NGERANO, GILBERT NJIRU, Ph.D. Optimal Characteristics of Anchor Tests in Vertical Scaling: A Special Case of Non Equivalent Groups with Anchor Test (NEAT) Design in Vertical Scaling. (2019)
Directed by Dr. Richard M. Luecht. 331 pp.

There are multiple empirical issues and complications associated with vertical scaling methods that have not been sufficiently explicated even though there has been scanty research conducted within the general framework of the nonequivalent group with anchor test (NEAT) design. Germane to any vertical scale study is the issue of optimal characteristics of anchor tests whenever the preferred data collection design is NEAT. The main focal point of this research study is to explore some of practical problems as well as complexities that frequently emerge in the context of vertical scaling methods under NEAT design. Specifically, the study investigated various study conditions and comparison of their performance with different equating methods.

This study used both real and simulated data. The real data were from a large-scale testing program for professionals. The simulated study was carried out using 162 conditions, where the major factors included: (1) total test length; (2) item *a*-discrimination parameter; (3) between-grade mean ability difference; (4) distribution of ability difference; and (5) anchor test mean difficulty difference. The results of the simulation indicate that small between-grade mean ability difficult when considered together with a short test length, a moderate item *a*-discrimination parameter, below average distribution of ability difference, and below average anchor test mean ability difference produce most reasonable results.

In addition, the results revealed that equating error somewhat depended on satisfaction of the underlying equating assumptions that are related to a specific equating method under each study condition. For instance, Braun/Holland, Frequency Estimation Equating, keNEATPSE linear, and keNEATPSE equipercentile methods performed almost similarly under all study conditions; however, a closer examination of the above equating methods corroborate that when the equating relationship was linear, keNEATPSE linear outperformed all linear-related equating methods considered in this study. Similarly, when the equating relationship was non-linear, keNEATPSE equipercentile was more accurate in terms of total error, because it produced the smallest RMSE values than all non-linear equating methods. Other results are summarized in greater depth in Chapter V.

OPTIMAL CHARACTERISTICS OF ANCHOR TESTS IN VERTICAL SCALING: A SPECIAL CASE OF NON EQUIVALENT GROUPS WITH ANCHOR TEST (NEAT) DESIGN IN VERTICAL SCALING

by

Gilbert Njiru Ngerano

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2019

Approved by

_____
Committee Chair

To my beloved wife Maria, my sons Dennis Mwaniki and Collins Kariuki, my daughter Joy-Chiara Wanjiru, my mom Jane Mururi, and my late grandparents Zechariah Njiru (aka Carani), Sarah Kiura (aka Ciarunji), and Betha Muthoni (aka Muringo)

APPROVAL PAGE

This dissertation, written by Gilbert Njiru Ngerano, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

The process of writing this dissertation was a special experience for me. The experience that made me develop feelings of utter abandonment and loneliness, but strangely enough, I came to understand and accept the fact that I was surrounded by an extraordinary team of people who cared, loved, and supported me along the way, both in small and big ways. This unique team comprised my dissertation committee members, colleagues, friends, and family members.

First of all, I owe a huge debt of gratitude to the former chair and current chair-cum-director of my committee: Dr. Terry Ackerman and Dr. Richard (Ric) Luecht. Until Dr. Ackerman left the University of North Carolina at Greensboro, he was the chair of my dissertation committee and Dr. Luecht as the director. Dr. Ackerman's support and open-door policy during his time in UNCG was valuable to me. I really appreciate that he kept his promise to serve in the committee as a member even when he was not teaching at UNCG. Dr. Luecht took over the mantle as a chair from Dr. Ackerman and combined the role with that of a director. I am forever grateful for Dr. Luecht`s efforts, advice, and encouragement from the time I conceived the dissertation topic. I am thankful too for his willingness to assume the dual role of the chair and director for my dissertation, a task he painstakingly undertook. His mentorship, immense experience, use of modern technology—special thanks for encouraging me to use Webex platform during dissertation defense—and penchant for high quality work and graphics shaped the direction this dissertation took. You threw down the gauntlet! I learned a lot from you in

class and outside class especially during the whole process of conducting my research study; it would be remiss of me not to express my sincere gratitude for your outstanding mentorship.

I am also grateful to all other dissertation committee members for their guidance, time, and encouragement. Dr. Alina von Davier made sure I got real data and LOGLIN software. Her comments and insight helped me not only clarify technical aspect of the dissertation but also the practical part of it. Dr. Devdass Sunnassee encouraged me to take more SAS programming skills, provided constructive comments, and feedback, which were tremendous in the long run; Dr. John Willse gave valuable comments, thoughts, and facilitated transitioning of my committee chairs. With their collaborative work, I was able to complete writing and defending this dissertation satisfactorily.

I am also thankful to ERM department administrative assistant, Jewell Pradier, for providing me with help whenever I needed it. Also, I am grateful to Sandra Hart for training me on how to use Webex video conferencing equipment, which made my dissertation defense run smoothly and effectively for committee members who were remotely connected.

Also, let me sincerely express my gratitude to all my friends and colleagues who supported and encouraged me in one way or the other throughout the entire process. Thank you, Dr. Lori McLeod, for organizing lunches meant to give an update on my progress and numerous emails of encouragement and hope, even when my spirits were at the lowest ebb. In the same vein, I would like to thank all Educational Research

Methodology graduate students in that department who I interacted with in various classes and for cheering me up to the finish line.

To my late grandparents—Carani, Sara, and Betha—for raising and providing me with formal education, even though you did not have Western education. I know you are smiling in the afterlife to see your first grandson finish writing this dissertation.

Lastly, to my beautiful and lovely wife, Maria, who is the mother of our three children, for standing by me during the entire period of graduate school. Specifically, her unconditional love for our family during the defining moment of writing this dissertation is unmatched. To Dennis, Collins, and Joy-Chiara, you have a bright future; you have no reason whatsoever not to excel in your school work and future endeavors—all of you have unlimited potential!

A big THANK YOU to all of you for your support, encouragement, and believe in me.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

This introductory chapter presents the backbone of this study. Specifically, it focuses on the context, nature and scope of the problem, importance of vertical equating, purpose of the study, key research questions to be answered, and significance of this study to test-score equating research and practice in constructing vertical scales.

1.1. Nature and Scope of Vertical Scales

There is a fundamental need to compare the test scores for different examinees across multiple test forms. When test forms differ in difficulty and/or reliability—which is almost always the case in practice to some extent—we need to equate the scores so that they can be used interchangeably (Kolen & Brennan, 2004, 2014). There are many ways to design and carry out equating studies; however, most fall under one of two basic paradigms: (1) equating with randomly equivalent groups or (2) equating with common persons or common items serving as data links between the test forms. Using randomly equivalent groups, where feasible, therefore provides a *sampling* solution to the equating problem. Using common persons (i.e., the same examinees taking both forms) or common items appearing on the different forms provides a *design* solution. As noted, there are multiple ways to actually design equating studies as well as there are many ways to carry out the actual statistical equating steps (Dorans, Moses, & Eignor, 2011; Holland & Dorans, 2006; Kolen & Brennan, 2004, 2014; von Davier, 2011b, 2011).

Intrinsic to equating is the notion of a score scale. In fact, virtually all classical and item response theory (IRT) equating methods are intended to obtain scores on a common scale to facilitate appropriate comparisons and other interpretations and uses. In educational measurement, the term *horizontal scale* is sometimes used to characterize a scale that is only used within a particular grade. Different grades would have different scales. The term *vertical scale* is used when a single score scale spans many grades. In a practical sense, the distinctions between these two types of tests are somewhat artificial since a horizontal scale could be developed for each of several designated grade bands (e.g., one English language arts or ELA scale that spans grades 4 and 5, another ELA scale covering grades 6 to 8, and a third ELA scale including all examinees in grades 9 to 12). If we put all of those three grade-band specific scores on a single scale, the grade 4 to 12 ELA scale would qualify as a *vertical* scale.

However, there are substantive differences between horizontal and vertical scales. A horizontal scale is typically preferred when the composite of knowledge, skills and abilities (KSAs)—that is, learning—changes across grades or grade-bands, perhaps due to maturation and emphasizing different KSAs within each grade. A vertical scale assumes that the KSAs measured are the same across grades—with the items simply incrementing in difficulty as we move from the lowest to the highest levels of proficiency. Said another way, a horizontal scale may be used when there is a change in the underlying construct across grades or grade bands. A vertical score scale may be desired when the same underlying construct is assumed to be measured across all of the

grades. Vertical scales are typically used for academic assessments that claim to measure student proficiency changes across grades or grade bands (Kolen & Brennan, 2014).

1.2. Practical Importance of Vertical Scaling

Developing and maintaining a vertical score scale requires some type of statistical mechanism for placing scores from students taking different test forms within different grades on a common metric. The mechanisms used fall under a general class of vertical equating methods. The tests to be equated are often of possibly somewhat different content and are usually of unequal difficulty even for adjacent grades.

Vertical scaling has been used in many large-scale educational testing situations that employ a multilevel battery of tests characterized by increasing difficulty across the grade levels. Examples include the Iowa Tests of Basic Skills (ITBS) (Hoover, Dunbar, & Frisbie, 2003), and the TerraNova (CTB/McGraw- Hill, 1997, 2001). The vertical scales are maintained within each content domain (mathematics, ELA, science, etc.) The scale may be used to report grade-level expectations as well as to assess so-called academic *growth*. Kolen and Brennan (2004, 2014) conceptualize growth in two dimensions—i.e., domain versus grade-to-grade definitions. On one hand, growth is discerned as spanning the entire range of test content and, on the other hand, growth is defined in terms of content appropriateness for a particular grade level. Further they contend that there is interplay between definition of growth and types of content domain. For instance, if test content is closely linked to curriculum, it is likely that there is more academic growth with grade-to-grade definition than it is with domain definition.

However, developing and interpreting a vertical scale is characterized by unresolved issues, as Briggs (2010) observes,

> There are some rather thorny issues that need to be resolved to reconcile the creation of vertical scales with the current operational perspectives deriving from Lord's imprint that dominate the research literature. First and foremost we need a better answer to the question of why it is a good idea for large-scale assessments to be placed onto a developmental score scale. If the purpose of vertical scaling is different from the one I defined at the outset of this paper, what is the purpose? It should be clear that any answer having to do with growth implicitly brings us back to the intuition of Figure 1, and that intuition is grounded in an assumption of interval scale properties. If the claim is that the purpose is to produce "quasi-interval" scales this just skirts the issue. Finally, the notion that it should be up to consumers to decide upon a conception of growth that must be met by a vertical scale a priori is little more than an invitation for chicanery. (p. 27)

To address the challenge quite often encountered when implementing domain and grade-to-grade conceptualization of a vertical scale using a common item-linking designs—like overwhelming examinees in lower grades with hard items from upper classes or boring examinees in upper class with too easy items—and adopting a learning progression (or learning trajectory; Confrey, 2012) as a foundation for a common item-linking design has been proposed by Briggs and Peck (2015). According to Briggs and Peck (2015) the strength of learning progressions as a basis of constructing a vertical scale is that they are developed by blending learning theories and empirical studies that are linked to how student reasoning evolves over learning continuum, space, and time.

Despite the many potential pitfalls associated with vertical scales, they continue to be used for largely pragmatic reasons. Patz and Yao (2007) contend that a properly constructed vertical scales facilitate estimation of scores and tracking of growth in those scores over time, allowing more robust comparisons (compared to horizontal scales), and

can lead to more efficient field testing of new content, because items targeted for one grade might be of more appropriate difficulty for an adjacent level. They also contend that vertical scales may make standard setting more reliable, specifically, due to a richer set of items that might be ordered as the density of the items increases. There are, however, many counter arguments to those claims (Briggs, 2013).

This study does not specifically take sides in the substantive debate about the development and use of vertical scales or vertical equating methods. Rather, this study explores some of the empirical issues and complications associated with vertical equating methods for a particular class of equating designs known as the non-equivalent groups with anchor test (NEAT) designs (von Davier, Holland, & Thayer, 2004). Also, this design is called the common items non-equivalent group (CINEG) design or anchor test design (Kolen & Brennan, 2004, 2014). Following some recent work (for example, von Davier et al., 2004; also see Kolen & Brennan, 2014), this type of the NEAT equating design is extended to apply the concept of vertical equating to multistage designs popularized as a type of efficient computerized adaptive testing (CAT) design (Luecht & Nungester, 1998; Zenisky, Hambleton, & Luecht, 2010; Yan, von Davier, & Lewis, 2014). Put differently, the special NEAT design is an amalgamation of some of the ideas or thoughts in the common test designs and equivalent group designs and their new versions. Ultimately, the goal is to merge the vertical scaling methodology with the test design common for multistage adaptive tests (MST). In addition, this idea can help in dealing with some of the missing data by design issues in vertical scaling or linking.

The strength of this research study is on the application of the MST design and the use of panels for the anchors and tests in vertical scaling. This is an area in test score equating, scaling, and linking that has not been adequately explored; therefore, this study has been motivated by the need to address this gap. It is also important to note that originally vertical scaling procedures were constructed primarily for use with the norm-reference elementary achievement test batteries. Similarly, they are used with a few standard-based state testing programs. Although the main goal of equating is to put scores on different test forms on a common metric to facilitate score interchangeability (or comparability for that matter), vertical scaling is not equating in the true sense of equating because the content of the test given across grade levels differ not only on content but also on item difficulties (and to some extent on other psychometric or measurement and statistical properties).

1.3 Purpose of the Study and Research Questions

1.3.1 Purpose

The primary purpose of this study is to explore some of the empirical issues and complications associated with vertical equating methods for a particular class of equating designs known as non-equivalent groups with anchor test (NEAT) designs—i.e., using real and generated data. Selected equating methods under NEAT design are:

1. Tucker linear method;
2. Levine true score method;
3. Braun & Holland linear;
4. Frequency estimation equipercentile equating method;

5. Chained equating linear method;
6. Chained equating equipercentile method;
7. Kernel NEAT post-stratification equating method with a large bandwidth (KeNEATPSE_Linear);
8. Kernel NEAT post-stratification equating method with optimal bandwidth (KeNEATPSE_Non-linear/equipercentile);
9. Kernel NEAT chained equating method with a large bandwidth (KeNEATCE_Linear); and
10. Kernel NEAT chained equating method with optimal bandwidth (KeNEATCE_Non-linear/equipercentile)

Even though the main focus of this dissertation is on NEAT design—and given the nature and design of constructing the vertical scale (see Figure 3.2) herein—it is inevitable not to integrate the Random Groups Design (or the Equivalent Groups design). For this reason, two additional linear and nonlinear equating procedures are considered under RG/EG design—that is, (1) linear equating and (2) equipercentile equating (more details in Chapter II).

Specifically, this study investigated the effect of different equating methods under a variety of simulation conditions on certain properties of a vertical scale and anchor test that was constructed under the NEAT design. For a comprehensive and practical understanding of the impact equating methods may have on vertical scales, the study utilized datasets from large-scale standardized tests for professionals. Further study of these equating methods could give practitioners some practical, useful guidelines, and in-

depth insights regarding which equating method could be preferred under different practical testing realities. The study used five different simulation conditions—(1) test length; (2) item discrimination parameter (*a*-parameter); (3) between-grade mean ability differences (θ, examinee proficiency on the theta scale or the separation of grade ability distributions); (4) distribution of ability difference (Pool information or grade-to-grade ability variability); and (5) anchor test mean difficulty differences or anchor test difficulty variability—to create nine test study designs that may influence the resulting vertical scale. By examining twelve (12) equating methods together with the five simulation study conditions and real data, this study can provide much-needed guidelines for practitioners as to what the consequences of the interpretation and use of these equating methods are on the vertical scales they construct—that is, where vertical scaling will simply work or breakdown. In sum, to evaluate the vertical scale developed, this study mainly focused on five fundamental properties of vertical scaling: test length, item discrimination parameter, between-grade mean ability differences, distribution of ability difference, and anchor test mean difficulty differences to investigate where there is small, medium or large bias, SEE and RMSE under different equating methods for the nine test designs.

1.3.2 Research Questions

In consideration of the preceding scenario, the aim of this study was to address three overarching research questions. These are:

1. How do variations of multiple study conditions (i.e., test length, test mean discrimination, between-grade mean ability difference, distribution of ability

difference, and anchor test mean difficulty differences) affect equating errors—i.e., bias, standard error, and root mean square error—for different equating methods when constructing a vertical scale using a special NEAT design? This main question is partitioned into two sub-questions:

(i) How does this variation affect the equating accuracy across the five study conditions?

(ii) How consistent are the results across the five study conditions?

2. How much difference between anchor test difficulty and the other four study conditions can be endured under each equating method?
3. Does the use of equating introduce more errors than it can be rationalized?

The first two questions were addressed by generated data while the last question was addressed by real data from a large-scale testing program for business professionals.

1.4 Significance of the Study

Given lack of enough research on characteristics of anchor tests in the context of vertical scaling and the scarcity of empirical studies for comparing anchor tests against full tests with equating methods in the NEAT design, and ultimate merger of the vertical scaling methodology with the test design common for MST, this study was motivated to fill that gap. More importantly, blending of NEAT design, equivalent group design and vertical scaling methodology with MST is a nascent idea that can contribute to discourse on dealing with some of the missing data by design issues in vertical scaling or linking. It is hoped that this study will make significant contributions in selecting common items to be used in equating and eventually in constructing a vertical scale. Additionally, results

from this study will provide more comprehensive guidance and insights for practitioners in order to select appropriate vertical scaling methods based on their purposes, goals and objectives. Finally, it is also expected that this study will inform equating practice by suggesting anchor test characteristics under diverse conditions (i.e., both realistic and extreme) that might lead to some equating procedures to either work or fail. Put differently, study of conditions that might significantly contribute to failure in simulation studies is a useful undertaking. This is because—in real world scenario—those failures are not only disastrous but also expensive to examinees and other stakeholders. This risk is not worth taking.

1.5 Description of Notations and Abbreviations

Table 1.1 provides a comprehensive listing of all possible variables in this research study. Additionally, the generic test forms notation and equating methods are shown at the beginning of the table.

Table 1.1

Comprehensive Notational Listing and Descriptions: Test Forms, Equating Methods, and Variables

Notations and Descriptions
F=Base test form (regular test + anchor test)
G=Comparative alternate total test form (regular + anchor test)
RT=Regular (on-grade) test items
AT=Anchor test/Common items
RG=Random groups equating design
NEAT=non-equivalent groups with anchor test design

Table 1.1

Cont.

Notations and Descriptions (cont.)
xt=Observed TOTAL test scores on the BASE form
xa=Observed anchor test scores on the BASE form
xr=Computed observed test scores, excluding anchor test, *xr=xt-xa* for the BASE form
yt=Observed TOTAL test scores on the ALTERNATE form
ya=Observed anchor test scores on the ALTERNATE form
yr=Computed observed test scores, excluding anchor test, *yr=yt-ya* for the ALTERNATE form
tt=True TOTAL test scores on the BASE form
ta=True anchor test scores on the BASE form
tr=Computed true test scores, excluding anchor test, *tr=tt-ta* for the BASE form
ut=True TOTAL test scores on the ALTERNATE form
ua=True anchor test scores on the ALTERNATE form
ur=True observed test scores, *ur=ut-ua* for the ALTERNATE form
eqxt=Equated TOTAL test scores on BASE form, *eqxt*=Equated_to_Y(*xt*)
eqxa=Equated anchor test scores on BASE form, *eqxa*=Equated_to_Y(*xa*)
eqxr=Equated computed observed test scores, excluding anchor tests, on the BASE form, *eqxr*=Equated_to_Y(*xr*)
eqyt=Equated TOTAL test scores on BASE form, *eqxt*=Equated_to_X(*yt*)
eqya=Equated anchor test scores on BASE form, *eqxt*=Equated_to_X(*ya*)
eqyr=Equated computed observed test scores, excluding anchor tests, on the BASE form, *eqxt*=Equated_to_X(*yr*)
Other Abbreviations
a=Discrimination parameter
ATMDD=Anchor Test Mean Difficulty Differences
b=Test Difficulty parameter

Table 1.1

Cont.

Other Abbreviations (cont.)
BH=Braun&Holland Linear Equating Method
BGMAD=Between-Grade Mean Ability Differences
CE=Chained Equating Method
Chained_E=Chained equating Equipercentile method
CINEG=Common items non-equivalent groups
Chained_L=Chained equating Linear Method
Corr=Correlation
DAD=Distribution of Ability Difference (Pool Information)
FEEE=Frequency Estimation Equipercentile Equating
Ke=Kernel Equating Method
KeNEATCE_E= Kernel NEAT Chained Equating (Equipercentile) Method
KeNEATCE_L=Kernel NEAT Chained Equating (Linear) Method
KeNEATPSE_E=Kernel NEAT Post-Stratification Equating (Equipercentile) Method
KeNEATPSE_L=Kernel NEAT Post-Stratification Equating (Linear) Method
NEAT=Non-equivalent groups with anchor test design
P=New Form (Alternate) Population
PSE=Post-Stratification Equating Method
Q=Old Form (Base) Population
RMSE=Root Mean Square Error
S=Synthetic population (or target population, which is combination of populations P&Q)
SEE=Standard Error of Equating
T=Target Population (or synthetic population of P&Q)
V=Anchor test/Common items

1.6 Operationalization of Terms

Alternate forms—only for grades 4 and 6, which is always RT(5.1), i.e., Form 1 of the grade 5 within-grade regular test.

