose', and 'property' have different meanings in the regular context, they are used interchangeably in this dissertation in the phrases such as 'equating definition', 'equating purpose', and 'equating property' to mean the same things. All of them mean what is supposed to be accomplished from equating.
- The term 'equating criterion' is frequently used in this dissertation. In general, it means a certain property of equating that should hold for the equating results to be considered accurate. For example, equity criterion means equity property proposed by Lord (1980). Note that the equating criterion used in this study is a different concept than the commonly used statistical criterion.

1.10. Overview of the dissertation

The rest of this dissertation is organized as follows.

- Chapter 2 presents theoretical background relevant to the study. Major topics that are

addressed include the NEAT design, equipercentile OSE and IRT equating methods used in the study, equating criteria, and a review of relevant research.

- Chapter 3 is reserved for presenting research design and methodology. Detailed steps are laid out including overall framework, research factors, procedures, and evaluation criteria.
- Results of the study are presented in Chapter 4 for all research conditions, focusing on addressing proposed research questions.
- The last chapter, Chapter 5, summarizes main findings and discusses their practical implications. Limitations of the study, current issues and future steps are also discussed in this closing chapter.

CHAPTER 2

LITERATURE REVIEW

In this chapter, theoretical issues relevant to this study are discussed. The chapter begins with the issue of test equating including its two main components: equating design and equating methods. Details about a specific design, the NEAT design which was used in this study, are discussed next. After that, four equating methods used in this study, frequency estimation equipercentile equating, chain equipercentile equating, IRT true score equating, and IRT observed score equating, are examined. The equipercentile definition and the equity definition along with their corresponding criteria are the next topics. The chapter concludes with a summary of prior research relevant to this study.

2.1. Test equating

Testing programs often use multiple forms of the same test for a variety of reasons. For example, in situations such as college admission, people can take the test at different times. If the same questions were used at each administration, they would become known and people taking the test at a later administration would have advantages. Thus, using multiple forms of a test maintains test fairness and security. Another example is a situation where it is necessary to use pretest and posttest (e.g., measuring growth). The main reason for using different forms of a test is to ensure that a test taker's score is a current measure of his or her competence and not a measure of ability to recall questions on the form previously administered. Furthermore, using multiple alternative forms of the same test also serves to satisfy broad content coverage.

Although multiple forms are created to have similar characteristics, it is unlikely that test forms are exactly equivalent. For this reason, some examinees may have advantages or

disadvantages by taking an easy or difficult form. To ensure test fairness, scores for different test forms must be adjusted by a process commonly referred to as equating. Whenever alternative forms are used, equating is performed to place scores from different test forms on the same scale. Equating is commonly defined as a statistical process for adjusting scores of different test forms to account for unintended form-to-form differences such that scores can be considered comparable (Kolen & Brennan, 2004).

When test forms are equated, a group (or population) of test takers to whom the equating relationship is supposed to be applied must be identified (Braun & Holland, 1982). This group is usually called the target population in the equating literature.

An equating process consists of two major components: equating design and equating method. The equating design is a framework for collecting equating data. Common equating designs include: (a) single group design where the two test forms being equated are given to a single group randomly drawn from a population which is also the target population; (b) random groups design where the two forms are administered to two groups of test takers randomly drawn from a target population; and (c) nonequivalent groups with anchor test (NEAT) design where two test forms which share a set of common item, called an anchor, are given to two groups from two populations which usually differ in level of ability measured by the test. This study focused on the third design which is discussed in the next section.

Under each equating design, various methods can be used. Equating methods can be classified into two categories: (a) observed-score equating (OSE) methods, and (b) IRT equating methods. OSE methods, which are also called traditional methods, are conducted on empirical observed scores and can be further grouped into two kinds depending on the hypothetical equating relationship. Linear methods specify a linear relationship between scores on the two

test forms being equated. Equipercentile methods determine a non-linear relationship between scores of the two test forms. This study used two common equipercentile OSE methods that are widely used in practice: the frequency estimation equating method and chain equating method. Details about these methods follow.

Unlike traditional OSE methods, IRT methods are not conducted on empirical scores. These methods are based on IRT models which hypothesize a relationship between a specific examinee's latent ability, represented by θ, and the probability of his or her getting a correct answer to a specific test item. IRT equating is either conducted on true score or observed scores generated by the adopted models. Two common IRT equating methods were investigated in this study: the IRT true score equating and the IRT observed score equating. Details of these methods are also reviewed in the subsequent sections.

Further details about equating designs and methods can be found in Holland and Dorrans (2006), Kolen and Brennan (2004), von Davier, Holland, and Thayer (2004a), and Petersen, Kolen, and Hooever (1989).

2.2. The nonequivalent groups with anchor test (NEAT) design

Various equating designs can be used to collect data for equating. One of the most popular designs is the nonequivalent groups with anchor test (NEAT) design (von Davier, Holland, & Thayer, 2004a). This design is also called the common-items non-equivalent group design (Kolen & Brennan, 2004). In this dissertation, the term 'NEAT' is adopted.

In this design (see Table 2.1), two test forms to be equated, Form X and Form Y, are administered to two groups (i.e., samples), group 1 and group 2, of test takers from two different populations *P* and *Q*, respectively. The two test forms share a subset of items which is usually called the anchor (denoted A in Table 2.1). The sets of non-common (i.e., unique) items of Form

X and Form Y are labeled X_U and Y_U, respectively. A is internal anchor if its score is included in the total score. Otherwise, it is external anchor. Note that in the NEAT design, as presented in Table 2.1, scores on X_U are not obtained for the population *Q* sample and scores on Y_U are not obtained for the population *P* sample.

Table 2.1. The NEAT design

Population	Sample	X_U	A	Y_U
P	1	√	√	Not observed
Q	2	Not observed	√	√

The anchor A is used to adjust for differences between the two groups in terms of abilities or skills relevant to the test. In other words, A serves to remove group differences to increase equating accuracy. It is recommended that the anchor should be a representative of the test forms being equated in content and statistical characteristics (see Sinharay & Holland, 2007; Kolen & Brennan, 2004). That is, the anchor should be a mini-version of the test forms. When groups differ substantially, the anchor may fail to adjust for group differences. In such situations, equating may not be accurate.

In the NEAT design, the target population *T* is the mixture of *P* and *Q* and can be formulated as

$$T = wP + (1\text{-}w)Q \tag{2.1}$$

The mixture is determined by the weight *w* given to population *P*. Theoretically, *w* can be

any number between 0 and 1. When $w = 1$, $T \equiv P$; and when $w = 0$, $T \equiv Q$. In most cases, w is the ratio of sample size of the group from P and the sum of the sample sizes of the two groups (Angoff, 1971). In the equating literature, the mixture is also called the synthetic population.

The NEAT design is widely used in many testing programs. There are some reasons for its popularity. The first reason is that this design requires only one test form to be administered per test date. In many testing situations, it is not possible to give more than one test form at the same administration because of the test security and disclosure concerns. In such situations, the NEAT design is a good choice. Another reason is that with external anchors, non-common items can be disclosed after the test date without compromising future test forms. The ability to disclose test items is important for many testing programs as some states require disclosure of test items. The ability to deal with groups of test takers with different abilities is another advantage of the NEAT design because the groups taking the test at different administrations tend to be self-selected so they usually differ in systematic ways (Petersen, Kolen, & Hooever, 1989).

Various equating methods can be employed in the NEAT design. This dissertation focused on four non-linear methods: two equipercentile OSE methods and two IRT equating methods. Details of these methods are discussed in the following sections.

2.3. Equipercentile OSE methods under the NEAT design

2.3.1. General framework

Equating methods can be classified into two major categories: observed score equating (OSE) methods, and IRT equating methods. Among OSE methods, equipercentile equating methods are the most important (von Davier, Holland, & Thayer, 2004a) and they are widely

used in testing practice (Brennan, 2010). This study focused on two popular equipercentile OSE methods.

Equipercentile OSE methods focus on the distributions of observed scores on the two test forms being equated. These methods equate the quantiles of those score distributions on the two forms. In other words, in equipercentile OSE, scores on two forms are considered to be equivalent if their corresponding percentile ranks in some groups are equal (Angoff, 1971). The equipercentile equivalence of a score from Form Y on the scale of Form X is calculated by first finding the percentile rank on Form Y of a score *y*, and then finding the score *x* on Form X associated with that percentile rank.

Formally, general equipercentile equating framework can be described as follows. Let *X* and *Y* represent scores on Form X and Form Y, respectively, and *Y* is equated to the scale of *X*. The equipercentile equating transformation is a function from the scale of possible values of *Y* to the scale of *X*, that is, from *y* to *x*. The transformation $\varphi(y)$ equates the quantiles of the two population distributions for *X* and *Y* using their cumulative distribution functions, $F_Y(y)$ and $F_X(x)$, on the target population *T*. The transformation function is

$$\varphi(y) = F_X^{-1}\left(F_Y(y)\right) \tag{2.2}$$

where F_X^{-1} represents the inverse function of F_X.

Because actual scores are discrete rather than continuous, a procedure is required to approximate a continuous distribution of score. This procedure is called continuization (von Davier, Holland, & Thayer, 2004a). Some methods have been used including linear interpolation and Gaussian kernel smoothing (for more details see Kolen & Brennan, 2004; von Davier, Holland, & Thayer, 2004a).

From Equation 2.2, in order to conduct equipercentile equating, two cumulative distribution functions of the scores on the two test forms being equated on the target population must be defined. In the NEAT design, because of the missing data of X_U on Q and Y_U on P (see Table 2.1), some assumptions have to be made to compute $F_X(x)$ and $F_Y(y)$. Two popular equipercentile equating methods in the NEAT design were investigated in this study. They are discussed next.

2.3.2. Frequency estimation equipercentile equating method (FE)

The frequency estimation equipercentile equating method (FE) consists of two steps. The first step is to estimate score distributions for each test form on a target population *T*, which is usually a synthetic population as defined in (2.1). The second step is to derive the equating function using the estimated score distributions obtained from the first step and the equipercentile equating framework (2.2).

The score distributions of the two test forms in the target population are estimated as

$$f_X(x) = w.f_{X.P}(x) + (1-w).f_{X.Q}(x) \quad (2.3)$$

$$f_Y(y) = w.f_{Y.P}(y) + (1-w).f_{Y.Q}(y) \quad (2.4)$$

where *f* represents the population frequency distribution and *w* represents the weight given to population *P*.

Because of the characteristics of the NEAT design, as seen in Table 2.1, $f_{X.Q}(x)$ and $f_{Y.P}(y)$ are not available from the observed data. Therefore, some statistical assumptions need to be made to obtain score distributions in the target population. The FE method assumes that the conditional distributions of *X* and *Y*, conditioning on the anchor score *A*, are population

independent. That is,

$$f_{X|A.P}(x \mid a) = f_{X|A.Q}(x \mid a) \tag{2.5}$$

$$f_{Y|A.P}(y \mid a) = f_{Y|A.Q}(y \mid a) \tag{2.6}$$

where *a* is a particular value of *A*.

Combining (2.3) with (2.5), (2.4) with (2.6), it follows that

$$f_X(x) = w.f_{X.P}(x) + (1-w).\sum_a f_{X|A.P}(x \mid a) f_{A.Q}(a) \tag{2.7}$$

$$f_Y(y) = w.\sum_a f_{Y|A.Q}(y \mid a) f_{A.P}(a) + (1-w).f_{Y.Q}(y) \tag{2.8}$$

where $f_{A.P}(a)$ and $f_{A.Q}(a)$ are the marginal distributions of *A* in *P* and *Q*, respectively. All quantities on the right hand sides of (2.7) and (2.8) are observable from the NEAT design. From $f_X(x)$ and $f_Y(y)$, the cumulative distributions $F_X(x)$ and $F_Y(y)$ can be derived for Form X and Form Y, respectively. Equipercentile equating is then applied to $F_X(x)$ and $F_Y(y)$ using (2.2).

Although the FE method is theoretically appealing, it was found to produce larger equating bias in comparison to the other methods in the NEAT design, especially when the group differences are substantial (e.g., see Holland, von Davier, Sinharay, & Han, 2008; Wang, Lee, Brennan, & Kolen, 2008). One reason for this disadvantage might be that the assumptions about missing data made in the FE method is too strong and does not always hold in practical situations (Sinharay & Holland, 2010).

2.3.3. Chain equipercentile equating method (CE)

The chain equipercentile equating method (CE) (Dorans, 1990; Kolen & Brennan, 2004)

is another popular OSE method used in the NEAT design. It is also called the Design V method (Angoff, 1971; Braun & Holland, 1982; Harris & Kolen, 1990).

The CE method consists of three sequential steps. In the first step, Form Y score *y* is equated to the anchor score *a* in population *Q* using the equipercentile equating method, resulting in an equating function

$$\varphi_{YA.Q}(y) = F_{A.Q}^{-1}\left(F_{Y.Q}(y)\right) \tag{2.9}$$

where $F_{A.Q}^{-1}$ represents the inverse cumulative function of *A* in *Q* and $F_{Y.Q}$ represents the cumulative function of *Y* in *Q*.

In the second step, the anchor score *A* is equated to Form *X* score *x* in population *P*, producing an equating function

$$\varphi_{AX.P}(a) = F_{X.P}^{-1}\left(F_{A.P}(a)\right) \tag{2.10}$$

where $F_{X.P}^{-1}$ represents the inverse cumulative function of *X* in *P* and $F_{A.P}$ represents the cumulative function of *A* in *P*.

Finally, *Y* is equated to *X* through a chain of the two equipercentile equating functions

$$\varphi(y) = \varphi_{AX.P}(\varphi_{YA.Q}(y)) = F_{X.P}^{-1}(F_{A.P}(F_{A.Q}^{-1}(F_{Y.Q}(y)))) \ . \tag{2.11}$$

In comparison with the FE method, the CE method is easier and less computationally intensive to implement because it does not require consideration of the joint distribution of total score and anchor score (Kolen & Brennan, 2004). It can use marginal distributions of *X* and *A* for the examinees taking Form X and the marginal distributions of *Y* and *A* for the examinees taking Form Y.

However, the CE method has theoretical shortcomings. The method involves

equipercentile equating between a long test (total test) and a short test (anchor). Theoretically, test forms of unequal lengths, thus unequal reliabilities, cannot be equated in the sense that their scores can be used interchangeably. Another problem is that the CE method does not clearly determine the target population (Braun & Holland, 1982). The CE method consists of two equating procedures performed on two different groups but it is not clear how the groups are combined. However, the CE method does not require equivalent groups so it can be helpful when group differences exist. Equating research has found that the CE method produces smaller bias than the FE method when group differences are large (e.g., see Holland, von Davier, Sinharay, & Han, 2008; Wang, Lee, Brennan, & Kolen, 2008).

2.4. Presmoothing score distributions using log-linear models

In the NEAT design, there are two observed bivariate score distributions, one for the pair (X, A) of Form X and the other for the pair (Y, A) of Form Y. Those distributions are obtained from samples of examinees taking the two forms. The sample score distributions are usually irregular, particularly at the extremes of the score range. The irregularities are primarily due to the random errors in sampling examinees from the population of the test takers. This may, especially when the sample sizes are small, result in unstable and inaccurate equating functions (Liou & Cheng, 1995). To mitigate these effects, smoothing sample score distributions prior to equating is often recommended (Hanson, 1991, Kolen & Brennan, 2004; Rosenbaum & Thayer, 1987; van der Linden & Wiberg, 2010). This process is called presmoothing since it is conducted prior to equating. The purpose of presmoothing is to smooth out some of sampling variability to produce more stable score distribution estimates. The resulting smoothed distributions are then used to equate test forms.

It has been found that for small samples presmoothing can reduce equating error. When

the samples are large, presmoothing may not produce a large improvement, but it may be a useful way to remove undesired roughness in the sample score distributions (Hanson, Zeng, & Colton, 1994; Livingston, 1993; Livingston & Feryok, 1987).

Various presmoothing methods are available to psychometricians. Among the popular models used in presmoothing are the log-linear models, the beta binomial models, and the four-parameter binomial models. The log-linear models were used in this study because they are very flexible in the sense that they can potentially fit a wider class of bivariate distributions. The log-linear models are discussed in more details in Holland & Thayer (2000).

The log-linear models considered in this study are those used to produce a smoothed version of a bivariate distribution of total test score and anchor score such as (*X*, *A*) for Form X or (*Y*, *A*) for Form Y.

Assume that possible values for *X* and *A* are x_i (*i*=1,…,*I*) and a_j (*j*=1,…,*J*) respectively. The vector of observed bivariate frequencies, $\boldsymbol{n} = (n_{11},...,n_{IJ})'$, sums to the total sample size, *N*. The following log-linear model can be used to fit a bivariate distribubion to the observed distribution of (*X*, *A*)

$$\log_e(p_{ij}) = \beta_0 + \sum_{c=1}^{C} \beta_{xc} x_i^c + \sum_{d=1}^{D} \beta_{ad} a_j^d + \sum_{e=1}^{E} \sum_{f=1}^{F} \beta_{xaef} x_i^e a_j^f \qquad (2.12)$$

where p_{ij} is the expected joint score probability of the pair (x_i, a_j) (x_i on *X*, a_j on *A*), β_0 is a normalizing constant that forces the sum of the expected probability p_{ij} to equal 1, and the remaining βs are free parameters to be estimated in the model-fitting process.

This model produces a smoothed bivariate distribution that preserves *C* moments in the

marginal (univariate) distribution of *X*; *D* moments in the marginal (univariate) distribution of *A*; and number of cross moments in the bivariate (*X*, *A*) distribution determined by *E* and *F*. For example, a model with *C*=*D*=2, *E*=*F*=1 (denoted as model 2211) will preserve the first two univariate moments (i.e., mean and standard deviation) of *X* and *A* as well as the first cross moment (i.e., covariance) between *X* and *A*.

The observed bivariate (*Y*, *A*) distribution can be fit by a log-linear model in a similar procedure.

2.5. Item response theory (IRT) equating methods under the NEAT design

Item response theory (IRT) equating methods are used in many testing programs. In this section, two commonly used IRT equating methods employed in this study are discussed.

2.5.1. Three-parameter logistic model

IRT consists of a family of probabilistic models that relate examinee's proficiency level θ to the probability of answering an item within a particular category (Lord, 1980). For dichotomously scored items, there are only two response categories, correct and incorrect.

Various IRT models have been developed for dichotomously scored items as well as for polytomously scored items. The general and commonly used IRT model for dichotomous items is Birnbaum's three-parameter logistic (3PL) model (Lord & Novick, 1968).

Under the 3PL model, the probability that an examinee, with latent ability θ_j, scores a correct response, $u_{ij} = 1$, to item i, is

$$p_{ij} = p(u_{ij} = 1 \mid \theta_j; a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i(\theta_j - b_i)\right]} \qquad (2.13)$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the item

guessing parameter, and *D* is the scaling constant equal to 1.7.

In practice, item and examinee parameters are estimated from data (i.e., examinees' responses to test items).

2.5.2. IRT scale linking

When using the NEAT design, IRT item and ability parameters are typically estimated separately for the two test forms, resulting in two different ability scales. However, in order to perform IRT applications, parameters must be on the same scale. This problem can be solved by a process called scale linking, or simply, linking.

In the 3PL model, the two scales *X* and *Y* have a linear relationship

$$\theta_X = S\theta_Y + I \tag{2.14}$$

If the 3PL model perfectly holds, parameters of common items have the following relationship

$$a_{Xi} = \frac{a_{Yi}}{S} \tag{2.15}$$

$$b_{Xi} = Sb_{Yi} + I \tag{2.16}$$

$$c_{Xi} = c_{Yi} \tag{2.17}$$

where i indices a common item. If the model holds perfectly and item parameters are known, the true linking coefficients *S* and *I* can be obtained from any one of the common items. In practice, equations (2.15), (2.16), and (2.17) are not satisfied for all common items. Thus, a linking process is needed to estimate *S* and *I*. Various linking procedures are available (see Kolen & Brennan, 2004, for more details about IRT linking methods). Four linking methods are often used in research and practice: mean/sigma (Macro, 1977), mean/mean (Loyd & Hoover, 1980), Haebara (Haebara, 1980), and Stocking-Lord (Stocking & Lord, 1983). In this study, the

Stocking-Lord method was used. This method estimates *S* and *I* by minimizing the difference between the test characteristic curves for the anchor associated with two sets of anchor item parameter estimates obtained from two separate calibrations, one for each form. After *S* and *I* are estimated, equations (2.14)-(2.17) can be applied to ability estimates of Form Y and to all non-common item parameter estimates of Form Y to place them on the scale of Form X.

Once item parameter estimates of the two forms are on the same scale, equating can be conducted. Details of IRT equating methods can be found in Kolen and Brennan (2004) and Lord (1980). The following sections briefly present two commonly used IRT equating methods which were used in this study.

2.5.3. IRT true score equating method (TS)

In the 3PL model, the number-correct true scores on Form X and Form Y associated with ability θ are defined through their test characteristic function (Lord, 1980) as, where the summations are over items in Form X and Form Y, respectively

$$\tau_X(\theta) = \sum_{i:X} p_i(\theta; a_i, b_i, c_i) \tag{2.18}$$

$$\tau_Y(\theta) = \sum_{j:Y} p_j(\theta; a_j, b_j, c_j) \tag{2.19}$$

where *p* represents the probability of getting the correct response as presented in (2.13)

In the 3PL model, very low true scores are not available because when $\theta \to -\infty$, $p(\theta) \to c$. The same problem occurs when true score equals to all-correct score because $\theta \to +\infty$. Therefore, the range of true scores on Form X and Form Y are defined as

$$\sum_{i:X} c_i < \tau_X < K_X \tag{2.20}$$

$$\sum_{j:Y} c_j < \tau_Y < K_Y \tag{2.21}$$

where K_X and K_Y are the total numbers of items on Form X and Form Y, respectively.

In IRT true score (TS) equating, the true number-correct score on one form associated with a given θ is considered to be equivalent to the true score on another form associated with the same θ (Kolen & Brennan, 2004). Mathematically, $\tau_X(\theta)$ and $\tau_Y(\theta)$, as computed from (2.18) and (2.19) where the same θ value is used, are considered to be equivalent.

TS equating can be conducted in three steps (Kolen & Brennan, 2004, p. 176):

1. Specify a true score τ_Y on Form Y , $\sum_{j:Y} c_j < \tau_Y < K_Y$

2. Find a value θ that correspond to τ_Y

3. Find the true score on Form X, τ_X , that is associated with θ obtained from step 2.

The second step, which requires solving (2.19) for θ, requires an iterative procedure such as Newton-Raphson as presented in Kolen and Brennan (2004).

Pairing two true scores associated with the same θ values across different θs produces a true-score equating table. This table is then applied in practice to observed number-correct scores. Since true score is not the same as observed score, this step does not have a sound theoretical justification (Lord, 1980).

When using TS equating with observed scores, a procedure is needed for equating scores outside the range of possible true scores described in equations (2.20) and (2.21). Lord (1980) and Kolen (1981) proposed ad hoc procedures to handle this problem. The Kolen's procedure, which was used in this study, is as follows:

1. Set a score of 0 on Form Y equal to a score of 0 on Form X

2. Set a score of $\sum_{j:Y} c_j$ on Form Y equal to a score of $\sum_{i:X} c_i$ on Form X

3. Apply linear interpolation to find equivalents between these points

4. Set of score of K_Y on Form Y equal to a score of K_X on Form X.

Because TS equating is theoretically population invariant and straightforward to implement, it has been used widely in equating research and practice.

2.5.4. IRT observed score equating method (OS)

IRT observed score equating (OS) method consists of two steps. The first step is to estimate the distributions of observed number-correct scores on Form X and Form Y on the target population *T*. The second step is to conduct traditional equipercentile equating on these estimated distributions.

For Form X, the recursion formula presented in Lord and Wingersky (1984) can be used to obtain the conditional distribution of observed scores at each θ value. Define $f_r(x|\theta)$ as the distribution of the number-correct scores over the first r items for examinees with ability θ, and p_r as the probability for those examinees getting r^{th} item correct. For r >1, the recursion formula is as follows (Kolen & Brennan, 2004):

$$\begin{aligned} f_r(x|\theta) &= f_{r-1}(x|\theta).(1-p_r) && x = 0 \\ &= f_{r-1}(x|\theta).(1-p_r) + f_{r-1}(x\text{-}1|\theta).p_r && 0 < x < r \\ &= f_{r-1}(x\text{-}1|\theta).p_r && x = r \end{aligned} \qquad (2.22)$$

To use this recursion formula, begin with r=1 and repeatedly apply the formula by

increasing r on each repetition. The process is stopped after r=K, the number of items on the test form. That is $f_X(x|\theta)$. These conditional observed score distributions are accumulated over the target population T to obtain the marginal distribution of observed scores, which, when the ability distribution is continuous, can be calculated as

$$f_X(x) = \int_{\theta} f_X(x \mid \theta) f_\theta(\theta) \tag{2.23}$$

where $f_\theta(\theta)$ is the distribution of θ. In the NEAT design, the target population is the synthetic population defined in (2.1).

In practice, a posterior distribution of θ, which is estimated from the calibration process, is often used. In such a situation, the marginal distribution can be calculated as

$$f_X(x) = \sum_{\theta} f_X(x \mid \theta) f_\theta(\theta) \tag{2.24}$$

where $f_\theta(\theta)$ represents the posterior weight at the quadrature point θ.

For Form Y, the same procedure is used to obtain $f_Y(y)$. From $f_X(x)$ and $f_Y(y)$, the cumulative distribution functions, $F_X(x)$ and $F_Y(y)$, can be obtained and a conventional equipercentile equating is conducted using framework (2.2).

An advantage of OS equating is that it defines the equating relationship for observed scores and can be applied directly to the observed score. However, OS equating function is population dependent because the target population score distribution has to be specified.

2.6. Equating criteria

One important step in any equating process is to determine whether the scores are really interchangeable and comparable after they were equated. When choosing an appropriate

procedure to evaluate equating results, it is very important to consider the operational definition of equating adopted by the testing program. The goal of equating evaluation should be to determine the degree to which the definition of equating holds. Any evaluation process requires selection of a criterion or criteria. In order to obtain a fair evaluation of equating results, the selected criteria must be closely linked to the adopted definition of equating.

Two basic operational definitions of equating have been proposed in the equating literature: equipercentile definition and equity definition. The following sections provide more details about the two definitions and related criterion.

2.6.1. Equipercentile criterion

The equipercentile definition of equating, also called the definition of observed-score equating, was proposed by Angoff (1971) as "two scores, one on form *X* and the other on form *Y* (where *X* and *Y* measure the same function with the same degree of reliability), may be considered equivalent if their corresponding percentile ranks in any given group are equal" (p. 563). This definition implies that score distributions on the two test forms to be equated must be identical after equating (Kolen & Brennan, 2004, p.12). The equipercentile definition is often referred to as the equipercentile equating property and its corresponding criterion is called equipercentile criterion. Algebraically, the equipercentile criterion requires that

$$F_{Y_e}(y_e) = F_X(x) \tag{2.25}$$

where F_{Ye} and F_X represents cumulative distribution functions (cdf) of Y_e and X, respectively.

In evaluating equating with equipercentile criterion, the cdf of Y_e is compared to the cdf of X. To quantify the difference between two cdfs, the Kolmogorov statistic (Conover, 1999), which is the largest difference between the two cdfs across all score levels, can be used. Another

approach is to use the area between the two cdf curves. The smaller the Kolmogorov statistic or the area between two cdfs, the more accurate the equating is.

2.6.2. Equity criteria

The equity definition of equating, also called the definition of true-score equating, was proposed by Lord (1980) as "if an equating of test *X* and *Y* is to be equitable to each applicant, it must be a matter of indifference to applicants at every given ability level θ whether they are to take test *X* or test *Y*" (p.195). Equity requires that for every θ in the range of interest, the conditional distributions of scores on the two test forms to be equated must be identical after equating.

The equipercentile and equity definitions are not unrelated. If one applies the equipercentile definition to a group of examinees with the same ability θ level, it is equivalent to the equity definition (Divgi, 1981).

The equity definition is commonly referred to as the equity property, or simply equity. Equity is closely related to test fairness. If the two conditional distributions differ, a test taker may be advantaged by taking one test rather than the other. For example, a low-ability test taker with a larger variance of his observed score on Form Y than on Form X has a better chance of passing a certain cutoff score on Form Y than on Form X.

Algebraically, equity property can be stated as

$$F_{Y_e}(y_e \mid \theta) = F_X(x \mid \theta), \text{ for all } \theta \tag{2.26}$$

where '$|\theta$' denotes 'conditioning on θ'.

However, Lord (1980) also showed that equity is never satisfied unless the tests being equated are perfectly reliable or strictly parallel, in which case equating is unnecessary. In

practice, test forms are never perfectly reliable nor strictly parallel. In other words, equity is unlikely to be fully satisfied in practice. Nevertheless, it can be considered as a gold standard for equating in the sense that it represents an ideal equating. Therefore, the issue is not one of determining whether the equity is fully satisfied, because it is never, but rather of describing the extent to which the equity holds, or fails to hold.

Like equipercentile criterion, equity criterion can be used to evaluate equating by comparing two score distributions, one for equated scores of Form Y and one for scores of Form X, but conditioning on θ and the evaluation must be conducted at every θ value in the range of interest. The most popular method for obtaining the conditional distribution is to apply the compound binomial method proposed by Lord and Wingersky (1984) presented in the equation (2.22). The difference between the two conditional cdfs can be quantified by the Kolmogorov statistic or the area between two cdf curves.

Since full equity is unlikely to be satisfied in practice, researchers have proposed some weaker versions focusing on the equivalence of some specific moments instead of on the equivalence of the full conditional distributions of scores. Two popular weakened versions of equity are the first-order equity (Divgi, 1981) and the second-order equity (Morris, 1982).

The first-order equity, also called weak equity as opposed to full equity as strong equity, requires only the first-order moment (i.e., expected value) of the two conditional distributions be the same across all levels of θ. In other words, first-order equity refers to the true scores on two test forms. Algebraically, the first-order equity can be stated as

$$E(Y_e \mid \theta) = E(X \mid \theta), \quad \text{for all } \theta \tag{2.27}$$

where E denotes expected value.

The first-order equity is satisfied when each examinee is expected to obtain the same test score (after equating) regardless of which test form is administered.

The second-order equity requires the second-order moment (i.e., variance or standard deviation) of the two conditional distributions be the same across all θs. That is,

$$\sigma_{Y_e|\theta} = \sigma_{X|\theta}, \qquad \text{for all } \theta \tag{2.28}$$

where σ denotes standard deviation, which is also called standard error of measurement (SEM).

The second-order equity holds if, conditional on ability, examinees have the same SEM on the two forms after equating (Morris, 1982). Kolen, Hanson, and Brennan (1992) argued that the second-order equity should be nearly satisfied in order for the two forms being equated to be used interchangeably.

The first-order equity can be assessed by comparing the means of two conditional distributions $F_{Ye}(y_e|\theta)$ and $F_X(x|\theta)$. Similarly, the standard deviations of these distributions are compared to evaluate the second-order equity.

2.7. Summary of related research

This section reviews prior research relevant to this dissertation. It is divided into two subsections. The first subsection reviews prior research on comparing equating methods used in this study. Although the NEAT design was the focus of this study, several studies conducted on other designs are also included. In the second subsection, research on using equipercentile and equity criteria to evaluate equating is reviewed.

2.7.1. Prior research on comparing equating methods

Research on comparing equating methods is fairly rich but not many studies

simultaneously comparing FE, CE, TS, and OS or comparing between IRT equating and OSE were found in the literature. Most studies reported in the literature compared two, usually FE vs. CE or TS vs. OS, or three methods.

Several studies compared FE and CE using real and simulated data. Braun and Holland (1982), Livingston, Dorans, and Wright (1990), and Marco, Petersen, and Stewart (1983) used real data from admission tests. Sinharay and Holland (2007) used both IRT-based simulated and real data from a certification test. Wang, Lee, Brennan, and Kolen (2008) used several simulated data sets. These studies found that FE and CE produced quite different results. CE method performs better than FE method in terms of smaller bias, especially when group differences are large. When two groups are similar, FE method performs slightly better than CE. However, those studies lacked a clear criterion for evaluation preventing these from providing a clear and conclusive comparison of FE and CE.

In another study, von Davier, Holland, and Thayer (2004b) showed that both FE and CE are special cases of the OSE framework and that they may lead to similar results under special conditions. It is not clear from their conclusions what conditions are deemed special.

Holland, von Davier, Sinharay, and Han (2008) compared FE (called post-stratification method in their study) and CE using a special data set. By manipulating a large data set, they mimicked the NEAT design to test the assumptions of the two methods. They found that CE performed slightly better than FE. However, this study lacked control over differences in group ability and test item difficulty.

Harris and Kolen (1990) used real data from a licensure test to compare FE, CE, and OS. They concluded that FE and OS provide better results than CE but the CE method required less computation.

Recently, Sinharay and Holland (2010) conducted a study comparing FE, CE, and OS methods in the NEAT design. They found that in general, the CE method is somewhat more satisfactory than the others.

Only a few studies comparing TS and OS were found in literature. Kolen (1981) used random samples of examinees to cross-validate his comparisons of IRT methods with traditional equipercentile and linear equating methods. His criterion was the mean squared difference between scores with equivalent percentile ranks in the cross-validation and equated distributions. He found that OS is the best method. However, using a circular equating design as their criterion, Lord and Wingersky (1984) found no difference between two methods. Tsai, Hanson, Kolen, and Forsyth (2001) compared TS and OS in the NEAT using bootstrap technique. Their criterion was standard errors of equating (SEE). They found that when item parameters were estimated separately, both methods produced similar SEE.

Research studies comparing IRT and traditional equating methods appear to be somewhat less prevalent in the literature. Using real data from a college admission test, Harris and Kolen (1986) compared equipercentile and TS methods under the 3PL model in the random groups design. They found that the two methods performed similarly, concluding that both methods were relatively invariant to the group proficiency levels.

Another study by Han, Kolen, and Pohlmann (1997) also used operational test forms in the random groups design. They compared IRT and equipercentile methods using results from equating a test to itself as the criterion. Among major findings are: TS produces most stable results, however the mean differences in equating stability are small; OS produces more stable results than equipercentile, larger equating differences exist when form differences are larger.

Lord (1977) also found that IRT and traditional equipercentile methods produced

different results.

2.7.2. Prior research using equipercentile and equity criteria

Of the two basic criteria, equipercentile criterion has been used widely in research. Because the equipercentile criterion requires the marginal distributions of scores be identical after equating, any study which used an equipercentile method on a large sample as its criterion in fact used the equipercentile criterion, although it may not have explicitly stated. Among studies that used the equipercentile criterion, either implicitly or explicitly stated, to evaluate equating are Holland et al. (2008); Kim et al. (2010); Kim, Brennan, and Kolen (2005); Livingston and Kim (2010); Sinharay and Holland (2007); Skaggs (2005); Tong and Kolen (2004); von Davier et al. (2006).

Although it is one of the most important equating criteria, equity has rarely been used in evaluating equating. This might be due to the fact that the equity is computed based on the conditional distributions of the test scores which are difficult to obtain and interpret. Most studies that used the equity criterion employed the first-order equity or second-order equity.

Harris (1991) used the first-order and the second-order equity criteria to compare Angoff's Design I and Design II in a vertical scaling situation.

In a simulation study with the NEAT design, Thomasson (1993) found that IRT equating methods are superior to traditional equating methods with respect to the first-order and second-order equity, provided that the test data are unidimensional.

In a study comparing IRT equating and beta 4 equating in the random groups design, Kim, Brennan, and Kolen (2005) found that when forms differ, the equipercentile method performs well relative to the equipercentile criterion and poorly relative to first-order equity criterion. They also concluded that the TS method performs well relative to first-order equity and

poorly relative to second-order equity.

In an equating study using random groups design, Tong and Kolen (2005) found that when test forms are similar, equipercentile methods and IRT methods lead to adequate equating regardless of the criterion used. When forms are different, however, they found that TS method performs best in terms of first-order equity and both equipercentile and OS methods perform well in terms of the equipercentile criterion and second-order equity.

In a study on equating evaluation, Bolt (1999) used the first-order and second-order equity criteria to investigate whether the TS method is affected by the presence of multidimensionality. He found that the TS method performs as well as the traditional methods when the correlation between dimensions is high ($\geq .7$) and slightly inferior to the equipercentile method when the correlation is moderate or low ($\leq .5$).

Currently, Wyse and Reckase (prepublished online) developed an index based on the first-order equity and used it to compare various IRT linking methods under the NEAT design. They found that the Stocking-Lord and fixed-parameter methods perform well and the use of concurrent calibration is not recommended.

2.7.3. Summary

It is increasingly obvious from the equating literature that the CE method may be preferable to the FE method when group differences exist. In case of equivalent groups, they may produce similar results. Contradictory results were found regarding the two IRT methods. Some studies found they may produce different results while others found no difference. There is not enough evidence from the literature about how OSE methods and IRT equating methods are compared to each other, especially in the NEAT design.

From a limited number of studies using equiprcentile and equity criteria to evaluate

equating results, the observed score equating methods have been found to perform well relative to the equipercentile criterion. The IRT methods, however, perform well relative to the first- and second-order equity. It is unknown if the full equity criterion has ever been used in evaluating equating results, either in research or in practice, but no research of that kind has been found in the literature, at least up to the time this dissertation was written.

CHAPTER 3

RESEARCH METHOD

This chapter describes the research method used in this study. Main topics include the overall research design, simulation conditions, test form and data generation, equating processes, and procedures for calculating evaluation indices.

3.1. Purpose of the study and research questions

As presented in Chapter 1, the main purpose of this study was to use equipercentile and equity-based criteria to evaluate the performances of four commonly used equating methods under the NEAT design. Those methods are:

- Presmoothed frequency estimation equipercentile method (FE)
- Presmoothed chain equipercentile method (CE)
- IRT true score equating method (TS)
- IRT observed score equating method (OS)

In addition, identity equating method was also employed to examine possible conditions when no equating is preferred. The performance of those equating methods was investigated in various conditions of test form and group differences. Particularly, the study aimed to address the following research questions:

Question 1: Overall, how do the equating methods compare to one another in terms of equipercentile and equity criteria?

Question 2: How do test form differences affect equating results for each method?

Question 3: How do group differences affect equating results for each method?

Question 4: Are there interaction effects between test form differences and group

differences for each method?

Question 5: In what conditions is the identity equating preferred to the others?

3.2. Overall research design

3.2.1. General framework

The general framework of this study consists of the following main points:

- The NEAT design was modeled. Two test forms with equal lengths, which share internal anchor items, were administered to two samples with equal sample sizes, randomly drawn from two different populations, *P* and *Q*.
- The old form, Form X, was fixed. The new form, Form Y, was varied to simulate form differences.
- Form Y number-correct (NC) scores were equated to Form X NC scores on the simulated sample data using the investigated methods. An equating function φ was obtained for each method.
- Marginal and conditional population distributions of NC scores on the two forms were generated using the adopted IRT model to be used in assessing criteria. The cumulative distribution functions of those distributions (on target population *T*) are denoted as $F_X(x)$, $F_Y(y)$, $F_X(x|\theta)$, and $F_Y(y|\theta)$.
- Each obtained equating function φ was applied to the population NC score distributions of Form Y to obtain the distributions $F_{Ye}(y_e)$ and $F_{Ye}(y_e|\theta)$ of the equated score y_e, where $y_e = \varphi(y)$.
- $F_{Ye}(y_e)$ was compared to $F_X(x)$, and $F_{Ye}(y_e|\theta)$ was compared to $F_X(x|\theta)$ to evaluate

equipercentile and equity criteria, respectively.

Details of those steps are presented in the following sections. Many factors may affect equating results in the NEAT design. Some factors were fixed while others were varied in this study.

3.2.2. Data source

Simulated data were used in this study to reduce possible effects of model misfit (Davey, Nering, & Thompson, 1997) and to facilitate manipulation of various experimental conditions. Data were simulated in such a way as to exhibit a wide range of form and group differences in the NEAT design.

3.2.3. IRT model

The 3PL model for dichotomous items was adopted for data simulation, IRT equating, and criterion evaluation. This model was chosen because it fits common testing situations with multiple-choice tests when guessing is possible and items may have different discrimination power. In the 3PL model, the probability of a correct response to item *i* by examinee *j* is given by

$$p_i(\theta_j) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta_j - b_i)]}{1 + \exp[Da_i(\theta_j - b_i)]} \tag{3.1}$$

where a_i, b_i, and c_i are item difficulty, discrimination, and guessing parameters, respectively, and *D* is the scaling constant equal to 1.7.

3.2.4. Fixed factors

Some factors were fixed to reduce the number of conditions and sources of random variations.

- *Test length*. Each form had 60 dichotomous items. This length is typical in practice

and research.

- *Anchor items*. The anchor was internal and had 20 items, which was one third of the total test. The anchor was created as a miniversion of the total test in terms of statistical characteristics. According to Kolen and Brennan (2004), the anchor should have at least 20% of the test items and should be a mini version of the test in order to function well.
- *Sample size*. Each group had 2,000 examinees, randomly drawn from the two populations. This sample size was large enough to provide stable estimation of all four equating methods (see Hulin, Lissak, & Drasgow, 1982; Jarjoura & Kolen, 1985). It was also within the range used in other studies (e.g., Hanson, & Beguin, 2002; Wang et al., 2008).

3.2.5. Varied factors

Two main factors were manipulated in this study to simulate common conditions in the NEAT design. The magnitude of variations was chosen to demonstrate a wide range of conditions, some of which might have been extreme. This allowed the examination of how equating methods performed in extreme cases rather than just in typical cases.

- *Test form difference*. Form X was fixed throughout the study. In order to simulate form difference, Form Y was varied. Since Form X was fixed and the anchor A was shared by the two forms, thus, fixed, the set of unique (non-common) items of Form Y, denoted as Y_U, was varied. At the beginning, Y_U was simulated to be similar to X_U, the set of unique items of Form X. To simulate the difference in item difficulty between the two forms, a constant Δb was added to the b-parameter of all items of the

initial Y_U. To simulate the difference in item discrimination power between two forms, a constant Δa was multiplied with the a-parameter of all items of the initial Y_U. Seven values of Δb (-1.2, -.8, -.4, 0, .4, .8, 1.2) and three values of Δa (.5, 1, 2) were used. Using these values made Y_U either easier or more difficult and either less or more discriminating relative to X_U. All levels of Δa and Δb were fully crossed, resulting in 21 (3x7) different conditions of form difference, denoted as (Δa,Δb).

- *Group difference.* Population P θ distribution was fixed standard normal N (0, 1) throughout the study. Population Q θ distribution was also normal with unit variance but mean $\Delta\mu$, that is N ($\Delta\mu$, 1). The ability difference between two populations was quantified by $\Delta\mu$. Five different values of $\Delta\mu$ were used: 0, .25, .5, .75, 1. Thus, Q was set to be as equally able as or more able than P. In test equating, mean differences larger than .5 are generally considered very large. Some values of $\Delta\mu$ used in this study were very large, but the purpose was not only to look at typical cases but also to see how equating performs in extreme cases.

3.2.6. Simulation conditions

Form and group difference conditions were fully crossed, resulting in 105 different (21x5) conditions. These conditions are referred to by a combination of three numbers ($\Delta\mu$,Δa,Δb). For example, in the condition (.5, 2, .-4), the mean difference between two populations was half of the standard deviation,Y_U a-parameters were twice as large as those of X_U, and Y_U b-parameters were .4 less than those of X_U.

3.2.7. Equating methods

Five equating methods were used in this study

- Presmoothed frequency estimation equipercentile equating (FE)
- Presmoothed chain equipercentile equating (CE)
- IRT true-score equating (TS)
- IRT observed-score equating (OS)
- Identity equating (IE)

All equating methods were conducted following the description of equating methods described in Chapter 2. In the FE and CE methods, NC scores were presmoothed by a loglinear model before equating.

Because the two groups had equal sample sizes (2,000), equal weights were given to each population in defining the target population *T*.

3.2.8. Replications

For each of 105 ($\Delta a, \Delta b, \Delta \mu$) simulation conditions, 50 replications were used. Note that within each condition, two forms were fixed across all replications. This served to eliminate possible random effects due to sampling different forms within each condition.

3.3. Test form generation

As presented previously, each test form consisted of two blocks. Form X consisted of common-item block A and unique-item block X_U. Similarly, Form Y consisted of block A and unique-item block Y_U. X_U and A were fixed throughout the study, Y_U was varied to simulate form differences. X_U and Y_U consisted of 40 items. Anchor A consisted of 20 items. The test forms were generated as follows

Table 3.1. Descriptive statistics for item parameters of three initial blocks

Item block	Parameter	*N*	*M*	*SD*	*Min*	*Max*
Anchor items (*A*)	*a*	20	0.936	0.368	0.443	1.632
	b	20	-0.022	0.837	-1.704	1.585
	c	20	0.163	0.054	0.058	0.273
Form *X* unique items (X_U)	*a*	40	0.928	0.261	0.368	1.651
	b	40	-0.020	0.763	-1.765	1.660
	c	40	0.147	0.044	0.026	0.263
Form *Y* unique items (Y_U)	*a*	40	0.907	0.253	0.392	1.591
	b	40	-0.077	0.678	-1.648	1.744
	c	40	0.138	0.054	0.052	0.275

- First, three initial blocks of X_U, Y_U, and A were generated. The *a*-, *b*-, and *c*-parameters were randomly sampled from lognormal LN(-.15, .30), normal N(0, .7), and beta BETA(7, 43) distributions, respectively. Note that although the same distributions were used, item parameters for the three initial blocks were generated independently. X_U and A were combined to produce Form X which was fixed throughout the study. The descriptive statistics of item parameters of the three initial blocks are presented in Table 3.1. As shown in Table 3.1, the three initial blocks were similar.

- Block Y_U was varied by adding a constant Δb to all b-parameters and multiplying a constant Δa with all a-parameters of the initial Y_U. Seven values of Δb (-1.2, -.8, -.4, 0, .4, .8, 1.2) and three values of Δa (.5, 1, 2) were used. All levels of Δa and Δb were fully crossed, resulting in 21 (3x7) different Y_U blocks. Each Y_U block was combined with block A to create one Form Y, which then was paired with Form X for equating. Therefore, there were 21 XY pairs representing 21 conditions of form difference.

3.4. Data simulation

Data were simulated using the 3PL model. Data for the group taking Form X were simulated in the following steps:

- 2,000 simulees were randomly drawn from the θ distribution of the population P, which was $N(0,1)$.
- The probability of a correct response to item i by simulee j, $p_i(\theta_j)$, was calculated by the equation (3.1) using item parameters of Form X. The calculated $p_i(\theta_j)$ was then compared to a randomly drawn number from a uniform [0,1] distribution. If $p_i(\theta_j)$ was greater than the random number, the answer to that item by that simulee was coded as 1 (correct). Otherwise, it was coded as 0 (wrong). This process was repeated for all items in the form and all simulees in the sample.
- The NC scores for the whole test and for the anchor items were calculated by summing item scores for all items and for the anchor items only, respectively. Therefore, each simulee had a string of binary responses to all items, a total score x and an anchor score a.

Similarly, data for the group taking Form Y were simulated using 2,000 simulees randomly drawn from the population *Q* and item parameters of Form Y.

For each group, simulated data were divided into two different data sets. One data set contained only binary item responses to be used for IRT equating. The other data set had total and anchor NC scores to be used in FE and CE methods.

Therefore, for each replication, four data sets were produced. Because there were 105 simulation conditions, each had 50 replications, 21,000 data sets were simulated (105x50x4).

3.5. Equipercentile equating procedures

Two equipercentile OSE equating methods commonly used in the NEAT design, FE and CE, were used in this study. Prior to equating, loglinear presmoothing techniques were applied to smooth out irregularities of the sample observed score distributions. Two bivariate distributions (X, A) and (Y, A) were smoothed using several different models in a trial. A loglinear model which preserved the first two univariate moments and the first cross-moment was selected due to good smoothing results. This model was used for all observed-score equating processes in this study to avoid possible effects due to using different loglinear models.

Since the two groups had equal sample sizes, equal weights were given to two populations in defining the target population.

FE and CE equating processes were conducted using the *R*-package *equate* (Albano, 2010). This package, which was written in *R* language, can perform various equating methods in the NEAT design with embedded loglinear presmoothing functions.

3.6. IRT equating procedures

Two IRT common equating methods, TS and OS, were used in this study. The IRT equating consisted of three sequential steps: calibration, scale linking, and equating.

3.6.1. Calibration

Two simulated binary data sets of a XY pair were used for IRT equating. Each data set was calibrated separately using the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). From the calibration, item parameter estimates were obtained. In addition, 40 equally spaced θ quadrature points were also obtained from each calibration. These quadrature points and their relative frequencies provided an estimated posterior distribution of ability, sometimes referred to as the true θ distribution, in the corresponding population. These quadrature points and their weights were needed for the OS method.

3.6.2. Scale linking

Because parameter estimates were obtained separately for Form X and Form Y, they were not on the same scale due to the model indeterminacy problem (see Chapter 2 for more details). Therefore, a linking process was performed to place item and ability estimates for Form Y on the scale of Form X. Particularly, parameter estimates of Form Y were linked to the scale of Form X using the common Stocking-Lord (Stocking & Lord, 1983) linking method. This process was conducted using the computer program ST (Hanson, Zeng, revised by Cui, 2004).

3.6.3. Equating

After parameter estimates were placed on the same scale, TS and OS methods were performed. Details about these methods were presented in Chapter 2.

For the TS method, Kolen's (1981) ad hoc procedure was used for the scores below the

limits set by the item guessing parameters and for the all-correct score.

For the OS method, the θ distribution for the target population had to be defined. As previously mentioned, equal weights were given to two populations *P* and *Q*, resulting in the posterior θ distribution for the target population *T* as

$$\psi(\theta) = .5 \times \psi_P(\theta) + .5 \times \psi_Q(\theta) \tag{3.2}$$

where $\psi_P(\theta)$, $\psi_Q(\theta)$ were the weights for the quadrature point θ obtained from BILOG-MG calibrations for *P* and *Q*, respectively, and $\psi(\theta)$ was the weight for the target population *T*. $\psi(\theta)$ was used in the place of $f_\theta(\theta)$ in equation (2.24) for OS equating.

Both TS and OS equating processes were conducted using the computer program PIE (Hanson & Zeng, revised by Cui, 2004)

3.7. Procedures for assessing criteria

3.7.1. Equating criteria

Four equating criteria were used to evaluate the accuracy of equating results in this study:

- The equipercentile criterion
- The full equity criterion
- The first-order equity criterion
- The second-order equity criterion

3.7.2. Population score distributions

In order to assess those criteria, the following distributions were required:

- The conditional population distributions of NC score on Form X and Form Y at each

predetermined θ value in the range of interest.

- The marginal population distributions of NC score on Form X and Form Y.

Those distributions were defined for populations. They were created by using the generating (true) item parameters, population θ distributions instead of the item and ability parameter estimates obtained from the calibration of the simulated sample data. For simplicity in the subsequent sections, the term 'population' is dropped when it does not cause any confusion. In this study, 41 equally spaced θ values in the interval [-4, 4] were used. The following sections describe the ways those distributions were produced and used to evaluate the criteria.

Conditional distributions.

Conditioning on a given θ, the NC score distribution of Form X, denoted as $f_X(x|\theta)$, was produced by applying the Lord and Wingersky's (1984) recursive formula (as presented in equation (2.22)), using the generating item parameters of Form X and the 3PL model. Similarly, the conditional NC score distribution of Form Y, $f_Y(y|\theta)$, was produced.

Marginal distributions.