Base forms—these are only for grade 4 and 6, i.e., the within-grade regular tests (RT), plus the corresponding anchor tests (AT) that link those grade-specific scores to the grade 5 scales.

Form—different set of test questions conforming to predefined content and statistical specifications or different editions of a test

Performance levels—this is categorization of students depending on their scores or proficiency categories (e.g., below basic, basic, proficient, and advanced)

Scaling—refers to the establishment of units for reporting measures of proficiency (scale score) and scaling occurs in conjunction with the identification of measurement models.

Score scale—these are scores produced by the process of scaling

Scaled score—these are scores used to reflect performance of an examinee or transformed test score obtained after statistical adjustment to insure consistent meaning, interpretation, and validity of test scores for all examinees.

Vertical scaling—this is the process of placing scores on tests that measure the same domains, but at different levels of education, onto a common metric. The resulting scale is called a vertical scale (developmental score scale). That means a vertical scale encourages monitoring of students' academic growth and achievement or it is a procedure

used to place test scores, across grades within a content area, on a common scale so that a student's progress can be compared over time.

CHAPTER II

LITERATURE REVIEW

This chapter is about review of literature that is relevant to the current study. To expand on this chapter, a general overview of vertical scaling is provided. Then a discussion on criteria for selecting anchor test and whether there is any consensus on what constitutes a best vertical scale follows. Next are the types of data collection designs in vertical scaling and appropriate test score equating methods (or procedures) under NEAT and EG/RG designs. The rest of the chapter delves into general observation on test score equating methods under NEAT design, perspectives on scaling, current research on vertical scaling, and a summary.

2.1 Overview

Johnson and Yi (2011) investigated common item stability check procedures to arrive at vertical linking item sets that would produce constants for computing vertical theta (ability or proficiency) estimates and scale scores on a vertical scale metric. In their research study, they noted that in the context of vertical linking, it is expected that the vertical linking items will display a difference in performance between on-level and off-level examinees, an expectation which is irrelevant in horizontal equating studies. In addition, they found that the presence of linking items that were remarkably easier at the lower level than at the upper level lead to patterns of increasing achievement growth starting at the lowest level to the highest level of the scale.

In vertical scaling literature, there are a number of factors to consider when researchers or practitioners are deriving vertical scale: (1) choice of scaling methodologies which includes statistical methods—(a) Hieronymus scaling (Petersen, Kolen, & Hoover, 1989); (b) Thurstone scaling (Gulliksen, 1950; Thurstone, 1925, 1938); and (c) IRT calibration and scaling—recent scalings have frequently applied IRT and tend to replace the Thurstonian scaling which has got a long history in educational and psychological testing; (2) vertical linking strategies across levels—(a) concurrent; (b) separate level-groups; and (c) level-by-level; and (3) types of vertical equating methods or scaling designs—(a) scaling test; (b) common items across levels; and (c) equivalent groups design. An excellent treatment of this topic is found in the work of Kolen and Brennan (2004, pp. 381–412). Other than considering scaling methods, strategies for vertical linking and different types of vertical equating methods, other factors that are important when designing a vertical scale have also been investigated; there are studies that have analyzed these factors—that is, cross-grade scale expansion/shrinkage (Ito, Sykes, & Yao, 2008), test content, subject area, IRT scoring procedures, and proficiency estimators (Tong & Kolen, 2007)—and demonstrated how multiple combinations of these variables can have an effect on resulting vertical scales. Although these vertical equating studies have tremendously enriched the equating literature, they have been criticized for failing to give concrete direction on factors to consider in order to construct a reliable and best vertical scale. Furthermore, practitioners or experts that are engaged in vertical scales are left to decide which factors to combine and analyze in relation to how

they affect the vertical scale within the general framework of unique testing and assessment program (Johnson & Yi, 2011).

Kolen and Brennan (2004) have pointed out that a number of factors might affect vertical scaling results in any of the scaling methodologies cited previously. Fundamentally, these factors include: (1) the data collection design, (2) dimensionality—the complexity of the subject matter area; (3) the curriculum dependence of the subject matter area; (4) test characteristics—average item difficulty and discrimination, and relationships of the item characteristics to group proficiency; (5) item type—multiple-choice (MC) and constructed response (CR); (6) grade level; and (7) nonlinear scale transformations following implementation of a scale method.

In the case of the common item approach, vertical linking items are assessed within on-level test forms and within off-level test forms. The next section examines in details anchor related studies.

2.2 Anchor Test

In the context of classical test theory (CTT), the common items are mainly meant for adjusting proficiency differences in the groups of examinees (e.g., Angoff, 1968, 1971; Gulliksen, 1950; Holland & Dorans, 2006; Kolen & Brennan, 2004; Petersen et al., 1989). An important aspect of the NEAT design is tied to the construction of an anchor set of items (common items).Three important properties of an anchor test are length, content, and statistical characteristics— these are some of the properties used as guidelines for linking items for horizontal equating; they are also applicable in the vertical scaling context with the goal of establishing a strong measurement link that

enhances a tenable vertical scale (scale of growth) across all grades (Johnson & Yi, 2011). These features are discussed in detail in the proceeding paragraphs.

It is rather well-known that score scale reliability is directly associated with test length; that is, adding more test items or measurement opportunities tends to increase the reliability of the test scores. Angoff (1968) observed that longer tests are more reliable than shorter ones that measure similar construct. Put differently, the statistical association between reliability and test length has an impact on the quality of the linking mechanism—in this case, anchor test or linking items for vertical equating. The impact of test length has been explored and explicated in equating literature and has been shown to have a direct effect on the reliability of test scores (Allen & Yen, 2002). Furthermore, it can be argued that the magnitude of equating error—that is, random error expressed in terms of the standard error of equating and systematic error decomposed into bias and measurement errors—can be evaluated to assess the degree of accuracy of any equating method when applied to test scores. Specifically, this can be done when observed test scores are included in the process of equating. An example of this application is tied to equating methods under NEAT design. The research literature recommends anchor test lengths in comparison to the operational test—that is, how many items are required for placing item parameters on the common scale. Most of the horizontal equating research suggests a rule of thumb of having the anchor test represent at least 20% of the test or at least 15 items in case of IRT equating framework (e.g., Kolen & Brennan, 2004). Fitzpatrick (2008) concluded that shorter anchor test lengths seriously compromised the integrity of the equating results under IRT equating methods. She suggests that instead of

lengthening the anchor test, we should use survey sampling techniques like optimal allocation procedure (Sudman, 1976). Optimal allocation procedure involves sampling more elements from strata with more sampling variability. When this technique is applied to sampling items to be included in the anchor test, items from subsets known to have more variability on the basis of content or statistical characteristics would be selected in bigger proportions than subsets showing less variability given these attributes (Deng, Sukin, & Hambleton, 2009).

Another important consideration for NEAT equating methods is the inclusion of both the variances and correlation between the base form and the anchor test scores whenever equating transformation functions are computed. For instance, as reliability increases, the variances of observed test score decreases as the correlation of these scores is somewhat strengthened. The equating literature further observes that wherever distributions of the observed test scores are manipulated during equipercentile equating or moments are used to approximate equating transformation constants like in the case of linear equating, the effect of differences in reliability is not predicable.

For typical NEAT designs, it is rather common wisdom to design the anchor test to be statistically similar and content proportional to the test specifications for the operational test is an important consideration (Cook & Eignor, 1991; Cook & Petersen, 1987; Dorans, Kubiak, & Melican, 1998; Hambleton, Swaminathan, & Rogers, 1991; Klein & Jarjoura, 1985; Kolen, 1988; Kolen & Brennan, 2004; Petersen et al., 1989; Petersen, Marco, & Stewart, 1982; Sinharay & Holland, 2006, 2007, 2008). This wisdom actually stems from a fundamental assumption about the equivalence of the regression of

the total-test observed scores on the anchor test for NEAT designs (Kolen & Brennan, 2014). That is, we assume that we can use the regression of the anchor test to essentially predict performance on the portion of the total test missing for each of the involved groups. When a content area is omitted, over-represented, or under-represented and growth actually occurs in this area; therefore, the amount of overall growth for the construct being measured may be incorrectly estimated (Deng et al., 2009). Furthermore, it can lead to threats to validity—construct underrepresentation and construct irrelevant variance (AERA, APA, & NCME, 2014; Downing, 2002, 2005; Downing & Haladyna, 2004; Messick, 1989) and subsequently invalidate equating inferences, conclusions, meaning, interpretation and use of test scores that are made. For this reason, the linking of tests may be incorrect because any change that occurs over time should be reflected only in the common items (Deng et al., 2009).

Supporting evidence to the recommendation that anchor and operational tests contain equivalent proportions of items representing multiple content areas is well documented in the equating literature. A widely cited work is that by Klein and Jarjoura (1985). These authors conducted a study to compare a content representative anchor against a long anchor without content representation. The result of their study was that the shorter anchor with content representation outperformed the long anchor without content representation under two classical test theory equating methods—Tucker linear equating and Levine equating. Another study examined four anchor item sampling designs and four equating methods—two of them used IRT designs (Yang, 2000). The findings of this study indicated that equating accuracy was best when using the item-

sampling scheme that chose items to be included in the anchor test in a manner that the anchor items proportionally matched specifications of the content for the entire test.

Recommendations from content matching equating research studies propose that the anchor test be made up of items that mimic the statistical characteristics of the operational test (Angoff, 1968; Cook & Eignor, 1991; Dorans et al., 1998; Kolen, 1988; Kolen & Brennan, 2004; Petersen et al., 1989; Petersen et al., 1982). In the equating literature, this is referred to as a "mini-test." The mini-test is made up of items with similar mean difficulty and similar range of difficulty. Scholastic Aptitude Test (SAT) and Test of Standard Written English (TSWE) were studied by using various equating methods—mean difficulty similarity, external vs. internal, and content similarity (Petersen et al., 1982). They concluded that matching the mean difficulty of test and anchor test items—that is, based on equating a test using equipercentile methods for example—was a more important factor to establish a reliable anchor test for equating test forms. On the same vein, Petersen et al. (1982) found that when there are differences in difficulty between the anchor and operational test forms the mini-test performs best as an anchor and that equipercentile equating outperforms linear equating.

Although the "mini-test" can be applicable when using an internal anchor, some researchers in test score equating have not agreed if the same ideas can be used when considering the external anchor design. In their study, Sinharay and Holland (2006, 2007, 2008) proposed the "semi-midi and midi-test" forms as anchors instead of the mini-test form. The semi-midi and midi-test are characterized by the spread of the item difficulties which are more constrained to preserve items that are very easy or very difficult. When

using post-stratification and equipercentile equating methods, these writers observed that the semi-midi and midi-test performed better than the mini-test—although, at times they might all perform reasonably well. When the mini-test and semi-midi and midi test were correlated to the complete test, they found that the latter has a higher anchor-test-to-complete test correlations. Another recommendation is that the anchor-test score should be a proxy of the proficiency measured by the test and the equating should be conditional on this score (van der Linden & Wiberg, 2010).

Linking item guidelines of horizontal equating, mentioned above, are applicable in the vertical linking context so that a strong measurement link can be established that will foster a reasonable scale of growth across all levels (Kolen & Brennan, 2004, 2014). Kolen and Brennan (2004, 2014) observed that vertical scaling is "a very complex process that is affected by many factors," which includes the design for data collection, the content area being studied, the test itself, a scaling method, and the computer program used (p. 418). The same sentiment is echoed by Harris (2007) when she noted that "vertical scaling is a complex process, involving philosophical, technical, and practical issues" (p. 251). Reviewed literature suggests that vertical scaling is design-dependent (Harris, 1991), group-dependent (Harris & Hoover, 1987; Skaggs & Lissitz, 1988; Slinde & Linn, 1979a), and method-dependent (Kolen, 1981; Skaggs & Lissitz, 1986b).

2.3 Is There a Best Vertical Scale?

Yen (1986) contends that there is no best vertical scale. In the same vein, Harris (2007) noted that despite the fact that it can be disconcerting that there is no agreement on the best way to construct a vertical scale, it is comforting at the same time. They

advise that “instead of arguing which single scaling method is the best, we might do better to see which slate of options work for which purpose, under which conditions” (p. 251). Similarly, Kolen and Brennan (2004) suggest that practitioners should embrace a scale that they consider to reflect the nature of growth for their tests. Certainly, such decisions will affect the nature of the scale construction; therefore, it behooves the test developer to informed examinees and other stakeholders about this potential ambiguity in scaling (Tong & Kolen, 2007). Although vertical scales are useful in tracking students’ academic growth and achievement from year to year and provide intervention where required (Harris, 2007), Tong and Kolen (2007) advise to be cautious whenever the interpretation of scores from a vertical scale is made.

2.4 Designs for Vertical Scaling: Types of Data Collection Designs

Three approaches to data collection for vertical scaling have been proposed in the equating literature (e.g., Holland & Doran, 2006; Kolen, 2006; Kolen & Brennan, 2004, 2014; Young, 2006). In general, a data collection design may use one of these approaches: (a) Common item or CINEG/NEAT design; (b) Equivalent group/Random group designs; and (c) Scaling test designs. Each of the three designs is summarized here for completeness; an in-depth and thorough treatment is provided by Kolen and Brennan (2004, 2014). The current study focuses on the first and second vertical scaling designs—common item and equivalent group designs in addressing issues and complications encountered in vertical scaling.

2.4.1 Common Item Design

In the common item design, each test level is administered to examinees at the appropriate grade. When the common item set scores contribute to the total test scores the common item set is said to be internal; otherwise, it is external if it doesn`t contribute to the total score (Kolen & Brennan, 2004, 2014). This design takes advantage of the overlapping content of adjacent levels. This feature makes it possible to conduct scaling in subjects like math and reading because some common or similar concepts are found in adjacent levels. Its application is also in achievement and aptitude test batteries administered in elementary schools in the United States.

It is important to note that item blocks that are common between adjacent grades are used for linking purposes. This follows a chaining process where scores from all grades are placed on the base grade. The design is easily implemented in standard administration conditions with the standard test batteries (Kolen & Brennan, 2004). One key issue associated with the common design is that it is affected by context effect. This is because at the lower level the common items between the adjacent grades are placed at the end of the test while they are placed at the beginning of the test for the higher grade (Kolen & Brennan, 2004). To go around the issue of context effect in this study, all anchor test items are put at the beginning of the test.

In summary, common item design produce vertical scale through a linking chain. Common items are sampled from adjacent grades which are level appropriate to each grade. In practice, selecting common items for this design is also based on: (1) content representativeness of a set of items from the lower as well as the upper grade levels

(Figure 2.1); (2) a range of grades—i.e., selection of items not necessarily from the adjacent grades (Kolen, 2011). It is an empirical question whether these various ways of selecting common items would produce different scaling results. However, this study adopted the first approach of selecting common items based on psychometric specifications like item difficulty and item discrimination parameters for adjacent grades rather than content representativeness.

2.4.2 Equivalent Group Design/Random Group Design

From methodological and philosophical perspectives, equivalent group design and random group design are the same; therefore, there is no distinction that has been made between the two in this study. In fact, the two terminologies are used interchangeably in this dissertation. It is important to note that the equivalent group or random group design is another approach used to gather data for building a vertical scale. The equivalent groups are obtained by spiraling, which results in groups that have a smaller variance than they would have if they were random. In this design, randomly equivalent groups of examinees are administered either the level appropriate test (on-level test) for their grade or the level just below or above (the off-level test) their grade. Although in vertical equating literature the off-level test is often associated with the test from the immediate lower grade level, in this dissertation it is also considered as a test just above the given grade. Specifically, random assignment using spiraling ensures that test questions administered are not too difficult or too easy for each grade.

Except for the lowest grade, each group of examines per grade is administered one of the two levels of the test. The data gathered for this administration is used to place

scores from all of the test grades on a common metric by using chaining across grades. The design does not use common items found in adjacent levels. In this study, equivalent group design is used for equating within grade forms—specifically, with reference to the base grade test forms—and to provide a linking mechanism to common item equating.

2.4.3 Scaling Test Designs

In the scaling test design, a special test is built that spans the content domain across all grade levels and puts all the items on one form. The scaling test is administered to all students across the grades alongside test level appropriate for their grade. Although this design is hard to implement in a practice, it outshines the other two designs because it ranks all students in all grades in one domain. This design has been criticized for lacking useful information when students are tested with too easy or too difficult items (Carlson, 2011).

Alternatives to the first two designs—common item and equivalent group designs—have been proposed, discussed and illustrated by Carlson (2011). In case of the common item design, a group of students at each grade level is identified to be administered blocks of items that are composed of (1) the anchor blocks (common items) shared with adjacent—that is, either below or above—grade levels, and (2) blocks of unique items in their grade level. The only feature that distinguishes the common item design postulated by Kolen and Brennan (2004) and the variant posited by Carlson (2011) is that the latter incorporates in his design on-grade item block for each grade level.

2.5 Equating Methods/Procedures

There are a number of equating procedures under NEAT design from which a practitioner or a researcher in vertical scaling can choose. In this dissertation, the rationale for selecting multiple test score equating methods, which were previously outlined in Chapter I, is based on the fact that they perform better when there are substantive disparate group abilities in the context of horizontal equating. This notion can be expanded and applied in vertical scaling and linking studies where non-equivalent of target populations is prevalent. In the world of vertical scaling, it is assumed that the group abilities (or even learner's ability) vary across grades and within grades. Additionally, Sinharay and Holland (2009) recommend that the operational testing programs to apply different test score equating methods and study the variation (or differences) among their equated score results. Also, research studies in vertical scaling are popular with the NEAT data collection design. Even though these methods under NEAT design are appropriate in vertical scaling situation, they have their faults. Further, some of these equating methods make indefensible underlying assumptions about missing data by design and score distribution, which often time are never tested in practice (Holland, von Davier, Sinharay, & Han, 2006). The test score equating methods in this subsection are revisited from Chapter I and re-classified according to data collection designs, which are NEAT and EG/RG designs, and on basis of their nature of the equating function—i.e., either linear or nonlinear. These are:

(a) Equating Methods Under NEAT Design

(i) Tucker linear method

(ii) Levine true score Method

(iii) Frequency estimation equipercentile equating (FEE) Method

(iv) Braun & Holland linear method

(v) Chained linear

(vi) Chained equipercentile

(vii) Kernel NEAT post-stratification equating (KeNEATPSE)

(a) Linear

(b) Non-linear

(viii) Kernel NEAT chained equating (KeNEATCE)

(a) Linear

(b) Non-linear

(b) Equating Methods Under Equivalent/Random Group Design

(a) Linear

(b) Equipercentile

Random groups and NEAT designs were used to compare and investigate performance of twelve different equating methods under different study conditions. These equating methods can be classified into two families—that is, linear and non-linear. In the equating literature curvilinear methods are also referred to as equipercentile or curvilinear. The equating methods under NEAT design that are linear are Tucker method, Levine-true method, Braun-Holland method, chained linear method, kernel NEATPSE linear method, and kernel NEATCE linear method. The equipercentile methods under NEAT design include frequency estimation equipercentile equating

method, chained equipercentile, kernel NEATPSE equipercentile method, and kernel NEATCE equipercentile method. Linear and equipercentile equating methods are also considered under equivalent groups design. Next is a description of linear and equipercentile procedures under random group design and then each of the other methods or procedures (outline above) are considered in the context of NEAT design.

2.5.1 Equivalent Groups Design/Random Groups Design (RG)

As noted previously, in the random group equating design, examinees are randomly assigned the test form to be operationalized. A spiraling process can be used to randomly assign different test forms under this design. This typically leads to comparability of randomly equivalent groups that take Form X and Form Y. Under this design, "the difference between group-level performance on the two forms is taken as a direct indication of the difference in difficulty between the forms" (Kolen & Brennan, 2004, pp. 13–15). More discussion on practical features and issues involved in random group equating design are explicated by Kolen and Brennan (2004, 2014).

2.5.1.1 Linear Equating Method. Linear and mean for the random groups design is extensively covered by Kolen and Brennan (2004, 2014). In this design, the equations use only the first two moments—mean and standard deviation—of the marginal distributions for Forms X (alternate Form) and Y (the base form).

For mean equating, the equation function that puts raw scores for the new Form X on the scale of the raw scores for the old Form Y is computed as follows:

$$my(x) = y = x - \mu(X) + \mu(Y). \qquad \text{(Eq. 2.1)}$$

Similarly, for linear equating the function is governed by:

$$ly(x) = y = [\sigma(Y)/\sigma(X)]x + [\mu(Y) - \{\sigma(Y)/\sigma(X)\}\mu(X)]. \quad \text{(Eq. 2.2)}$$

$$= A + Bx, \quad \text{(Eq. 2.3)}$$

where

$$\text{slope} = B = \sigma(Y)/\sigma(X) \text{ and} \quad \text{(Eq. 2.4)}$$

$$\text{intercept} = A = \mu(Y) - B\mu(X) \quad \text{(Eq. 2.5)}$$

Remarkably, Equation 2.1 is similar to Equation 2.2 if and only if the slope is 1, i.e., $\sigma(Y)/\sigma(X)$. That means $ly(x) = my(x)$ give exactly the same results when $\sigma(Y)/\sigma(X)$, $B = 1.0$. Linear equating adjusts one set of scores so that the first and second moments of the score distribution are equal; therefore, it involves an adjustment to the center or location of the scale and the unit size. For realized or observed scores, x on Form X and y on Form Y are standardized—i.e., centered at the mean and normalized to the standard deviation—and set equal. Under certain conditions, linear equating is no different than linear regression. This is because when X and Y are perfectly correlated, linear equating and regression produce similar results. Again, in linear regression, the slope is given by:

$$\beta = \rho(X,Y)\sigma(Y)/\sigma(X), \quad \text{(Eq. 2.6)}$$

but

$$\beta = \rho B = B \quad \text{(Eq. 2.7)}$$

when X and Y correlation is a unit.