The conditional NC score distributions of x were integrated across the θ distribution to produce the marginal NC score distribution of x, denoted as $f(x)$. The marginal NC score distribution for Form Y, $f(y)$, was produced in a similar process. That is,

$$f(x) = \int_{\theta} f(x \mid \theta) f(\theta) d\theta \tag{3.3}$$

$$f(y) = \int_{\theta} f(y \mid \theta) f(\theta) d\theta \tag{3.4}$$

where $f(\theta)$ represents the probability distribution function of θ in the target population T. These

integrations were done numerically using 41 quadrature points θ with corresponding weight $\psi(\theta)$ computed from the θ distribution for the target population.

$$f(x) = \sum_{\theta} f(x \mid \theta)\psi(\theta) \tag{3.5}$$

$$f(y) = \sum_{\theta} f(y \mid \theta)\psi(\theta) \tag{3.6}$$

The term $\psi(\theta)$ in (3.5) and (3.6) were computed in two steps:

- Calculating the probability distribution function (pdf) for θ in the target population T with equal weights given to two populations P and Q as

$$f(\theta) = .5 f_P(\theta) + .5 f_Q(\theta) \tag{3.7}$$

  where $f_P(\theta)$ and $f_Q(\theta)$ represents the pdfs of θ on P and Q as specified in each simulation condition.

- Calculating $\psi(\theta)$ using the following formula

$$\psi(\theta) = \frac{f(\theta)}{\sum_{\theta} f(\theta)} \tag{3.8}$$

3.7.3. Evaluation indices

As mentioned previously, each equating method produced an equating function $\varphi(y)$ to equate Form Y to Form X. Those equating functions were applied to the conditional and marginal NC score distributions of Form Y, as described in Section 3.8.2, to obtain the marginal and conditional distributions for the equated Form Y score, denoted as Y_e. Those distributions of Y_e were compared to the distribution of X to evaluate equating results with different criteria.

Particularly, the marginal distribution of Y_e was compared to the marginal distribution of X to assess equipercentile criterion. The conditional distribution of Y_e at a specific θ was compared to the conditional distribution of X at the same θ to assess equity criteria. Details on evaluation indices follow.

Equipercentile index.

Equipercentile property requires that marginal distributions of X and Y_e be identical. To determine how this requirement was satisfied, the area between the two cumulative distribution function (cdf) curves of X and Y_e was used the evaluation index for equipercentile criterion. This index was denoted as EP. The equating method that produced the smallest area was considered to best preserve the equipercentile property. Because X and Y_e were discrete, their cdfs were step functions. Therefore, the area between the two cdf curves was the sum of the areas of the rectangles between the two curves.

To illustrate how the index EP was calculated, an example similar to the one used by Kim (2000) is presented here. Let's assume Form X and Form Y each consist of three items. Therefore, the possible NC scores on the two forms are 0, 1, 2, and 3. Table 3.2 shows scores x, y, and y_e along with cumulative distributions of X and Y_e. Note that the scales of X and Y are equal. In this example, there are eight discrete scores (four scores for X and four scores for Y_e) because none of y_e and x are identical. Figure 3.1 shows the cdfs of X and Y_e and the area between them. The EP index can be calculated as

$$\text{EP} = \sum_{i=1}^{7} A_i = (.3)(.2) + (.7)(.1) + (.4)(.2) + (.6)(.1) + (.5)(.1) + (.5)(.1) + (.4)(.2) = .45$$

Table 3.2. Illustrative example: x, y, y_e, and cumulative distributions

x or y	y_e	$F(X)$	$F(Y)$ or $F(Y_e)$
0	0.3	0.2	0.3
1	1.4	0.5	0.6
2	2.5	0.7	0.8
3	3.4	1.0	1.0

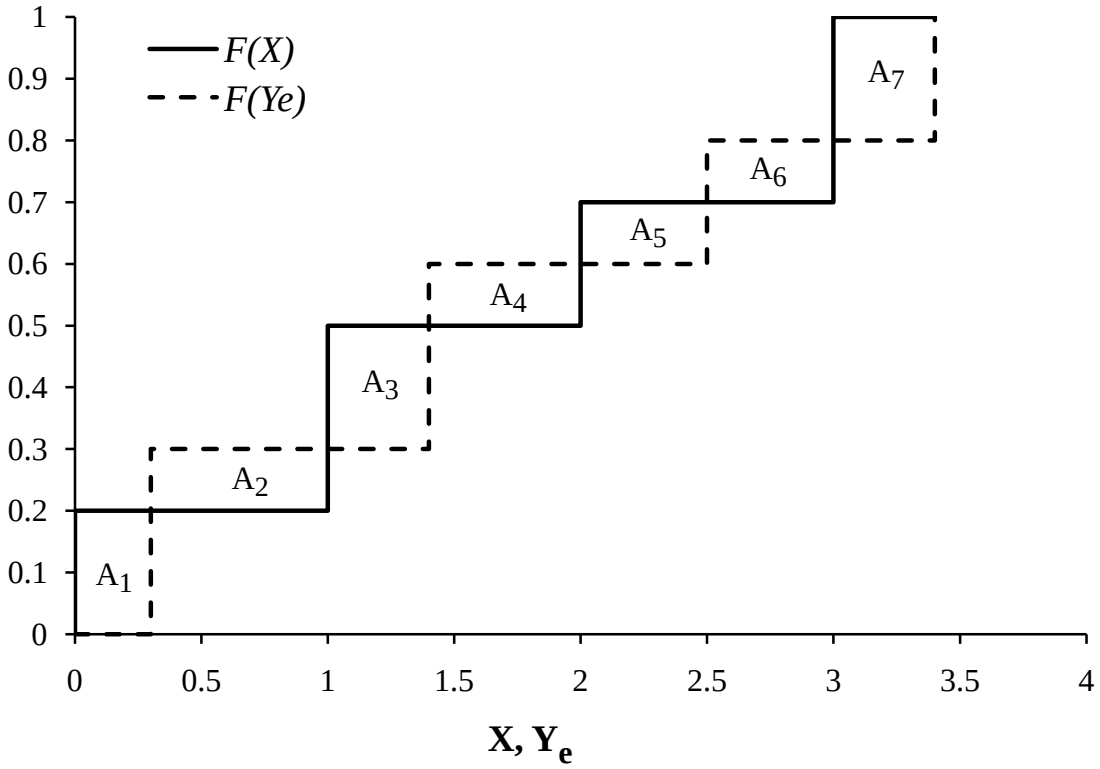

Figure 3.1: Illustrative example of area between cumulative distribution functions of X and Y_e

Equity indices.

Full equity property requires that conditioning on θ, distributions of X and Y_e be identical. Like in the case of equipercentile criterion, the area between two (conditional) cdf curves of X and Y_e was calculated, denoted as EP_θ. The EP_θ was calculated at each of 41 quadrature points θ and then weighted across all θ to create the evaluation index for the full equity criterion, denoted as E. That is,

$$\mathrm{E} = \sum_{\theta} EP_\theta \times \psi(\theta) \tag{3.9}$$

where $\psi(\theta)$ was defined previously.

For the first-order equity criterion, the difference between the expected values of X and Y_e at each θ was calculated, then weighted across all θ to create the index for the first-order equity criterion, denoted as E_1. That is,

$$\mathrm{E}_1 = \sum_{\theta} \left| E(Y_e \mid \theta) - E(X \mid \theta) \right| \times \psi(\theta) \tag{3.10}$$

Similarly for the second-order equity criterion, the index E_2 was calculated as

$$\mathrm{E}_2 = \sum_{\theta} \left| \sigma(Y_e \mid \theta) - \sigma(X \mid \theta) \right| \times \psi(\theta) \tag{3.11}$$

where σ denotes standard deviation.

3.8. Simulation steps within each condition

In each of 105 simulation conditions, after test forms were produced, the following steps were carried out:

1. Calculating 41 conditional NC score distributions of Form X and 41 conditional NC score distributions of Form Y at 41 values of θ.

2. Calculate marginal NC score distributions of Form X and Form Y using the conditional distributions obtained from step 1.

3a. Simulating data for two groups from *P* and *Q* populations, taking Form X and Form Y, respectively.

3b. Performing equating using FE, CE, TS, and CE methods, resulting in four equating functions.

3c. Applying the equating functions obtained from step 3b and the identity equating to the conditional and marginal distributions of *Y* to produce the distributions of equated score Y_e.

3d. Calculating evaluation indices EP, E, E_1, and E_2.

4. Repeating steps 3a through 3d 50 times.

CHAPTER 4

RESULTS

This chapter presents the results obtained from the study. For completeness and convenience, the results are presented in both numerical and graphical forms.

The chapter begins with a quick review of the research purpose and research questions, followed by a review of the evaluation indices used to evaluate equating results based on the equipercentile and equity criteria. A general framework for presenting the results is described next. After that, the results are presented. The chapter concludes with a brief summary of the main findings.

4.1. Review of research purpose and questions

This study evaluated the performance of four commonly used equating methods in the NEAT design, using evaluation criteria related to the equipercentile and equity definitions of equating as presented in previous chapters. The methods under investigation are the presmoothed frequency estimation equipercentile (FE), the presmoothed chain equipercentile (CE), the IRT true score equating (TS), and the IRT observed score equating (OS). In particular, the study aimed to address several research questions on how those methods perform across all studying conditions, how difference between test forms being equated and difference between groups taking the test forms affect the equating results, and whether no equating should be done in some certain circumstances because doing equating in those cases may introduce more errors than it may removes.

4.2. Review of evaluation indices

The evaluation indices used in this study are based on two common definitions of equating. Let X and Y represent scores on the two forms to be equated. Y is equated to the scale of X, resulting in equated score Y_e. The equipercentile definition requires the marginal distributions of X and Y_e be identical. The equity definition requires that conditioning on ability θ, the distributions of X and Y_e be identical. Two weaker forms of equity are the first-order equity and the second-order equity which require the equivalence of expected values and of standard deviations of those two conditional distributions.

Based on those definitions, four evaluation indices were developed and used in this study. They are the equipercentile index EP, the full equity index E, the first-order equity index E_1, and the second-order equity index E_2, each reflects the degree the corresponding property holds. The smaller the index value, the better the associated property holds. Note that although equity is defined at θ levels, the equity-based indices (E, E_1, E_2) are weighted across θ values for reporting. Analysis of variance (ANOVA) was performed to determine if investigated factors had significant effects. Index means for each condition over replications (i.e., cell means) were also computed and reported, both numerically and graphically.

4.3. General framework for presenting the results

The results are presented centering on two main themes: (1) the overall comparison among the investigated methods, and (2) the effects of form and group factors on the performance of each method. For each theme, the ANOVA results are presented first do indicate if the effects were statistically significant. Then, the cell means are presented, in both numerical

and graphical forms, for further discussion. The cell means are presented graphically in the main text. Their numerical values are presented in tables in the Appendix. On those tables, for each index in each specific condition, the smallest index value is boldfaced to indicate the best method (i.e., producing the smallest index value).

Graphical presentations are used throughout this chapter to facilitate the presentation of the results for a specific focus. The same format is used in the figures (for an example, see Figure 4.1). The vertical axis displays the mean index value (averaged over replications within each condition). The horizontal axis presents 105 studying conditions (21 conditions of form difference are fully crossed with five conditions of group difference). The group difference (reflected on the mean difference between the two generating populations *P* and *Q*) increases from left to right, from the smallest (0) to the largest (1.0). Within each group condition, the ratio Δa between the *a*-parameters of Form Y unique items (Y_U) and of Form X unique items (X_U) increases from left to right, form the smallest (.5) to the largest (2.0). Within each Δa section, moving from the left to the right, the difference in *b* parameters (Δb) between Y_U and X_U changes from the smallest (-1.2) to the largest (1.2). Due to limited space in the figures, the values of Δb are not presented in the figures.

Except when the comparison is needed, the scale for the vertical axis may not be the same. They are modified according to the values of the index in discussion to facilitate clearer presentation, especially when a pattern is of interest.

4.4. Overall comparison among methods

In this section, the overall comparison on the performance of the investigated methods is presented. Repeated measures ANOVA was used for each index to compare results from

different methods. Repeated ANOVA was used because different equating methods were applied to the same data in each replication. Accordingly, equating method was treated as within-subject factor (or repeated factor) while group and form factors were treated as between-subject factors. Since the Mauchly's sphericity test was very significant ($p < .0001$) for all indices, either multivariate test results or ANOVA tables with adjusted significance levels could be used. Results from these two approaches have the same significance level. The ANOVA tables are presented here mainly because they allow an easier way to interpret the results, especially difference among methods.

Tables A1 through A4 in the Appendix display the repeated measures ANOVA results for each index. The terms 'between conditions' and 'within conditions', instead of 'between subjects' and 'within subjects', are used in those tables to reflect the context of the current analyses. As shown in those tables, difference among methods was significant for all indices ($p < .0001$). Results from Tukey's post-hoc analyses revealed that the methods performed (significantly) differently from one another for all indices. For the EP and E indices, the OS method performed best, followed by the TS, CE, FE, and IE methods in that order. For the E_1 index, the order is TS, OS, CE, FE, and IE with the TS as the best. For the E_2 index, the best-to-worst order is (IE, OS, TS, CE, FE). Those results can also be seen from Tables A5 through A8 in the Appendix, which contain values for all indices in all conditions for five methods, including the identity equating (IE) (i.e., no equating). The graphical presentations for the cell means, separately for each index, are presented next for more detailed results. All the cell means presented in Tables A5 through A8 in the Appendix have positive values. Accordingly, the corresponding curves presented in the figures do not touch the zero line although in some cases they are very close to the zero line. The main reason is that the index value is the absolute value

of difference as presented in Section 3.7.3. Therefore, the cell mean of index values is always positive.

4.4.1. Index EP

The results for EP are presented in Figure 4.1 and Table A5 in the Appendix. The results for IE are excluded from the figure because its index values are much larger than those from other methods across all conditions and it would not be meaningful to present those in the figure. Some patterns can be seen from the figure.

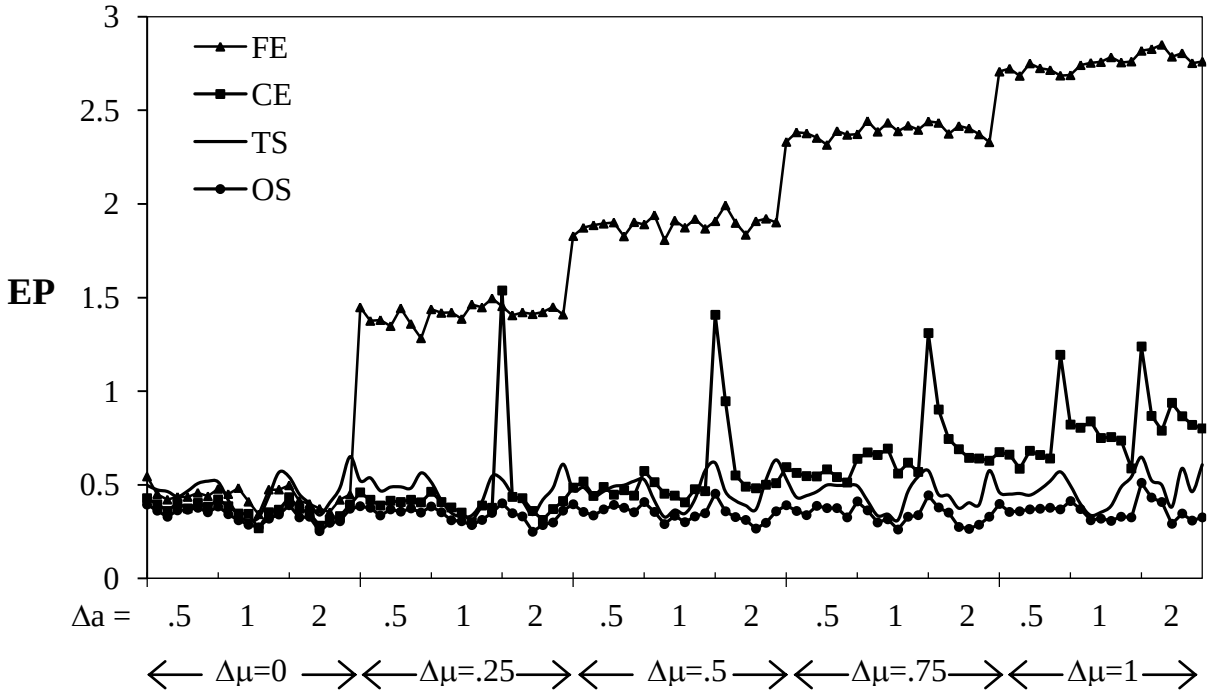

Figure 4.1: Means of index EP for FE, CE, TS, and OS methods in all conditions

Across all conditions, the OS performance was the best. This observation can also be seen in Table A5 where the values for the OS are all boldfaced. Except for some conditions, especially when group difference was small, the TS method was the second best. In most of

conditions, the CE method came in third, fairly close to the OS and the TS. When group difference existed, the FE method was the worst, producing much large index values. In the extreme cases when group difference was one standard deviation, the FE led to index EP values five times larger than those from the other methods.

When groups were similar, those curves stay close to one another, indicating that all methods performed similarly. When groups became different, those curves separate, especially the FE curve. The two IRT methods maintained similar difference when the groups became more different. The CE method slightly departed from the two IRT methods but its deviation was very small compared to that of the FE method.

There are some conditions where the CE suddenly produced much larger EP values. Those cases are associated with the easiest Y_U (Δb = -1.2). Another interesting observation is that the FE curve appears stepped while crossing different group conditions.

4.4.2. Index E

The results for index E are presented in Figure 4.2 and Table A6. Again, the results for IE are excluded from the figure because of its large index values. Some patters can be seen from the figure.

A clear observation seen from Figure 4.2 is that the OS curve is almost always the lowest one which means that the OS performed best among the four methods when evaluated with the E index. This observation can also be noted from Table A6 where all values for the OS are boldfaced. The TS and CE methods came in very close to each other and to the OS. When group difference was small, it appears that the CE performed slightly better than the TS. However, when group became more distinct, the TS was better.

As in the case of the index EP, when group difference existed, the FE method produced

much larger E index values. Those were almost three times larger than those resulted from the other methods.

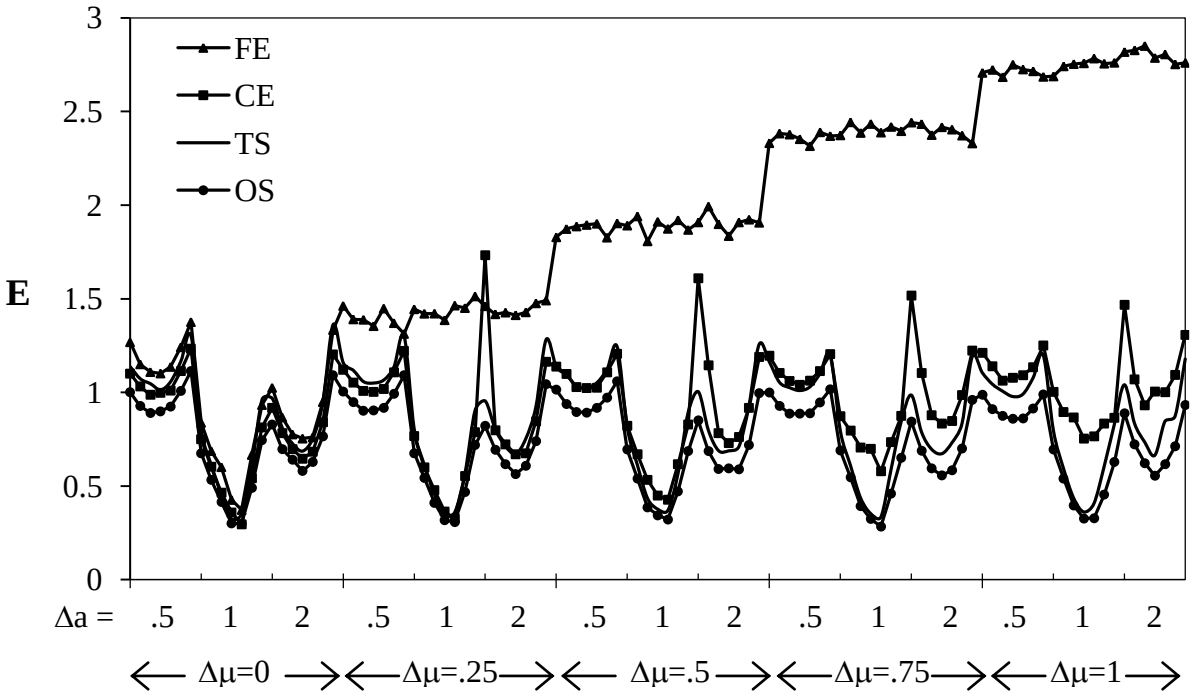

Figure 4.2: Means of index E for FE, CE, TS, and OS methods in all conditions

When groups were similar, all four curves stay together, which means that when the groups did not differ, results from the four methods were very similar. When groups deviated from each other, the difference among the method increased. However, the increment among the CE, TS and OS was not very large in comparison to the difference between the FE and the rest.

There are some conditions with the easiest Y_U (Δb = -1.2) where the CE suddenly produced much larger index E values as in the case of the index EP mentioned previously. Also, a step appearance for the FE curve is observed again while it crosses group difference conditions.

4.4.3. Index E_1

The results for index E_1 are presented in Figure 4.3 and Table A7. Again, the results for IE are excluded from the figure because of its large index values.

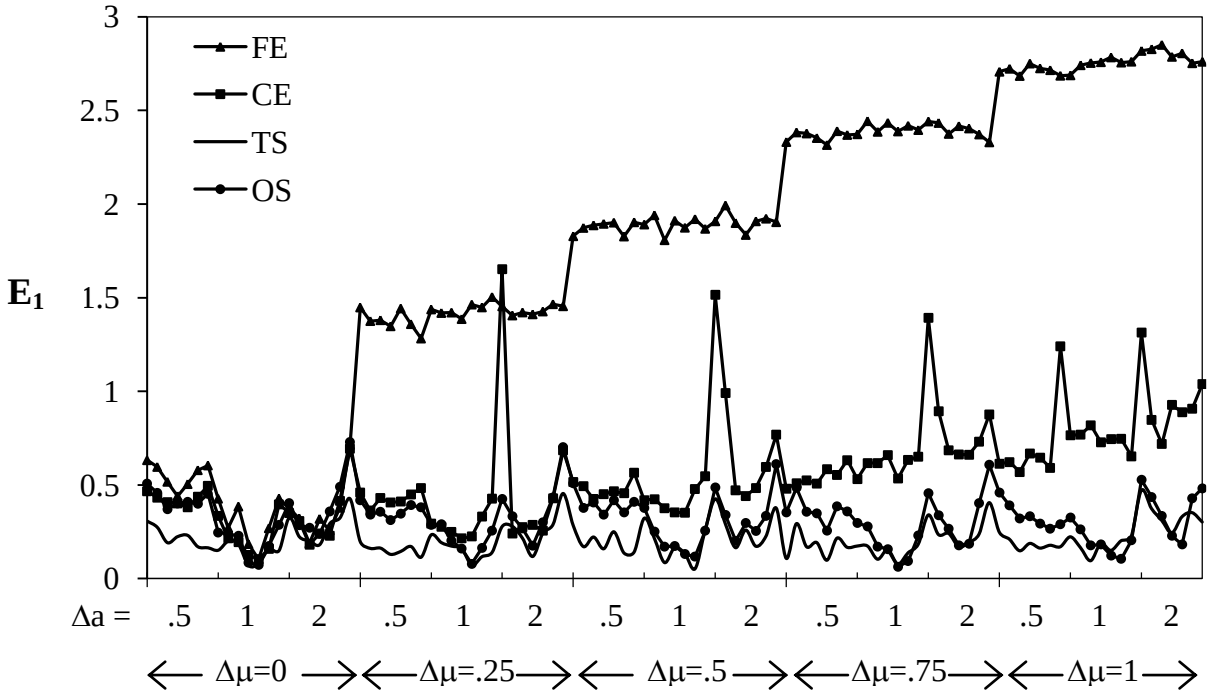

Figure 4.3: Means of index E_1 for FE, CE, TS, and OS methods in all conditions

In almost all conditions, the TS performance was the best. In Table A7, most of values associated with the TS are boldfaced. The OS came in second, fairly close to the TS. In a few conditions, particularly when Form Y was more difficult than Form X, the OS outperformed the TS. The CE stayed close to the two IRT methods when group difference was small but produced larger index values when groups became more distinct.

Once again, except when groups were similar, the FE produced the largest index values among the four methods. In the extreme cases when group difference was one standard

deviation, the FE led to index E_1 values fives time larger than those created by the other methods.

When groups were similar, those curves stay close to one another, indicating that all methods performed similarly. When groups become different, those curves separate, especially the FE curve. The two IRT methods maintained similar difference when the groups became more different. The CE method departed from the two IRT methods and its departure was larger when groups became more dissimilar.

Two other observations that were obtained for EP and E were also noted here. The CE suddenly produced much larger index E_1 values in some conditions with the easiest Form Y (Δb=-1.2). The FE curve appears stepped while crossing different group conditions.

4.4.4 Index E_2

The results for index E_2 are presented in Figure 4.4 and Table A8. There is a noted difference between Figure 4.4 and the previous figures. The curve for IE is included in the Figure 4.4 because its index E_2 values were comparable to those of the other methods. Therefore, there are five curves instead of four curves in the Figure 4.4. Interestingly, the IE curve appears to have different patterns compared to the other curves.

First of all, unlike the other indices, there was not a single method that performed the best under the E_2 index across all conditions. If the OS was clearly the best under the EP and E indices, and the TS was dominant under the E_1 index, the situation here was more complex although the post-hoc analysis showed that the IE method was the best. The difference among the methods were smaller than those for the other indices. A careful look at Table A8 in the Appendix reveals that the IE and OS methods were dominant in majority of the conditions.