In other words, in the equating literature, it has been shown that ρ(X,Y), the correlation between X and Y, impacts on both the slope (β) and intercept (α) in case of regression, but does not affect the slope (A) and intercept (B) for linear equating.

2.5.1.2 Equipercentile Equating Method. What sets equipercentile equating apart from mean and linear equating counterparts under random group design is the fact that it adjusts the shape of the cumulative score distribution of Form X to match the cumulative score distribution of Form Y in the target population. In fact, it allows for differential changes across the score scale, rather than merely adjusting the first two moments like it is the case with linear equating. The great challenge for adoption of equipercentile equating fundamentally lies on its requirement for very stable distributions which should essentially be truly randomly sampled groups from a common target population.

Braun and Holland (1982; see also Kolen & Brennan, 2004, 2014) have demonstrated that a symmetric equipercentile equating function, ey, is defined to be so if G* = G and that x and y are continuous random variables or continuized, thus:

$$ey(x) = G^{-1}[F(x)], \qquad \text{(Eq. 2.8)}$$

where G* is the cumulative distribution function (cdf) of score on Form X converted to the Form Y scale;

G is the cumulative distribution function of Y in the same population;

F is the cumulative distribution function of X in the same population; and

G^{-1} is the inverse of the cumulative distribution function, G.

Stated differently, ey(x) is the score on the Y-scale associated with the percentile rank of F(x).

2.5.1.3 Score Discreteness, Continuization Process, and Smoothing in Equipercentile Equating.

In practical equating realities, the x and y test scores are often non-negative integers that correspond to the number of correct items scored by a test taker. Score discreteness somewhat presents difficulties in obtaining percentile points on the scale of Y. This is because it is problematic if not impossible to get an integer score on Y that has a percentile rank exactly equal to F(x). The equating literature recommends continuization of the densities for X and Y. Two popular methods of continuization are in use: (1) linear interporation (Angoff, 1971; Kelly, 1923; Kolen & Brennan, 2004, 2014; Otis, 1916; Petersen et al., 1989); (2) Gaussian kernel smoothing—to continuize the discrete distributions (Holland & Thayer, 1989; von Davier et al., 2004).

Smoothing can be done before (presmoothing) or after (postsmoothing) calculating the equipercentile equivalents, êy(x); the focus is to try to preserve the moments after smoothing—this is an important consideration because it relates to one of the properties of smoothing. That is accuracy. Other smoothing properties discussed by Kolen and Brennan (2014) are flexibility, statistical framework and empirical research base. In presmoothing, the scores are smoothed while in postsmoothing the equipercentile equivalents are smoothed directly. Presmoothing methods include 2 or 4 parameter beta (compound) binomial and log-linear. Commonly used postsmoothing method is cubic-spline (Kolen & Brennan, 2004, 2014). Although the main purpose of smoothing in

equipercentile equating is to reduce the equating error, it has been shown in the equating literature that it can also introduce the same.

2.5.2 NEAT Design: Missing Data by Design

NEAT design involves administering Forms X and Y which share a set of common items (anchor test) to a target population T, which is composed of two different populations—population P and Q (see Eq. 2.9). Table 2.1 displays a visual pattern of the data for the NEAT design (Sinharay & Holland, 2008).

Table 2.1

An Illustration of the Non-equivalent Groups with Anchor Test (NEAT) Design

Target Population	Population/Test Form	X	AT(A or V)	Y
T	P	√	√	—
	Q	—	√	√

Note. √-symbol indicates a test form administered to a sample of population. A dash (—) shows a test form was never taken by either P or Q, hence missing data by design.

If Population P takes Form X, Population Q is administered Form Y; both Populations will take a common set of tests (AT or A or V) which is used for equating purposes. That means when P and Q are different or non-equivalent the statistical role of the common groups of items is: (i) to remove bias; (ii) increase precision in the estimation of the equating function (Holland, Dorans, & Peterson, 2007); (ii) to adjust for population differences or to account for any differences in ability between non-equivalent groups taking the new and old test forms (Kolen & Brennan, 2014); and (iv) to adjust for the differences in overall difficulty between X and Y (Ricker & von Davier, 2007; von

Davier et al., 2004). Other uses of the information gleaned from the anchor test item scores mentioned in the literature are: (i) it allows a new test to be used and equated at each successive operational test administration; (ii) it facilitates formulation of untestable, missing-data assumptions needed to interpret the linking results as constituting an equating; (ii) it is used as a conditioning variable, for instance in the case of the Tucker method and poststratification equating; (iv) it is used as a middle link, such as in chained equating; (v) it is used together with classical test theory. In this case, examples are Levine observed-score equating, hybrid Levine equipercentile equating and poststratification equating for true anchor scores (von Davier & Chen, 2013).

In this design, population P will never take Form Y. Conversely, Form X scores are never observed in population Q. For this reason, the NEAT design is a special case of missing data by design—i.e., data are not missing due to examinees skipping questions or any other type of testing situations (Sinharay & Holland, 2008; von Davier et al., 2004). Similarly, Liou, Cheng, and Li (2001) pointed out that the NEAT design is a case of missing data that are missing at random (MAR) in the technical usage advanced by Little and Rubin (2002). Missing data assumptions under NEAT design are essentially untestable in practical equating realities. For more details about missing data by design in NEAT, assumptions under poststratification equating (PSE), chained equipercentile equating (CE), item-response-theory observed-score equating (IRT OSE) and the concept of synthetic population (Braun & Holland, 1982) the reader is referred to the studies conducted by Holland and Dorans (2006), Sinharay and Holland (2000), and Holland et al. (2007).

Braun and Holland (1982) define synthetic population (S) as a target population (typically, S is never observed) for the NEAT design that is created by weighting populations P and Q. Thus,

$$T = wP + (1-w)Q, \qquad \text{(Eq. 2.9)}$$

where the sum of w + (1-w) = 0, i.e., the weights must function as proper density (Gulliksen, 1950); and their values greater than zero (w, 1-w ≥ 1). Various choices of weights, w and (1-w) include use of 1 and 0, equal weights like 0.5, sampling weights for the two populations and proportional probability weights. Considerable evidence has been shown that the choice of w has a relatively insignificant impact on the equating results (von Davier et al., 2004). This insensitivity to w has been cited as an example of upholding the population invariance assumption—a requirement in equating (Lord, 1980; Holland et al., 2007).

2.5.2.1 Tucker Linear Method. The Tucker method uses means and variances (or standard deviation scores) to convert observed test scores on Form X to the scale of observed scores on Form Y by use of the following linear function.

$$ly_s(x) = y_s = [\sigma_s(Y)/\sigma_s(X)]x + [\mu_s(Y) - \{\sigma_s(Y)/\sigma_s(X)\}\mu_s(X)] \qquad \text{(Eq. 2.10)}$$

This linear function is exactly the same as Equation 2.2 except that the former has a subscript s to denote synthetic population and that the four parameters— $\sigma_s(Y)$, $\sigma_s(X)$, $\mu_s(Y)$ and $\mu_s(X)$—are unobserved; they can be estimated from the parameters computed in Population P and Q (see Kolen & Brennan, 2014, Eqs. 4.2–4.5, p. 104).

This method makes two types of assumptions—(1) linear regression assumptions and (2) conditional variance assumptions—so that the four parameters can be estimated; they are not directly observable.

Assumption 1:

The regression of X on V (or Y on V) is assumed to be the same linear function for Populations P and Q. Setting α and β to represent regression slopes and intercept respectively,

$$\alpha_P(X|V) = \sigma_P(X, V)/\sigma_P{}^2(V) \quad \text{(Eq. 2.11)}$$

$$\beta_P(X|V) = \mu_P(X) - \alpha_P(X|V)\mu_P(V) \quad \text{(Eq. 2.12)}$$

The regression slope and intercept for the regression of Y on V can be computed in a similar way as in Equation 2.11 and 2.12. The two quantities are observed because they are calculated from realized data. Because Population Q never took Form X, the slopes and intercepts can be estimated as:

$$\alpha_Q(X|V) = \sigma_Q(X, V)/\sigma_Q{}^2(V) \quad \text{(Eq. 2.13)}$$

$$\beta_Q(X|V) = \mu_Q(X) - \alpha_Q(X|V)\mu_Q(V) \quad \text{(Eq. 2.14)}$$

Similarly, because Population P never took Form Y, the slope and intercepts can be calculated as in Equation 2.13 and 2.14. In summary, the regression assumption for X and V (or Y and V) is

$$\alpha_Q(X|V) = \alpha_P(X|V) \quad \text{(Eq. 2.15)}$$

and

$$\beta_Q(X|V) = \beta_P(X|V) \quad \text{(Eq. 2.16)}$$

Assumption 2:

The conditional variance of X given V (or Y given V) is assumed to be the same for Populations P and Q (see Kolen & Brennan, 2014, Eq. 4.12, p. 106).

The rationale for Tucker equating method is based on the fact that the means and standard deviations (variances) are observed-score parameter estimates adjusted in the synthetic population based on the anchor test—that is, test scores based on common items given to different Populations P and Q. Furthermore, if $\mu_P(V) = \mu_Q(V)$ and $\sigma_P(V) = \sigma_Q(V)$, the corresponding synthetic parameter estimates would equal the observed test score moment. Finally, the Tucker method works equally well with both internal and external anchor tests.

2.5.2.2 Levine True Score Method. Under Levine true-score equating, three assumptions are made about *true* test scores for Forms X and Y and the anchor test, V. These assumptions are the same for Levine observed score equating method (Levine, 1955). The assumptions of classical congeneric model are added to the other three assumptions such that the γ, or (λ_X/λ_V), the effective test length, for Levine observed-score equating with an external anchor is

$$\gamma_P = [\sigma^2_P(X) + \sigma_P(X, V)]/[\sigma^2_P(V) + \sigma_P(X, V)]; \quad \text{(Eq. 2.17)}$$

and

$$\gamma_Q = [\sigma^2{}_Q(Y) + \sigma_Q(Y, V)]/[\sigma^2{}_Q(V) + \sigma_Q(Y, V)] \quad \text{(Eq. 2.18)}$$

The effective test length, γ, or λ_X / λ_V, is proportional to both the reliability and error variances. For the internal anchor case with Levine's observed score method under the classical congeneric model, see Kolen and Brennan's (2014) Equation 4.53 and 4.54, p. 114.

Under the classical congeneric equating model—and to be consistent with Feldt and Brennan (1989)—we assume that X and V (or Y and V) are linearly related with slope, λ, and intercept, δ, such that

$$X = T_X + E_V = (\lambda_X T + \delta_X) + E_X \quad \text{(Eq. 2.19)}$$

$$V = T_V + E_V = (\lambda_V T + \delta_V) + E_V \quad \text{(Eq. 2.20)}$$

$$\sigma^2(E_X) = \lambda_X \sigma^2(E) \text{ and } \sigma^2(E_V) = \lambda_V \sigma^2(E) \quad \text{(Eq. 2.21)}$$

Assumption 1:

There is a perfect correlation between T_X and T_V (or T_Y and T_V) in Population P and Q.

Assumption 2:

The regression of T_X onT_V (or T_Y on T_V) is assumed to be the same linear function for both Populations P and Q.

Assumption 3:

The measurement error variance for X (or Y) is the same for Populations P and Q under the classical test theory model.

Although the Levine observed score method makes assumptions on true scores on T_X, T_Y and T_V it uses Equation 2.10 to relate observed test scores on Form X to the scale of observed test scores on Form Y (Kolen & Brennan, 2014).

Therefore, under classical test theory, observed scores are taken to be the same as true scores and the following equation is used for Levine-true score equating with observed scores (Kolen & Brennan, 2014).

$$ly_s(t_x) = \sigma_s(T_y)/\sigma_s(T_x)[t_x - \mu_s(X)] + \mu_s(Y), \qquad \text{(Eq. 2.22)}$$

where T=true score and s=synthetic population.

2.5.2.3 Braun and Holland Linear Method. Braun-Holland linear method, as the name suggests, was first proposed by Braun and Holland (1982). The method uses the first two moments (or mean and standard deviation) to conduct linear equating under the frequency estimation method (frequency estimation method is discussed next after Braun and Holland method). The resulting synthetic population means and standard deviations are substituted into the following general linear equating function for the NEAT design.

$$\hat{I}y_s(x) = \hat{\sigma_s}(Y)/\ \hat{\sigma_s}(X)[x - \hat{u}_s(X)] + \hat{u}_s(Y) \qquad \text{(Eq. 2.23)}$$

An equating that results from using Braun-Holland linear method is similar to the Tucker linear method if the regressions are strictly linear and homoscedastistic—i.e., if regressions of X on V and Y on V are linear; and if the regressions of X on V and Y on V are homogeneity such that $\sigma^2(X|v)$ and $\sigma^2(Y|v)$ are identical for all v (Braun & Holland, 1982). In other words, Braun-Holland method is a special case (or generalized form) of

the Tucker method that works whether the regressions of the total test on anchor test items are linear or nonlinear (Kolen & Brennan, 2014).

2.5.2.4 Frequency Estimation Equipercentile Equating Method (Frequency Estimation). Frequency estimation can be defined as an equipercentile (nonlinear) method of estimating the cumulative test score distribution for two or more forms within the synthetic population, using a group of common items without using the moments of the two forms (Angoff, 1971; Braun & Holland, 1982; Kolen & Brennan, 2004, 2014). Percentile ranks are calculated from the cumulative frequency distributions and then the forms are equated by equipercentile methods. The common items, V, is used to estimate the distribution of Population Q taking Form X and Population P taking Form Y. Table 1 shows that Population P and Q never took Form Y and Form X, respectively. Therefore, a key assumption—though tautological, but unavoidable in practice—is that the conditional distribution of x on v (or y on v) are the same across the groups.

The underlying assumption for the FEEE method is that the conditional distribution of the test score given the anchor test score is similar in the two test taker groups doing the test. The probability of x given v in Population P is equal to probability of x given v in Population Q, for all v. Conversely, the probability of y given v in Population Q is equal to probability of y given v in Population P, for all v regardless of internal or external anchor. This assumption can be expressed as

$$f_P(x|v) = f_Q(x|v), \text{ for all v and } g_P(y|v) = g_Q(y|v), \text{ for all v.} \qquad \text{(Eq. 2.24)}$$

Synthetic population distributions are used to put X on the scale of Y whenever FE is conducted under equipercentile equating. Thus,

$$f_S(x) = w_P f_P(x) + (1-w_Q) f_Q(x) \quad \text{(Eq. 2.25)}$$

$$g_S(y) = w_P g_P(y) + (1-w_Q) f_Q(y), \quad \text{(Eq. 2.26)}$$

where s stands for synthetic population, $f_P(x)$ and $f_Q(x)$ represent distributions for Form X in Population P and Q respectively while $g_P(y)$ and $f_Q(y)$ denote distribution for Form Y in Population P and Q; but, $f_Q(x)$ and $g_P(y)$ are unobservable in Populations Q and P, respectively.

The equipercentile function for the synthetic population (subscript, s) is

$$eys(x) = G\text{-}1s[Fs(x)] \quad \text{(Eq. 2.27)}$$

2.5.2.5 Chained Equating (CE) Linear Method. The chained equating linear method (Angoff, 1971; Holland & Dorans, 2006) involves a scaling of the total-to-anchor scores in the base form and the alternate form and then chaining these scores together. The method assumes that the anchor-to-total test correlation is perfectly. When this assumption is violated—for example, in testing situations where the anchor test score is weakly correlated to the total test score—then chained equating leads to a less accurate equating results. According to Kolen and Brennan (2014), chained equating method involves three underlying procedures. These key techniques are: first, transform X to the scale of V to create ly(x); second, transform V to the scale of Y to create ly(v); and third obtain Y-equivalents such that

$$ly(x) = ly[lv(x)] \quad \text{(Eq. 2.28)}$$

2.5.2.6 Chained Equipercentile (CE) Equating Method. In chain equipercentile equating (Angoff, 1971; Doran, 1990; Livingston, Dorans, & Wright, 1990; Marco, Petersen, & Stewart, 1983), Form X test scores are converted to test scores on anchor test using examinees from Population P. Then test scores on the anchor test are converted to Form Y test scores using examinees from Population Q. This process of chain produces a conversion of Form X test scores to Form Y test scores (Kolen & Brennan, 2014). Therefore, the Form Y equipercentile equivalent of Form X test scores is a function of:

$$e_{y(chain)} = e_{y2}\,[e_{v1}\,(x)], \quad \text{(Eq. 2.29)}$$

where, $e_{v1}\,(x)$ is the equipercentile transformation for converting test scores on X to the scale of V in Population P while $e_{y2}\,(v)$ (not directly visible in the chain) is the equipercentile transformation for converting test scores on V to the scale of Y in Population Q. In addition, the CE equating method assumes that the equipercentile functions equating the test score to the anchor test score are similar in the two test taker groups doing the test.

Equating literature (for example Harris & Kolen, 1990; Livingston et al., 1990; Marco et al., 1983; Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008) indicated that CE methods have a propensity to produce less equating bias than that of PSE methods when groups ability substantially differ. Although Harris and Kolen (1990) proposed use of PSE methods because they have a better theoretical appeal vis-à-vis CE methods, Marco et al. (1983) and Livingston et al. (1990) advocated the application of

CE methods in testing situations where a large ability difference existed in the groups that took both test forms. When groups differ in ability and the correlation between the total test scores and anchor test scores is moderate, the PSE method adjusts form difficulty so that the two groups are more similar than they should be; therefore, leading to a biased equating (Livingston, 2004). But the CE method uses a symmetric scaling approach that is not much affected by the size of the correlation between the anchor test scores and the total test scores. For this reason, the CE method tends to produce less biased results particularly when the groups differ in ability.

2.5.2.7 Kernel Equating (KE) Method. Kernel method of test score equating (KE) can be conceptualized as a modified classical equipercentile observed-score equating that uses a normal or Gaussian kernel—rather than using linear interpolation as is the case in the traditional equipercentile equating method—for continuization of the discrete observed score distributions (Holland & Thayer, 2000; von Davier et al., 2004; von Davier, 2011a). It is a unified observed-score equating framework to test score equating based on a flexible group of equipercentile equating functions that considers the linear equating function as a special case (von Davier et al., 2004). Thus, the KE test score equating is governed by the following equation.

$$eY(x) = G^{-1}[F(x)], \qquad \text{(Eq. 2.30)}$$

where $e_y(x)$ is the equating function for equating test form X to Y—which means the test score on test form Y that corresponds to the test score value x of test form X, while $F(x)$ and $G(y)$ represent the cumulative distribution functions for test forms X and Y

respectively. G^{-1} is the inverse function of G after re-arranging the equation G(y)=F(x)—i.e., after making y the subject of the equation.

As demonstrated in the excellent work of von Davier, Holland, and Thayer (2004) and von Davier (2011b), KE is a sequential standard technique that encompasses five fundamental steps. To summarize, these key procedures are: (i) pre-smoothing the data using log-linear models; (ii) computing the marginal score probabilities for X, Y, and A, in-case of for chained equipercentile; (iii) continuization of the frequency distributions using the Gaussian kernel; (iv) computing the equipercentile equating function using these continuous distribution functions; and (v) computing the accuracy measures—the standard errors of equating (SEE) and the standard errors of equating differences (SEED). The current simulation study did not focus on the fifth step in the framework—a general formula for estimating the accuracy measures (SEE and SEED)—as conceived in the KE equating methodology. Rather after applying step (i) through step (iv), the measures of equating accuracy were calculated based on the assumption that *truth* or *criterion* of equating is known (see Chapter III under sub-section titled: Evaluation of Equating Results and Accuracy). However, the real data study embraced all the procedures in KE framework and the criterion equating was constructed on the same Population T as the equating functions of interest.

Table 2.2 juxtaposes KE and the traditional equating methods by the type of equating function—that is, either linear or curvilinear that are considered in this study under the general framework of NEAT design. Apart from Levine true score equating method, the other traditional equating methods are matched with the KE equating

methods to show their consanguinity. For example, the kernel version of PSE with large bandwidth approximates the Tucker linear method when Tucker assumption about the linearity of the regression holds—i.e., the Tucker method requires that the regression of the test and the anchor is linear. This assumption is not met most of the time. Specifically, in the vertical scaling scenario because the anchor test may be from a different grade; therefore, this regression is probably going to be curvilinear. The violation of linearity assumption would have profound consequences on the equating results and accuracy.

Table 2.2

NEAT Design: KE and Traditional Equating by Linear and Non-linear Equating Procedures

KE Method Type of Equating Function	Traditional Equating Method
Linear Functions	
PSE with large bandwidth	Braun & Holland linear
	Tucker[a] linear
	Levine True Score
CE with large bandwidth	Chained linear
Non-linear Functions	
PSE with optimal bandwidth (curvilinear)	Frequency estimation (FEEE)
CE with optimal bandwidth (curvilinear)	Chained Equipercentile

Note. [a]The kernel version of PSE with large bandwidth approximates the Tucker linear method if Tucker assumption about the linearity of the regression holds.

Research studies in KE have shown that there are multiple ways of selecting bandwidth. But before proceeding with bandwidth selection, it is noteworthy to provide

two equations to put the concept of bandwidth across. According to von Davier et al. (2004), when using a Gaussian kernel the continuized cumulative distribution function for a score *x* (this is true for a score value of y in form Y) is given by

$$F_{hx}(x) = \sum_{j=1}^{nx} r_j \, \phi \left(\frac{x - a_x x_j - (1 - a_x)\mu_x}{a_x h_x} \right), \qquad \text{(Eq. 2.31)}$$

where *nx* is the number of items on the test plus one, r_j is the probability of obtaining the score x_j, $\phi(\cdot)$ represents the standard normal cumulative distribution function, μ_x is the mean test score, σ_x is the standard deviation of the test scores (or σ^2_x is the variance of the test scores), and h_x is the bandwidth such that a_x –a scaling factor to ensure the variance of the original distributions is the same even after continuization of discrete distribution (this is also the case for form Y where the subscript x will be replaced by y)—is defined by

$$a_x = \sqrt{\frac{\sigma_x^2}{\sigma_x^2 + h_x^2}} \qquad \text{(Eq. 2.32)}$$

Some of the approaches for selection of bandwidth are (1) minimizing penalty functions; (2) plug-in methods; (3) Silverman`s rule of thumb; (4) cross-validation; (5) adaptive kernels (6) to achieve a particular goal—for example, linearity or not (equipercentile). In this dissertation the first technique to bandwidth selection—i.e., minimizing penalty function— was considered in order to obtain both linear and equipercentile functions.

2.5.2.7.1 Kernel NEAT post-stratification equating using linear method (KeNEATPSE_L). von Davier et al. (2004) have demonstrated that the selection of

bandwidth (h_x or h_y) somewhat determines the equating method under KE framework. The kernel NEATPSE linear is achieved by selecting large bandwidths. When this is done the kernel NEATPSE linear with bandwidths approximates the Braun and Holland (1982) linear method of score equating. Further, the kernel NEATPSE linear method of score equating approaches a linear method of score equating when using large bandwidth values that are larger than 10 times the standard deviation of the continuized distribution. Similarly, the larger the bandwidth parameter is the more likely the density at each discrete score point spreads out.

2.5.2.7.2 Kernel NEAT post-stratification equating using equipercentile method (KeNEATPSE_E). The procedure to achieve kernel NEAT poststratification equating with optimal bandwidths (or keNEATPSE equipercentile method) has also been outlined by von Davier et al. (2004). Research has demonstrated that the kernel NEAT post-stratification equating equipercentile method is equivalent to the frequency estimation equipercentile score equating method. In this case, the keNEATPSE optimal (equipercentile) equating method selects optimal values for h_x (or h_y) are automated by reducing the difference between the probability distributions of X (or Y) before and after continuization (and by using some additional penalty functions—for more details, see von Davier et al., 2004).

2.5.2.7.3 Kernel NEAT chained equating using linear method (KeNEATCE_L). Chained equating methods are described by Angoff (1984), Livingston (2004), and Kolen and Brennan (2004). The kernel version of chained equating approximates the chained linear method when large bandwidths are used (von Davier et al., 2004). The chained

equating represents a chain of linking from test form X to anchor test form A and then from anchor test form A to test form Y. In other words, chained linear equating assumes that the linking relationship between X and A would be the same if it were observed on population Q. Likewise, it assumes that the linking relationship between Y and A would be similar if it were observed on population P. In general, if each of the two links is linear, then the final equating is also linear (see Eq. 2.28).

2.5.2.7.4 Kernel NEAT chained equating using equipercentile method (KeNEATCE_E). The kernel version of chained equating will approximate the chained equipercentile method when the optimal bandwidths are used. It represents a chain of linking from test form X to the anchor test form V and from the anchor test form V to test form Y such that if each of the two links is equipercentile function, then the final equating is equipercentile too. The equating function with a nonlinear equipercentile equating function is derived using the same poststratification equating (PSE) assumptions stated previously and then applied to the KE NEAT framework (von Davier et al., 2004). To equate test form X to test form Y, it is presumed that the equipercentile equating relationship between test form X and the anchor test form V (or between test form Y and the anchor test form V) would be similar if it were observed on population Q (or on population P). Then the method converts test form X to the anchor test form V and then equates the resulting score for anchor test form V to the test form Y using equation 2.29 (Kolen & Brennan, 2004).

2.6 General Observation on Equating Methods under NEAT Design

Equating methods used with NEAT design can be categorized into two main types depending on the way they use the information from the anchor (Holland et al., 2007) and the missing data in the design. First, poststratification equating (PSE) or frequency equating is a type of missing data assumption. The PSE types of assumption is that the conditional distribution of X given anchor (or Y given anchor) is the same for any S, T = wP + (1-w)Q. According to PSE type of equating, it is assumed that the relationship that generalizes from each equating sample to the target population is in fact a conditional relationship. This means that conditioned on the anchor test score, A, the distribution of X in Q, where it is missing and unobserved, is similar to P, where it is not missing, but it is realized. Second, the chain equating (CE) assumption all have the form that a linking function from X to anchor (or from Y to anchor) is the same for any S, T = wP + (1-w)Q. In CE approach, the test scores on the new form are equated to test scores on the old form through a chain created by these two linear equating links/functions—$Lin_{xv;p}(x)$ and $Lin_{vy;Q}(v)$. The CE linear function is given by:

$$CE_{XY}(x) = Lin_{VY;Q}(Lin_{XV;P}(x)) \qquad \text{(Eq. 2.33)}$$

In sum, PSE and CE approaches hypothesize that an important distributional property that connects scores on X or Y to scores on the anchor test is invariant for any S, T = wP + (1-w)Q—i.e., is population invariance (Holland et al., 2007). von Davier et al. (2004) have shown that when P and Q are substantially different, PSE and CE assumptions can result in equating functions that are different.

In practice, the common items are assumed to be a representative of the whole form in both content and statistical characteristics. Section 2.2 provides a thorough albeit inexhaustive treatment of anchor studies in the context of NEAT equating design. The forms are administered to different groups of examinees which may have a considerable difference in their knowledge, skills and abilities. This design is most appropriate in vertical scaling because the different test forms are constructed that include common items sampled from either one of the adjacent grade levels or both grade levels (Tong & Kolen, 2007). In vertical scaling literature, it is assumed that in theory student progression (or growth and development) across grades "underlies a collection of test items that have been written for the purpose of creating a vertical scale" (Briggs & Domingue, 2013, p. 553). Figure 2.1 demonstrates a conceptual framework or a hypothetical scenario of the distribution of ability across the three grades with overlapping portions in a proficiency scale; grade 5 is designated as a base grade scale and adjacent grades 4 and 6 are linked to this base scale. The sections marked common items indicate the area assumed for sampling anchor test items—that is, common items can be selected from the test for the grade below or for the grade above, or from both combinations. The unique test items are sampled from the area where the graphs do not intersect.

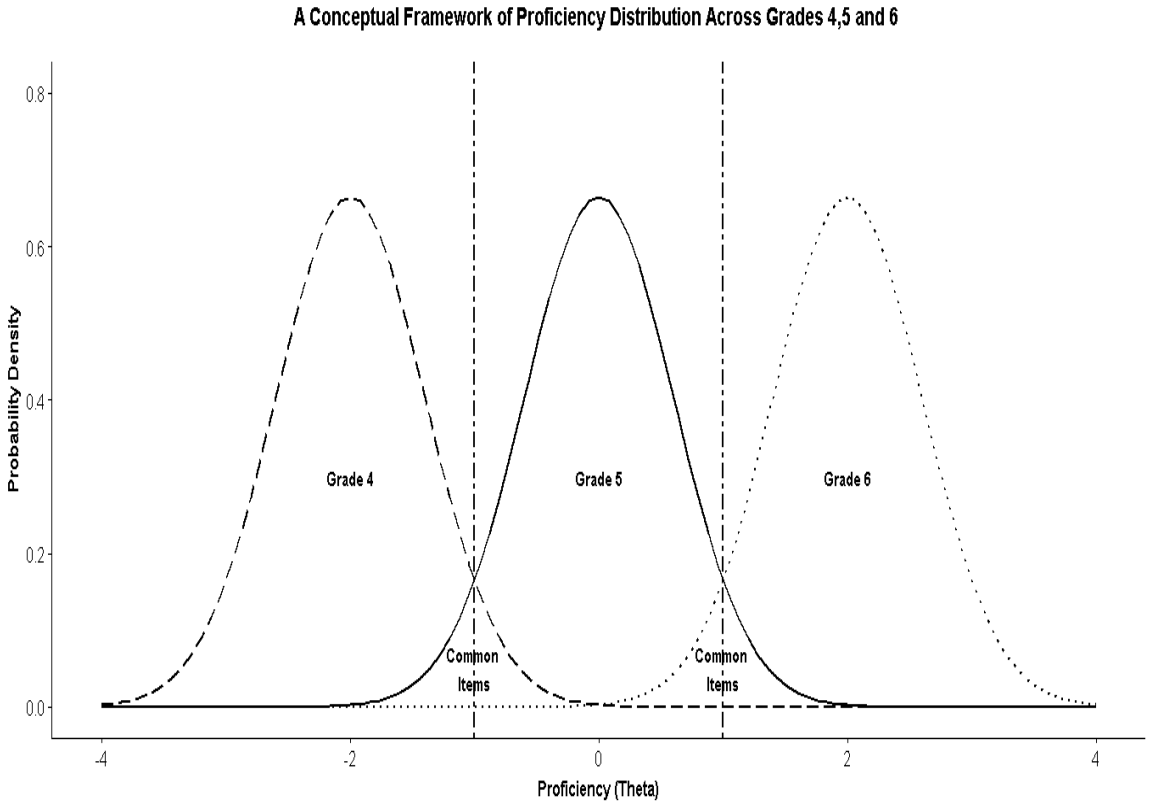

Figure 2.1. Demonstration of a Hypothetical Scenario of the Distribution of Ability across the Three Grades with Overlapping Portions in a Proficiency Scale.

Test takers performance in these anchor test items is crucial because they are used to statistically adjust for any differences in ability between nonequivalent groups taking the two forms; therefore, after a successful scaling or linking a common metric is constructed that spans across grades. While there is a general consensus and assumption that examinees in higher grade levels will outperform examinees in lower grade levels on the anchor test items, there is an exception to this belief particularly in a scenario where there is comparatively little or no curriculum overlap from grade to grade; this means lower-grade students may perform better than higher-grade students on lower-grade items probably due to the fact that they have been taught the curriculum more lately. In other

words, when the content area that is tested is strongly curriculum-dependent, the choice of anchor test items and students` performance on those items can have far-reaching consequences to the measurement, meaning, and interpretation of constructed vertical scale. Next subsection delves into different viewpoints on scaling and linking.

2.7 Perspectives on Scaling

Equating literature and scaling theory over the decades seem sharply divided on the meaning of a scale and its properties. This has created multiple perspectives on scaling (Kolen, 2011; Kolen & Brennan, 2014). Table 2.3 summarizes some of the predominant viewpoints on scaling theory and practice.

Table 2.3

Divergent Viewpoints on Scaling

Proponent	Perspective
(Angoff, 1971; Lord, 1975, 1980)	Proposes equal interval property of a scale.
(Coombs, Dawes, & Tversky, 1970; Stevens, 1946; Suppes & Zinnes, 1963)	Scale Classification: Nominal, ordinal, interval & ratio. Scale attributes should be clearly defined
(Guttman, 1944; Thurstone, 1925; Wright, 1977)	Scaling should be based on psychometric models
(Lindquist, 1953)	The scaling method should not influence the content of the test or change the meaning of objectives in a test.
(Petersen et al., 1989)	The main goal of scaling is to facilitate interpretation of a test score
(Yen, 1986)	Choice of a scale should be driven by a specific application. Choosing a scale and using it is a must.

2.8 Current Research on Vertical Scaling

Studies in vertical scaling can be classified into two main groups. The first group deals with examination of the results from vertical scaling methods and designs to compare and contrast the results. Research in this direction investigates whether general differences in the scaling results exist or not and has produced different results and conclusions. Vertical scaling literature—from the first group—suggests that vertical scaling results: (1) depends on examinee groups; (2) are sensitive to linking design; and (3) differ considerably depending on different statistical methods employed to construct the scale. The second aspect is more specific because it delves into comparison of methods and designs with emerging issues and themes like the pattern and meaning of grade-to-grade growth, grade-to-grade variability, separation of grade distributions, sensitivity of results to scale transformation, multidimensionality and IRT scaling methods and factors that influence vertical scaling results (Kolen & Brennan, 2004, 2014; Skaggs & Lissitz, 1986c). This dissertation is a merger of some of these thoughts and ideas in vertical scaling. Table 2.4 summarizes the research on vertical scaling. The table is divided into three columns: (1) researchers and related areas of vertical scaling studied, (2) aspect(s)/method(s) of vertical scaling investigated, and (3) results/ conclusions reached under each category of researchers in vertical scaling. For example, under researchers there are six areas of vertical scaling commonly examine—i.e., (i) general differences in scaling (ii) grade-to-grade growth, (ii) grade-to-grade variability, (iv) separation of grade distributions, (v) sensitivity of results to scale transformation, and (iv) multidimensionality and IRT vertical scaling methods.

Table 2.4

Summary of Contemporary Research on Vertical Scaling

Researchers	Aspect(s)/Method(s) of vertical scaling studied	Results/Conclusions
(1) General differences in scaling results related group of scholars:	(1) Comparison of vertical scaling results on methods and designs	Generally, results are:
(Forsyth, Saisangjan, & Gilmer, 1981; Gustafsson, 1979; Harris & Hoover, 1987; Holmes, 1982; Loyd & Hoover, 1980; Skaggs & Lissitz, 1988; Slinde & Linn, 1977, 1978, 1979a, 1979b; Tong & Kolen, 2007)	(a) different examinee groups	(i) examinee groups dependent
(Briggs & Weeks, 2009a, 2009b; Custer, Omar, & Pomplun, 2006; Guskey, 1981; Harris, 1991; Hendrickson, Kolen, & Tong, 2004; Hendrickson, Wei, & Kolen, 2005; Ito et al., 2008; Jodoin, Keller, & Swaminathan, 2003; Kolen, 1981; Lei & Zhao, 2012; Li & Lissitz, 2012; Paek & Young, 2005; Phillips, 1983, 1986; Pomplun, Omar, & Custer, 2004; Skaggs & Lissitz, 1986a)	(b) different statistical methods	(ii) found to differ depending on statistical methods used
(Harris, 1991; Hendrickson et al., 2004, 2005; Tong & Kolen, 2007)	(c) different linking designs	(ii) found to be sensitive to linking design
(2) Grade-to-Grade Growth Related Group of Scholars	(2) Grade-to-grade growth: Hieronymus, Thurstone & IRT scaling	
(Andrews, 1995; Bock, 1983; Briggs & Weeks, 2009a, 2009b; Hendrickson et al., 2004, 2005; Seltzer, Frank, & Bryk, 1994; Tong & Kolen, 2007; Williams, Pommerich, & Thissen, 1998; Yen, 1985, 1986)	_____	(i) there is decelerating growth from grade to grade—i.e., grade-to-grade differences in averages decreases as grade increases
(Hoover, 1984a)	_____	(ii) that anomalies exist—i.e., grade-to-grade growth scaling produced irregularities
(Becker & Forsyth, 1992)	_____	(iii) no evidence of decelerating growth
(3) Grade-to-Grade Variability Related Group of Scholars	(3) Grade-to-grade variability: increasing versus decreasing	
(Andrews, 1995; Thurstone, 1925, 1927, 1928; Thurstone & Ackerman, 1929; Tong & Kolen, 2007; Yen, 1986)	(a) Thurstone scaling	(i) that score variability increases with age;

Table 2.4

Cont.

Researchers	Aspect(s)/Method(s) of vertical scaling studied	Results/Conclusions
(Williams et al., 1998)	_____	(ii) that the extent of increase is affected by how scaling method was implemented (iii) there`s evidence of decreasing grade-to-grade variability
(Andrews, 1995)	(b) Hieronymus scaling	(iv) that there is increasing grade-to-grade variability
(Andrews, 1995; Hoover, 1984a; Omar, 1996, 1997, 1998; Yen, 1986) (Yen, 1985) (Camilli, 2005) (Camilli, 2005) (Camilli, 2005) (Williams et al., 1998)	(c) IRT scaling	(v) that score variability decreases over grades. Justification for the decrease: (a) multidimensionality (b) measurement error differences at different grade level (c) due to estimation of IRT proficiency for extremely (very high and very low) scoring individuals (d) old procedures for IRT parameter estimation (e) use of old version of LOGIST for joint maximum likelihood (JML) method
(Becker & Forsyth, 1992)	_____	(vi) that there`s increase in grade-to-grade variability (no linking was involved; the same test was administered to each grade)
(Seltzer et al., 1994)	_____	(vii) that no evidence of decrease in grade-to-grade variability (used Rasch scaling)
(Bock, 1983)	_____	(viii) there was a homogenous variance across age
(Camilli, Yamamoto, & Wang, 1993)	_____	(ix) that there is little or no evidence of decrease in grade-to-grade variability

Table 2.4

Cont.

Researchers	Aspect(s)/Method(s) of vertical scaling studied	Results/Conclusions
(Hendrickson et al., 2004, 2005; Tong & Kolen, 2007; Williams et al., 1998; Yen & Burket, 1997)	_____	(x) that there is evidence of scale shrinkage—they combined test & statistical procedures
(Hoover, 1984a)	_____	-argued that the grade-to-grade differences in score variability should increase over grades instead of decreasing
(Phillips & Clarizio, 1988a)	_____	-demonstrated implications of vertical scaling for placement of children with special needs in education
(Burket, 1984; Clemans, 1993, 1996; Hoover, 1984b, 1988; Phillips & Clarizio, 1988b; Yen, 1988; Yen, Burket, & Fitzpatrick, 1996)		-debate on the plausibility and practicality of vertical scaling results in educational and psychological testing
(4) Separation of grade distributions related group of scholars	(4) Separation of grade distributions	
(Andrews, 1995)		(i) there is less separation—more grade-to-grade overlap— between distributions for tests using the scaling test design than for tests using IRT NEAT design, Thurstone or Hieronymus scaling methods
(Mittman, 1958)	Hieronymus Scaling	(ii) the results are opposite the findings by Andrews (1995)
(Yen, 1986)		(iii) that the IRT and Thurstone scaling methods performed similarly for the separation of grades distributions when the differences are put in a z-score scale.
(5) Sensitivity of results to scale transformation group of scholars	(5) Sensitivity of results to scale transformation	

Table 2.4

Cont.

Researchers	Aspect(s)/Method(s) of vertical scaling studied	Results/Conclusions
(Schulz & Nicewander, 1997; Zwick, 1992)		(i) that nonlinear monotonic transformations of the score scale can alter the pattern of grade-to-grade growth or grade-to grade variability from decreasing to increasing and vice versa
(Braun, 1988)	_____	(ii) percentile ranks comparing two distributions are not affected by nonlinear monotonic transformations of scale; effect size is affected by nonlinear scale transformation
(6) Multidimensionality and IRT vertical scaling methods group of scholars	Multidimensionality and IRT vertical scaling methods	

2.9 Summary

Previous research has established that kernel equating is a sound and stable test score equating method, which leads to an improvement of the results of traditional test score equating methods; however, no simulation studies have been published—by the time this simulation study was conducted—to compare kernel test score equating to its traditional analogs particularly in the context of vertical scaling. The benefits of a simulation study are great: the researcher is allowed full control over the difficulties of the test forms, the ability levels of the examinees, the reliability, length, and difficulty of the anchor test, the relationship of those test forms, and ultimately investigate where equating works or fails. This dissertation attempts to remedy this lack of information (or

existing gaps in current literature) by creating situations in which truth is known and several test score equating methods under NEAT design, including kernel equating, are compared and investigated for accuracy and applicability to real-life testing situations. Also, this kind of simulation study has not been applied to operational vertical scaling in large-scale testing programs.

CHAPTER III

DATA AND METHODOLOGY

The main purpose of this study is to explore some of the empirical issues and complications associated with vertical scaling methods under NEAT design and to a less extent RG design. This in turn will give us new insights into how multiple test designs and different sampling factors affect the accuracy of vertical scaling for different AT conditions. This chapter outlines the design of a large-scale simulation study to examine the impact of total test length, discrimination and difficulty item parameters, between-grade differences, between-grade ability differences, distribution of ability differences, anchor test difficulty differences, and equating methods on equating error. Succinctly stated, this study subtly investigates the extent the accuracy of different equation methods under NEAT design under various study conditions can be tenable when constructing a vertical scale. An equally important portion of this chapter is the real-data analysis, which constitutes the second component of the study. Next is a description of simulated data, item generation and calibration, simulation conditions, 3-parameter logistic (3PL) model, vertical scale construction, real data and their analysis procedures. Evaluation of equating accuracy is also provided; in addition, analysis methods employed in the study are discussed. The results of the analysis are presented in the proceeding chapter—i.e., Chapter IV.