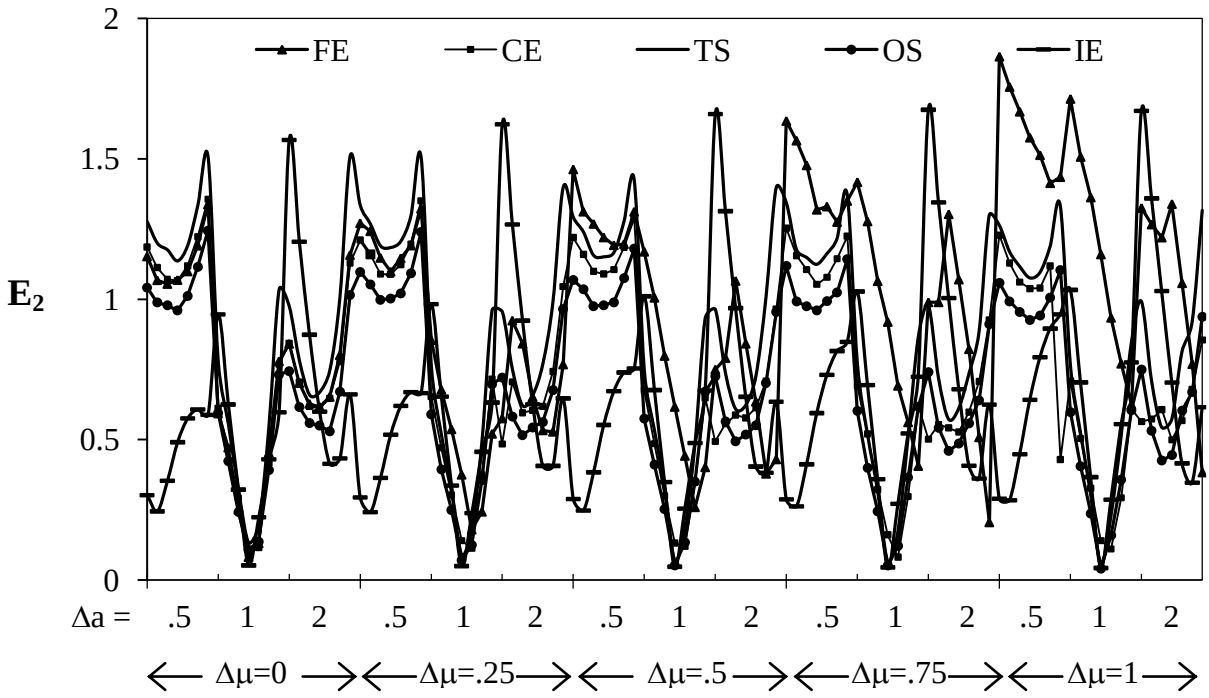

Figure 4.4: Means of index E_2 for FE, CE, TS, OS, and IE methods in all conditions

The IE tended to produce smallest index E_2 values when Form Y had smaller *a*-parameters than those of Form X. When Form Y *a*-parameters were equal to or larger than those of Form X, the performance of those two methods depended on the difficulty of Form Y. If Form Y was equal to or easier than Form X, the OS performed the best. If Form Y was more difficult, the IE led to smaller E_2 index values.

In general, the four investigated methods either performed equally well or worse than the IE method under the E_2 index. The four investigated methods performed similarly across all conditions. Unlike in the case of the other indices, the FE method did not lead to worse E_2 values compared to the CE, TS, and OS methods.

From this point on, effects of group and form differences on equating results are presented for each method separately.

4.5. Effects of group and form factors on the performance of the FE method

To investigate the effects of group and form factors on the indices, the following three-way ANOVA was used:

$$I_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} \quad (4.1)$$

where

I_{ijkl}: index value in the condition ($\Delta\mu_i$, Δa_j, Δb_k) for replication l

μ: overall mean of the index

α_i: effect of being in $\Delta\mu_i$ condition

β_j: effect of being in Δa_j condition

γ_k: effect of being in Δb_k condition

$(\beta\gamma)_{jk}$: interaction effect of (Δa_j, Δb_k) condition

$(\alpha\beta\gamma)_{ijk}$: interaction effect of ($\Delta\mu_i$, Δa_j, Δb_k) condition (group*form interaction)

ε_{ijkl}: error term

The ANOVA results for the FE method are displayed in Table 4.1. In order to examine the results in more details, cell means are displayed, in the numeric form, in the Appendix and in Figures 4.5 through 4.8. Note that those figures are the graphical presentation of the values presented in Tables A5 through A8 in the Appendix. They are used here for a clearer presentation. The effects of the group and form factors are examined next.

Table 4.1. ANOVA results for the FE method for each index

Index	Source	SS	df	MS	F	Sig.
EP	$\Delta\mu$	3425.90	4	856.47	2939.68	< .0001
	Δa	1.17	2	0.59	2.01	0.1343
	Δb	0.48	6	0.08	0.28	0.9486
	$\Delta a*\Delta b$	0.80	12	0.07	0.23	0.9971
	$\Delta\mu*\Delta a*\Delta b$	7.27	80	0.09	0.31	0.9978
	Error	1498.99	5145	0.29		
	Total	4934.61	5249			
E	$\Delta\mu$	2250.93	4	562.73	1753.97	< .0001
	Δa	6.61	2	3.31	10.31	< .0001
	Δb	4.82	6	0.80	2.51	0.0201
	$\Delta a*\Delta b$	1.06	12	0.09	0.27	0.9931
	$\Delta\mu*\Delta a*\Delta b$	76.57	80	0.96	2.98	< .0001
	Error	1650.70	5145	0.32		
	Total	3990.69	5249			
E_1	$\Delta\mu$	3519.69	4	879.92	10033.83	< .0001
	Δa	0.63	2	0.32	3.60	0.0274
	Δb	1.54	6	0.26	2.93	0.0074
	$\Delta a*\Delta b$	0.80	12	0.07	0.76	0.6954
	$\Delta\mu*\Delta a*\Delta b$	31.68	80	0.40	4.52	< .0001
	Error	451.19	5145	0.09		
	Total	4005.53	5249			
E_2	$\Delta\mu$	184.46	4	46.12	30.05	< .0001
	Δa	377.39	2	188.70	122.97	< .0001
	Δb	109.84	6	18.31	11.93	< .0001
	$\Delta a*\Delta b$	50.99	12	4.25	2.77	0.0009
	$\Delta\mu*\Delta a*\Delta b$	165.32	80	2.07	1.35	0.0221
	Error	7895.07	5145	1.53		
	Total	8783.07	5249			

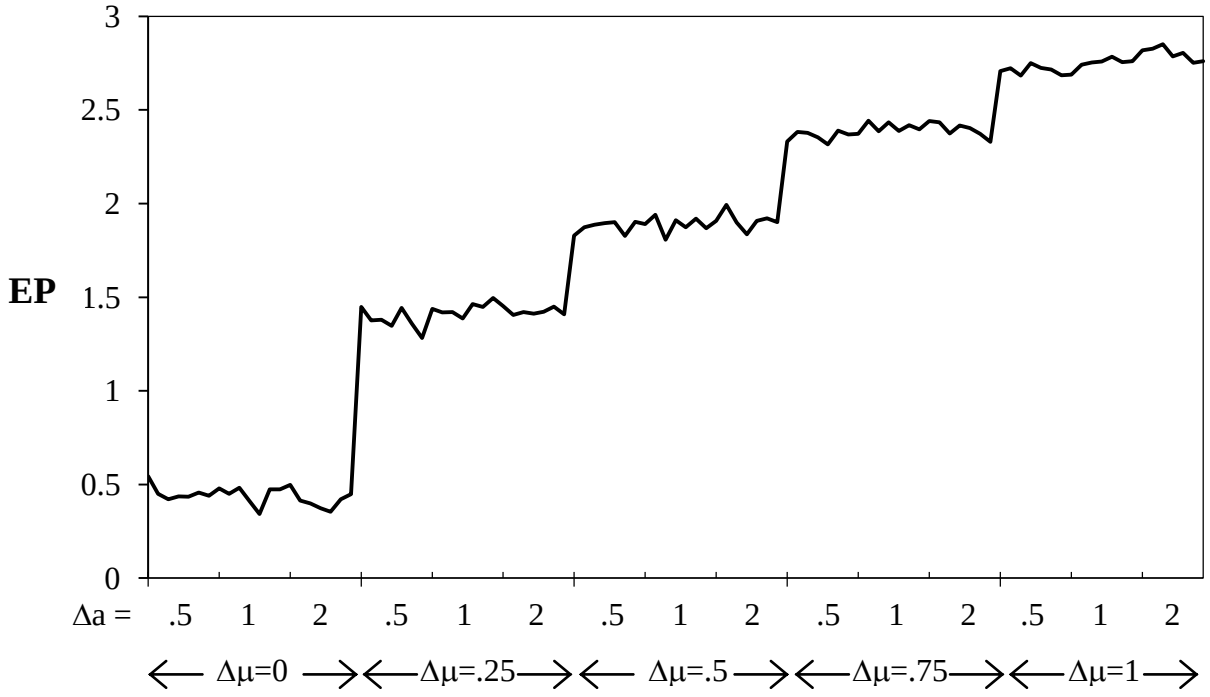

Figure 4.5: Means of index EP for FE method in all conditions

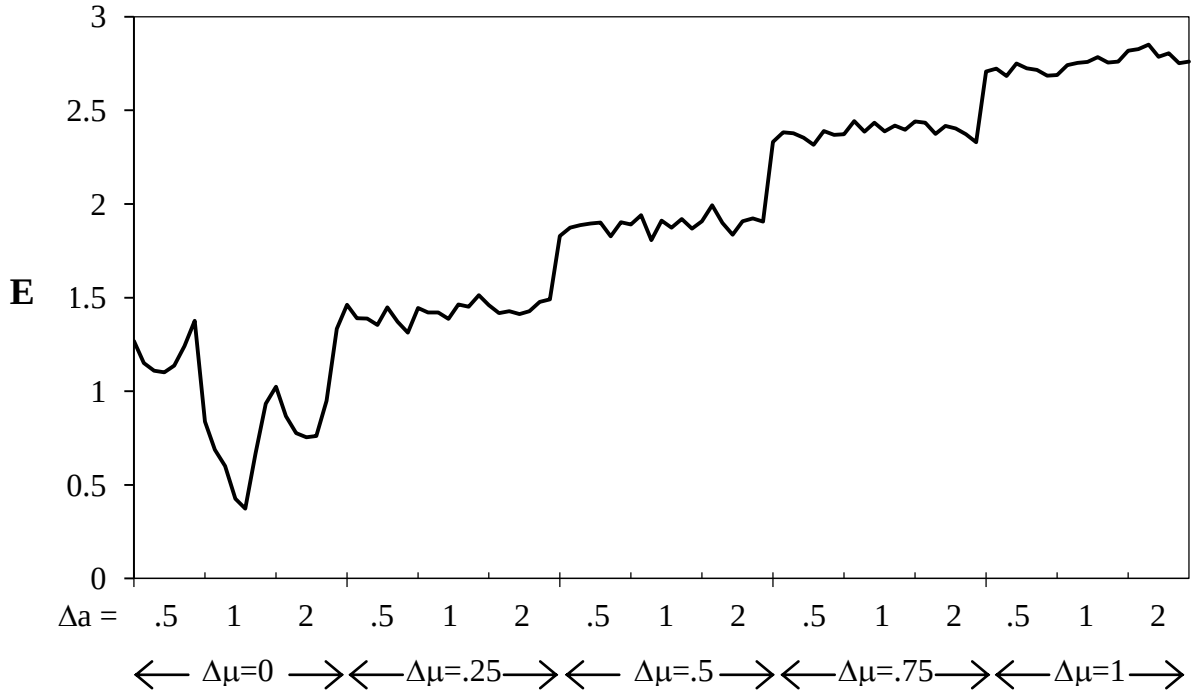

Figure 4.6: Means of index E for FE method in all conditions

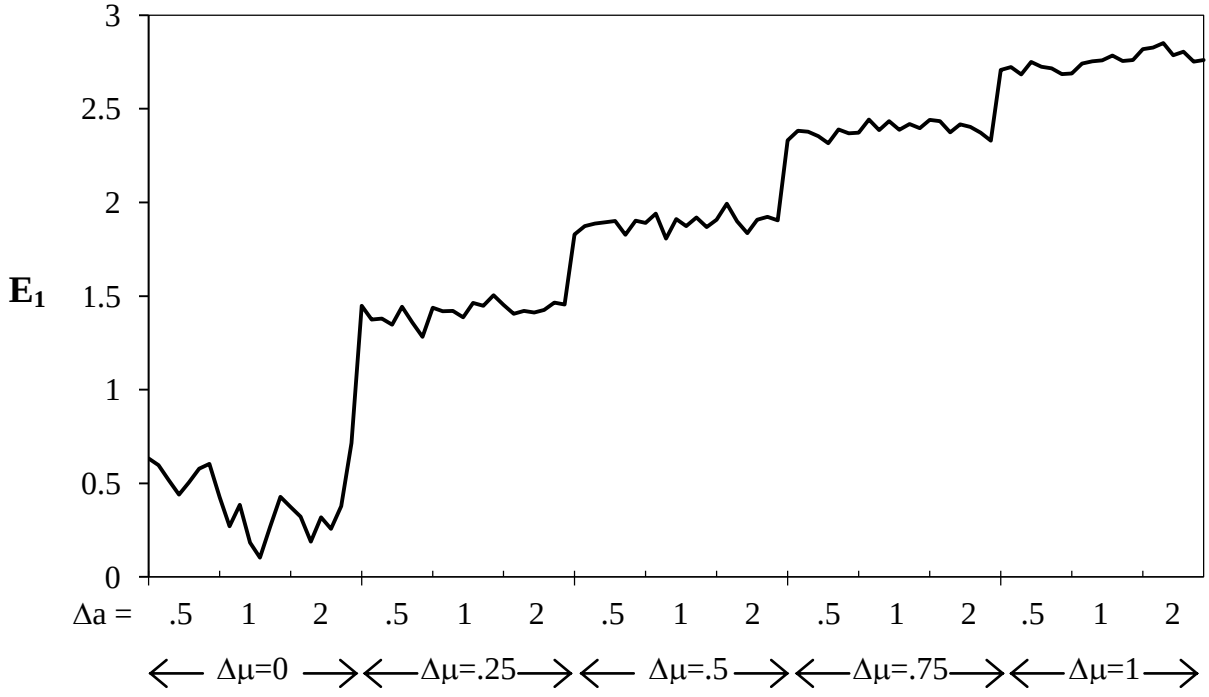

Figure 4.7: Means of index E_1 for FE methods in all conditions

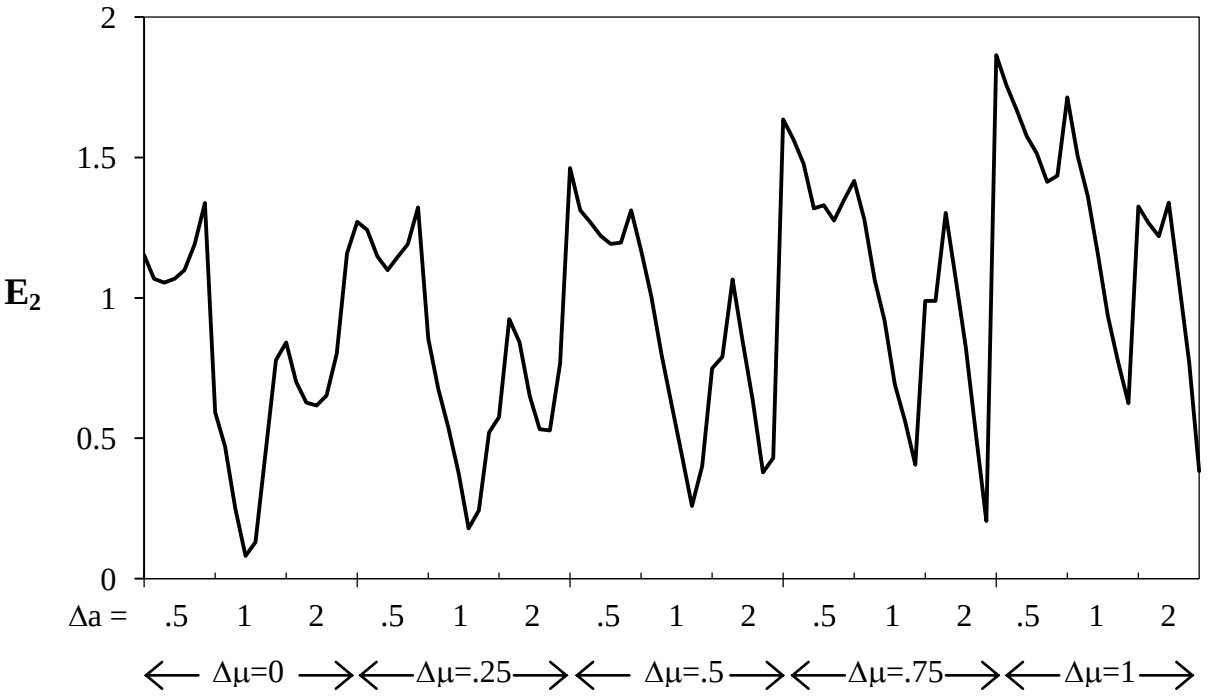

Figure 4.8: Means of index E_2 for FE method in all conditions

4.5.1. Group effects for the FE method

From Table 4.1, it is clear that the group factor had statistically significant effects on all four indices. A closer look at the graphic reveals more details.

Figures 4.5, 4.6, and 4.7, which present results for the EP, E, and E_1 indices, respectively, look similar. Those curves have stepped shapes and they jump where the group condition changes, reflecting strong group effects. The more different the groups were, the larger the index values were. In other words, when groups became more distinct, the equating results became worse under those indices. Except when the groups were similar, the index values were fairly stable within each group condition.

However, a different story can be said about the E_2 index as seen in Figure 4.8. Although the group factor still had effects on E_2, those effects were much smaller. The E_2 curve has completely different shape compared to those for the other indices and does not go up steeply when groups became more different. In other words, group difference did not have much impact on the values of E_2. The E_2 values even went down to some very small values in some conditions. This phenomenon will be discussed later when effects of form difference are examined.

4.5.2. Form effects for the FE method

The form effects can be examined by looking at the effects by Δa (i.e., *a*-parameter difference), by Δb (i.e., *b*-parameter difference), and by their interaction $\Delta a*\Delta b$ in Table 4.1. As seen from the table, the form factor had no significant effects on EP but had strong effects on the other indices, both by *a*- and *b*-parameter differences. The interaction effect of Δa and Δb were found on E_2, but not on E and E_1.

More details are revealed from Figures 4.5 through 4.8. Note that the differences between Form X and Form Y are determined by the differences between two sets of the unique items, X_U and Y_U, of Form X and Form Y, respectively. Those differences were made by changing *a*-parameters and *b*-parameters of the items in Y_U. Also note that in the figures, within each Δa section, the *a* parameters of Form Y were unchanged, only the *b*-parameters changed making Form Y more difficult when moving from the left to the right. When moving to the next Δa section, the same pattern repeats.

Effects of form difference can be examined by a two-step process. The first step is to look at the portion of the curve in each Δa section to see the trend when *b*-parameters change. The second step is to see if the portions in different Δa sections (but within a specific group condition or a $\Delta\mu$ section) differ. The effects of form difference are examined for each method and each index.

For the EP index (see Figure 4.5), there were some small fluctuations within a Δa section meaning that form difficulty difference did not have much impact on EP. When moving from one Δa section to another Δa section, the curve remains the same, which means difference in the *a* parameters did not have effects, either. These results are consistent with those seen from the ANOVA table (e.g., Table 4.2).

From Figures 4.6 and 4.7, a similar pattern can be observed for E and E_1 as for EP presented above, except when groups were similar (i.e., $\Delta\mu = 0$). For each Δa section within the '$\Delta\mu = 0$' section, the curve has a local minimum in the middle where the $\Delta b = 0$ (i.e., no difference in form difficulty). Comparing among the three Δa sections, the middle one (where $\Delta a = 1$) was lowest and the left one (where $\Delta a = .5$) was highest. Therefore, when groups were

similar, equating similar forms led to the best results and a better result was obtained by using higher rather than lower *a*-parameters in the new form.

For E_2, form effect was clearer. When groups did not differ much, using similar forms (in both *a*- and *b*-parameters) led to the best results. However, when group difference was large, using similar *b*-parameters with higher *a*-parameters in the new form was the best choice.

4.5.3. Group and form interaction effects for the FE method

As seen from Table 4.1, the interaction effects of the group and form factors were significant for all but the EP index. Again, more detailed can be obtained from the graphics.

The interactive effects can be examined by checking whether the pattern changed from one group condition to another. If the pattern remained the same or similar across group conditions, there was no interaction between form and group effects.

From Figure 4.5 (for EP), the pattern appeared similar across all five group conditions. Therefore, there was no interaction between form and group effects for EP.

For E (Figures 4.6) and E_1 (Figure 4.7), the pattern did not change in the conditions where group difference existed ($\Delta\mu \neq 0$) but the pattern where groups were similar ($\Delta\mu = 0$) was different from the other conditions. Form difference had effects when groups were similar but did not have effects when groups became different.

For E_2, presented in Figure 4.8, form effects changed when group difference increased, especially after $\Delta\mu = .25$. When $\Delta\mu \leq .25$, using similar forms led to the best results. When $\Delta\mu > .25$, using Form Y with similar *b*-parameters as Form X and higher *a*-parameters resulted in better results.

4.6. Effects of group and form factors on the performance of the CE method

All procedures that were conducted for the FE method were also used with the other methods, including the ANOVA model presented in (4.1).

The ANOVA results for the CE method are displayed in Table 4.2. The cell means are presented in Figures 4.9 thorough 4.12 and in Tables A5 through A8 in the Appendix.

4.6.1. Group effects for the CE method

From Table 4.2, it is clear that the group factor had statistically significant effects on all four indices. A closer look at the graphic reveals more details.

Results for indices EP, E, E_1, and E_2 for the CE method are presented in Figures 4.9, 4.10, and 4.11, respectively. Those figures have different shapes from those for the FE method. They show that the curves go up slightly from the left to the right. That means when groups became more different, the equating results by the CE method became worse. However, the effects of group difference on EP and E_1 were slightly stronger than on E, and much stronger than on E_2.

In summary, although the effects of the group factor were found to be significant, they were smaller for the CE method than for FE method.

4.6.2. Form effects for the CE method

The ANOVA table (Table 4.2) clearly indicates that the form factors have statistically significant effects on all indices for the CE method. A review of Figures 4.9 through 4.12 provides more information.

For EP (Figure 4.9) and E_1 (Figure 4.11), there are some high index values indicating unsatisfactory results in some conditions. Those conditions were when group difference existed, and Form Y was much easier than Form X (Δb=-1.2) and at the same time had higher *a*

Table 4.2. ANOVA results for the CE method for each index

Index	Source	SS	df	MS	F	Sig.
EP	$\Delta\mu$	120.54	4	30.14	338.23	< .0001
	Δa	24.20	2	12.10	135.83	< .0001
	Δb	39.75	6	6.62	74.35	< .0001
	$\Delta a*\Delta b$	51.51	12	4.29	48.18	< .0001
	$\Delta\mu*\Delta a*\Delta b$	59.03	80	0.74	8.28	< .0001
	Error	458.41	5145	0.09		
	Total	753.45	5249			
E	$\Delta\mu$	34.04	4	8.51	38.98	< .0001
	Δa	177.44	2	88.72	406.43	< .0001
	Δb	101.92	6	16.99	77.81	< .0001
	$\Delta a*\Delta b$	32.25	12	2.69	12.31	< .0001
	$\Delta\mu*\Delta a*\Delta b$	35.55	80	0.44	2.04	< .0001
	Error	1123.11	5145	0.22		
	Total	1504.31	5249			
E_1	$\Delta\mu$	146.89	4	36.72	394.57	< .0001
	Δa	50.05	2	25.03	268.90	< .0001
	Δb	55.45	6	9.24	99.29	< .0001
	$\Delta a*\Delta b$	67.21	12	5.60	60.17	< .0001
	$\Delta\mu*\Delta a*\Delta b$	90.00	80	1.12	12.09	< .0001
	Error	478.86	5145	0.09		
	Total	888.45	5249			
E_2	$\Delta\mu$	6.34	4	1.59	3.45	0.0081
	Δa	489.32	2	244.66	531.69	< .0001
	Δb	66.91	6	11.15	24.23	< .0001
	$\Delta a*\Delta b$	50.74	12	4.23	9.19	< .0001
	$\Delta\mu*\Delta a*\Delta b$	37.11	80	0.46	1.01	0.4597
	Error	2367.49	5145	0.46		
	Total	3017.91	5249			