3.1 Sources of Data

The main sources of data for analysis were twofold in this dissertation. These were (1) simulated data and (2) real data. The real data were from a large-scale assessment involving common items that were used as a link between the two test forms that have been constructed to the same content specification and psychometric properties. The generated data sets were created from GENEQUATE software (Luecht, v45 2014), which assumed random sampling of test takers performance (or proficiency scores) from a normal underlying ability distribution, $\theta \sim (\mu, \delta^2)$—i.e. mean ability of 0 and 1 standard deviation—with $N = 3000$ for every test form across grades 4, 5 and 6. Item response theory (IRT) was used to generate item parameters and theta (or proficiency) parameters. Item response theory (IRT) is a probabilistic model which makes predictions about probability that examinees at different scale (trait, ability or proficiency) levels will correctly answer each item. An example of IRT model is a 3PL model (Birnbaum, 1968). It assumes that the probability that an examinee with proficiency value, θ_i, equal to the ability of person j will get an item *i* correct. This probabilistic relationship is governed by the equation below:

$$P_{ij}(\theta) = p_{ij}(\theta_i; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp[1.7a_j(\theta_i - b_j)]}{[1+\exp(1.7a_j(\theta_i - b_j))]}, \qquad \text{(Eq. 3.1)}$$

where θ_i is the underlying ability parameter for examinee *i* ranging between $-\infty < \theta < +\infty$, a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the item lower asymptote (guessing) parameter of item *I*, -1.7 is a scaling factor, and *exp* is equal to 2.71828…, which is the exponent value of *e*. In using 3PL model, item parameters and

trait values are estimated from examinee`s response pattern (correct or incorrect) to a set of dichotomously score items.

3.2 Importance of Stimulation Studies

There are many benefits accruing from simulation studies that cannot be overemphasized. For example, the researcher has considerable latitude to control over the difficulties of the test forms, the ability levels of the examinees, the reliability and test length, difficulty of the anchor test, the relationship of those test forms or create any other study conditions that not only mimic real testing situations but also extreme situations which might look unrealistic to testing practitioner or policy maker. This dissertation attempts to create diverse testing situations in which truth is somehow known and different equating methods in the context of NEAT design are compared and investigated for measurement accuracy and their application to real-life testing realities. This aspect is extremely important in contemporary educational measurement, theory, and practice because both extreme (unrealistic) and realistic testing circumstances are factored in and taken care of when designing the simulation research study.

The data for this dissertation are generated from an IRT model—3PL model—as alluded to previously (see Eq. 3.1 above); however, classical test score equating (or observed-score equating under classical test theory) and vertical scaling were chosen for this dissertation rather than IRT test score equating methods—i.e., IRT observed score equating and IRT true score equating methods. The fact of the matter is that IRT equating was not considered as being of any interest in this research study. Even though this is the case, there is a possibility of future research in this area. Importantly, the usefulness of

IRT equating framework cannot be gainsaid. A case in point where IRT equating is not only appropriate but also beneficial is in computerized adaptive testing (CAT) or testing programs that employ multiple test forms within a specific time frame; and when the main purpose is to calibrate item banks instead of form-to-form equating—that is, less than three test forms does not warrant IRT test score equating (personal conversation with Dr. Luecht). Furthermore, it can be argued that some difficulties and perhaps extra financial expenses could be incurred to develop a stable, IRT-calibrated item banks for a testing program that has got at least two or three test forms. It is a common practice among testing and equating practitioners to use form-to-form test score equating whenever they have two or three test forms.

Remarkably, given practical considerations for form-to-form equating, there is no general consensus in the equating literature that IRT test score equating methods are superior to classical test score equating methods specifically when from-to-form test score equating is used under a non-equivalent groups with anchor test (NEAT) design. Although the two equating frameworks—IRT and classical equating methods—offer no practical advantages over each other, the latter is generally considered least complicated vis-à-vis the former; therefore, it makes few underlying assumptions (Petersen, 2007). Germane to this discussion is the understanding that IRT test score equating puts stringent conditions that all items on the new form, the old form and the anchor test must measure exactly the same underlying hypothetical construct. This underlying assumption under IRT test score equating is, however, considerably relaxed under classical test score equating when the concept of classical congeneric is introduced in the equating

enterprise. In sum, it is important to clarify that although IRT generation modus operandi was employed to simulate the dataset for this dissertation, IRT test score equating was not of primary concern in this study.

3.3 Design of Vertical Scale Panels

In this study missing-by-design configuration in a vertical scaling context is designated as "VS panels." This creates a loose tie to ca-MST, but also somewhat ties the design to the notion of cross sectional "panel data" as used in statistical and experimental design studies. Instead of building a vertical scale to represent learning progressions across grades 3 through 8, for example, this study is designed to create a vertical scale that spans only across grades 4 through 6. A description of how this panel data is used in this study is shown by Figure 3.1—the eight test forms (spanning from grade 4 through grade 6) per panel are constructed as below:

Form #1: RT(4.1) + AT(4.1)
Form #2: RT(4.2) + AT(5.1)
Form #3: RT(5.1) + AT(4.1))←BASE FORM (if external anchors)
Form #4: RT(5.2) + AT(5.1)
Form #5: RT(5.1)* + AT(5.2)←BASE FORM (same base RT form as Form #3)
Form #6: RT(5.2)** + AT(6.1)
Form #7: RT(6.1) + AT(5.2)
Form #8: RT(6.2) + AT(6.1)

* Same RT as Form #3.
** Same RT as Form #4.

Figure 3.1. Construction of a Vertical Scaling Panel.

For convenience, it is assumed that 5th grade is the "base" grade for the vertical scale in this study. A panel is comprised of exactly 8 test forms (RT=regular test,

AT=anchor test). Table 3.1 displays *figure 3.1* in a NEAT design and randomly equivalent group context. VS panels (vertical scaling panels) are defined as multi-grade test form configurations. Each panel is comprised of multiple test forms representing unique combinations of operational or regular test (RT) forms and common item anchor tests (AT). The AT items are treated as external anchors since no off-grade items would normally count in student scores—in reality as a matter of assessment policy. Therefore, RT(5.1) contains the same items for Forms #3 and #5. Similarly, RT(5.2) has exactly the same items for Forms #4 and #6. This allows the score data for RT(5.1) and RT(5.2) to be combined into two larger data sets: RT(5.1) combines the scored data for Forms #3 and #5; RT(5.2) combines the scored data for Forms #4 and #6. RT(5.2) can then be equated to RT(5.1) using a randomly equivalents groups strategy.

Also note there are four sets of AT [AT(4.1), AT(5.1), AT(5.2), and AT(6.1)] across the three grades (i.e., grade 4-6). Grade 5, the base form, has got all AT. Although Form #3 RT(5.1) and Form #5 RT(5.1) have the same items under regular test, they have different AT items—AT(4.1) and AT(5.2), respectively. The same observation can made for Form #4 RT(5.2) and Form #6 RT(5.2) with their respective AT items: AT(5.1) and AT(6.1).

Table 3.1

A NEAT Design with On-Grade, Off-Grade and Anchor Items Blocks

Grade	Form #	Regular (RT)	Anchor (AT)	Regular (RT)	Anchor (AT)	Regular (RT)
4	1	RT4.1	AT4.1			
	2	RT4.2	AT5.1			
5	3		AT4.1	RT5.1		
	4		AT5.1	RT5.2		
	5			RT5.1	AT5.2	
	6			RT5.2	AT6.1	
6	7				AT5.2	RT6.1
	8				AT6.1	RT6.2

Legend	Anchor Test	Regular Test Form
AT=Anchor Test	AT4.1=1st AT only grade 4 items	RT4.1=1st of one grade 4 forms
RT=Regular Test	AT5.1=1st AT only grade 5 items	RT4.2=2nd of two grade 4 forms
	AT5.2=2nd AT only grade 5 items	RT5.1=1st of one grade 5 forms
	AT6.1=1st AT only grade 6 items	RT5.2=2nd of two grade 5 forms
		RT5.2=6th grade items
		RT6.1=6th grade items

3.4 Vertical Equating Design and Description of Study Conditions

This subsection provides a summary of simulation conditions, a special NEAT design, and description of simulation conditions. Figure 3.2 illustrates the linkages and equating for each VS panel design used in this study—random group equating and NEAT design. The linkages within a panel occur via either random assignment of forms to students, within grades—resulting in randomly equivalent groups—or by having the shared common items across grades (e.g., items shared by grades 4 and 5 test forms).

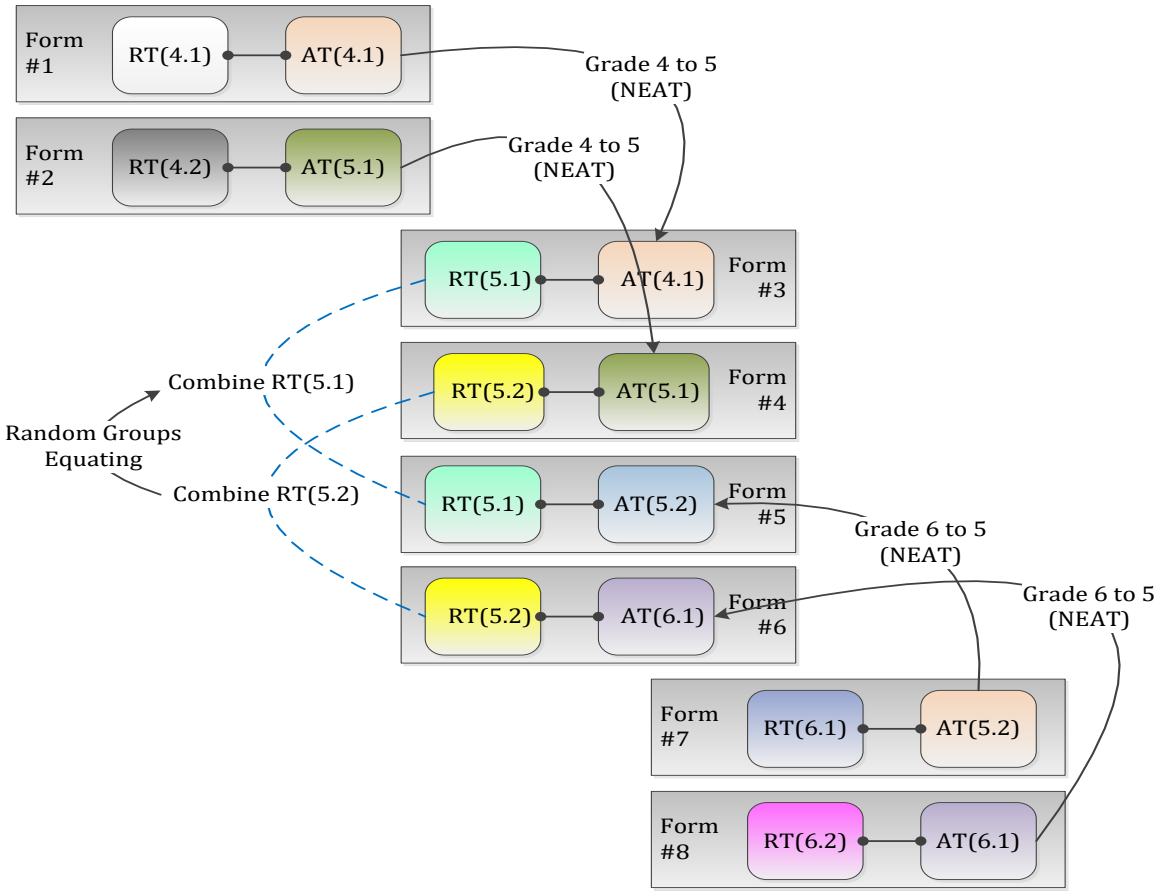

Figure 3.2. An Illustrative Diagram Depicting Vertical Scale Panel with Multiple Linkages and Equating Designs across Grades and Forms with Grade 5 (Form # 3) as a Base Form.

Random groups links are for forms # (1 & 2), (3, 4, 5 & 6) and (7 & 8). Nonequivalent AT links are for forms # (1 & 3), (2 & 4), (5 & 7), and (6 & 8). Because all factors examined were completely crossed with each other, a total of 162 conditions—that is, 1 sample size × 3 total test length × 2 total test mean discrimination × 3 between-grade mean ability differences × 3 pool information/distribution of ability difference × 3 anchor test/mean difficulty differences—were investigated. Moreover, there were 8 test forms in each of the 162 study conditions and 10 replications per study,

which gave an overall of 12,960 data sets for analysis—8 forms × 162 conditions × 10 replications. Table 3.2 is a summary of simulation conditions studied in this dissertation.

Table 3.2

Factors Controlled in the Simulation Study

Name of Factor	Study Condition Levels	Number of Counts
1. Total test length n_{Total}	n = (30, 60, 120)	3
2. Total test mean discrimination	mean(a) or $\mu(a)$ = (.6, 1)	2
3. Between-grade mean ability differences (BGMAD) $\Delta[\mu(\theta_g),\mu(\theta_{g+1})]$	$\Delta[\theta]$ = (.5, 1.0, 1.5)	3
4. Pool Information: Distribution of ability difference (DAD) $\Delta[\mu(\theta_g),\mu(b_{g,RT})]$	Δ[RT.b] = (-1.0, 0.0, 1.0)	3
5. Anchor test: mean difficulty differences (ATDD) $\Delta[\mu(b_{g,RT}),\mu(b_{g,AT})]$	Δ[AT.b] = (-1.0, 0.0, 1.0)	3
Total Conditions 3 x 2 x 3 x 3 x 3	=	162

Note. 8 forms x 162 = 1,296 test forms; 10 replications x 1,296 = 12,960 total data sets
a=discrimination; AT=anchor test; b=difficulty; g=group; RT=Regular test; Δ(delta)=differences/change; θ_g=adjacent lower grade; θ_{g+1}=adjacent upper grade; μ=mean

Broadly speaking, these study conditions can be dichotomized. The first category falls under study conditions related to group characteristics, which include sample size (in this study sample size is treated as a constant, i.e., 3,000 examinees), between-grade mean ability differences, and pool information or distribution of ability difference. The second category encompasses study conditions closely aligned to test measurement information characteristics like test length (in this dissertation proportion of anchor test to total test is considered a constant, 20%), test mean discrimination, and anchor test mean

difficulty differences. Next is an in-depth treatment of each of the factors and their summary.

(a) Test Length (*n*) and Anchor Test (AT/A/V)

Generally speaking, it has been shown that a longer anchor test is considered desirable. Oftentimes, such a test is more reliable and tends to generate fewer random equating errors (Budescu, 1985). In this study, the length of the total test was varied to three sizes. This meant that short, medium, and long tests were operationally defines as consisting 30, 60, and 120 items respectively. The total number of common anchor test items was held constant at 20% of the total test to produce 6, 12, and 24 common anchor items from 30, 60, and 120 total test lengths respectively. In equating literature, anchor test can be either internal or external. While the internal anchor test means that the examinee`s test score on the anchor test counts, the external anchor test score is not counted as part of the score for the examinee. Most equating research under NEAT design utilizes the external anchor test scores for scientific purposes. Unlike the internal common anchor test items, the external common items are never released to test takers or any other stakeholder after the test is done. In equating studies, it has been shown that internal anchors are advantageous over external anchors because the former tends to have high correlations with the total test score, which is attributed to the fact that the internal anchor test scores contribute to the total test score (Dorans, Moses, & Sinharay, 2010), unlike the external anchor test score which yields not high correlation with the total test score due to its exclusion from the computation of the total test score. Although the choice of the internal versus external anchor test is influenced by both federal

requirement and to some extent by the testing program, this dissertation assumed that all common items were external and were placed at the beginning of each test. For instance, for the total test length of 30 items, there were 6 common anchor items forming the first set of questions and the rest 24 items constituted unique or regular test. This is extended to the medium and long total tests as well. The three test forms assumed that the anchor test was external.

(b) Test Mean Discrimination

The relationship between item discrimination and the precision of test scores is well studied and documented aspect in psychometrics studies. For example, smaller measurement error, which means high measurement precision, is closely related with high values of item discrimination; but the converse is also true that larger measurement errors are attributed to lower item discrimination parameters, which results in lower measurement precision. In this study, two characteristics of item discrimination parameters were examined. These are: (i) a = .6 and (ii) a = 1.0. For practical considerations, *a*-item discrimination parameter of value .6 is presumed to be moderate while its counterpart—*a*-item discrimination parameter of value 1.0—represents a high discrimination. Although measurement precision is predominately affected by item discrimination parameters, to some extent it is also affected by *b*-item difficulty and pseudo-guessing parameter (or *c*-parameter). For instance, the location of the measurement precision is highest close to the mean of item difficulty distribution and the size of the standard deviation determines the extend of the spread (or the variability) of the measurement precision across the underlying proficiency scale. The last two study

conditions—i.e., distribution of ability difference (DAD) and anchor test with mean difficulty differences (ATDD)—blend the concept of *b*-item difficulty parameter when manipulating these variables (DAD and ATDD).

(c) Between-grade Mean Ability Differences (BGMAD)

In the context of vertical scaling, this is the magnitude of group separation or group effect which can be understood as a mean ability differences between adjacent grades that took the alternate test form in comparison with the base test form. The values of between-grade ability difference—denoted as delta theta or $\Delta[\theta]$, where θ is a variable to represent the hypothetical underlying proficiency of examinees from two IRT θ distribution—were offsets from a starting point relative to grade 5; hence, impacting between-grade differences. Between-grade ability differences (or $\Delta[\theta]$) were studied under three distinct levels: .5, 1, and 1.5. These values were calculated relative to mean (theta.5) = 0.0. For instance, when $\Delta[\theta]$ =.5 it meant means (theta.4) = -.5 and mean (theta.6) = +.5. Also, when $\Delta[\theta]$ =1.0, mean (theta.5) was still 0.0, but mean (theta.4)= -1.0 and mean (theta.6)=+1.0. The same concept applied when $\Delta[\theta]$ =1.5, mean (theta.5) was still 0.0, but mean (theta.4)=-1.5 and mean (theta.6)=+1.5—i.e., these three different degrees of between-grade ability differences were computed using $\Delta[\theta] = \mu(\theta_{g+1})—\mu(\theta_g)$ formula.

Mean for between-grade ability differences of .5 was considered small while the value of 1 was medium and large when the value was 1.5. Correlations can be computed between total test score and anchor test scores within each group. If those correlations are strong enough, then the anchor test is considered a good indicator of the within-group

difference between individual test takers in the knowledge, skills and abilities that the test purports to measure. Equating literature corroborates the fact that population differences in ability may explain the issue of a large amount of residual variance when dealing with nonequivalent groups—this situation is expected in vertical scaling where there is remarkable group mean ability differences between adjacent grades. In this simulation study, between-grade mean ability differences were manipulated relative to a starting point in grade 5 as previously stated; therefore, the effects of examinee between-grade mean ability differences were reflected in the equating results.

(d) Pool Information: Distribution of Ability Differences (DAD)

In the current study, distribution of ability differences (or pool information) was represented by Δ[RT.*b*]. This implied that the Δ[RT.*b*] or the delta RT.*b*mean was set relative to the mean(theta.grade), with values of -1.0, 0.0, and 1.0—i.e., this indirectly impacts reliability. These values have an operational meaning—that is, -1 means below average (or -1 unit below the mean [theta.grade]), 0 means no difference between the two means (or mean of the *b*-item difficulty parameter for regular test in a specific grade is the same as the mean of underlying ability for that particular grade) and 1 stands for above average (or 1 unit above the mean [theta.grade]). When Δ[RT.*b*] = 0, the implication was mean(*b*.RT.grade)=mean(theta.grade). But, when Δ[RT.*b*] = -1.0, it implied that mean(theta.4)=-1.5, and mean(*b*.RT.4)=-2.5 (that is, -1 unit below the mean grade 4 theta). Similarly, when Δ[RT.*b*] = 1.0, it implied that mean(theta.4)=1.5, and mean(*b*.RT.4)=2.5 (that is, 1 unit above the mean grade 4 theta)—i.e., Δ[RT.*b*] = $\mu(b_{g,RT})$—$\mu(\theta_g)$ equation was applied to obtain the three levels of this condition.

(e) Anchor Test: Mean Difficulty Differences (ATMDD)

Anchor test difficulty differences can be defined as anchor test difficulty variability. This variability—denoted as delta.*b* anchor or Δ[AT.*b*], where item difficulty was measured by IRT *b* parameter—values were calculated relative to the means of the within-grade *b*-parameters for both the anchor test and regular test—hence it impacts AT characteristics. The three values (or levels) manipulated for the anchor test mean difficulty differences condition was: -1.0, 0.0, and 1.0. These three levels were operationally defined as first, below average for -1.0, which meant that average *b*-difficulty parameter for the regular or unique test was greater than the average *b*-difficulty parameter for the anchor test; second, average for 0.0 indicated that there were no differences between mean of *b*-difficulty parameters for anchor test and *b*-difficulty parameters for the regular test; and third, above average for 1 meant that the mean *b*-difficulty parameter for regular test was less than the mean *b*-difficulty parameter for anchor test. So, if mean (*b*.RT.4) = -2.5 and mean (*b*.AT.4) = -3.5 then Δ[AT.*b*] = -1.0. Similarly, if mean (*b*.RT.4) = -3.5 and mean (*b*.AT.4) = -3.5 then Δ[AT.*b*] = 0.0. The same computation applied when mean (*b*.RT.4) = 2.5 and mean (*b*.AT.4) = 3.5 to get a difference of 1 (or Δ[AT.*b*] = 1.0)—i.e., using this formula $\Delta[AT.b] = \mu(b_{g,AT})—\mu(b_{g,RT})$ the three levels -1.0, 0.0, and 1.0 were obtained.