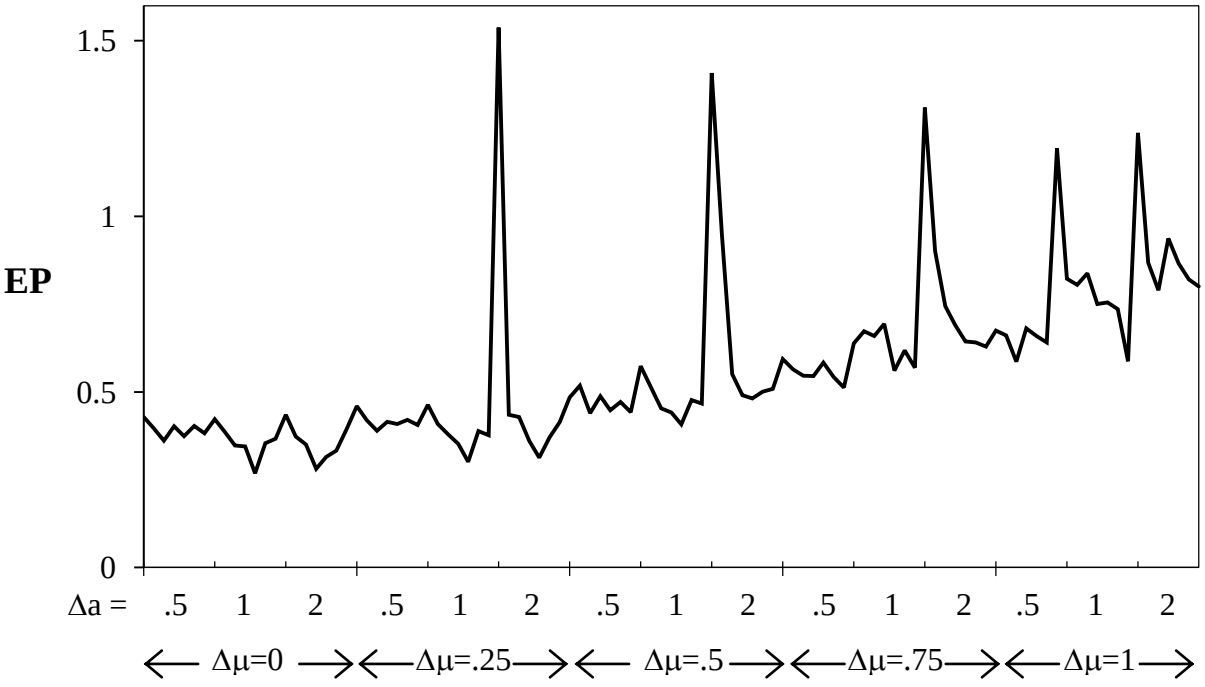

Figure 4.9: Means of index EP for CE method in all conditions

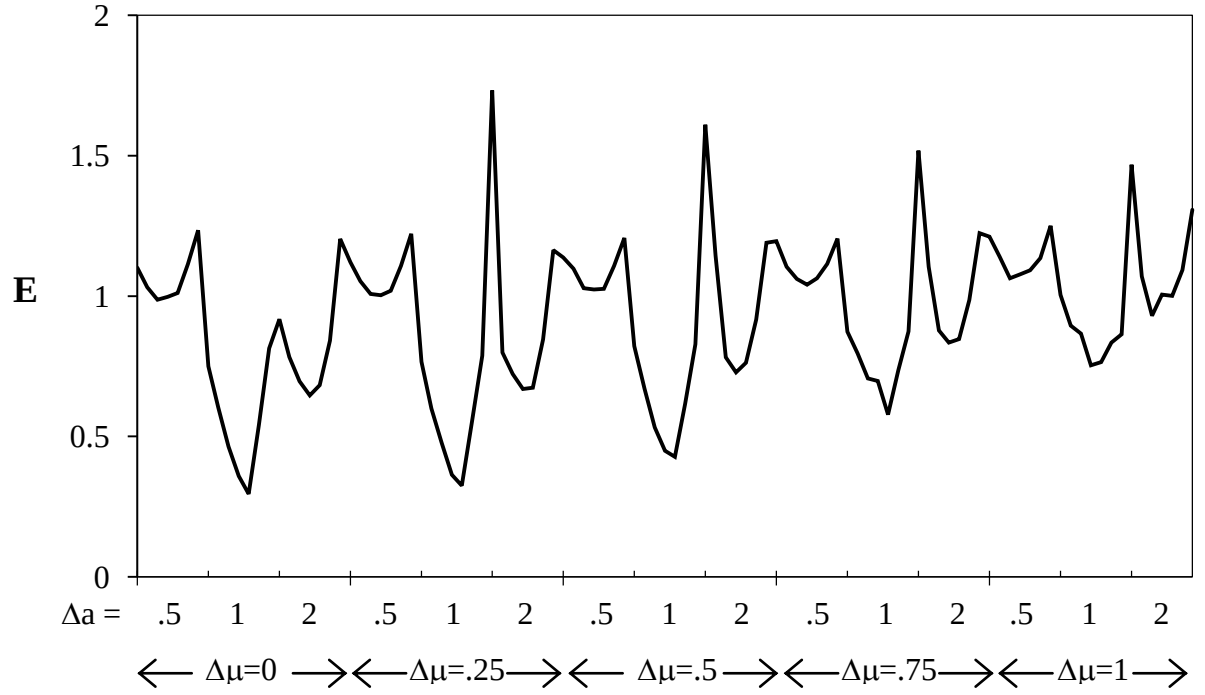

Figure 4.10: Means of index E for CE method in all conditions

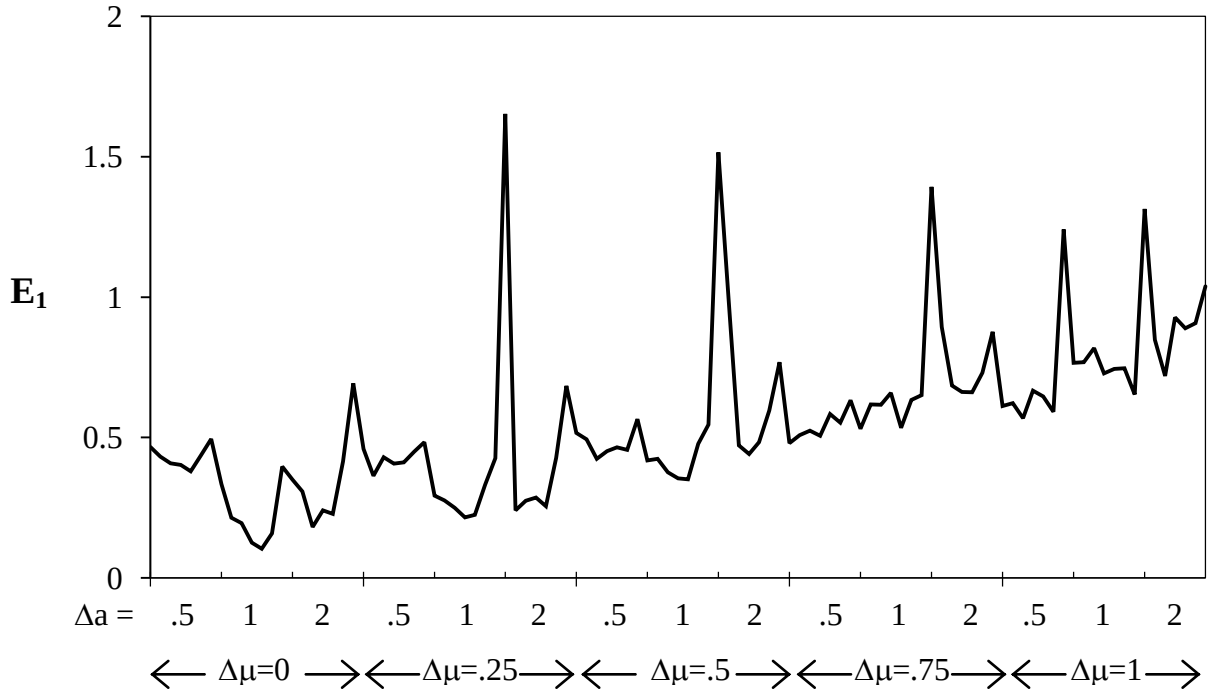

Figure 4.11: Means of index E_1 for CE method in all conditions

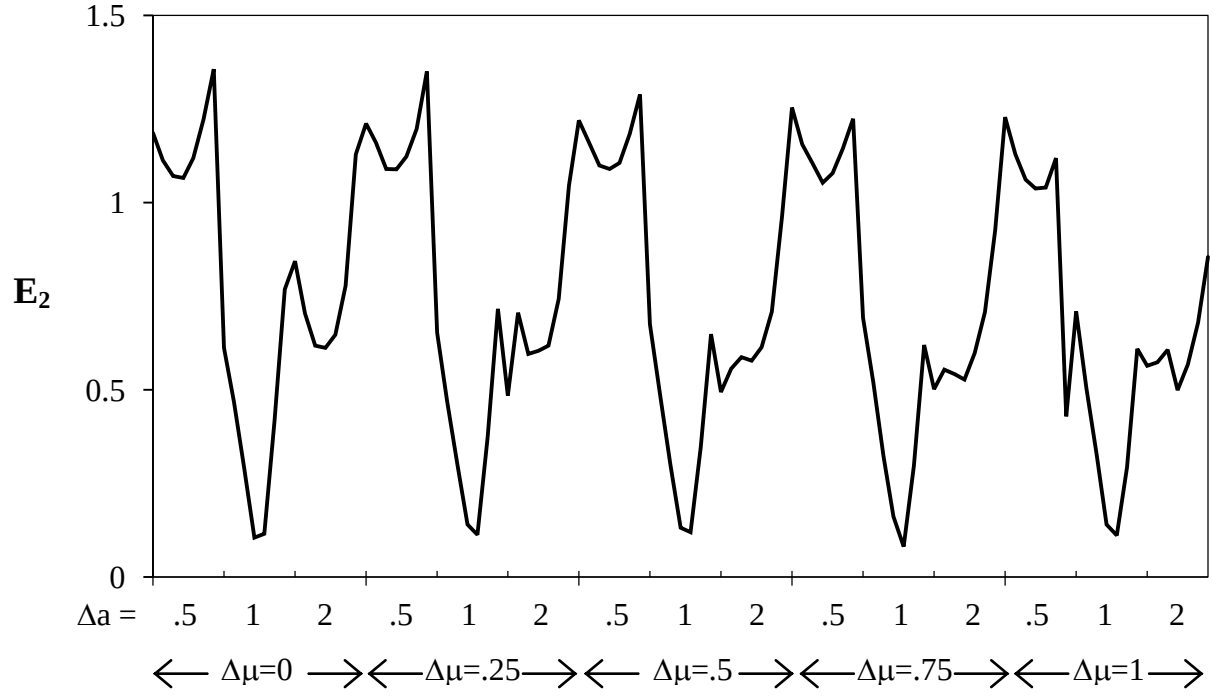

Figure 4.12: Means of index E_2 CE method in all conditions

parameters ($\Delta a = 2$). Except for those conditions, the form effects on EP and E_1 were not very strong and clear. However, from those figures, equating similar forms (in terms of *a*- and *b*-parameters) appeared to produce good results.

For E (Figure 4.10) and E_2 (Figure 4.12), the same observation can be made. For both indices, when forms were similar ($\Delta a = 1$, $\Delta b = 0$), the best results were obtained. When Form Y *a*-parameters were higher than those on Form X, the results were better than when Form Y had lower *a*-parameters.

4.6.3. Group and form interaction effects for the CE method

The ANOVA table (Table 4.2) indicates that significant group-form interaction effects were found for all but E_2 index.

However, from Figures 4.9 through 4.12, except for some surprisingly high index values mentioned previously, it appears that the group-form interaction effects were not strong. The interaction effects found may have been produced by those spikes.

4.7. Effects of group and form factors on the performance of the TS method

The ANOVA results for the TS method are displayed in Table 4.3. The cell means are presented in Tables A5 through A8 in the Appendix. They are also graphically displayed in Figures 4.13 thorough 4.16.

4.7.1. Group effects for the TS method

The ANOVA table (Table 4.3) clearly indicates that the group factor did not have statistically significant effect on all indices for the TS method. This result can be verified by the graphical presentation in Figures 4.13 through 4.16. Although the shapes and the index values

Table 4.3. ANOVA results for the TS method for each index

Index	Source	SS	df	MS	F	Sig.
EP	$\Delta\mu$	0.52	4	0.13	1.18	0.3195
	Δa	4.75	2	2.38	21.42	< .0001
	Δb	19.45	6	3.24	29.22	< .0001
	$\Delta a*\Delta b$	4.73	12	0.39	3.56	< .0001
	$\Delta\mu*\Delta a*\Delta b$	4.20	80	0.05	0.47	0.9981
	Error	570.62	5145	0.11		
	Total	604.28	5249			
E	$\Delta\mu$	0.84	4	0.21	1.61	0.1699
	Δa	237.21	2	118.61	901.90	< .0001
	Δb	116.31	6	19.39	147.41	< .0001
	$\Delta a*\Delta b$	19.40	12	1.62	12.30	< .0001
	$\Delta\mu*\Delta a*\Delta b$	3.98	80	0.05	0.38	0.9975
	Error	676.60	5145	0.13		
	Total	1054.35	5249			
E_1	$\Delta\mu$	1.31	4	0.33	1.82	0.1223
	Δa	15.24	2	7.62	42.45	< .0001
	Δb	6.92	6	1.15	6.43	< .0001
	$\Delta a*\Delta b$	4.77	12	0.40	2.22	0.0089
	$\Delta\mu*\Delta a*\Delta b$	10.58	80	0.13	0.74	0.9618
	Error	923.20	5145	0.18		
	Total	962.02	5249			
E_2	$\Delta\mu$	3.26	4	0.82	1.48	0.2066
	Δa	507.11	2	253.55	458.98	< .0001
	Δb	216.22	6	36.04	65.23	< .0002
	$\Delta a*\Delta b$	47.89	12	3.99	7.22	< .0003
	$\Delta\mu*\Delta a*\Delta b$	4.93	80	0.06	0.11	0.9999
	Error	2842.26	5145	0.55		
	Total	3621.67	5249			

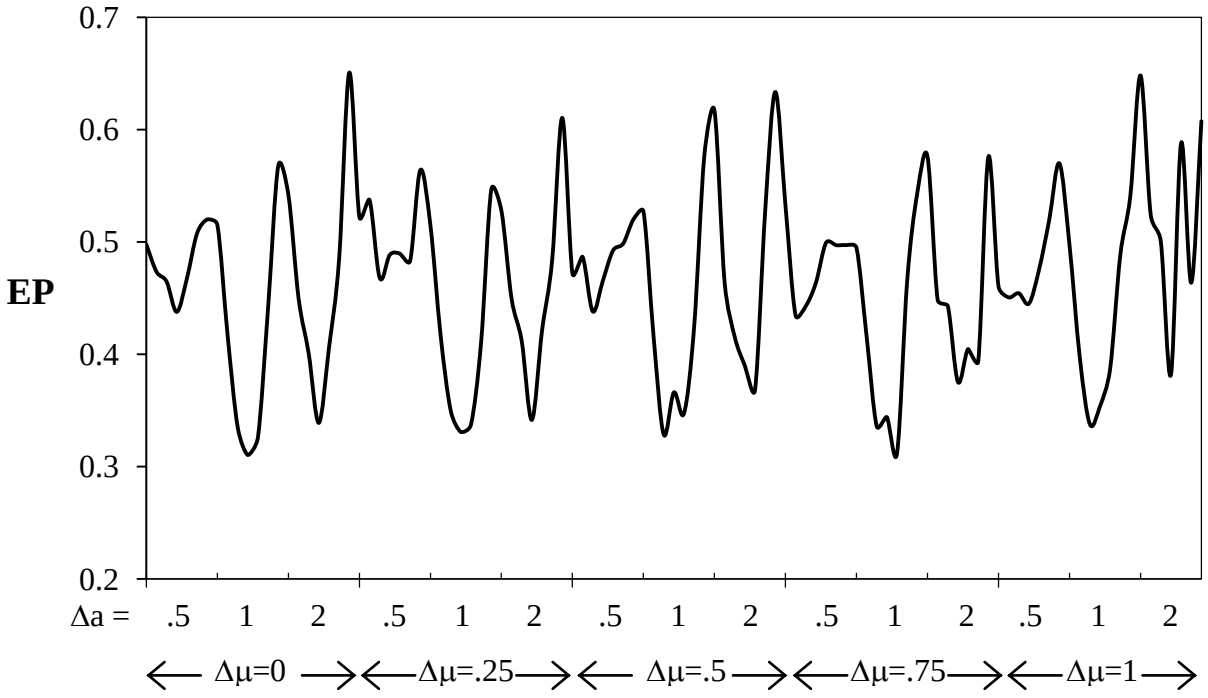

Figure 4.13: Means of index EP for TS method in all conditions

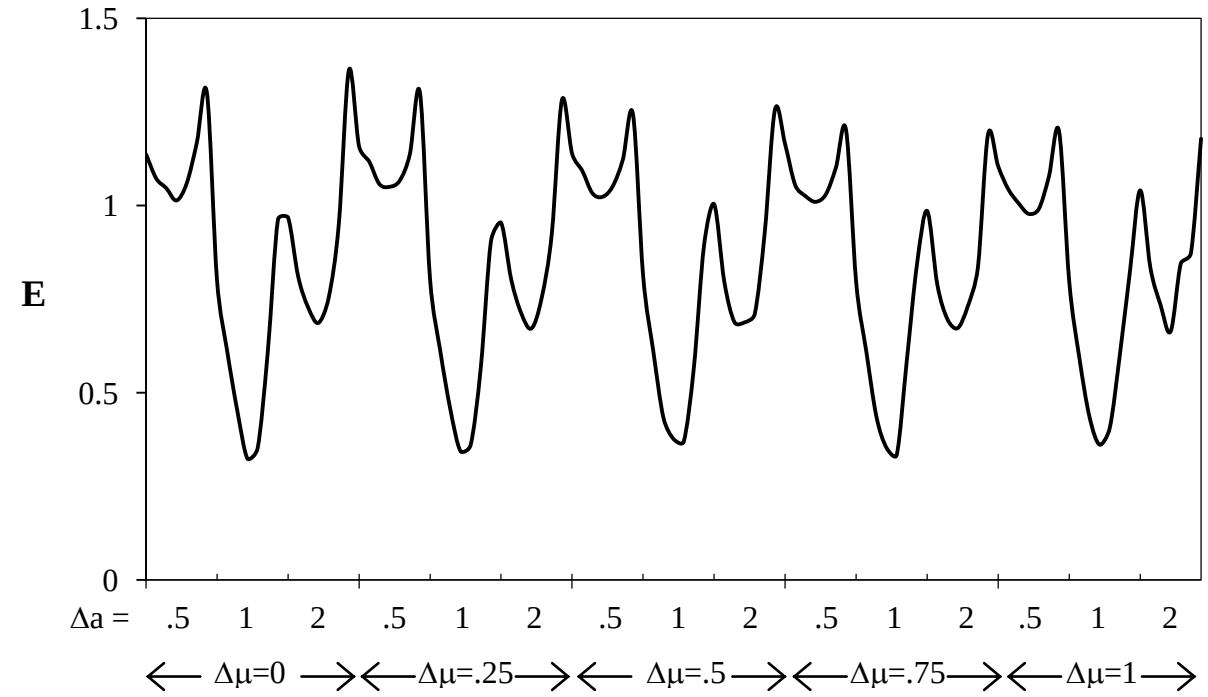

Figure 4.14: Means of index E for TS method in all conditions

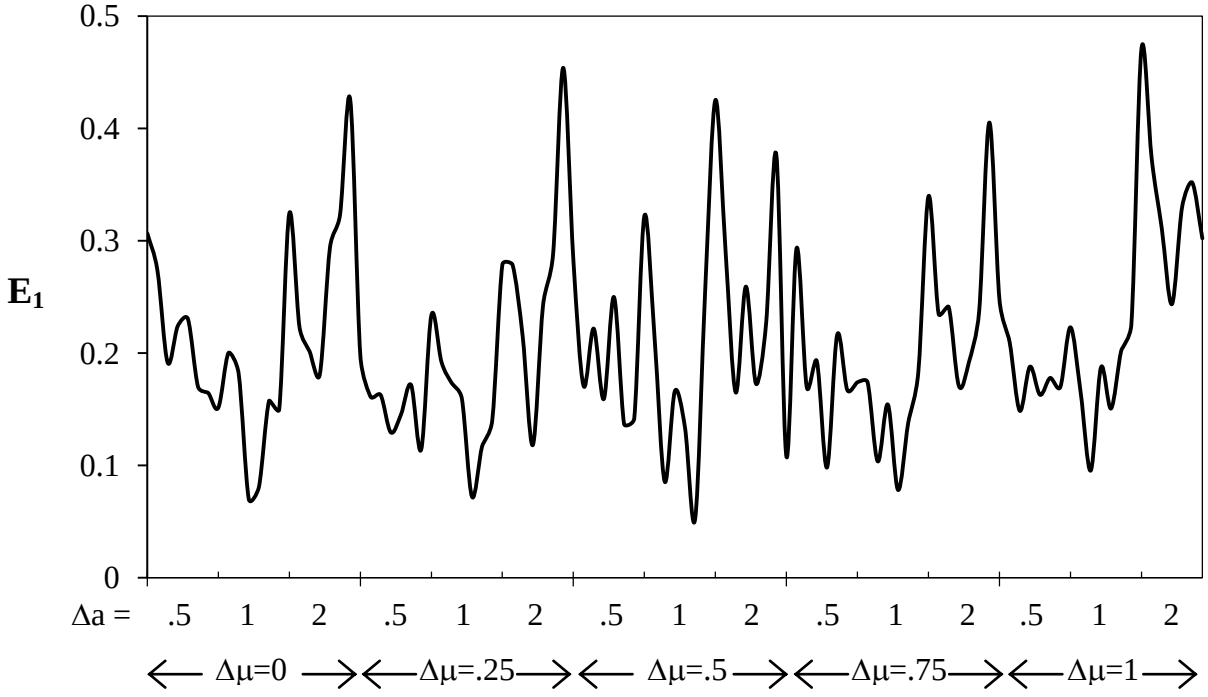

Figure 4.15: Means of index E_1 for TS method in all conditions

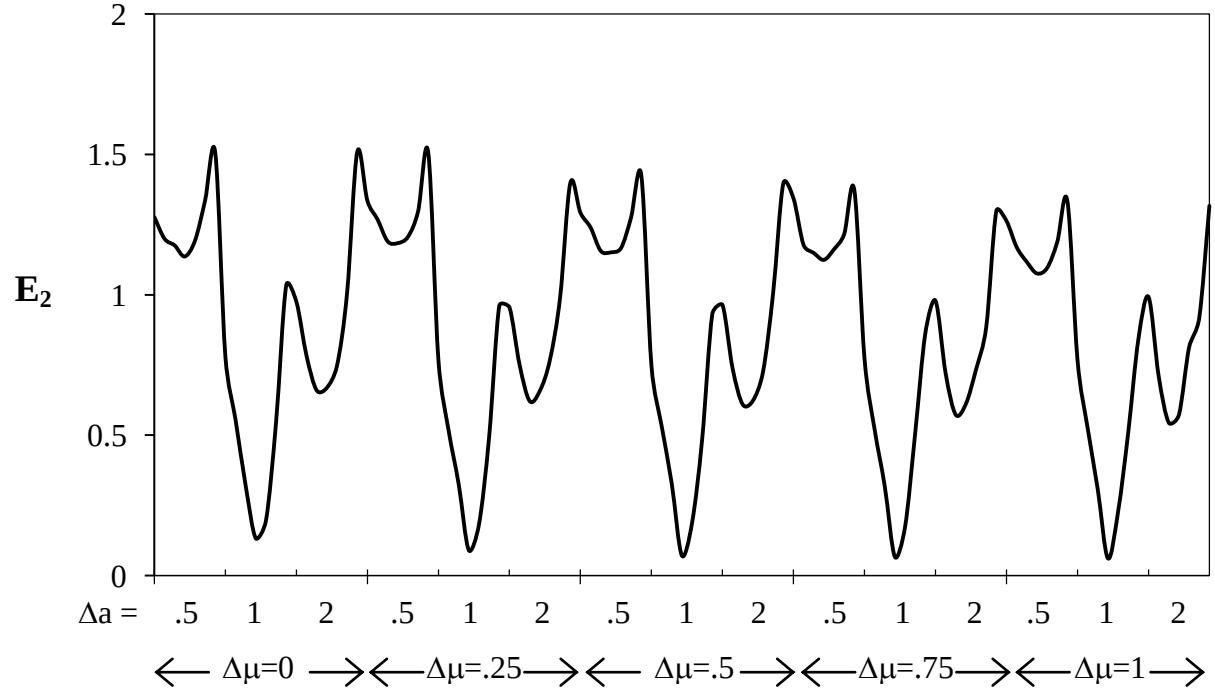

Figure 4.16: Means of index E_2 for TS method in all conditions

for the curves in those figures are not the same, they share one common pattern. That is the curves do not change when group difference changed. In other words, group difference had no effects on equating results of the TS methods evaluated by the four indices.

4.7.2. Form effects for the TS method

As seen from Table 4.3, the form effects were statistically significant. The interaction effects of Δa and Δb were also significant. More details are obtained from the figures.

Similar conclusions can be made for E and E_2 (see Figures 4.14 and 4.16). The best results were obtained when forms were similar (in terms of both *a*- and *b*-parameters). When Form Y *a*-parameters were higher than those on Form X, the results were better than otherwise.

For EP (Figure 4.13), the best results were obtained if the two forms had similar *b*-parameters unless Form Y *a*-parameters were smaller than those of Form X.

For E_1 (Figure 4.15), the pattern was more complicated. Nevertheless, using similar forms still led to the best results. When *a*-parameters on the two forms differed, the results were worse no matter which form had higher *a*-parameters.

4.7.3. Group and form interaction effects for the TS method

The ANOVA did not find significant group-form interaction effects (see Table 4.3). Figures 4.13 through 4.16 display similar patterns in all group conditions, indicating that for the TS methods, there were no interaction effect between the group and form factors.

4.8. Effects of group and form factors on the performance of the OS method

The ANOVA results for the OS method are displayed in Table 4.4. The cell means are presented in Tables A5 through A8 in the Appendix and in Figures 4.17 thorough 4.20.

Table 4.4. ANOVA results for the OS method for each index

Index	Source	SS	df	MS	F	Sig.
EP	$\Delta\mu$	0.41	4	0.10	1.81	0.1233
	Δa	1.15	2	0.58	10.18	< .0001
	Δb	4.81	6	0.80	14.13	< .0001
	$\Delta a*\Delta b$	2.58	12	0.21	3.79	< .0001
	$\Delta\mu*\Delta a*\Delta b$	1.44	80	0.02	0.32	0.9979
	Error	291.70	5145	0.06		
	Total	302.09	5249			
E	$\Delta\mu$	0.76	4	0.19	1.78	0.1292
	Δa	185.12	2	92.56	866.21	< .0001
	Δb	67.61	6	11.27	105.45	< .0001
	$\Delta a*\Delta b$	11.84	12	0.99	9.23	< .0001
	$\Delta\mu*\Delta a*\Delta b$	2.34	80	0.03	0.27	0.9981
	Error	549.78	5145	0.11		
	Total	817.45	5249			
E_1	$\Delta\mu$	1.12	4	0.28	1.55	0.1836
	Δa	36.79	2	18.39	102.37	< .0001
	Δb	25.15	6	4.19	23.33	< .0001
	$\Delta a*\Delta b$	16.43	12	1.37	7.62	< .0001
	$\Delta\mu*\Delta a*\Delta b$	10.52	80	0.13	0.73	0.9651
	Error	924.48	5145	0.18		
	Total	1014.49	5249			
E_2	$\Delta\mu$	1.42	4	0.35	1.56	0.1832
	Δa	413.51	2	206.76	909.30	< .0001
	Δb	100.22	6	16.70	73.46	< .0001
	$\Delta a*\Delta b$	23.52	12	1.96	8.62	< .0001
	$\Delta\mu*\Delta a*\Delta b$	3.07	80	0.04	0.17	0.9999
	Error	1169.87	5145	0.23		
	Total	1711.61	5249			

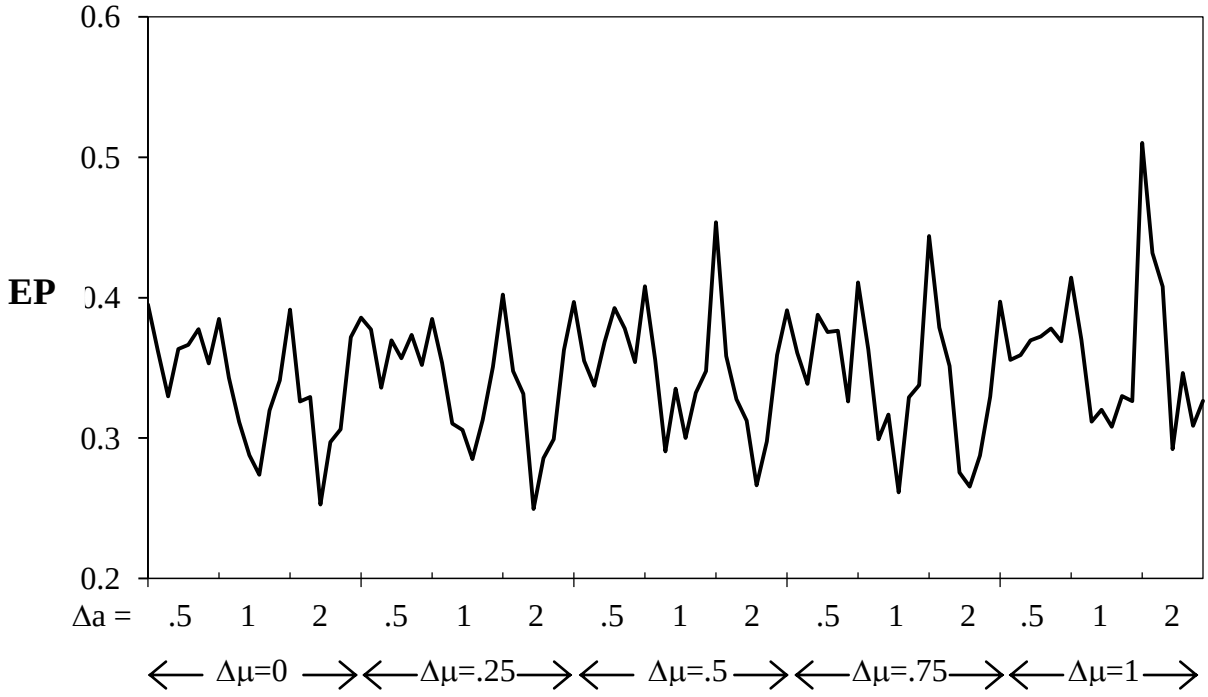

Figure 4.17: Means of index EP for OS method in all conditions

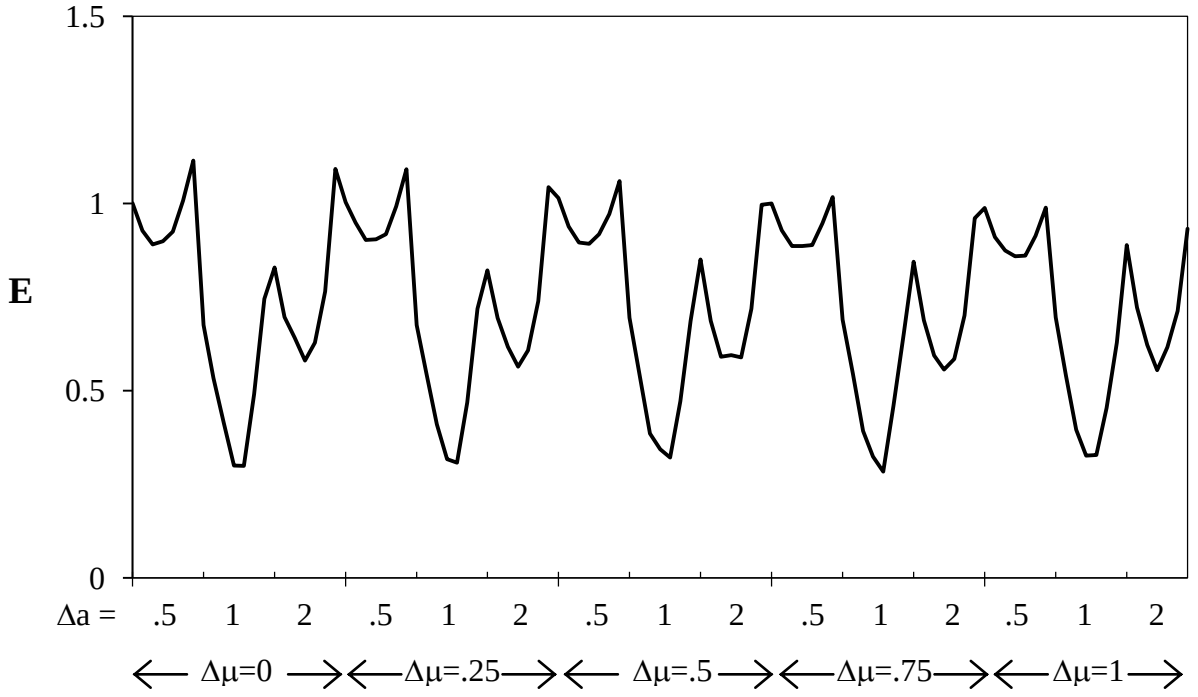

Figure 4.18: Means of index E for OS method in all conditions

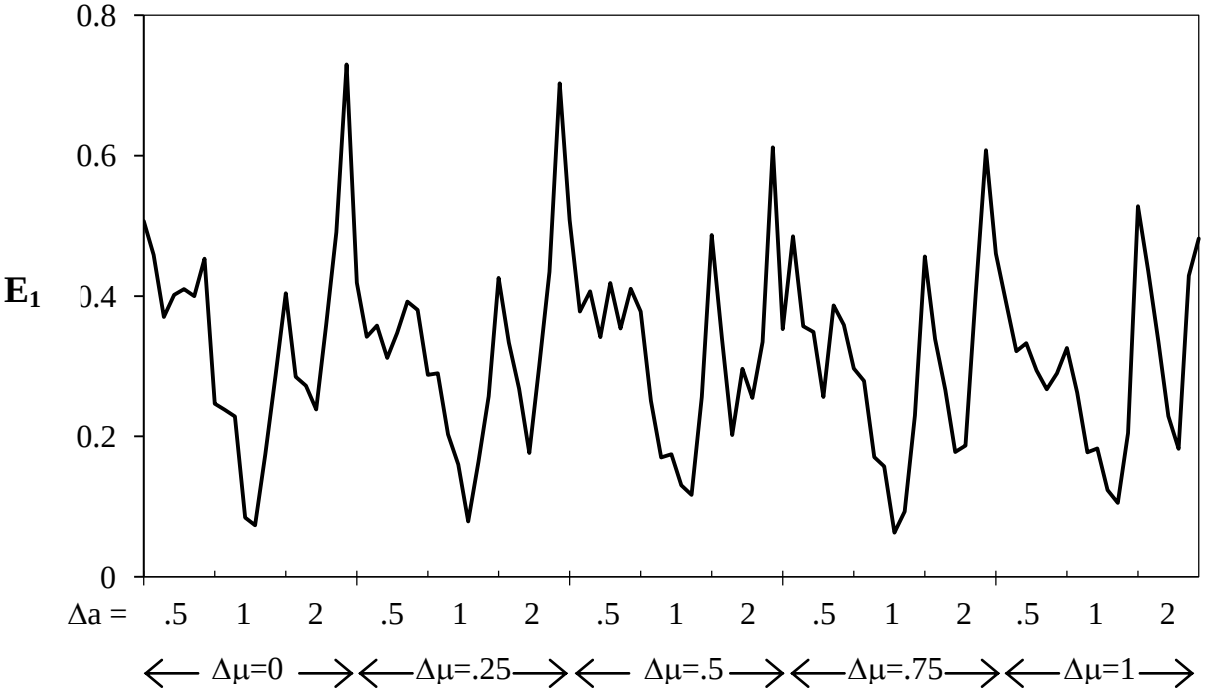

Figure 4.19: Means of index E_1 for OS method in all conditions

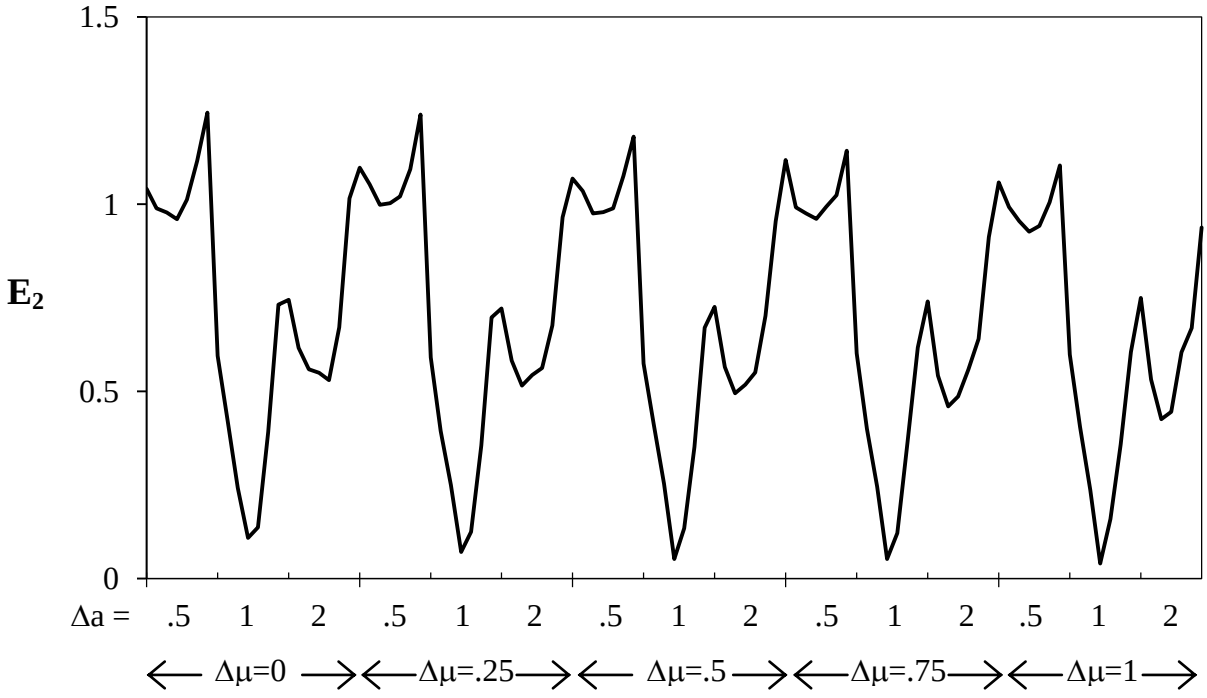

Figure 4.20: Means of index E_2 for OS method in all conditions

4.8.1. Group effects for the OS method

Like the TS method, results from the OS method were not affected by the group factor. Table 4.5 clearly shows that the group effects were not statistically significant.

In Figures 4.17 through 4.20, which present results for the OS method, the curves do not change their shape when moving from one group condition to another. Therefore, it is obvious that group difference did not have effects on equating results of the OS methods evaluated by the four indices.

4.8.2. Form effects for the OS method

As seen from Table 4.4, the form effects were statistically significant. The interaction effects of Δa and Δb were also significant. More details are obtained from the figures.

A review of Figures 4.18, 4.17 and 4.20 reveals more information. When similar forms were used, the best results were produced. When Form Y *a*-parameters were higher than those on Form X, the results were better than otherwise.

For EP (Figure 4.17), using forms with similar *b*-parameters resulted in the best results as long as Form Y *a*-parameters were not smaller than those on Form X.

4.8.3. Group and form interaction effects for the OS method

It is clear from Table 4.4 that the group-form interaction effects were not statistically significant. Figures 4.17 through 4.20 display similar patterns in all group conditions, indicating that for the OS methods, there were no interaction effect between the group and form factors.

4.9. To equate or not to equate?

The IE method (i.e., identity equating or no equating) was used in this study to determine if there was any condition among those studied when no equating would be the best choice. It

came from a concern that sometimes, equating might introduce more errors than it can remove. In such a case, using the directly observed score Y instead of the equated score Y_e would be better.

In order to address this issue, the IE was used as a regular equating and its index values were calculated for all conditions. Those results are included in Tables A1 through A4 in the Appendix. The IE results were also used in the ANOVA procedure comparing equating methods presented in Table 4.1. As previously mentioned, the IE performed poorly in terms of the EP, E, and E_1 indices. Therefore, in all conditions used in this study, if equating results were evaluated using either EP, E, or E_1 index, the IE should not have been recommended. In other words, doing equating was absolutely better than not equating at all. Even using the FE method, which was the worst method among those used in this study, would have led to better results than using the IE method.

However, there is an exception. The IE surprisingly produced comparable values for E_2 as discussed previously. In fact, the IE method produced the best in terms of the E_2 index. A closer look at the cell means, presented in Table A1 through A4, reveals that the IE outperformed the other methods when Form Y had smaller a-parameters than Form X. Therefore, if the equating purpose is to satisfy the second-order equity, which is associated with E_2, then not equating would be preferred.

4.10. Summary

In the Table 4.5, some main results are presented. When groups were similar, all four methods performed similarly. When groups became distinct, results produced by different methods were different to various degrees. The overall performance of the four investigated

methods are ranked for each index. The best method is ranked number 1, the next is number 2, and so on. Based on the results presented above, the OS was ranked the best in three indices EP, E, and E_2 while the TS method outperformed the others in terms of the index E_1. The FE method was ranked last due to its higher index values. The CE method, although ranked third in all indices, performed fairly well, coming close to the two IRT methods.

Table 4.5. Summary of major results

					Factor Effects (**)											
	Overall Performance (*)				Form				Group				Form * Group			
Index	FE	CE	TS	OS	FE	CE	TS	OS	FE	CE	TS	OS	FE	CE	TS	OS
EP	4	3	2	**1**	N	Y	Y	Y	Y	Y	N	N	N	Y	N	N
E	4	3	2	**1**	Y	Y	Y	Y	Y	Y	N	N	Y	Y	N	N
E_1	4	3	**1**	2	Y	Y	Y	Y	Y	Y	N	N	Y	Y	N	N
E_2	4	3	2	**1**	Y	Y	Y	Y	Y	Y	N	N	Y	N	N	N

Note. (*) number representing the ranking where 1 is the best, 4 is the worst
(**) Y: yes, N: no

Table 4.5 also summarizes the results on the effects of form difference and group difference as well as their interactive effects. In the table under the Factor Effect heading, a Y (yes) indicates a significant effect and a N (no) indicates non-significant effect. Form difference was found to have effects on equating results under different indices except for EP in the case of the FE method. In general, equating similar forms tended to produce the best results. Effects of

group difference were found for the OSE methods (FE and CE) but not for the IRT methods (TS and OS). Small interactive effects of form and group difference were found for the FE method for the equity indices (E, E_1 and E_2).

This chapter presents the results obtained from this study. Those results will be discussed in more details in the next chapter where practical implications of the obtained results will also be addressed. In addition, the next chapter will present some perceived limitations of the study. Some related issues will also be presented along with recommendations for further studies.

CHAPTER 5

SUMMARY AND DISCUSSION

In this final chapter, the overall structure of the study is briefly reviewed, followed by the summary of major findings. The discussion of the results will be provided next. The subsequent sections are reserved for the recommendations, limitations of the study, and ideas for further research.

5.1. Brief overview of the study

Testing programs often use multiple test forms of a single test due to test security and exposure concerns. Despite the efforts to make them parallel, these forms are usually not parallel and their sores cannot be used directly before being adjusted to be comparable. Equating is a statistical process of making scores from different test forms of the same test comparable. However, this definition does not explicitly state what it means for the scores to be comparable. Therefore, an operational definition of equating is needed and the equating results must be evaluated by the criteria directly linked to the adopted operational definition.

This study used the criteria derived from two common operational definitions of equating to evaluate results from some equating methods in the NEAT design. The two operational definitions of equating used in this study are called the equipercentile and the equity definitions in the literature. The equipercentile definition requires that the distributions of scores on the two test forms being equated be the same after equating. The equity definition requires that conditioning on ability θ, the distributions of scores on the two forms be the same after equating. Four evaluation indices based on the two definitions were used: (1) the equipercentile index EP,

(2) the full equity index E, (3) the first-order equity index E_1, and (4) the second-order equity index E_2.

Four commonly used equating methods were evaluated: (1) the presmoothed frequency estimation equipercentile equating (FE), (2) the chain equipercentile equating (CE), (3) the IRT true score equating (TS), and (4) the IRT observed score equating (OS). In addition, the identity equating (i.e., no equating) (IE) was also employed to determine if there is any situation where not equating at all is even better than equating.

The IRT 3PL model was used to simulate the data and to compute the evaluation indices. Two factors were varied in this study. The first factor was the difference between the test forms, which was manipulated by changing the *a*- and *b*-parameters of the new form. The second factor was the difference in ability of the two groups taking the two test forms. This factor was manipulated by changing the population mean of the ability θ for the group who took the new form.

A summary and discussion of major results, centering on the research questions presented in Chapter 1, are provided next.

5.2. Summary of major findings

Detailed results were presented in the previous chapters. Some major results are presented here.

5.2.1. Overall performance

- When groups were similar in the ability measured by the test, the four methods produced similar results, evaluated by the values of all four indices. When group difference increased, the results produced by different methods diverged, especially in terms of the

EP, E, and E_1 indices. The difference in terms of the E_2 index was not large even when groups became dissimilar.

- In general, the OS method outperformed the others in regarding to the EP and E indices across all studying conditions. The TS method produced the smallest values of the E_1 index in almost all conditions. However, the difference between these two IRT methods was small. Surprisingly, the IE method produced the best results in terms of the E_2 index although the results from the OS method were close.
- Between the two OSE methods, the CE method produced much better results and they were close to those from the two IRT methods.
- The FE method produced the worst results. Its values for the EP, E, and E_1 indices were far higher than those from the other three methods.

5.2.2. Effects of form difference

- Form difference had clear effects on all methods with all indices. When test forms were dissimilar (the new form was either easier or more difficult than the old form), the equating results became worse.

5.2.3. Effects of group difference

- For the two IRT methods, group difference did not have clear effects. When the groups became dissimilar, the index values produced by those methods did not change.
- The two observed OSE methods (FE and CE) were clearly affected by group difference. Larger group differences led to larger index values. The impacts of group difference were much stronger for the FE method than for the CE method.

5.2.4. Interaction effects of form difference and group difference

- No group-form interaction effects were found for the IRT methods.
- Although significant group-form interaction effects were found for the FE and CE methods, those had small magnitudes.

5.2.5. To equate or not to equate?

- The identity equating (IE) method produced huge values of the EP, E, and E_1 indices compared to those from the four investigated methods.
- For the E_2 index, the IE method produced values either equal to or better than those from the other methods.

5.3. Discussion of the results

In this section, the obtained results are discussed in more details. Again, the discussion is organized around the research questions.

5.3.1. Overall performance

The finding that when groups were similar all four equating methods produced similar results was not unexpected. Research has already shown that different equating methods tend to lead to comparable results when the groups taking the forms come from the same or similar populations (Kolen & Brennan, 2004; Sinharay & Holland, 2007; Wang et al., 2008). When groups are distinct, group difference may produce confounding effects with form difference making equating, which is supposed to adjust for form difference, more complicated. Different methods behave differently in these situations. Each method makes, either implicitly or

explicitly, some assumptions which may be violated to various degrees when groups are different, leading to various results.

It is reasonable that the two IRT methods were found to perform well compared to the two OSE methods. The same IRT model was used for data simulation, IRT equating, and for producing the population distributions of scores, which were used to compute the index values. This gave the two IRT methods advantages.

It was also expected that the TS method outperformed the others in regarding to the E_1 index. The TS method is based on matching the true scores on the two forms which share the same θ. The true score is in fact the expected score at a given θ. Matching two expected values with the same θ is the first-order equity property, which is evaluated by the index E_1. In other words, the purpose of the TS equating is perfectly matched with the property associated with the E_1 index. That explains why the TS was found to be the best method to satisfy the first-order equity.

The OS method is the equipercentile equating on two distributions which were produced in the same way the population distributions of scores on two forms were produced for computing the index EP. Therefore, it is explainable why the OS was found the best method in regarding to the EP index. Perhaps for the same reason, the OS performed well in regarding to the full equity index E, which was calculated based on the same model used by the OS method.

Between the two OSE methods, research has found that the CE tends to produce better results than the FE, especially when groups differ (Wang et al., 2008). This result was confirmed again in this study.

5.3.2. Effects of form difference

Equating is supposed to adjust for unintended form difference. The degree to which this adjustment can be made depends on how large the difference is. The form difference cannot be as large as possible and the equating is still able to adjust for it. There must be some point where the equating can no longer adjust for the form difference, simply because the difference is too large to be adjusted. Research has found that although equating is used to adjust for form difference, it works best when the forms are similar, and larger form difference tends to result in larger equating errors (Kolen & Brennan, 2004; von Davier et al., 2004b). This was confirmed again in this study. In an equating study using evaluation indices associated with the equipercentile and equity definitions, Tong and Kolen (2005) also found that the evaluation index values increased when form difference increased.

5.3.3. Effects of group difference

In the IRT framework, the item parameters are assumed to be population invariant. In other words, they are assumed to remain unchanged across different examinee populations. The TS method is conducted using only item parameters. Thus, its results are not affected by group difference.

The OS method uses the estimated θ distribution (from empirical data) of the target population and the assumed IRT model to produce two marginal score distributions and then conducts a regular equipercentile equating on those distributions. The evaluation indices were calculated using the distributions of X and Y as presented in Section 3.7 of Chapter 3. Those distributions are also produced from the assumed IRT model and the theoretical θ distribution of the target population. The difference between the two θ distributions, one used in the OS method and one used in computing the indices, is that the former is estimated from empirical data and the

latter is theoretically hypothesized. The latter was also used to simulated data in this study. Therefore, is it reasonable to assume that those two θ distributions are close. This and the fact that the same IRT model was used in the OS method and in the computation of the indices may be one reason why group differences had no clear effect on the OS results. More research is needed to shed more lights on this issue.

For the two OSE methods, the results were affected by group differences but to different degrees. The CE method was affected less than the FE method. Although the CE method consists of two equating steps, from *Y* to *A* and from *A* to *X*, it does not make any strong assumption. The FE, on the other hand, makes a strong assumption about the equality of the conditional distributions in the two involved populations (Section 2.3.2 in Chapter 2). When group difference is substantial, this assumption may not hold. This can be illustrated as follows.

Let $f_G(\theta)$ be the distribution of θ and $f_{XA|\theta.G}(x,a \mid \theta)$ be the joint conditional distribution of *X* and *A* in a population *G*. The distribution of *X* conditioning on *A* in *G* is

$$f_{X|A.G}(x \mid a) = \frac{\int f_{XA|\theta.G}(x,a \mid \theta) df_G(\theta)}{\int f_{A|\theta.G}(a \mid \theta) df_G(\theta)} \tag{5.1}$$

It is obvious from the equation (5.1) that $f_{X|A.G}(x \mid a)$ depends on $f_G(\theta)$, which means that $f_{X|A.G}(x \mid a)$ is not likely to be population invariant. In other words it is likely that

$$f_{X|A.P}(x \mid a) \neq f_{X|A.Q}(x \mid a) \tag{5.2}$$

$$f_{Y|A.P}(y \mid a) \neq f_{Y|A.Q}(y \mid a) \tag{5.3}$$

if *P* and *Q* are different. Therefore, the assumptions made in the FE method (i.e., equations (2.5) and (2.6) in Chapter 2) may not hold. This may explain why the FE performed poorly compared to the CE method when groups were different.

5.3.4. Interactive effects of form difference and group difference

In the NEAT design, there are two sources of differences that need to be adjusted. The equating process is supposed to adjust for form difference. The group difference is supposed to be adjusted by a set of common items (or anchor). Those two sources of difference are confounded and may create interactive effects on equating results. In this study, the interactive effects were not found. That can be explained by looking at the quality of the anchor. The anchor was created in such as way that it was a mini-version of the two test forms in terms of statistical characteristics with a reasonable length (i.e., one third of the total test). In other words, the anchor used in this study was fairly ideal. As a result, it performed well in adjusting group differences. This may explain why the interactive effects between form difference and group difference were either small (in case of the OSE methods) or not founded (in case of the IRT methods) in this study.

5.3.5. To equate or not to equate?

The IE method was used in this study to see if there was any condition when not doing equating would be a good solution. The IE produced large values for the indices EP, E, and E_1. Therefore, if those indices are used to evaluate equating results, the IE is not recommended. In other words, doing equating is always, in the conditions of this study, better than not doing equating at all.

When it comes to using the index E_2 to evaluate the equating results, the IE method may be a good choice because it produced E_2 values either equal to or better than those from other methods. However, as recommended by Harris and Crouse (1993), the second-order equity should not be used alone but in combination with the first-order equity. If this recommendation is followed, the IE method is no longer preferred because it produced huge values for E_1 index, which is associated with the first-order equity. Combining E_1 and E_2 would render the IE method unacceptable.

5.3.6. Order effect of *a*-parameter difference

In several cases, it appears that the direction of *a*-parameter difference between the two forms has effects on equating results. Particularly, in those cases, index values tend to be smaller when Form Y *a*-parameters are larger than those of Form X. To investigate this issue further, two special cases were selected and equating was conducted in both directions: from Y to X, and from X to Y. Results are presented in Figure B in the Appendix. For each case, two figures are presented, one for each equating direction. Those figures seem to be mirrored to each other. From these results, it seems that equating a form with larger *a*-parameters (i.e., more reliable) to a form with smaller *a*-parameters (i.e., less reliable) would results in smaller errors than the conducting equating in the opposite direction. Apparently, more research needs to be conducted to shed more lights in this issue.