3.5 Summary of Study Conditions

There are 162 unique design conditions manipulated in this dissertation. Each condition contains specifications for one panel; however, because of the eight test forms within the panel, there were actually 1,296 test forms ($8 \times 162 = 1{,}296$) in play. This

implied 1,296 item files, one per test form. In addition, since there are 10 replication data sets per panel, there were a total of 12,960 data sets to analyze.

3.6 Data Generation Procedures and Output

To execute VSPANELREPS. BAT (Window Batch file) the following are required: first, GENEQUATE_ v45 –EXE; second, VSPANELS_1296ItemFiles or item files containing number of items for both unique and common items and item [*a, b, c*] parameters; and third, 162 control [.CON] files with 10 replications =1620 control files (VSPNL0001_01.CON to VSPNL0162_10.CON). Figure 3.3 shows the first panels while Figure 3.4 displays the first and last control file.

VSPNL0001_01_RT_4-1_AT4-1ni030_mna06_dt05_RTdb-1_ATdb-1
VSPNL0001_02_RT_4-2_AT5-2ni030_mna06_dt05_RTdb-1_ATdb-1
VSPNL0001_03_RT_5-1_AT4-1ni030_mna06_dt05_RTdb-1_ATdb-1
VSPNL0001_04_RT_5-2_AT5-2ni030_mna06_dt05_RTdb-1_ATdb-1
VSPNL0001_05_RT_5-1_AT5-2ni030_mna06_dt05_RTdb-1_ATdb-1
VSPNL0001_06_RT_5-2_AT6-1ni030_mna06_dt05_RTdb-1_ATdb-1
VSPNL0001_07_RT_6-1_AT5-2ni030_mna06_dt05_RTdb-1_ATdb-1
VSPNL0001_08_RT_6-2_AT6-1ni030_mna06_dt05_RTdb-1_ATdb-1

Figure 3.3. Panel No. 1 Showing 8 Forms and Conditions.

VSPNL1620_01_RT_RT4-1_AT4-1ni120_mna10_dt15_RTdb10_ATdb10
VSPNL1620_02_RT_RT4-2_AT5-2ni120_mna10_dt15_RTdb10_ATdb10
VSPNL1620_03_RT_RT5-1_AT4-1ni120_mna10_dt15_RTdb10_ATdb10
VSPNL1620_04_RT_RT5-2_AT5-2ni120_mna10_dt15_RTdb10_ATdb10
VSPNL1620_05_RT_RT5-1_AT5-2ni120_mna10_dt15_RTdb10_ATdb10
VSPNL1620_06_RT_RT5-2_AT6-1ni120_mna10_dt15_RTdb10_ATdb10
VSPNL1620_07_RT_RT6-1_AT5-2ni120_mna10_dt15_RTdb10_ATdb10
VSPNL1620_08_RT_RT6-2_AT6-1ni120_mna10_dt15_RTdb10_ATdb10

Figure 3.4. Panel No. 1,620 Showing 8 Forms and Conditions.

VSPANELREPS.BAT created all files for the entire study. Each of the 1620 control files created 8 set of files (observed RT and AT raw scores, theta scores, true RT and AT scores on the BASE form, and summary file for each analysis, *.OUT"). In short, there were 12,960 data files to actually equate. Put differently, each control file—each run (i.e., each .CON control file)—created EIGHT sets of data corresponding to one VS panel (2 sets for grade 4, 4 sets for grade 5, and 2 sets for grade 6). For example, first output showed observed score for the eight groups/forms; the second output showed response pattern (0 and 1); the third output showed true score for the eight groups/forms and a summary file. For this study only observed score and true score files were relevant.

3.7 Test Forms and Equating Methods under NEAT and RG/EG Designs

This subsection is a comprehensive description of how base forms and alternate forms were created, type of test score equating applied, equated scores, comparison of scores on the alternate form to equated scores and residual of variable applied in the computation of measures of accuracy—i.e., Bias, SEE, and RMSE as illustrated in Appendix C. Each of the tables in Appendix C is briefly outlined below.

Table C.1 provides a complete listing of the test forms for this study. RT(5.1) is presumed to be the alternate form for all grade 4 and 6 forms. Table C.2 shows the equating needed for each grade and group. As noted above, this equating paradigm assumes that RT(5.1) is specified as the alternate form for all grade 3 and grade 6 test forms. If RT(5.2) is instead used as the alternate form, the true scores, *tr*, will need to be linearly transformed to the RT(5.1) scale and the observed scores on the base form, *xr*, will need to be double-equated: first to the RT(5.2) scale and then to the RT(5.1) scale.

Using the test forms described in Table C.1, Table C.3 depicts the regular test variables to which the equating (see Table C.2) should be applied to get the scores on the RT(5.1) scale. These observed scores are converted to *eqxr* scores (see Table 1.1 Notations and Descriptions) after the equating is applied. After equating, the equated regular test score variable for the base form, denoted "*xr*", would become "*eqxr*" variables.

Table C.4 displays the equated regular test observed scores used in this study. Some variables shown in Table 1.1 are not used here, but they are included in Table 1.1 for completeness. Table C.5 shows the corresponding true scores required to compute residuals from estimated equated scores for each equating method across the grades. These are the true scores on the alternate form and are denoted as *ur* (see Table 1.1). Note that the comparative true scores assume that the alternate form for grades 4 and 6 include RT(5.1) as shown earlier in Table C.1.

Finally, Table C.6 shows the residual variables of interest. Note that there is only one set of residuals per grade and group. These residuals are computed by subtracting the corresponding regular test true score variable "*ur*" (see Table E.4) from the appropriate equated regular test observed score, "*eqxr*" (see Table C.3). These residuals are summarized as bias statistics (Bias), error variance statistics (SEE) or root mean-square errors (RMSE).

3.8 Equating Steps

The following is a summary of the equating steps for each replication and panel in this study.

i. Combine the Grade 5 data into two larger data sets, where one set comprises only RT(5.1) data and the other set contains only RT(5.2) data. Equate RT(5.2) to RT(5.1) by randomly equivalent groups methods (Levine observed score and equipercentile equating). Retain these <u>two</u> equating functions, $lin.eq_{RT(5.1)}[RT(5.2)]$ and $ep.eq_{RT(5.1)}[RT(5.2)]$ to apply to all form #2 results below; that is, to put all 5.2 equated form #2s on the RT(5.1) scale.

ii. Use AT(4.1) and AT(5.1) to equate RT(4.*f*) to RT(5.*f*) by form, *f*=(1,2) using NEAT design equating methods

iii. Use AT(5.1) and AT(6.1) to equate RT(6.*f*) to FT(6.*f*) by form, *f*=(1,2) using NEAT design equating methods

iv. For all form #2 results, *g*=(4,5,6), apply the random groups equating functions from Step #1—that is, $lin.eq_{RT(5.1)}[RT(5.2)]$ and $ep.eq_{RT(5.1)}[RT(5.2)]$—to the NEAT equated scores to put everything on the RT(5.1) scale.

v. Compare equated Form 5.1 scores relative to the $T(\theta_j)_{RT5.1}=\sum P_i(\theta_j)$, $i \in$ RT(5.1) for each panel replication with j=1,..,*N* examinees.

3.9 Evaluation of Equating Results and Accuracy

One of the major advantages of using generated data to evaluate test score equating methods is that the true item parameters and any equating relationships are well known; therefore, the precision and accuracy of equating results can be appropriately evaluated (Harris & Crouse, 1993). To compare the results of different equating methods under the five different study conditions, the simulation study was designed with an assumption that a *true*, or *criterion*, equating was available—i.e., the underlying ability

trait of the third-grade students; therefore, it was possible to investigate the closeness of the multiple equating methods to the criterion equating. In educational measurement and theory, it is a common practice in simulation and re-sampling studies for equating to be typically compared and evaluated on the basis of random and systematic error (or difference). The first is estimated by the standard error of equating (SEE or SE) while the second by the bias. On one hand, SEE is the random error that is introduced by an equating method; on the other hand, bias or systematic error is closely associated with the equating method and is the difference between the estimated equated relationship and a criterion equating relationship—this means smaller absolute values of bias indicate less biased estimates; therefore, more accuracy. Another type of equating error is root mean square error (RMSE)—this is total error; smaller RMSEs values suggest greater accuracy. It is the sum of SEE and bias.

In this research study, bias and RMSE will be calculated (van der Linden, 2006; Wang et al., 2008). For example, let d_j be the residual of interest where it is defined by this function: $d_j = X_j^{*} - Y_j$ where $Y_j = T_Y(\theta_j) = \sum_i P(\theta_j)$, as the true equivalent score (or the equivalent score that results from the IRT calibration) for the i=1,2,3,…,n test items on the alternate form—that is, Form 1 or RT(5.1) of the grade 5 within-grade regular test—and X* denotes the equated alternate form score based on sample N using any equating method. Table C.6 in Appendix C shows residuals that were analyzed for every equating method under both NEAT and RG/EG designs. Generally, the tables in Appendix C have been summarized in subsection 3.6 under "Test Forms and Equating Methods under NEAT and RG/EG Designs." Bias index is computed by averaging d_j. Thus,

$$Bias_{eqx} = \frac{1}{N}\sum_{j=1}^{N} d_j, \quad \text{(Eq. 3.2)}$$

where *N* represents the number of samples. Mostly expectation of bias is zero. The RMSE or root mean square difference (RMSD) statistic is calculated by averaging d^2_j and taking the square root. Therefore, the total error variability is represented by

$$RMSE_{eqx} = \sqrt{\frac{1}{N}\sum_{j=1}^{N} d_j^2}\ . \quad \text{(Eq. 3.3)}$$

It is important to note that each condition in this dissertation provides a prospective study to analyze where equating will naturally or automatically breakdown or simply work. Furthermore, there are only ten replications to the 162 conditions hence a small sampling distribution of means, and standard deviations of both bias and RMSE.

3.10 Real Data

The second component of this study covers empirical analysis of real data, which was from a large-scale international language assessment. For the purpose of this study, the two test forms are labeled as "Form X and Form Y," where the first form is considered new and the second one as old. The main purpose of these two tests was to measure people's language skills in an international business context. Each test form composed of 100 multiple-choice items that were scored dichotomically (that is, 1 for correct response and 0 for incorrect response), and there was a total of 47,289 examinees. Forty incomplete sentence item types, 48 reading composition item types, and 12 text-complete item types—i.e., a total of 100 items, which included the unique and anchor items in each form—are analyzed to investigate the extent equating can introduce errors

that can be tolerated. In the context of NEAT design there are fundamentally two sets of items which broadly constitute either a unique test or an anchor (common) test. The common set of questions can be either internal or external depending on the policy of the testing organization. There were twenty common items that were seen by all examinees taking the two test forms. The choice of number of common items were based on twenty per cent (20%), a popular rule of thumb in the equating literature and advocated by Kolen and Brennan (2014). The common items were treated as external test, which means the score on these common items did not count towards the total score. In some testing occasions, the anchor test score is added to the score of unique items to get an aggregated score. In such situations, the common item test is referred to as an internal anchor. There were eighty unique set of questions which means these items were uncommon (or were different) in the two test forms. Any missing response, for whatever reason, was treated as an incorrect answer, hence labeled as zero.

Although multiple considerations were put in place to construct and assembly both operational and field test items for the anchor test items in Form X and Form Y by the testing company/or test developers—for example, content, statistical, and psychometric properties of the common items, embedding field items to ensure examines do no detect operational and non-operational/or field items, use of classical test theory, IRT calibration, and differential item functioning to assess the quality of items, where poorly performing items were flagged and discontinued from operational use, approval by both subject matter expert and psychometricians—two fundamental aspects seemed to have been ignored. First, minority of items (only 4 out of 20 common items) were found

to be in the same serial position across the two test forms; a majority of the items (16 out of 20) were found to have changed the serial positions across the two test forms. Notably, the maximum position change recorded for some items was 23 position slots (i.e., the difference between item position in Form X and Form Y; second, whether indeed lack of motivation could have impacted on the equating error. These two setbacks are not the focus of this study; therefore, they are not investigated. Rather, they are accentuated for completeness and perspicuity.

3.11 Analysis of Real Data

The computer program LOGLIN/KE version 3.1 (Chen, Yan, Hemat, Han, & von Davier, 2011) was employed to perform pre-smoothing and equating procedures using the real data. The following three steps were followed. Step one involved LOGLIN procedure. Basically, LOGLIN is an independent computer program that fits loglinear models to either univariate or bivariate score distributions to smooth a variety of discreet empirical distributions (Holland & Thayer, 2000). In this study, bivariate test score distributions were applied because the dataset was split into two—i.e., the unique and the anchor test scores and their frequencies were computed using the program. The default method for converting test scores into Loglin input data was used in the analysis of the test scores from Form X and Form Y. The second step involved KE procedure. The KE software package is designed to perform observed score equating within different types of equating designs. For the purpose of this study, two equating methods with pre-smoothing option—i.e., Chain Equating and Post-Stratification Equating—were selected and conducted within the NEAT design. The third step entailed a merger of the results

from Chained Equating and Post-Stratification Equating. In this step, the equated scores and SEE were equated across the x-score scale to investigate if equating methods indeed introduced more equating error. The results are reported in Chapter IV.

Because the major goal of computation of SEE in the context of real data analysis was done to examine the extent equating methods could have introduced random error, this was achieved by not knowing the "truth" about the underlying hypothetical construct of the examinees who took both Form X and Form Y. This means that it was impossible to investigate how close the equating methods were relative to the truth.

CHAPTER IV

RESULTS

4.1 Overview

The data visualization approach to data analysis and presentation of the results was adopted in the current research study to depict both trend and pattern of the three common types of equating error or measures of accuracy—i.e., bias, SEE, and RMSE—for different equating methods used to construct a common vertical scale across multiple simulated conditions. The findings of these three measures of accuracy were presented in each of the study designs of this dissertation. As discussed in the previous chapters I and III, this study investigated five (5) factors: (i) test length; (ii) item discrimination (*a*-parameter); (iii) between-grade mean ability differences (θ, examinee proficiency on the theta scale) or the separation of grade ability distributions; (iv) distribution of ability difference (Pool Information) or grade-to-grade ability variability; and (v) anchor test mean difficulty differences or anchor test difficulty variability. Further, the results of the study were also tied to the nine study designs, which formed the bedrock of the research questions about the impact of each of the five (5) study conditions on the equating bias, SEE and RMSE. Each of the nine study designs were unique in the sense that total test length (anchor test proportion) and between-grade mean ability differences were held constant while the other three factors—i.e., (i) item discrimination (*a*-parameter); (ii) distribution of ability difference (Pool Information) or grade-to-grade ability variability;

and (iii) anchor test mean difficulty differences or anchor test difficulty variability—were conditioned on the two variables to produce bias, SEE and RMSE for each study design. For this reason, the bias, SEE, and RMSE were presented for each study design by equating method as follows:

30_0.5(6) Test Design by Equating Method (**small** BGMAD)

Bias for small BGMAD conditions by equating method for 30_0.5(6) Test Design
SEE for small BGMAD conditions by equating method for 30_0.5(6) Test Design
RMSE for small BGMAD conditions by equating method for 30_0.5(6) Test Design

30_1.0(6) Test Design by Equating Method (**medium** BGMAD)

Bias for medium BGMAD conditions by equating method for 30_1.0(6) Test Design
SEE for medium BGMAD conditions by equating method for 30_1.0(6) Test Design
RMSE for medium BGMAD conditions by equating method for 30_1.0(6) Test Design

30_1.5(6) Test Design by Equating Method (**large** BGMAD)

Bias for large BGMAD conditions by equating method for 30_1.5(6) Test Design
SEE for large BGMAD conditions by equating method for 30_1.5(6) Test Design
RMSE for large BGMAD conditions by equating method for 30_1.5(6) Test Design

60_0.5(12) Test Design by Equating Method (**small** BGMAD)

Bias for small BGMAD conditions by equating method for 60_0.5(12) Test Design
SEE for small BGMAD conditions by equating method for 60_0.5(12) Test Design
RMSE for small BGMAD conditions by equating method for 60_0.5(12) Test Design

60_1.0(12) Test Design by Equating Method (**medium** BGMAD)

Bias for small BGMAD conditions by equating method for 60_1.0(12) Test Design
SEE for small BGMAD conditions by equating method for 60_1.0(12) Test Design
RMSE for small BGMAD conditions by equating method for 60_1.0(12) Test Design

60_1.5(12) Test Design by Equating Method (**large** BGMAD)

Bias for small BGMAD conditions by equating method for 60_1.5(12) Test Design
SEE for small BGMAD conditions by equating method for 60_1.5(12) Test Design

4.2 Results of Simulated Data: Bias and RMSE

Bias statistic is used to measure the extent to which the equated score estimates align with those of the alternate form ability estimates calibrated from the 3PL model. If there is no difference between alternate form ability and equated score estimates for each equating method then the results do not prove anything. Differences between equated scores and ability estimates suggest existence of considerable impact necessitated by the study conditions under different equating methods. The study hypothesized that the 3PL model used to generate underlying ability in all test forms across all conditions somewhat

differs from the models used to approximate equating relationship when constructing the vertical scale. This difference witnessed in this study could be interpreted as a reflection of noise oftentimes experienced in real life testing.

In this study, the first research question examines the equating accuracy and consistency of the results across the five study conditions in order to assess potentially where an equating within the context of vertical scaling can be considered successful or not within and across various test study designs. The second research question attempts to clarify the amount of variation between anchor test item difficulty and the other four study conditions that can be tolerated under each equating method. In order to address these twin research questions, the total test items, the anchor test items and between-grade mean ability difference (BGMAD) are held constant and then the other three study conditions are manipulated. Bias and RMSE under similar research conditions are discussed for each test study design. The next nine subsections discuss the results of each test study design and then answer the two research questions.

4.2.1 30_0.5_6 Test Study Design

In this subsection, the results for test study design 30_0.5_6 are presented. This design had 30 total items and 6 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of group separation (or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large respectively. Medium (1.0) and large (1.5) BGMAD will be discussed in the subsequent subsections. This subsection mainly focuses on small (1.0) BGMAD,

which means 30 total items and 6 anchor test items were held constant and the other three study conditions were varied. Tables A.1-A.9 in Appendix A display average descriptive statistics and Figure B.1 in Appendix B shows the standard error of equating (SEE) for test study design 30_0.5_6. Table 4.1 represents bias, SEE, and RMSE for this test study design. Figure 4.1 demonstrates that there is almost zero bias for all conditions under all equating methods for small (0.5) between-grade mean ability difference (BGMAD). Both negative and positive values of bias are very close to zero apart from a few study conditions where results are inconsistent. For example, when distribution of ability difference (DAD) and anchor test mean difficulty difference (ATMDD) are below average (-1) and average (0) and item discrimination is moderate (0.6) the equating methods show inconsistency. This pattern of inconsistency is also repeated when item discrimination is high (1) where DAD is below average (-1) and above average (1) and especially where ATMDD is below average (-1). However, where BGMDD is small (0.5) and DAD is average (0), the equating methods perform similarly across the five study conditions with bias about zero.

Figure 4.2 shows test study design 30_0.5_6, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (30), small (0.5) BGMAD, moderate (0.6) and high (1) item discriminations are invariant. The RMSE values fall between 4 and 8. Interesting, when all conditions are held constant and manipulate item discrimination (moderate versus high), RMSE displays a clear consistency when item discrimination is moderate. Also, the RMSE values for moderate discrimination when other conditions are varied are lower than RMSE values when item

discrimination is high (1).There was one trend that stood out in this design that all equating methods consistently produced almost similar values of RMSE when magnitude of group separation or BGMAD was considerably small and *b*-item parameter for a grade was the same as the mean ability for that grade [or DAD was average (0)].

Addressing the research question number 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.1 revealed three results. First, the bias was consistent and very close to zero for all equating methods when small (0.5) BGMAD and moderate (0.6) item discrimination were held constant and DAD varied across below (-1), average (0), and above average (1) and when ATMDD was average (0) and above average (1). Second, the equating results were inaccurate and underestimated accuracy for all equating methods, as evidenced by negative bias, under small (0.5) BGMAD where DAD was below average (-1) for both moderate (0.6) and high (1) item discrimination and ATMDD was below average (-1) and average (0). Third, when small (0.5) BGMAD, high (1) item discrimination, and above average (1) ATMDD were held constant and manipulated DAD from below average (-1), average (0), and above average (1) the bias overestimated, as indicated by positive bias values, the accuracy of the equating results for all equating methods.

Overall the equating results in Figure 4.2 show that there was no significant difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods. That is, all equating methods seemed to have an indistinguishable performance without any discernible pattern apart

from slight differences where item discrimination was moderate (0.6) and high (1) for all conditions. Comparatively, though, moderate (0.6) item discrimination produced rather more accurate overall results than high (1) item discrimination under all conditions.