5.3.7. Unusual high index values for CE method

As seen from Figures 4.9 and 4.11, there are some unusually high values of EP and E_1 in the CE method. Those spikes are associated with $\Delta\mu > 0$, $\Delta a = 2$, and $\Delta b = -1.2$. In other words, in those conditions, the new form was much easier and more reliable than the old form but was

taken by a more able group (group *Q*). That might have produced a huge difference between two score distributions of the two forms. This may be a reason for those spiked values. It is desirable that more research needs to be done to further understand the reasons of those unusual values.

5.4. Recommendations

Results from this study have some practical implications on equating, especially in the NEAT design. Some recommendations on selecting appropriate equating methods in the NEAT design and on communicating equating results are made as follows.

5.4.1. Recommendation on the selection of equating methods in the NEAT design

- Group difference should be assessed before the selection of equating methods. The magnitude of group difference can be determined by comparing scores of the two groups on the common items.
- If groups are similar, either FE, CE, TS, or OS method can be used.
- When groups are different, the FE method is not recommended. The results obtained in this study show that even when the group difference is one fourth of the standard deviation, the index values are more than one score point for the FE method which suggests that it should not be used. The IRT methods, especially the OS method, are highly recommended. If satisfying the first-order equity is the priority, the TS method is the best choice. The CE method can also be used if the group difference is not too large. The use of IRT method may require some strong assumptions such as unidimensionality. Some researchers argued that in practical situations, tests are multidimensional (Reckase, 1985; Reckase & McKinley, 1991). However, the unidimensional model is believed to be somewhat robust to some violations of the unidimensionality assumption (Reckase, 1979;

Thissen, Wainer, & Thayer, 1994). Therefore, unless there are strong reasons to switch to multidimensional models, unidimentional IRT equating is suitable.

- When form difference is substantial, it is recommended that various methods should be used and the results be compared before the final decision is made.

5.4.2. Recommendation on the communication of equating results

- Equating reports should explicitly state the operational definitions of equating that was adopted and how it was determined if the results were accurate relative to the selected definitions.
- The *Standards* (AERA, APA, NCME) should clarify how the equating accuracy is evaluated and require this information be reported to the clients by the organizations that conduct equating.

5.5. Limitations

Any study has limitations and this dissertation is not an exception. Several limitations are perceived and listed as follows:

- Only simulated data were used in this study. Although using simulated data allows the factor manipulation, the applicability of the obtained results to real data remains somewhat unclear.
- Some important factors which were found to have significant impacts on equating, especially in the NEAT design, were not studied. Among those are anchor characteristics, test length, sample size, IRT linking method, and presmoothing technique.
- The study depended heavily on an IRT model. The 3PL model was used to simulate data and to produce the population distributions of scores on two test forms for calculating the

evaluation indices. This may have given the IRT methods advantages. In addition, the selected IRT model was assumed to be true so the conclusions are limited to the situations in which the IRT model fits the data.

- Within each condition, two test forms were fixed across replications while in practice test forms are constantly changed. The main purpose for fixing the test forms in each condition was to eliminate random errors due to sampling items for test forms in each replication. To determine if using randomly generated forms would have led to different results, additional simulations were conducted for four extreme conditions where form and group differences were largest. For each replication, Form X and Form Y were randomly created by sampling their item parameters from the corresponding distributions according to the condition specifications. The results from using random forms, along with those from using fixed forms, are presented in Table A9 in the Appendix. The notable differences in the results from the two approaches suggest that using random forms should be considered in future research.
- This study assumed that tests are unidimensional (i.e., measuring a single latent ability). In practice, tests tend to be multidimensional. For example, a mathematical test can measure both mathematical and verbal abilities. When tests are multidimensional, an equating framework which takes into account the nature of multimendionality should be used.

5.6. Directions for future research

Some directions for future research have been planned as follows

- Extend the current study to examine effects of other important factors such as anchor characteristics, test length, sample size, IRT linking method, presmoothing technique,

and other equating methods such as linear equating, local equating (van der Linden, 2010), kernel equating (von Davier et al., 2004a), and modified FE (Wang & Brennan, 2009).

- Investigate effects of equating direction to see if equating a more reliable form to a less reliable form results in smaller errors than the opposite direction.
- Investigate the possible relationship between the framework used in this study and the concept of population invariance in equating (Holland & Dorans, 2006).
- Extend the study to multidimentional tests (Reckase, 2009).
- Apply the current research framework to real data.

APPENDIX

Table A1. Repeated ANOVA results for index EP [(a)]

Source	SS	df	MS	F	Sig.	Adjusted Sig. G-G[(b)]	Adjusted Sig. H-F[(c)]
Between conditions							
Δμ	707.45	4	176.86	439.72	<.0001		
Δa	1012.00	2	506.00	1258.04	<.0001		
Δb	6674.74	6	1112.46	2765.82	<.0001		
Δμ*Δa	5.21	8	0.65	1.62	0.1141		
Δμ*Δb	55.32	24	2.31	5.73	<.0001		
Δa*Δb	793.24	12	66.10	164.35	<.0001		
Δμ*Δa*Δb	46.74	48	0.97	2.42	<.0001		
Error	2069.40	5145	0.40				
Within conditions							
method	89735.51	4	22433.88	58849.10	<.0001	<.0001	<.0001
method∗Δμ	2967.17	16	185.45	486.47	<.0001	<.0001	<.0001
method*Δa	3395.39	8	424.42	1113.36	<.0001	<.0001	<.0001
method∗Δb	23259.46	24	969.14	2542.28	<.0001	<.0001	<.0001
method*Δμ*Δa	66.70	32	2.08	5.47	<.0001	<.0001	<.0001
method*Δμ*Δb	203.52	96	2.12	5.56	<.0001	<.0001	<.0001
method*Δa*Δb	2452.11	48	51.09	134.01	<.0001	<.0001	<.0001
method*Δμ*Δa*Δb	118.52	192	0.62	1.62	<.0001	0.0004	0.0003
Error (method)	7845.31	20580	0.38				
Total	141407.80	26249					

Note. (a) The multivariate tests are significant at the same level.
(b) Greenhouse-Geisser Epsilon = 0.4225
(c) Huynh-Feldt Epsilon = 0.4312

Table A2. Repeated ANOVA results for index E [(a)]

Source	SS	df	MS	F	Sig.	Adjusted Sig. G-G[(b)]	Adjusted Sig. H-F[(c)]
Between conditions							
Δμ	327.89	4	81.97	65.13	<.0001		
Δa	950.82	2	475.41	377.73	<.0001		
Δb	8079.97	6	1346.66	1069.97	<.0001		
Δμ*Δa	10.59	8	1.32	1.05	0.3941		
Δμ*Δb	50.05	24	2.09	1.66	0.0231		
Δa*Δb	911.60	12	75.97	60.36	<.0001		
Δμ*Δa*Δb	34.07	48	0.71	0.56	0.9935		
Error	6475.51	5145	1.26				
Within conditions							
method	77813.63	4	19453.41	15631.10	<.0001	<.0001	<.0001
method∗Δμ	2087.78	16	130.49	104.85	<.0001	<.0001	<.0001
method*Δa	3837.49	8	479.69	385.43	<.0001	<.0001	<.0001
method∗Δb	21392.96	24	891.37	716.23	<.0001	<.0001	<.0001
method*Δμ*Δa	108.89	32	3.40	2.73	<.0001	0.0021	0.002
method*Δμ*Δb	217.11	96	2.26	1.82	<.0001	0.0038	0.0035
method*Δa*Δb	2410.33	48	50.22	40.35	<.0001	<.0001	<.0001
method*Δμ*Δa*Δb	104.49	192	0.54	0.44	1.0000	1.0000	1.0000
Error (method)	25612.55	20580	1.24				
Total	150425.73	26249					

Note. (a) The multivariate tests are significant at the same level.
(b) Greenhouse-Geisser Epsilon = 0.3195
(c) Huynh-Feldt Epsilon = 0.3260

Table A3. Repeated ANOVA results for index E_1 [(a)]

Source	SS	df	MS	F	Sig.	Adjusted Sig. G-G[(b)]	Adjusted Sig. H-F[(c)]
Between conditions							
Δμ	715.01	4	178.75	417.35	<.0001		
Δa	1142.13	2	571.07	1333.32	<.0001		
Δb	6922.09	6	1153.68	2693.60	<.0001		
Δμ*Δa	8.08	8	1.01	2.36	0.0157		
Δμ*Δb	52.14	24	2.17	5.07	<.0001		
Δa*Δb	824.59	12	68.72	160.44	<.0001		
Δμ*Δa*Δb	61.92	48	1.29	3.01	<.0001		
Error	2203.63	5145	0.43				
Within conditions							
method	95408.05	4	23852.01	64200.70	<.0001	<.0001	<.0001
method∗Δμ	3083.11	16	192.69	518.66	<.0001	<.0001	<.0001
method*Δa	3143.25	8	392.91	1057.56	<.0001	<.0001	<.0001
method∗Δb	22349.48	24	931.23	2506.52	<.0001	<.0001	<.0001
method*Δμ*Δa	88.74	32	2.77	7.46	<.0001	<.0001	<.0001
method*Δμ*Δb	215.94	96	2.25	6.05	<.0001	<.0001	<.0001
method*Δa*Δb	2522.67	48	52.56	141.46	<.0001	<.0001	<.0001
method*Δμ*Δa*Δb	122.78	192	0.64	1.72	<.0001	<.0001	<.0001
Error (method)	7645.93	20580	0.37				
Total	146509.53	26249					

Note. (a) The multivariate tests are significant at the same level.
(b) Greenhouse-Geisser Epsilon = 0.4033
(c) Huynh-Feldt Epsilon = 0.4116

Table A4. Repeated ANOVA results for index E_2 [(a)]

Source	SS	df	MS	F	Sig.	Adjusted Sig. G-G[(b)]	Adjusted Sig. H-F[(c)]
Between conditions							
Δμ	22.98	4	5.75	3.99	0.0031		
Δa	1393.21	2	696.61	483.39	<.0001		
Δb	418.29	6	69.71	48.38	<.0001		
Δμ*Δa	13.97	8	1.75	1.21	0.2872		
Δμ*Δb	45.26	24	1.89	1.31	0.1432		
Δa*Δb	182.99	12	15.25	10.58	<.0001		
Δμ*Δa*Δb	8.27	48	0.17	0.12	1		
Error	7414.32	5145	1.44				
Within conditions							
method	379.63	4	94.91	280.15	<.0001	<.0001	<.0001
method∗Δμ	179.06	16	11.19	33.03	<.0001	<.0001	<.0001
method*Δa	531.04	8	66.38	195.94	<.0001	<.0001	<.0001
method∗Δb	244.42	24	10.18	30.06	<.0001	<.0001	<.0001
method*Δμ*Δa	24.27	32	0.76	2.24	<.0001	0.0032	0.003
method*Δμ*Δb	84.73	96	0.88	2.61	<.0001	<.0001	<.0001
method*Δa*Δb	370.52	48	7.72	22.79	<.0001	<.0001	<.0001
method*Δμ*Δa*Δb	41.39	192	0.22	0.64	1.000	0.9978	0.998
Error (method)	6972.07	20580	0.34				
Total	18326.42	26249					

Note. (a) The multivariate tests are significant at the same level.
(b) Greenhouse-Geisser Epsilon = 0.4964
(c) Huynh-Feldt Epsilon = 0.5066

Table A5. Means of index EP for five equating methods in all conditions

Δa	Δb	Δμ = 0					Δμ = 0.25				
		FE	CE	TS	OS	IE	FE	CE	TS	OS	IE
0.5	-1.2	0.54	0.43	0.50	**0.40**	6.35	1.45	0.46	0.52	**0.39**	5.91
	-0.8	0.45	0.40	0.47	**0.36**	4.59	1.38	0.42	0.54	**0.38**	4.25
	-0.4	0.42	0.36	0.46	**0.33**	3.10	1.38	0.39	0.47	**0.34**	2.91
	0	0.44	0.40	0.44	**0.36**	2.34	1.35	0.41	0.49	**0.37**	2.34
	0.4	0.44	**0.37**	0.47	**0.37**	2.72	1.44	0.41	0.49	**0.36**	2.91
	0.8	0.46	0.40	0.51	**0.38**	4.19	1.36	0.42	0.48	**0.37**	4.50
	1.2	0.44	0.38	0.52	**0.35**	6.12	1.28	0.41	0.56	**0.35**	6.49
1	-1.2	0.48	0.42	0.52	**0.38**	9.08	1.44	0.46	0.51	**0.38**	8.77
	-0.8	0.45	0.39	0.42	**0.34**	6.54	1.42	0.41	0.42	**0.35**	6.36
	-0.4	0.48	0.35	0.34	**0.31**	3.63	1.42	0.38	0.35	**0.31**	3.56
	0	0.41	0.34	0.31	**0.29**	0.51	1.39	0.35	0.33	**0.31**	0.50
	0.4	0.34	**0.27**	0.33	**0.27**	2.77	1.46	0.30	0.34	**0.29**	2.75
	0.8	0.47	0.35	0.44	**0.32**	5.89	1.45	0.39	0.41	**0.31**	5.93
	1.2	0.47	0.37	0.57	**0.34**	8.73	1.50	0.38	0.55	**0.35**	8.86
2	-1.2	0.50	0.43	0.54	**0.39**	10.85	1.45	1.54	0.53	**0.40**	10.57
	-0.8	0.41	0.37	0.45	**0.33**	7.99	1.41	0.44	0.45	**0.35**	7.92
	-0.4	0.40	0.35	0.40	**0.33**	4.65	1.42	0.43	0.41	**0.33**	4.74
	0	0.37	0.28	0.34	**0.25**	2.15	1.41	0.36	0.34	**0.25**	2.17
	0.4	0.35	0.31	0.41	**0.30**	3.60	1.42	0.31	0.42	**0.29**	3.39
	0.8	0.42	0.33	0.48	**0.31**	7.24	1.45	0.37	0.48	**0.30**	7.07
	1.2	0.45	0.39	0.65	**0.37**	10.51	1.41	0.41	0.61	**0.36**	10.53

Note. **Boldface** represents an equating method that produces the smallest value of EP in a specific condition.

Table A5. (cont'd)

Δa	Δb	Δμ = 0.50					Δμ = 0.75				
		FE	CE	TS	OS	IE	FE	CE	TS	OS	IE
0.5	-1.2	1.83	0.48	0.47	**0.40**	5.47	2.33	0.59	0.53	**0.39**	5.05
	-0.8	1.87	0.52	0.49	**0.36**	3.93	2.38	0.56	0.43	**0.36**	3.64
	-0.4	1.89	0.44	0.44	**0.34**	2.75	2.38	0.55	0.44	**0.34**	2.64
	0	1.89	0.49	0.47	**0.37**	2.38	2.35	0.55	0.46	**0.39**	2.45
	0.4	1.90	0.45	0.49	**0.39**	3.11	2.32	0.58	0.50	**0.38**	3.30
	0.8	1.83	0.47	0.50	**0.38**	4.79	2.39	0.54	0.50	**0.38**	5.03
	1.2	1.90	0.44	0.52	**0.35**	6.81	2.37	0.51	0.50	**0.33**	7.07
1	-1.2	1.89	0.57	0.53	**0.41**	8.39	2.37	0.64	0.49	**0.41**	7.98
	-0.8	1.94	0.51	0.42	**0.36**	6.12	2.44	0.67	0.42	**0.36**	5.85
	-0.4	1.81	0.45	0.33	**0.29**	3.45	2.39	0.66	0.34	**0.30**	3.31
	0	1.91	0.44	0.37	**0.34**	0.48	2.43	0.69	0.34	**0.32**	0.46
	0.4	1.87	0.41	0.35	**0.30**	2.72	2.39	0.56	0.31	**0.26**	2.66
	0.8	1.92	0.48	0.42	**0.33**	5.90	2.42	0.62	0.47	**0.33**	5.81
	1.2	1.87	0.47	0.58	**0.35**	8.91	2.40	0.57	0.54	**0.34**	8.85
2	-1.2	1.91	1.41	0.62	**0.45**	10.19	2.44	1.31	0.58	**0.44**	9.73
	-0.8	1.99	0.95	0.46	**0.36**	7.75	2.43	0.90	0.45	**0.38**	7.49
	-0.4	1.90	0.55	0.42	**0.33**	4.75	2.37	0.74	0.44	**0.35**	4.70
	0	1.84	0.49	0.39	**0.31**	2.19	2.42	0.69	0.38	**0.28**	2.21
	0.4	1.91	0.48	0.37	**0.27**	3.16	2.40	0.64	0.40	**0.27**	2.95
	0.8	1.92	0.50	0.53	**0.30**	6.82	2.37	0.64	0.39	**0.29**	6.49
	1.2	1.90	0.51	0.63	**0.36**	10.40	2.33	0.63	0.58	**0.33**	10.13

Note. **Boldface** represents an equating method that produces the smallest value of EP in a specific condition.

Table A5. (cont'd)

Δa	Δb	Δμ = 1.00				
		FE	CE	TS	OS	IE
0.5	-1.2	2.71	0.67	0.46	**0.40**	4.67
	-0.8	2.72	0.66	0.45	**0.36**	3.40
	-0.4	2.68	0.59	0.45	**0.36**	2.57
	0	2.75	0.68	0.45	**0.37**	2.53
	0.4	2.72	0.66	0.48	**0.37**	3.47
	0.8	2.72	0.64	0.52	**0.38**	5.21
	1.2	2.69	0.80	0.57	**0.37**	7.25
1	-1.2	2.69	0.82	0.50	**0.41**	7.55
	-0.8	2.74	0.80	0.40	**0.37**	5.55
	-0.4	2.75	0.84	0.34	**0.31**	3.15
	0	2.76	0.75	0.35	**0.32**	0.43
	0.4	2.78	0.75	0.39	**0.31**	2.58
	0.8	2.76	0.74	0.49	**0.33**	5.67
	1.2	2.76	0.59	0.54	**0.33**	8.70
2	-1.2	2.82	1.24	0.65	**0.51**	9.23
	-0.8	2.83	0.87	0.52	**0.43**	7.17
	-0.4	2.85	0.79	0.50	**0.41**	4.58
	0	2.79	0.94	0.38	**0.29**	2.22
	0.4	2.81	0.87	0.59	**0.35**	2.75
	0.8	2.75	0.82	0.46	**0.31**	6.11
	1.2	2.76	0.80	0.61	**0.33**	9.74

Note. **Boldface** represents an equating method that produces the smallest value of EP in a specific condition.

Table A6. Means of index E for five equating methods in all conditions

Δa	Δb	Δμ = 0					Δu = 0.25				
		FE	CE	TS	OS	IE	FE	CE	TS	OS	IE
0.5	-1.2	1.27	1.10	1.14	**1.00**	6.38	1.46	1.12	1.16	**1.00**	5.96
	-0.8	1.15	1.03	1.07	**0.93**	4.67	1.39	1.05	1.12	**0.95**	4.34
	-0.4	1.11	0.99	1.05	**0.89**	3.24	1.39	1.01	1.06	**0.90**	3.05
	0	1.10	1.00	1.01	**0.90**	2.51	1.35	1.00	1.05	**0.90**	2.52
	0.4	1.14	1.01	1.06	**0.92**	2.87	1.45	1.02	1.07	**0.92**	3.05
	0.8	1.24	1.11	1.17	**1.01**	4.27	1.37	1.11	1.14	**0.99**	4.57
	1.2	1.38	1.23	1.30	**1.11**	6.14	1.31	1.22	1.30	**1.09**	6.51
1	-1.2	0.84	0.75	0.80	**0.67**	9.08	1.44	0.77	0.80	**0.68**	8.77
	-0.8	0.69	0.60	0.61	**0.53**	6.54	1.42	0.60	0.61	**0.54**	6.36
	-0.4	0.60	0.46	0.45	**0.41**	3.63	1.42	0.48	0.46	**0.41**	3.56
	0	0.43	0.36	0.32	**0.30**	0.53	1.39	0.36	0.34	**0.32**	0.52
	0.4	0.37	**0.30**	0.35	**0.30**	2.77	1.46	0.33	0.36	**0.31**	2.75
	0.8	0.67	0.54	0.61	**0.49**	5.89	1.45	0.55	0.57	**0.47**	5.93
	1.2	0.93	0.81	0.96	**0.75**	8.73	1.51	0.79	0.90	**0.72**	8.86
2	-1.2	1.02	0.92	0.97	**0.83**	10.85	1.46	1.73	0.95	**0.82**	10.57
	-0.8	0.87	0.78	0.81	**0.70**	8.01	1.42	0.80	0.80	**0.69**	7.94
	-0.4	0.78	0.70	0.73	**0.64**	4.72	1.43	0.72	0.71	**0.62**	4.80
	0	0.75	0.65	0.69	**0.58**	2.27	1.41	0.67	0.67	**0.56**	2.28
	0.4	0.76	0.68	0.75	**0.63**	3.65	1.43	0.67	0.75	**0.61**	3.45
	0.8	0.95	0.84	0.95	**0.77**	7.24	1.48	0.85	0.93	**0.74**	7.08
	1.2	1.33	1.20	1.36	**1.09**	10.51	1.49	1.16	1.28	**1.04**	10.53

Note. **Boldface** represents an equating method that produces the smallest value of E in a specific condition.

Table A6. (cont'd)

Δa	Δb	Δμ = 0.50					Δu = 0.75				
		FE	CE	TS	OS	IE	FE	CE	TS	OS	IE
0.5	-1.2	1.83	1.14	1.14	**1.01**	5.53	2.33	1.20	1.16	**1.00**	5.12
	-0.8	1.87	1.10	1.09	**0.94**	4.03	2.38	1.10	1.05	**0.93**	3.75
	-0.4	1.89	1.03	1.03	**0.90**	2.90	2.38	1.06	1.03	**0.89**	2.79
	0	1.89	1.02	1.02	**0.89**	2.55	2.35	1.04	1.01	**0.89**	2.61
	0.4	1.90	1.03	1.05	**0.92**	3.24	2.32	1.06	1.03	**0.89**	3.42
	0.8	1.83	1.11	1.12	**0.97**	4.85	2.39	1.11	1.10	**0.95**	5.08
	1.2	1.90	1.21	1.24	**1.06**	6.82	2.37	1.20	1.20	**1.02**	7.08
1	-1.2	1.89	0.82	0.81	**0.70**	8.39	2.37	0.87	0.79	**0.69**	7.98
	-0.8	1.94	0.67	0.61	**0.54**	6.12	2.44	0.80	0.61	**0.55**	5.85
	-0.4	1.81	0.53	0.43	**0.39**	3.45	2.39	0.71	0.44	**0.39**	3.31
	0	1.91	0.45	0.38	**0.34**	0.49	2.43	0.70	0.35	**0.32**	0.47
	0.4	1.87	0.43	0.37	**0.32**	2.72	2.39	0.58	0.33	**0.28**	2.66
	0.8	1.92	0.62	0.57	**0.47**	5.90	2.42	0.74	0.59	**0.46**	5.81
	1.2	1.87	0.83	0.89	**0.69**	8.91	2.40	0.87	0.84	**0.65**	8.85
2	-1.2	1.91	1.61	1.00	**0.85**	10.19	2.44	1.52	0.99	**0.84**	9.73
	-0.8	1.99	1.14	0.80	**0.69**	7.76	2.43	1.10	0.79	**0.69**	7.50
	-0.4	1.90	0.78	0.69	**0.59**	4.81	2.37	0.88	0.70	**0.59**	4.75
	0	1.84	0.73	0.69	**0.59**	2.30	2.42	0.83	0.67	**0.56**	2.31
	0.4	1.91	0.76	0.71	**0.59**	3.23	2.40	0.85	0.73	**0.58**	3.02
	0.8	1.92	0.92	0.93	**0.72**	6.83	2.37	0.99	0.83	**0.70**	6.50
	1.2	1.91	1.19	1.26	**1.00**	10.40	2.33	1.22	1.19	**0.96**	10.13

Note. **Boldface** represents an equating method that produces the smallest value of E in a specific condition.

Table A6. (cont'd)

Δa	Δb	Δμ = 1.00				
		FE	CE	TS	OS	IE
0.5	-1.2	2.71	1.21	1.10	**0.99**	4.75
	-0.8	2.72	1.14	1.04	**0.91**	3.52
	-0.4	2.68	1.06	1.01	**0.87**	2.71
	0	2.75	1.08	0.98	**0.86**	2.68
	0.4	2.72	1.09	0.99	**0.86**	3.57
	0.8	2.72	1.13	1.08	**0.91**	5.26
	1.2	2.69	1.25	1.20	**0.99**	7.26
1	-1.2	2.69	1.00	0.80	**0.70**	7.55
	-0.8	2.74	0.90	0.59	**0.54**	5.55
	-0.4	2.75	0.87	0.44	**0.40**	3.15
	0	2.76	0.75	0.36	**0.33**	0.44
	0.4	2.78	0.76	0.41	**0.33**	2.58
	0.8	2.76	0.83	0.60	**0.45**	5.67
	1.2	2.76	0.86	0.83	**0.63**	8.70
2	-1.2	2.82	1.47	1.04	**0.89**	9.23
	-0.8	2.83	1.07	0.83	**0.72**	7.18
	-0.4	2.85	0.93	0.74	**0.62**	4.63
	0	2.79	1.00	0.66	**0.55**	2.30
	0.4	2.81	1.00	0.85	**0.62**	2.83
	0.8	2.75	1.09	0.87	**0.71**	6.13
	1.2	2.76	1.31	1.18	**0.93**	9.74

Note. **Boldface** represents an equating method that produces the smallest value of E in a specific condition.