Table 4.1

BIAS, SEE, and RMSE Statistics for Test Study Design 30_0.5_6 by Equating Method Under All Conditions

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
1	30	0.5	-1	0.6	-1	Tucker	-0.10	4.57	4.61
1	30	0.5	-1	0.6	-1	Levine True	-0.21	4.58	4.80
1	30	0.5	-1	0.6	-1	Braun	-0.10	4.58	4.61
1	30	0.5	-1	0.6	-1	FEEE	-0.11	4.56	4.60
1	30	0.5	-1	0.6	-1	Chain_L	-0.16	4.59	4.70
1	30	0.5	-1	0.6	-1	Chain_E	-0.15	4.56	4.66
1	30	0.5	-1	0.6	-1	keNEATPSE_L	-0.10	4.57	4.60
1	30	0.5	-1	0.6	-1	keNEATPSE_E	-0.10	4.57	4.61
1	30	0.5	-1	0.6	-1	keNEATCE_L	-0.16	4.59	4.70
1	30	0.5	-1	0.6	-1	keNEATCE_E	-0.15	4.58	4.68
1	30	0.5	-1	0.6	-1	Linear	0.00	4.59	4.59
1	30	0.5	-1	0.6	-1	Equipercentile	0.00	4.59	4.59
82	30	0.5	-1	1	-1	Tucker	-0.20	4.76	4.81
82	30	0.5	-1	1	-1	Levine True	-0.37	5.15	5.34
82	30	0.5	-1	1	-1	Braun	-0.19	4.75	4.80
82	30	0.5	-1	1	-1	FEEE	-0.19	4.73	4.78
82	30	0.5	-1	1	-1	Chain_L	-0.29	4.82	4.93
82	30	0.5	-1	1	-1	Chain_E	-0.24	4.74	4.83
82	30	0.5	-1	1	-1	keNEATPSE_L	-0.18	4.72	4.77
82	30	0.5	-1	1	-1	keNEATPSE_E	-0.18	4.74	4.79
82	30	0.5	-1	1	-1	keNEATCE_L	-0.29	4.82	4.93
82	30	0.5	-1	1	-1	keNEATCE_E	-0.25	4.76	4.86
82	30	0.5	-1	1	-1	Linear	0.00	4.73	4.73
82	30	0.5	-1	1	-1	Equipercentile	0.00	4.73	4.73
2	30	0.5	-1	0.6	0	Tucker	-0.03	5.10	5.14

Table 4.1

Cont.

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
2	30	0.5	-1	0.6	0	Levine True	-0.08	5.10	5.33
2	30	0.5	-1	0.6	0	Braun	-0.04	5.11	5.14
2	30	0.5	-1	0.6	0	FEEE	-0.03	5.09	5.13
2	30	0.5	-1	0.6	0	Chain_L	-0.06	5.11	5.22
2	30	0.5	-1	0.6	0	Chain_E	-0.06	5.10	5.20
2	30	0.5	-1	0.6	0	keNEATPSE_L	-0.03	5.09	5.13
2	30	0.5	-1	0.6	0	keNEATPSE_E	-0.03	5.10	5.14
2	30	0.5	-1	0.6	0	keNEATCE_L	-0.06	5.11	5.22
2	30	0.5	-1	0.6	0	keNEATCE_E	-0.06	5.11	5.22
2	30	0.5	-1	0.6	0	Linear	0.02	5.09	5.09
2	30	0.5	-1	0.6	0	Equipercentile	0.02	5.09	5.09
83	30	0.5	-1	1	0	Tucker	-0.12	5.67	5.81
83	30	0.5	-1	1	0	Levine True	-0.19	5.62	5.93
83	30	0.5	-1	1	0	Braun	-0.12	5.68	5.82
83	30	0.5	-1	1	0	FEEE	-0.12	5.67	5.81
83	30	0.5	-1	1	0	Chain_L	-0.16	5.64	5.88
83	30	0.5	-1	1	0	Chain_E	-0.17	5.67	5.89
83	30	0.5	-1	1	0	keNEATPSE_L	-0.12	5.67	5.81
83	30	0.5	-1	1	0	keNEATPSE_E	-0.12	5.68	5.82
83	30	0.5	-1	1	0	keNEATCE_L	-0.16	5.64	5.88
83	30	0.5	-1	1	0	keNEATCE_E	-0.17	5.66	5.90
83	30	0.5	-1	1	0	Linear	0.00	5.73	5.73
83	30	0.5	-1	1	0	Equipercentile	0.00	5.72	5.72
3	30	0.5	-1	0.6	1	Tucker	-0.06	5.51	5.59
3	30	0.5	-1	0.6	1	Levine True	-0.10	5.50	5.77
3	30	0.5	-1	0.6	1	Braun	-0.07	5.51	5.59
3	30	0.5	-1	0.6	1	FEEE	-0.07	5.50	5.57
3	30	0.5	-1	0.6	1	Chain_L	-0.08	5.51	5.68
3	30	0.5	-1	0.6	1	Chain_E	-0.10	5.50	5.68
3	30	0.5	-1	0.6	1	keNEATPSE_L	-0.07	5.50	5.57
3	30	0.5	-1	0.6	1	keNEATPSE_E	-0.07	5.51	5.58
3	30	0.5	-1	0.6	1	keNEATCE_L	-0.08	5.51	5.68

Table 4.1

Cont.

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
3	30	0.5	-1	0.6	1	keNEATCE_E	-0.09	5.51	5.69
3	30	0.5	-1	0.6	1	Linear	-0.01	5.45	5.45
3	30	0.5	-1	0.6	1	Equipercentile	0.00	5.44	5.44
84	30	0.5	-1	1	1	Tucker	0.07	5.84	5.91
84	30	0.5	-1	1	1	Levine True	0.10	5.86	6.16
84	30	0.5	-1	1	1	Braun	0.04	5.84	5.91
84	30	0.5	-1	1	1	FEEE	0.06	5.82	5.89
84	30	0.5	-1	1	1	Chain_L	0.09	5.85	6.02
84	30	0.5	-1	1	1	Chain_E	0.03	5.83	6.00
84	30	0.5	-1	1	1	keNEATPSE_L	0.06	5.81	5.88
84	30	0.5	-1	1	1	keNEATPSE_E	0.06	5.83	5.91
84	30	0.5	-1	1	1	keNEATCE_L	0.09	5.85	6.02
84	30	0.5	-1	1	1	keNEATCE_E	0.04	5.84	6.01
84	30	0.5	-1	1	1	Linear	0.02	5.83	5.82
84	30	0.5	-1	1	1	Equipercentile	0.02	5.82	5.82
4	30	0.5	0	0.6	-1	Tucker	-0.10	6.01	6.10
4	30	0.5	0	0.6	-1	Levine True	-0.21	5.97	6.31
4	30	0.5	0	0.6	-1	Braun	-0.09	6.00	6.09
4	30	0.5	0	0.6	-1	FEEE	-0.08	5.99	6.08
4	30	0.5	0	0.6	-1	Chain_L	-0.16	6.00	6.21
4	30	0.5	0	0.6	-1	Chain_E	-0.13	5.97	6.17
4	30	0.5	0	0.6	-1	keNEATPSE_L	-0.08	5.99	6.08
4	30	0.5	0	0.6	-1	keNEATPSE_E	-0.08	6.00	6.09
4	30	0.5	0	0.6	-1	keNEATCE_L	-0.16	6.00	6.21
4	30	0.5	0	0.6	-1	keNEATCE_E	-0.13	5.98	6.18
4	30	0.5	0	0.6	-1	Linear	0.00	5.99	5.99
4	30	0.5	0	0.6	-1	Equipercentile	0.00	5.98	5.98
85	30	0.5	0	1	-1	Tucker	-0.03	6.47	6.59
85	30	0.5	0	1	-1	Levine True	-0.09	6.40	6.77
85	30	0.5	0	1	-1	Braun	-0.04	6.47	6.59
85	30	0.5	0	1	-1	FEEE	-0.02	6.45	6.57
85	30	0.5	0	1	-1	Chain_L	-0.06	6.44	6.68

Table 4.1

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
85	30	0.5	0	1	-1	Chain_E	-0.06	6.43	6.66
85	30	0.5	0	1	-1	keNEATPSE_L	-0.02	6.44	6.57
85	30	0.5	0	1	-1	keNEATPSE_E	-0.02	6.46	6.58
85	30	0.5	0	1	-1	keNEATCE_L	-0.06	6.44	6.68
85	30	0.5	0	1	-1	keNEATCE_E	-0.05	6.44	6.68
85	30	0.5	0	1	-1	Linear	-0.03	6.52	6.52
85	30	0.5	0	1	-1	Equipercentile	-0.03	6.52	6.52
5	30	0.5	0	0.6	0	Tucker	0.03	6.25	6.35
5	30	0.5	0	0.6	0	Levine True	0.00	6.19	6.54
5	30	0.5	0	0.6	0	Braun	0.03	6.25	6.35
5	30	0.5	0	0.6	0	FEEE	0.04	6.23	6.33
5	30	0.5	0	0.6	0	Chain_L	0.01	6.22	6.45
5	30	0.5	0	0.6	0	Chain_E	0.01	6.21	6.43
5	30	0.5	0	0.6	0	keNEATPSE_L	0.04	6.23	6.33
5	30	0.5	0	0.6	0	keNEATPSE_E	0.04	6.24	6.34
5	30	0.5	0	0.6	0	keNEATCE_L	0.01	6.22	6.45
5	30	0.5	0	0.6	0	keNEATCE_E	0.01	6.23	6.45
5	30	0.5	0	0.6	0	Linear	0.01	6.25	6.25
5	30	0.5	0	0.6	0	Equipercentile	0.01	6.25	6.25
86	30	0.5	0	1	0	Tucker	-0.04	6.89	7.07
86	30	0.5	0	1	0	Levine True	-0.07	6.84	7.24
86	30	0.5	0	1	0	Braun	-0.05	6.89	7.07
86	30	0.5	0	1	0	FEEE	-0.04	6.87	7.05
86	30	0.5	0	1	0	Chain_L	-0.06	6.87	7.17
86	30	0.5	0	1	0	Chain_E	-0.08	6.87	7.16
86	30	0.5	0	1	0	keNEATPSE_L	-0.04	6.87	7.05
86	30	0.5	0	1	0	keNEATPSE_E	-0.04	6.88	7.07
86	30	0.5	0	1	0	keNEATCE_L	-0.06	6.87	7.17
86	30	0.5	0	1	0	keNEATCE_E	-0.07	6.88	7.18
86	30	0.5	0	1	0	Linear	-0.01	6.92	6.92
86	30	0.5	0	1	0	Equipercentile	-0.01	6.92	6.92
6	30	0.5	0	0.6	1	Tucker	0.03	6.13	6.22

Table 4.1

Cont.

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
6	30	0.5	0	0.6	1	Levine True	0.06	6.07	6.39
6	30	0.5	0	0.6	1	Braun	0.04	6.13	6.21
6	30	0.5	0	0.6	1	FEEE	0.04	6.11	6.20
6	30	0.5	0	0.6	1	Chain_L	0.05	6.10	6.31
6	30	0.5	0	0.6	1	Chain_E	0.03	6.09	6.29
6	30	0.5	0	0.6	1	keNEATPSE_L	0.04	6.11	6.20
6	30	0.5	0	0.6	1	keNEATPSE_E	0.04	6.12	6.21
6	30	0.5	0	0.6	1	keNEATCE_L	0.05	6.10	6.31
6	30	0.5	0	0.6	1	keNEATCE_E	0.04	6.11	6.31
6	30	0.5	0	0.6	1	Linear	0.00	6.14	6.14
6	30	0.5	0	0.6	1	Equipercentile	0.01	6.14	6.14
87	30	0.5	0	1	1	Tucker	-0.04	6.65	6.77
87	30	0.5	0	1	1	Levine True	-0.01	6.62	6.98
87	30	0.5	0	1	1	Braun	-0.05	6.64	6.76
87	30	0.5	0	1	1	FEEE	-0.04	6.62	6.74
87	30	0.5	0	1	1	Chain_L	-0.02	6.63	6.87
87	30	0.5	0	1	1	Chain_E	-0.07	6.63	6.86
87	30	0.5	0	1	1	keNEATPSE_L	-0.04	6.62	6.73
87	30	0.5	0	1	1	keNEATPSE_E	-0.03	6.64	6.76
87	30	0.5	0	1	1	keNEATCE_L	-0.02	6.63	6.87
87	30	0.5	0	1	1	keNEATCE_E	-0.05	6.64	6.88
87	30	0.5	0	1	1	Linear	-0.01	6.69	6.68
87	30	0.5	0	1	1	Equipercentile	0.00	6.68	6.68
7	30	0.5	1	0.6	-1	Tucker	-0.04	6.23	6.33
7	30	0.5	1	0.6	-1	Levine True	-0.07	6.17	6.52
7	30	0.5	1	0.6	-1	Braun	-0.03	6.23	6.32
7	30	0.5	1	0.6	-1	FEEE	-0.01	6.21	6.31
7	30	0.5	1	0.6	-1	Chain_L	-0.06	6.20	6.43
7	30	0.5	1	0.6	-1	Chain_E	-0.04	6.19	6.40
7	30	0.5	1	0.6	-1	keNEATPSE_L	-0.01	6.21	6.31
7	30	0.5	1	0.6	-1	keNEATPSE_E	-0.01	6.22	6.31
7	30	0.5	1	0.6	-1	keNEATCE_L	-0.06	6.20	6.43

Table 4.1

Cont.

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
7	30	0.5	1	0.6	-1	keNEATCE_E	-0.04	6.20	6.42
7	30	0.5	1	0.6	-1	Linear	0.00	6.26	6.25
7	30	0.5	1	0.6	-1	Equipercentile	0.00	6.25	6.25
88	30	0.5	1	1	-1	Tucker	-0.10	7.74	7.95
88	30	0.5	1	1	-1	Levine True	-0.15	7.66	8.11
88	30	0.5	1	1	-1	Braun	-0.06	7.73	7.94
88	30	0.5	1	1	-1	FEEE	-0.04	7.71	7.92
88	30	0.5	1	1	-1	Chain_L	-0.13	7.70	8.04
88	30	0.5	1	1	-1	Chain_E	-0.08	7.69	8.02
88	30	0.5	1	1	-1	keNEATPSE_L	-0.05	7.71	7.92
88	30	0.5	1	1	-1	keNEATPSE_E	-0.05	7.72	7.93
88	30	0.5	1	1	-1	keNEATCE_L	-0.13	7.70	8.04
88	30	0.5	1	1	-1	keNEATCE_E	-0.08	7.70	8.04
88	30	0.5	1	1	-1	Linear	0.01	7.83	7.83
88	30	0.5	1	1	-1	Equipercentile	0.01	7.83	7.83
8	30	0.5	1	0.6	0	Tucker	-0.01	6.33	6.41
8	30	0.5	1	0.6	0	Levine True	-0.02	6.29	6.61
8	30	0.5	1	0.6	0	Braun	0.00	6.33	6.41
8	30	0.5	1	0.6	0	FEEE	0.01	6.32	6.40
8	30	0.5	1	0.6	0	Chain_L	-0.01	6.31	6.51
8	30	0.5	1	0.6	0	Chain_E	-0.01	6.31	6.50
8	30	0.5	1	0.6	0	keNEATPSE_L	0.01	6.31	6.39
8	30	0.5	1	0.6	0	keNEATPSE_E	0.01	6.32	6.40
8	30	0.5	1	0.6	0	keNEATCE_L	-0.01	6.31	6.51
8	30	0.5	1	0.6	0	keNEATCE_E	-0.01	6.32	6.52
8	30	0.5	1	0.6	0	Linear	-0.01	6.32	6.32
8	30	0.5	1	0.6	0	Equipercentile	0.00	6.32	6.32
89	30	0.5	1	1	0	Tucker	0.04	6.60	6.73
89	30	0.5	1	1	0	Levine True	0.09	6.54	6.90
89	30	0.5	1	1	0	Braun	0.05	6.60	6.73
89	30	0.5	1	1	0	FEEE	0.06	6.59	6.71
89	30	0.5	1	1	0	Chain_L	0.06	6.57	6.82

Table 4.1

Cont.

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
89	30	0.5	1	1	0	Chain_E	0.07	6.56	6.81
89	30	0.5	1	1	0	keNEATPSE_L	0.06	6.58	6.71
89	30	0.5	1	1	0	keNEATPSE_E	0.06	6.60	6.72
89	30	0.5	1	1	0	keNEATCE_L	0.06	6.57	6.82
89	30	0.5	1	1	0	keNEATCE_E	0.07	6.58	6.83
89	30	0.5	1	1	0	Linear	0.00	6.61	6.61
89	30	0.5	1	1	0	Equipercentile	0.00	6.61	6.61
9	30	0.5	1	0.6	1	Tucker	0.07	6.27	6.33
9	30	0.5	1	0.6	1	Levine True	0.17	6.23	6.53
9	30	0.5	1	0.6	1	Braun	0.08	6.27	6.34
9	30	0.5	1	0.6	1	FEEE	0.09	6.26	6.33
9	30	0.5	1	0.6	1	Chain_L	0.12	6.26	6.43
9	30	0.5	1	0.6	1	Chain_E	0.11	6.24	6.41
9	30	0.5	1	0.6	1	keNEATPSE_L	0.09	6.26	6.32
9	30	0.5	1	0.6	1	keNEATPSE_E	0.09	6.27	6.33
9	30	0.5	1	0.6	1	keNEATCE_L	0.12	6.26	6.43
9	30	0.5	1	0.6	1	keNEATCE_E	0.12	6.26	6.43
9	30	0.5	1	0.6	1	Linear	-0.01	6.27	6.27
9	30	0.5	1	0.6	1	Equipercentile	-0.01	6.27	6.27
90	30	0.5	1	1	1	Tucker	0.14	6.00	6.04
90	30	0.5	1	1	1	Levine True	0.26	6.07	6.33
90	30	0.5	1	1	1	Braun	0.14	6.01	6.05
90	30	0.5	1	1	1	FEEE	0.17	6.00	6.04
90	30	0.5	1	1	1	Chain_L	0.20	6.03	6.15
90	30	0.5	1	1	1	Chain_E	0.18	6.01	6.13
90	30	0.5	1	1	1	keNEATPSE_L	0.17	6.00	6.04
90	30	0.5	1	1	1	keNEATPSE_E	0.17	6.01	6.05
90	30	0.5	1	1	1	keNEATCE_L	0.20	6.03	6.15
90	30	0.5	1	1	1	keNEATCE_E	0.19	6.03	6.15
90	30	0.5	1	1	1	Linear	0.00	5.97	5.97
90	30	0.5	1	1	1	Equipercentile	0.00	5.97	5.97

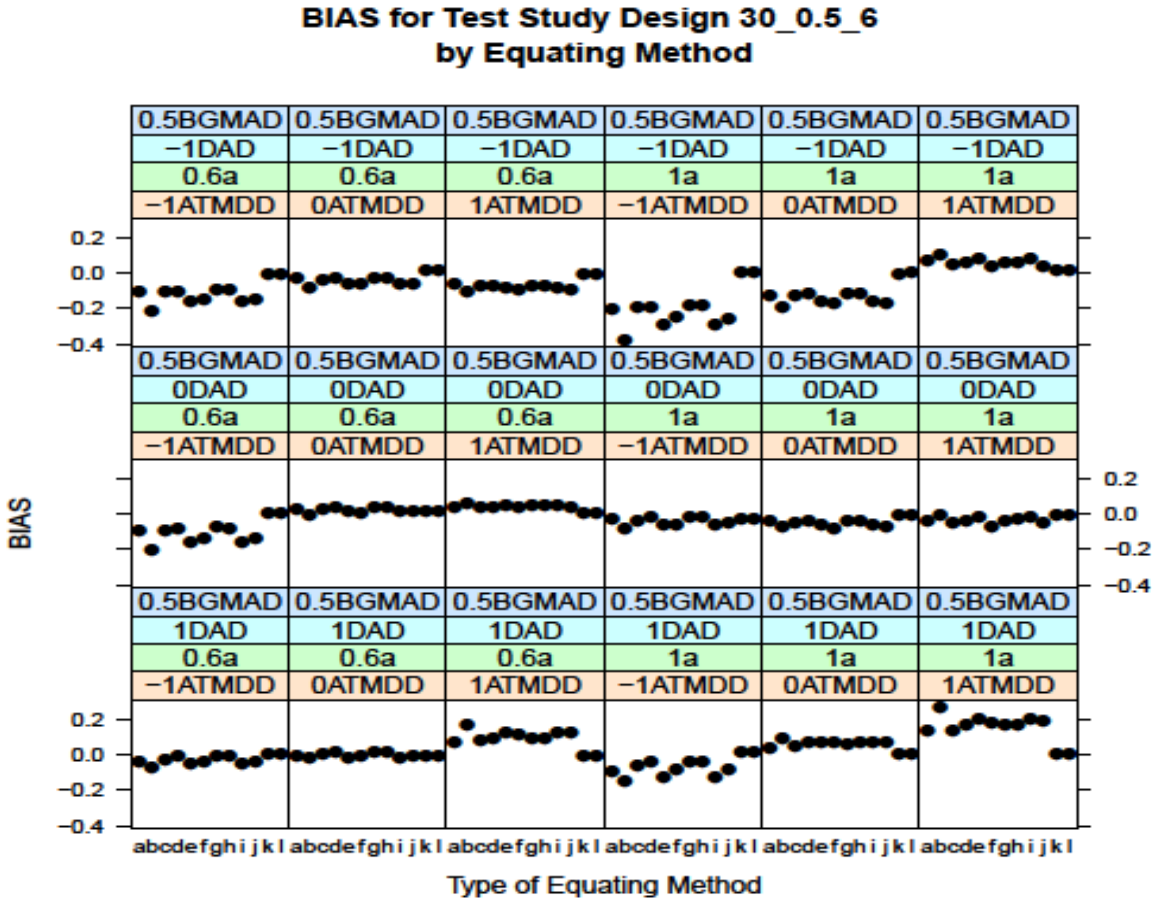

Figure 4.1. Bias for Test Study Design 30_0.5_6 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