Table A7. Means of index E_1 for five equating methods in all conditions

Δa	Δb	Δu = 0					Δu = 0.25				
		FE	CE	TS	OS	IE	FE	CE	TS	OS	IE
0.5	-1.2	0.63	0.47	**0.31**	0.51	6.38	1.45	0.46	**0.20**	0.42	5.96
	-0.8	0.60	0.43	**0.27**	0.46	4.67	1.38	0.36	**0.16**	0.34	4.34
	-0.4	0.52	0.41	**0.19**	0.37	3.23	1.38	0.43	**0.16**	0.36	3.05
	0	0.44	0.40	**0.22**	0.40	2.51	1.35	0.41	**0.13**	0.31	2.52
	0.4	0.50	0.38	**0.23**	0.41	2.87	1.44	0.41	**0.15**	0.35	3.05
	0.8	0.58	0.44	**0.17**	0.40	4.27	1.36	0.45	**0.17**	0.39	4.57
	1.2	0.60	0.49	**0.16**	0.45	6.14	1.28	0.48	**0.11**	0.38	6.51
1	-1.2	0.43	0.33	**0.15**	0.25	9.08	1.44	0.29	**0.23**	0.29	8.77
	-0.8	0.27	0.21	**0.20**	0.24	6.54	1.42	0.28	**0.19**	0.29	6.36
	-0.4	0.38	0.19	**0.18**	0.23	3.63	1.42	0.25	**0.17**	0.20	3.56
	0	0.18	0.13	**0.07**	0.08	0.53	1.39	0.22	**0.16**	**0.16**	0.52
	0.4	0.10	0.10	0.08	**0.07**	2.77	1.46	0.22	**0.07**	0.08	2.75
	0.8	0.27	0.16	**0.16**	0.17	5.89	1.45	0.33	**0.12**	0.16	5.93
	1.2	0.43	0.40	**0.15**	0.29	8.73	1.50	0.43	**0.14**	0.26	8.86
2	-1.2	0.37	0.35	**0.32**	0.40	10.85	1.45	1.65	**0.28**	0.43	10.57
	-0.8	0.32	0.31	**0.22**	0.28	8.01	1.41	**0.24**	0.28	0.33	7.94
	-0.4	0.19	**0.18**	0.20	0.27	4.72	1.42	0.28	**0.21**	0.27	4.80
	0	0.32	0.24	**0.18**	0.24	2.27	1.41	0.29	**0.12**	0.18	2.28
	0.4	0.26	**0.23**	0.29	0.36	3.65	1.43	0.26	**0.24**	0.30	3.44
	0.8	0.38	0.42	**0.32**	0.49	7.24	1.47	0.43	**0.29**	0.43	7.08
	1.2	0.71	0.69	**0.43**	0.73	10.51	1.45	0.68	**0.45**	0.70	10.53

Note. **Boldface** represents an equating method that produces the smallest value of E1 in a specific condition.

Table A7. (cont'd)

Δa	Δb	Δμ = 0.50					Δμ = 0.75				
		FE	CE	TS	OS	IE	FE	CE	TS	OS	IE
0.5	-1.2	1.83	0.52	**0.28**	0.51	5.53	2.33	0.48	**0.11**	0.35	5.12
	-0.8	1.87	0.49	**0.17**	0.38	4.03	2.38	0.51	**0.29**	0.49	3.75
	-0.4	1.89	0.42	**0.22**	0.41	2.90	2.38	0.52	**0.17**	0.36	2.78
	0	1.89	0.45	**0.16**	0.34	2.55	2.35	0.51	**0.19**	0.35	2.61
	0.4	1.90	0.47	**0.25**	0.42	3.24	2.32	0.58	**0.10**	0.26	3.42
	0.8	1.83	0.46	**0.14**	0.35	4.85	2.39	0.55	**0.22**	0.39	5.08
	1.2	1.90	0.57	**0.14**	0.41	6.82	2.37	0.63	**0.17**	0.36	7.08
1	-1.2	1.89	0.42	**0.32**	0.38	8.39	2.37	0.53	**0.17**	0.30	7.98
	-0.8	1.94	0.42	**0.21**	0.25	6.12	2.44	0.62	**0.17**	0.28	5.85
	-0.4	1.81	0.38	**0.09**	0.17	3.45	2.39	0.62	**0.10**	0.17	3.31
	0	1.91	0.35	**0.17**	**0.17**	0.49	2.43	0.66	**0.15**	0.16	0.47
	0.4	1.87	0.35	**0.13**	**0.13**	2.72	2.39	0.53	0.08	**0.06**	2.66
	0.8	1.92	0.48	**0.05**	0.12	5.90	2.42	0.63	0.14	**0.09**	5.81
	1.2	1.87	0.55	**0.26**	**0.26**	8.91	2.40	0.65	**0.18**	0.23	8.85
2	-1.2	1.91	1.52	**0.43**	0.49	10.19	2.44	1.39	**0.34**	0.46	9.73
	-0.8	1.99	0.99	**0.29**	0.34	7.76	2.43	0.89	**0.23**	0.34	7.50
	-0.4	1.90	0.47	**0.16**	0.20	4.81	2.37	0.68	**0.24**	0.27	4.75
	0	1.84	0.44	**0.26**	0.30	2.30	2.42	0.66	**0.17**	0.18	2.31
	0.4	1.91	0.48	**0.17**	0.26	3.23	2.40	0.66	**0.19**	**0.19**	3.02
	0.8	1.92	0.60	**0.23**	0.33	6.83	2.37	0.73	**0.24**	0.40	6.50
	1.2	1.90	0.77	**0.38**	0.61	10.40	2.33	0.88	**0.41**	0.61	10.13

Note. **Boldface** represents an equating method that produces the smallest value of E1 in a specific condition.

Table A7. (cont'd)

Δa	Δb	Δu = 1.00				
		FE	CE	TS	OS	IE
0.5	-1.2	2.71	0.61	**0.25**	0.46	4.75
	-0.8	2.72	0.62	**0.21**	0.39	3.51
	-0.4	2.68	0.57	**0.15**	0.32	2.71
	0	2.75	0.67	**0.19**	0.33	2.68
	0.4	2.72	0.65	**0.16**	0.29	3.57
	0.8	2.72	0.59	**0.18**	0.27	5.26
	1.2	2.69	0.78	**0.17**	0.29	7.26
1	-1.2	2.69	0.77	**0.22**	0.33	7.55
	-0.8	2.74	0.77	**0.17**	0.26	5.55
	-0.4	2.75	0.82	**0.10**	0.18	3.15
	0	2.76	0.73	0.19	**0.18**	0.44
	0.4	2.78	0.74	0.15	**0.12**	2.58
	0.8	2.76	0.75	0.20	**0.11**	5.67
	1.2	2.76	0.65	0.23	**0.20**	8.70
2	-1.2	2.82	1.31	**0.47**	0.53	9.23
	-0.8	2.83	0.85	**0.37**	0.44	7.18
	-0.4	2.85	0.72	**0.31**	0.33	4.63
	0	2.79	0.93	0.24	**0.23**	2.30
	0.4	2.81	0.89	0.33	**0.18**	2.83
	0.8	2.75	0.91	**0.35**	0.43	6.13
	1.2	2.76	1.04	**0.30**	0.48	9.74

Note. **Boldface** represents an equating method that produces the smallest value of E1 in a specific condition.

Table A8. Means of index E_2 for five equating methods in all conditions

Δa	Δb	Δu = 0					Δμ = 0.25				
		FE	CE	TS	OS	IE	FE	CE	TS	OS	IE
0.5	-1.2	1.15	1.19	1.28	1.04	**0.30**	1.27	1.21	1.33	1.10	**0.29**
	-0.8	1.07	1.11	1.20	0.99	**0.24**	1.24	1.16	1.27	1.05	**0.24**
	-0.4	1.05	1.07	1.18	0.98	**0.35**	1.15	1.09	1.19	1.00	**0.36**
	0	1.07	1.07	1.14	0.96	**0.49**	1.10	1.09	1.18	1.00	**0.52**
	0.4	1.10	1.12	1.19	1.01	**0.58**	1.15	1.12	1.21	1.02	**0.62**
	0.8	1.19	1.22	1.34	1.12	**0.61**	1.19	1.20	1.30	1.09	**0.67**
	1.2	1.34	1.36	1.51	1.24	**0.59**	1.32	1.35	1.51	1.24	**0.67**
1	-1.2	0.59	0.61	0.78	**0.59**	0.95	0.85	0.65	0.77	**0.59**	0.98
	-0.8	0.47	0.47	0.56	**0.42**	0.63	0.68	0.47	0.52	**0.39**	0.65
	-0.4	0.25	0.29	0.32	**0.24**	0.32	0.54	0.30	0.33	**0.25**	0.34
	0	0.08	0.11	0.13	0.11	**0.05**	0.37	0.14	0.09	0.07	**0.05**
	0.4	0.13	**0.12**	0.20	0.14	0.22	0.18	**0.11**	0.18	0.13	0.24
	0.8	0.45	0.42	0.55	**0.39**	0.43	**0.24**	0.37	0.49	0.36	0.46
	1.2	0.78	0.77	1.03	0.73	**0.60**	**0.52**	0.72	0.96	0.70	0.63
2	-1.2	0.84	0.84	0.98	**0.74**	1.57	0.58	**0.48**	0.96	0.72	1.62
	-0.8	0.70	0.70	0.78	**0.62**	1.20	0.92	0.71	0.75	**0.58**	1.27
	-0.4	0.63	0.62	0.66	**0.56**	0.87	0.84	0.60	0.62	**0.52**	0.92
	0	0.62	0.61	0.67	**0.55**	0.60	0.65	0.60	0.66	**0.54**	0.63
	0.4	0.65	0.65	0.75	0.53	**0.41**	0.53	0.62	0.77	0.56	**0.41**
	0.8	0.80	0.78	1.01	0.67	**0.43**	0.53	0.74	0.99	0.68	**0.41**
	1.2	1.16	1.13	1.51	1.02	**0.66**	0.77	1.05	1.40	0.96	**0.65**

Note. **Boldface** represents an equating method that produces the smallest value of E2 in a specific condition.

Table A8. (cont'd)

Δa	Δb	Δμ = 0.50					Δu = 0.75				
		FE	CE	TS	OS	IE	FE	CE	TS	OS	IE
0.5	-1.2	1.46	1.22	1.29	1.07	**0.29**	1.63	1.25	1.34	1.12	**0.29**
	-0.8	1.31	1.16	1.24	1.04	**0.25**	1.56	1.15	1.18	0.99	**0.26**
	-0.4	1.27	1.10	1.16	0.98	**0.38**	1.48	1.11	1.15	0.98	**0.41**
	0	1.22	1.09	1.15	0.98	**0.55**	1.32	1.05	1.12	0.96	**0.59**
	0.4	1.19	1.11	1.17	0.99	**0.67**	1.33	1.08	1.16	0.99	**0.73**
	0.8	1.20	1.18	1.28	1.08	**0.74**	1.27	1.14	1.22	1.02	**0.82**
	1.2	1.31	1.29	1.43	1.18	**0.75**	1.35	1.22	1.37	1.14	**0.85**
1	-1.2	1.17	0.68	0.75	**0.57**	1.01	1.42	0.69	0.77	**0.60**	1.03
	-0.8	1.01	0.49	0.53	**0.41**	0.68	1.28	0.52	0.52	**0.40**	0.69
	-0.4	0.80	0.30	0.33	**0.25**	0.35	1.06	0.32	0.32	**0.25**	0.36
	0	0.62	0.13	0.07	**0.05**	**0.05**	0.92	0.16	0.07	**0.05**	**0.05**
	0.4	0.44	**0.12**	0.19	0.13	0.25	0.69	**0.08**	0.17	0.12	0.27
	0.8	**0.26**	0.35	0.49	0.35	0.49	0.56	**0.30**	0.51	0.37	0.52
	1.2	**0.40**	0.65	0.93	0.67	0.67	**0.41**	0.62	0.86	0.62	0.72
2	-1.2	0.75	**0.49**	0.96	0.73	1.66	0.99	**0.50**	0.98	0.74	1.67
	-0.8	0.79	**0.56**	0.74	0.57	1.31	0.99	0.55	0.72	**0.54**	1.34
	-0.4	1.07	0.59	0.61	**0.49**	0.97	1.30	0.54	0.57	**0.46**	1.00
	0	0.84	0.58	0.62	**0.52**	0.65	1.07	0.53	0.61	**0.49**	0.68
	0.4	0.64	0.61	0.73	0.55	**0.40**	0.82	0.60	0.73	0.56	**0.41**
	0.8	**0.38**	0.71	1.01	0.70	**0.38**	0.51	0.71	0.88	0.64	**0.36**
	1.2	**0.43**	0.96	1.40	0.95	0.63	**0.21**	0.93	1.30	0.91	0.62

Note. **Boldface** represents an equating method that produces the smallest value of E2 in a specific condition.

Table A8. (cont'd)

Δa	Δb	Δu = 1.00				
		FE	CE	TS	OS	IE
0.5	-1.2	1.86	1.23	1.26	1.06	**0.29**
	-0.8	1.76	1.13	1.17	0.99	**0.28**
	-0.4	1.67	1.06	1.12	0.95	**0.45**
	0	1.58	1.04	1.08	0.93	**0.64**
	0.4	1.51	1.04	1.09	0.94	**0.79**
	0.8	1.41	1.12	1.19	1.01	**0.90**
	1.2	1.43	0.43	1.33	1.10	**0.95**
1	-1.2	1.71	0.71	0.77	**0.60**	1.03
	-0.8	1.51	0.50	0.52	**0.41**	0.70
	-0.4	1.36	0.33	0.31	**0.24**	0.37
	0	1.16	0.14	0.06	**0.04**	**0.04**
	0.4	0.93	**0.11**	0.23	0.16	0.29
	0.8	0.77	**0.29**	0.51	0.36	0.56
	1.2	0.62	0.61	0.84	**0.60**	0.77
2	-1.2	1.32	**0.56**	0.99	0.75	1.67
	-0.8	1.27	0.57	0.71	**0.53**	1.36
	-0.4	1.22	0.61	0.54	**0.43**	1.03
	0	1.34	0.50	0.57	**0.45**	0.70
	0.4	1.06	0.57	0.81	0.60	**0.42**
	0.8	0.77	0.68	0.92	0.67	**0.35**
	1.2	**0.38**	0.85	1.32	0.94	0.62

Note. **Boldface** represents an equating method that produces the smallest value of E2 in a specific condition.

Table A9. Comparison of results obtained from using fixed and random test forms.

Condition (Δμ = 1.00; Δa = 0.5, Δb = -1.2)

Index	Fixed forms				Random forms			
	FE	CE	TS	OS	FE	CE	TS	OS
EP	2.71	0.67	0.46	0.40	2.56	0.76	0.54	0.46
E	2.71	1.21	1.10	0.99	2.62	1.09	1.02	1.10
E_1	2.71	0.61	0.25	0.46	2.89	0.73	0.31	0.43
E_2	1.86	1.23	1.26	1.06	1.76	1.44	1.50	1.23

Condition (Δμ = 1.00; Δa = 0.5, Δb = 1.2)

Index	Fixed forms				Random forms			
	FE	CE	TS	OS	FE	CE	TS	OS
EP	2.69	0.80	0.57	0.37	2.90	0.85	0.65	0.32
E	2.69	1.25	1.20	0.99	2.92	1.35	1.12	1.04
E_1	2.69	0.78	0.17	0.29	2.54	0.84	0.21	0.33
E_2	1.43	0.43	1.33	1.10	1.35	0.52	1.23	1.22

Table A9. (cont'd)

Condition ($\Delta\mu$ = 1.00; Δa = 2, Δb = -1.2)

Index	Fixed forms				Random forms			
	FE	CE	TS	OS	FE	CE	TS	OS
EP	2.82	1.24	0.65	0.51	2.92	1.34	0.68	0.49
E	2.82	1.47	1.04	0.89	2.76	1.42	1.09	0.94
E_1	2.82	1.31	0.47	0.53	2.74	1.47	0.54	0.60
E_2	1.32	0.56	0.99	0.75	1.25	0.61	0.92	0.79

Condition ($\Delta\mu$ = 1.00; Δa = 2, Δb = 1.2)

Index	Fixed forms				Random forms			
	FE	CE	TS	OS	FE	CE	TS	OS
EP	2.76	0.80	0.61	0.33	2.79	0.78	0.64	0.37
E	2.76	1.31	1.18	0.93	2.84	1.40	1.13	0.91
E_1	2.76	1.04	0.30	0.48	3.01	1.12	0.34	0.56
E_2	0.38	0.85	1.32	0.94	0.41	0.79	1.45	0.89

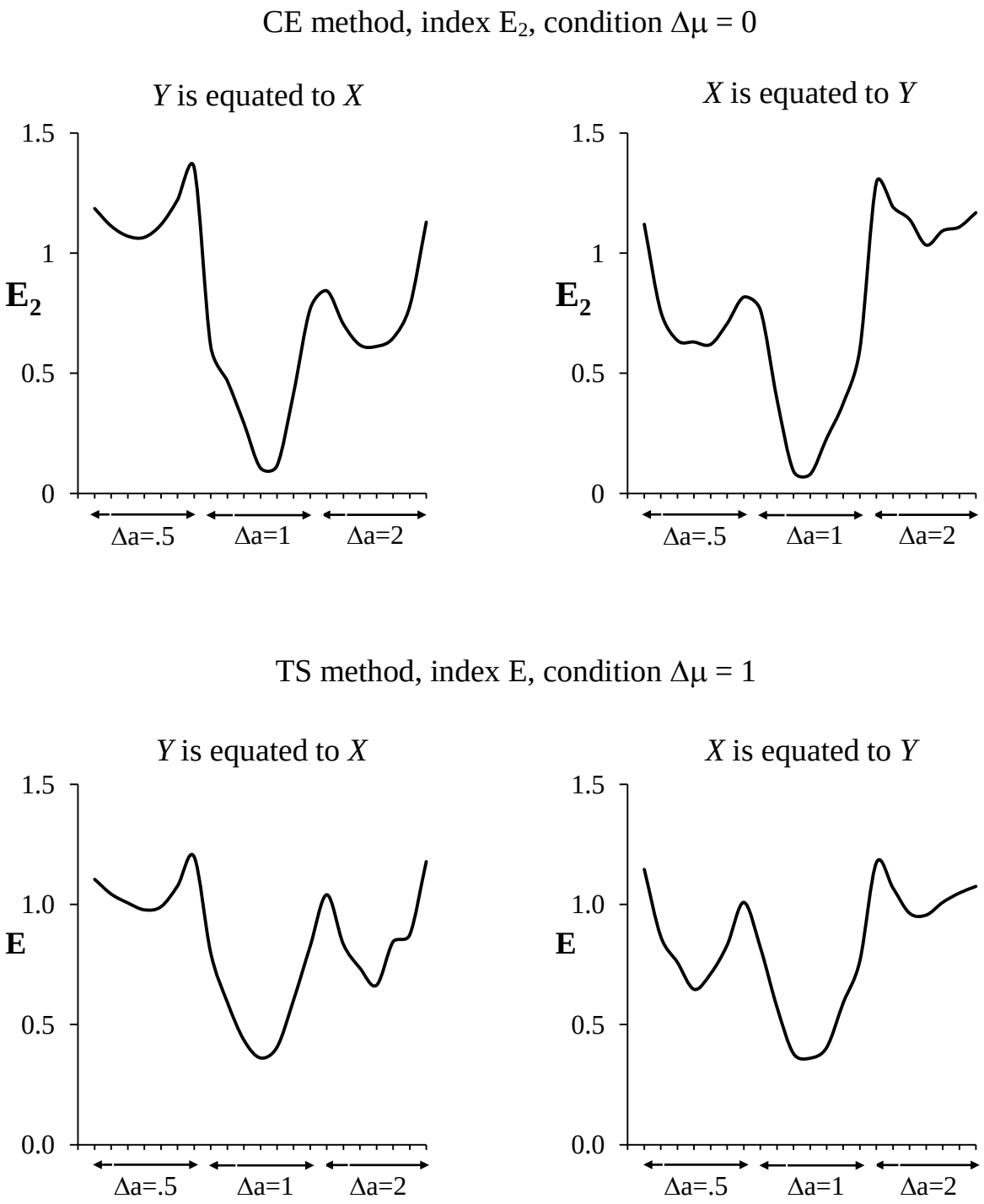

Figure B: Comparing equating results from two directions in two selected cases

REFERENCES

REFERENCES

Albano, A. (2010). Equate: statistical methods for test score equating (R package version 1.1-1).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education, 12*(4), 383-407.

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

Brennan, R. L. (2010). Assumptions about true-scores and populations in equating. *Measurement: Interdisciplinary Research and Perspectives, 8*(1), 1-3.

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

Cui, Z., & Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement, 32*(4), 334-347.

Davey, T., Nering, M. L., & Thompson, T. (1997). Realistic simulation of item response data. ACT Research Report Series.

Divgi, D. R. (1981). *Two direct procedures for scaling and equating tests with item response theory*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

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

Gafni, N., & Melamed, E. (1990). Using the circular equating paradigm for comparison of linear equating models. *Applied Psychological Measurement, 14*(3), 247-256.

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

Han, T., Kolen, M. J., & Pohlmann, J. (1997). A comparison among IRT true- and observed-

score equating and traditional equipercentile equating. *Applied Measurement in Education, 10*(2), 105-121.

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15*, 391-408.

Hanson, B. A., & Beguin, 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), 2-24.

Hanson, B. A., & Zeng, L. (revised by Cui, Z.) (2004). PIE. A computer program for IRT equating.

Hanson, B. A., & Zeng, L. (revised by Cui, Z.) (2004). ST. A computer program for IRT scale transformation.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (Research Report No. 94-4). Iowa City, IA: ACT, Inc.

Hanson, B. A., Zeng, L., & Kolen, M. J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17*(3), 225-237.

Harris, D. J. (1991). A comparison of Angoff's design I and design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement, 28*(2), 221-235.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*(3), 195-240.

Harris, D. J., & Kolen, M. J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement, 50*, 61-71.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*(1), 35-43.

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

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.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*(2), 133-183.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19*, 139-147.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte marlo study. *Applied Psychological Measurement, 6*, 249-260.

Kim, D. I. (2000). A comparison of IRT equating and beta 4 equating. Unpublished Doctoral Dissertation, University of Iowa.

Kim, D. I., Brennan, R., & Kolen, M. (2005). A comparison of IRT equating and beta 4 equating. *Journal of Educational Measurement, 42*(1), 77-99.

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement, 47*(3), 286-298.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47*(1), 36-53.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement, 45*(4), 325-342.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22*(3), 197-206.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education , 3*, 97-104.

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

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

Kolen, M. J., & Brennan, R. L. (1987). A reply to Angoff. *Applied Psychological Measurement, 11*(3), 301-306.

Kolen, M. J., Hanson, B. A, & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*, 285-307.

Liou, M., & Cheng, P. E. (1995). Asymptotic standard of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20*(3), 259-286.

Liu, Y., Schulz, E. M., & Yu, L. (2008). Standard error estimation of 3PL IRT true score equating with an MCMC method. *Journal of Educational and Behavioral Statistics, 33*(3), 257-278.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of*

Educational Measurement, 30(1), 23-39.

Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement, 47*(2), 175-185.

Livingston, S. A.,Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73-95.

Livingston, S. A., & Feryok, N. J. (1987) Univariate versus bivariate smoothing in frequency estimation equating (Research Report 87-36). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1982a). Standard error of an equating by item response theory. *Applied Psychological Measurement, 6*(4), 463-472.

Lord, F. M. (1982b). The standard error of equipercentile equating. *Journal of Educational Statistics, 7*(3), 165-174.

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

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*(4), 453-461.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test score*. Menlo Park, CA: Addison-Wesley.

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

Marco, G. L. (1977). Item characteristic curve solution 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 methods. In D. Weiss (Ed.), *New horizons in testing* (pp. 147-176). New York: Academic Presss.

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

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.221-262). New York: Macmillan.

Philips, S. E. (1985). Quantifying equating errors with item response theory methods. *Applied*

Psychological Measurement, 9(1), 59-71.

Puhan, G. (2010). A comparisons of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*(1), 54-75.

Puhan, G., Moses, T. P., Grant, M. C., & McHale, F. (2009). Small-sample equating using single-group nearly equivalent test (SiGNET) design. *Journal of Educational Measurement, 46*(3), 344-362.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reckase, M. D., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.

Rosenbaum, P. R., & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating . *British Journal of Mathematical and Statistical Psychology, 40*, 43-49.

Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47*(3), 261-285.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309-330.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3*, 105-113.

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

Thissen, D., Wainer, H., Thayer, D. T. (1994). Are tests comprising both multiple-choice and free response items necessarily less unidimensional than multiple-choice tests? *Journal of Educational Measurement, 31*, 113-123.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29*(6), 418-432.

Thomasson, G. (1993). *The asymptotic equating methodology and other test equating evaluation procedures*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Tsai, T-H., Hanson, B. A., Kolen, M. J., & Forsyth, R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education, 14*(1), 17-30.

van der Linden, W. J. (2010). Local observed-score equating. In A. A. von Davier (Ed.) *Statistical models for equating*. New York: Springer.

van der Linden, W. J., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement, 34*(8), 620-640.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudotests constructed from real test data*. ETS Research Report. Princeton, NJ: Educational Testing Service.

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

von Davier, A. A., & Holland, P. W., & Thayer, D. T. (2004b). The chain and post-stratification methods for observed-score equating: Their relationsip to population invariance. *Journal of Educational Measurement, 41*, 15-32.

Wang, T., & Brennan, R. L. (2009). A modified frequency estimation equating method for the common-item non-equivalent groups. *Applied measurement in education, 33(2),* 118-132.

Wang, T., Lee, W-C., Brennan, R. L., & Kolen, M. J. (2008). A Comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632-651.

Wang, T., Hanson, B. A., & Harris, D. J. (2000). The effectiveness of circular equating as a criterion for evaluating equating. *Applied Psychological Measurement, 24*(3), 195-210.

Wyse, A. E., & Reckase, M. D. A graphical approach to evaluating equating using test characteristic curves. *Applied Psychological Measurement.* Prepublished October 7, 2010, DOI:10.1177/0146621610377082.

Zeng, L., Hanson, B. A., & Kolen, M. J. (1994). Standard errors of a chain of linear equatings. *Applied Psychological Measurement, 18*(4), 369-378.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software and manual]. Chicago: Scientific Software International.