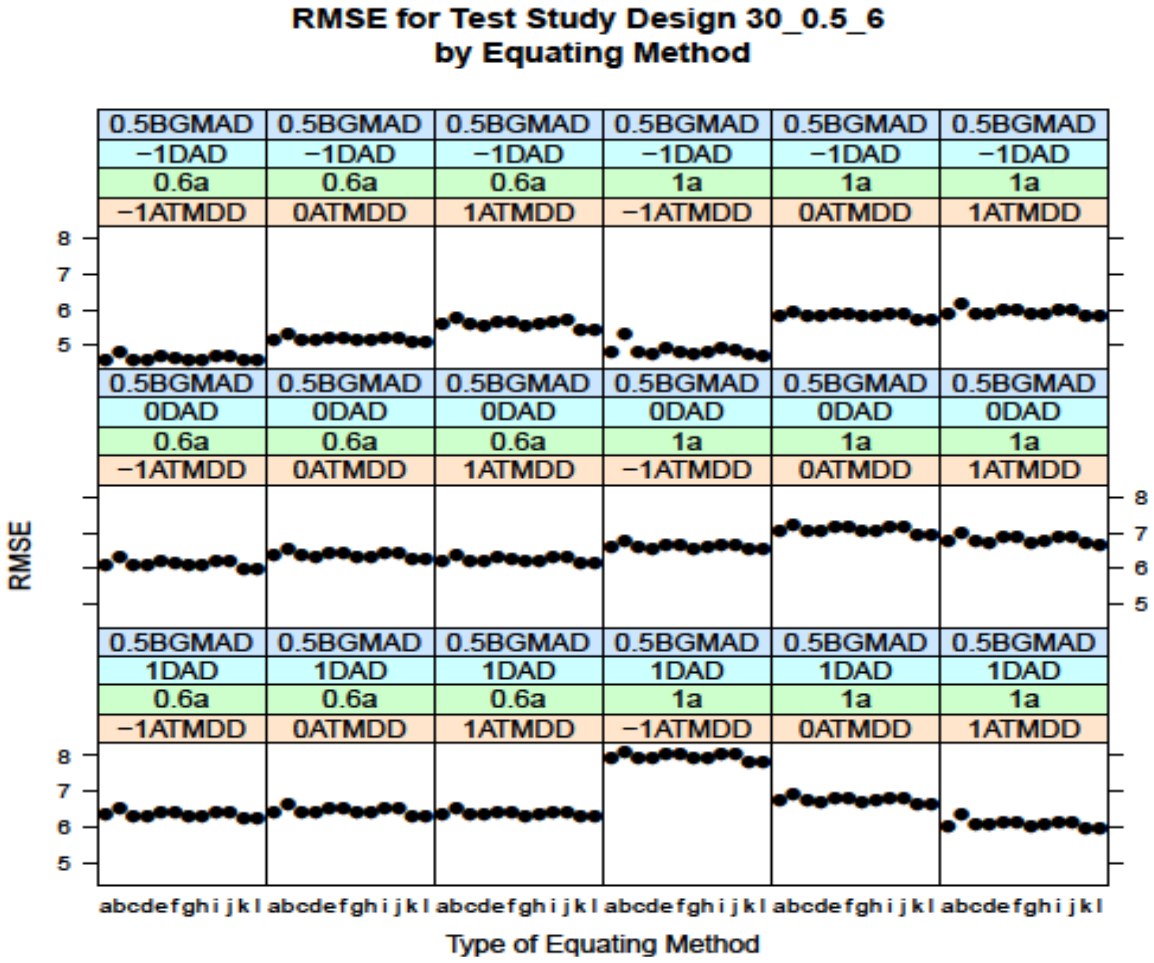

Figure 4.2. Root Mean Square Error (RMSE) for Test Study Design 30_0.5_6 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.2.2 30_1.0_6 Test Study Design

In this subsection, the results for test study design 30_1.0_6 are presented. This design had 30 total items and 6 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in

subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of group separation (or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large respectively. Large (1.5) BGMAD will be discussed in the subsequent subsections. This subsection mainly focuses on medium (1.0) BGMAD, which means 30 total items and 6 anchor test items were held constant and the other three study conditions were varied. Tables A.10-A.18 in Appendix A display average descriptive statistics and Figure B.2 in Appendix B shows the standard error of equating (SEE) for test study design 30_1.0_6. Table 4.2 represents bias, SEE, and RMSE for this test study design. Figure 4.3 demonstrates that there is zero bias for all conditions under linear and equipercentile equating methods for medium (1.0) between-grade mean ability difference (BGMAD). Other equating methods show both negative and positive values of bias which are very close to zero apart from a few study conditions where results are inconsistent. For example, when distribution of ability difference (DAD) and anchor test mean difficulty difference (ATMDD) are below average (-1), average (1.0), and above average (1) and item discrimination is moderate (0.6) the equating methods show inconsistency. This pattern of inconsistency is also repeated when item discrimination is high (1) where DAD is below average (-1), average (0), and above average (1). However, where BGMDD is medium (1.0) and DAD is average (0), the equating methods perform similarly across the five study conditions with bias about zero except when ATMDD is average (0) and item discrimination is high (1) resulting to negative values. Similarly, where BGMDD is medium (1.0) and DAD is above average (1), the equating methods

yield similar bias results for all the study conditions apart from when ATMDD is above average (1), which shows positive bias values.

Figure 4.4 shows test study design 30_1.0_6, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (30), medium (1.0) BGMAD, moderate (0.6) and high (1) item discriminations are invariant. The RMSE values fall between 6 and 8.5. When all conditions are held constant and manipulate item discrimination (moderate versus high), RMSE displays a clear consistency where moderate (0.6) item discrimination across the other four study conditions results in smaller (more accurate) RMSE values than RMSE values for high (1) item discrimination varied over the other four study conditions. Therefore, two patterns are discernible in this design based on conditions varied under either moderate item discrimination or high item discrimination with the former performing better than the latter in terms of accuracy.

Addressing the research question number 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.3 revealed the following results. First, the bias was consistent and very close to zero for all equating methods when medium (1.0) BGMAD for both moderate (0.6) and high (1.0) item discrimination and DAD was below average (-1) and average (0), and when ATMDD was below average (-1) and average (0). Second, the equating methods performed similarly when BGMAD was medium (1.0) under average DAD with moderate (0.6) item discrimination for average and above average ATMDD and when item discrimination was high for above average ATMDD

and also for medium (1.0) BGMAD when DAD was below average and moderate (0.6) item discrimination for above average (1) ATMDD. The rest of the results for other study conditions under this design were inaccurate and underestimated or overestimated accuracy for all equating methods, as evidenced by negative and positive bias.

Overall the equating results in Figure 4.4 show that the smallest difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods was under below average (-1), average (0), and above average (1) DAD when item discrimination was moderate (0.6) and ATMDD below average (-1) conditions. However, largest difference between anchor test mean difficulty and the other four study conditions when DAD varied across its three levels with a high (1) item discrimination and ATMDD was above average (1). Moderate (0.6) item discrimination produced rather more accurate overall results than high (1) item discrimination under all conditions across all equating methods.

Table 4.2

BIAS, SEE, and RMSE Statistics for Test Study Design 30_1.0_6 by Equating Method Under All Conditions

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
10	30	1	-1	0.6	-1	Tucker	-0.23	4.18	4.29
10	30	1	-1	0.6	-1	Levine True	-0.62	4.68	5.34
10	30	1	-1	0.6	-1	Braun	-0.25	4.21	4.32
10	30	1	-1	0.6	-1	FEEE	-0.23	4.18	4.29
10	30	1	-1	0.6	-1	Chain_L	-0.42	4.24	4.58
10	30	1	-1	0.6	-1	Chain_E	-0.40	4.20	4.51
10	30	1	-1	0.6	-1	keNEATPSE_L	-0.23	4.17	4.28

Table 4.2

Cont.

Panel	Study Condition: Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	Statistic: BIAS	SEE	RMSE
10	30	1	-1	0.6	-1	keNEATPSE_E	-0.23	4.19	4.30
10	30	1	-1	0.6	-1	keNEATCE_L	-0.42	4.24	4.58
10	30	1	-1	0.6	-1	keNEATCE_E	-0.41	4.22	4.56
10	30	1	-1	0.6	-1	Linear	0.02	4.17	4.17
10	30	1	-1	0.6	-1	Equipercentile	0.02	4.17	4.17
91	30	1	-1	1	-1	Tucker	-0.51	4.79	4.93
91	30	1	-1	1	-1	Levine True	-1.40	6.47	7.13
91	30	1	-1	1	-1	Braun	-0.48	4.74	4.87
91	30	1	-1	1	-1	FEEE	-0.44	4.71	4.83
91	30	1	-1	1	-1	Chain_L	-0.90	5.14	5.50
91	30	1	-1	1	-1	Chain_E	-0.67	4.78	5.04
91	30	1	-1	1	-1	keNEATPSE_L	-0.42	4.66	4.79
91	30	1	-1	1	-1	keNEATPSE_E	-0.44	4.71	4.84
91	30	1	-1	1	-1	keNEATCE_L	-0.90	5.14	5.50
91	30	1	-1	1	-1	keNEATCE_E	-0.71	4.88	5.26
91	30	1	-1	1	-1	Linear	0.00	4.56	4.56
91	30	1	-1	1	-1	Equipercentile	0.01	4.55	4.55
11	30	1	-1	0.6	0	Tucker	-0.33	4.45	4.66
11	30	1	-1	0.6	0	Levine True	-0.58	4.39	5.22
11	30	1	-1	0.6	0	Braun	-0.35	4.49	4.69
11	30	1	-1	0.6	0	FEEE	-0.35	4.45	4.65
11	30	1	-1	0.6	0	Chain_L	-0.47	4.45	4.96
11	30	1	-1	0.6	0	Chain_E	-0.51	4.48	4.96
11	30	1	-1	0.6	0	keNEATPSE_L	-0.34	4.44	4.64
11	30	1	-1	0.6	0	keNEATPSE_E	-0.34	4.47	4.67
11	30	1	-1	0.6	0	keNEATCE_L	-0.47	4.45	4.96
11	30	1	-1	0.6	0	keNEATCE_E	-0.48	4.47	4.98
11	30	1	-1	0.6	0	Linear	-0.01	4.47	4.47
11	30	1	-1	0.6	0	Equipercentile	-0.01	4.47	4.47
92	30	1	-1	1	0	Tucker	-0.55	5.15	5.50
92	30	1	-1	1	0	Levine True	-0.89	5.21	6.10
92	30	1	-1	1	0	Braun	-0.56	5.19	5.53
92	30	1	-1	1	0	FEEE	-0.53	5.15	5.50

Table 4.2

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
92	30	1	-1	1	0	Chain_L	-0.74	5.13	5.77
92	30	1	-1	1	0	Chain_E	-0.75	5.16	5.77
92	30	1	-1	1	0	keNEATPSE_L	-0.54	5.11	5.46
92	30	1	-1	1	0	keNEATPSE_E	-0.53	5.16	5.51
92	30	1	-1	1	0	keNEATCE_L	-0.74	5.13	5.77
92	30	1	-1	1	0	keNEATCE_E	-0.76	5.17	5.82
92	30	1	-1	1	0	Linear	0.00	5.19	5.19
92	30	1	-1	1	0	Equipercentile	0.00	5.19	5.19
12	30	1	-1	0.6	1	Tucker	-0.14	4.45	4.72
12	30	1	-1	0.6	1	Levine True	-0.23	4.35	5.20
12	30	1	-1	0.6	1	Braun	-0.15	4.47	4.73
12	30	1	-1	0.6	1	FEEE	-0.14	4.44	4.70
12	30	1	-1	0.6	1	Chain_L	-0.20	4.41	5.00
12	30	1	-1	0.6	1	Chain_E	-0.23	4.43	5.00
12	30	1	-1	0.6	1	keNEATPSE_L	-0.14	4.44	4.70
12	30	1	-1	0.6	1	keNEATPSE_E	-0.14	4.45	4.71
12	30	1	-1	0.6	1	keNEATCE_L	-0.20	4.41	5.00
12	30	1	-1	0.6	1	keNEATCE_E	-0.20	4.42	5.02
12	30	1	-1	0.6	1	Linear	0.01	4.50	4.50
12	30	1	-1	0.6	1	Equipercentile	0.01	4.50	4.50
93	30	1	-1	1	1	Tucker	-0.34	4.80	5.30
93	30	1	-1	1	1	Levine True	-0.48	4.67	5.62
93	30	1	-1	1	1	Braun	-0.38	4.86	5.35
93	30	1	-1	1	1	FEEE	-0.37	4.83	5.32
93	30	1	-1	1	1	Chain_L	-0.44	4.72	5.53
93	30	1	-1	1	1	Chain_E	-0.51	4.84	5.60
93	30	1	-1	1	1	keNEATPSE_L	-0.37	4.83	5.32
93	30	1	-1	1	1	keNEATPSE_E	-0.37	4.85	5.34
93	30	1	-1	1	1	keNEATCE_L	-0.44	4.72	5.53
93	30	1	-1	1	1	keNEATCE_E	-0.49	4.82	5.61
93	30	1	-1	1	1	Linear	-0.01	4.93	4.93
93	30	1	-1	1	1	Equipercentile	-0.01	4.92	4.92
13	30	1	0	0.6	-1	Tucker	-0.18	5.19	5.35

Table 4.2

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
13	30	1	0	0.6	-1	Levine True	-0.47	5.27	6.27
13	30	1	0	0.6	-1	Braun	-0.20	5.20	5.36
13	30	1	0	0.6	-1	FEEE	-0.17	5.16	5.32
13	30	1	0	0.6	-1	Chain_L	-0.32	5.25	5.74
13	30	1	0	0.6	-1	Chain_E	-0.29	5.19	5.64
13	30	1	0	0.6	-1	keNEATPSE_L	-0.17	5.14	5.30
13	30	1	0	0.6	-1	keNEATPSE_E	-0.17	5.18	5.33
13	30	1	0	0.6	-1	keNEATCE_L	-0.32	5.25	5.74
13	30	1	0	0.6	-1	keNEATCE_E	-0.28	5.21	5.69
13	30	1	0	0.6	-1	Linear	-0.01	5.17	5.16
13	30	1	0	0.6	-1	Equipercentile	0.00	5.16	5.16
94	30	1	0	1	-1	Tucker	-0.09	6.05	6.41
94	30	1	0	1	-1	Levine True	-0.22	5.76	7.02
94	30	1	0	1	-1	Braun	-0.10	6.06	6.42
94	30	1	0	1	-1	FEEE	-0.05	6.02	6.38
94	30	1	0	1	-1	Chain_L	-0.15	5.94	6.71
94	30	1	0	1	-1	Chain_E	-0.22	6.01	6.72
94	30	1	0	1	-1	keNEATPSE_L	-0.05	5.98	6.35
94	30	1	0	1	-1	keNEATPSE_E	-0.04	6.04	6.40
94	30	1	0	1	-1	keNEATCE_L	-0.15	5.94	6.71
94	30	1	0	1	-1	keNEATCE_E	-0.20	5.98	6.76
94	30	1	0	1	-1	Linear	0.00	6.15	6.15
94	30	1	0	1	-1	Equipercentile	0.00	6.15	6.15
14	30	1	0	0.6	0	Tucker	-0.16	5.48	5.86
14	30	1	0	0.6	0	Levine True	-0.29	5.25	6.40
14	30	1	0	0.6	0	Braun	-0.17	5.48	5.86
14	30	1	0	0.6	0	FEEE	-0.15	5.45	5.83
14	30	1	0	0.6	0	Chain_L	-0.24	5.37	6.17
14	30	1	0	0.6	0	Chain_E	-0.26	5.39	6.16
14	30	1	0	0.6	0	keNEATPSE_L	-0.15	5.44	5.82
14	30	1	0	0.6	0	keNEATPSE_E	-0.15	5.46	5.84
14	30	1	0	0.6	0	keNEATCE_L	-0.24	5.37	6.17
14	30	1	0	0.6	0	keNEATCE_E	-0.23	5.39	6.19

Table 4.2

Cont.

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
14	30	1	0	0.6	0	Linear	0.00	5.62	5.62
14	30	1	0	0.6	0	Equipercentile	0.00	5.62	5.62
95	30	1	0	1	0	Tucker	-0.54	6.60	7.14
95	30	1	0	1	0	Levine True	-0.80	6.48	7.79
95	30	1	0	1	0	Braun	-0.56	6.59	7.12
95	30	1	0	1	0	FEEE	-0.49	6.54	7.07
95	30	1	0	1	0	Chain_L	-0.69	6.56	7.51
95	30	1	0	1	0	Chain_E	-0.63	6.51	7.41
95	30	1	0	1	0	keNEATPSE_L	-0.48	6.49	7.03
95	30	1	0	1	0	keNEATPSE_E	-0.48	6.55	7.08
95	30	1	0	1	0	keNEATCE_L	-0.69	6.56	7.51
95	30	1	0	1	0	keNEATCE_E	-0.62	6.50	7.45
95	30	1	0	1	0	Linear	-0.02	6.64	6.63
95	30	1	0	1	0	Equipercentile	-0.02	6.63	6.63
15	30	1	0	0.6	1	Tucker	-0.01	5.60	5.89
15	30	1	0	0.6	1	Levine True	0.00	5.48	6.56
15	30	1	0	0.6	1	Braun	-0.04	5.60	5.89
15	30	1	0	0.6	1	FEEE	-0.01	5.57	5.85
15	30	1	0	0.6	1	Chain_L	0.00	5.55	6.24
15	30	1	0	0.6	1	Chain_E	-0.09	5.55	6.21
15	30	1	0	0.6	1	keNEATPSE_L	-0.01	5.55	5.84
15	30	1	0	0.6	1	keNEATPSE_E	-0.01	5.58	5.87
15	30	1	0	0.6	1	keNEATCE_L	0.00	5.55	6.24
15	30	1	0	0.6	1	keNEATCE_E	-0.04	5.55	6.24
15	30	1	0	0.6	1	Linear	0.00	5.65	5.65
15	30	1	0	0.6	1	Equipercentile	0.00	5.65	5.65
96	30	1	0	1	1	Tucker	0.00	7.20	7.98
96	30	1	0	1	1	Levine True	0.01	6.89	8.40
96	30	1	0	1	1	Braun	-0.01	7.16	7.95
96	30	1	0	1	1	FEEE	0.01	7.12	7.91
96	30	1	0	1	1	Chain_L	0.01	7.02	8.25
96	30	1	0	1	1	Chain_E	-0.13	7.08	8.26
96	30	1	0	1	1	keNEATPSE_L	-0.01	7.08	7.87

Table 4.2

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
96	30	1	0	1	1	keNEATPSE_E	0.01	7.13	7.92
96	30	1	0	1	1	keNEATCE_L	0.01	7.02	8.25
96	30	1	0	1	1	keNEATCE_E	-0.10	7.06	8.27
96	30	1	0	1	1	Linear	0.03	7.59	7.58
96	30	1	0	1	1	Equipercentile	0.03	7.58	7.58
16	30	1	1	0.6	-1	Tucker	-0.05	6.06	6.34
16	30	1	1	0.6	-1	Levine True	-0.13	5.90	7.12
16	30	1	1	0.6	-1	Braun	-0.02	6.05	6.33
16	30	1	1	0.6	-1	FEEE	0.03	6.02	6.30
16	30	1	1	0.6	-1	Chain_L	-0.09	6.00	6.72
16	30	1	1	0.6	-1	Chain_E	-0.02	5.97	6.66
16	30	1	1	0.6	-1	keNEATPSE_L	0.03	5.99	6.28
16	30	1	1	0.6	-1	keNEATPSE_E	0.03	6.03	6.31
16	30	1	1	0.6	-1	keNEATCE_L	-0.09	6.00	6.72
16	30	1	1	0.6	-1	keNEATCE_E	-0.03	5.99	6.70
16	30	1	1	0.6	-1	Linear	0.00	6.08	6.07
16	30	1	1	0.6	-1	Equipercentile	0.00	6.07	6.07
97	30	1	1	1	-1	Tucker	-0.02	6.96	7.59
97	30	1	1	1	-1	Levine True	-0.08	6.64	8.16
97	30	1	1	1	-1	Braun	0.04	6.94	7.57
97	30	1	1	1	-1	FEEE	0.09	6.90	7.54
97	30	1	1	1	-1	Chain_L	-0.06	6.80	7.90
97	30	1	1	1	-1	Chain_E	-0.02	6.83	7.89
97	30	1	1	1	-1	keNEATPSE_L	0.09	6.87	7.52
97	30	1	1	1	-1	keNEATPSE_E	0.09	6.91	7.55
97	30	1	1	1	-1	keNEATCE_L	-0.06	6.80	7.90
97	30	1	1	1	-1	keNEATCE_E	0.00	6.82	7.92
97	30	1	1	1	-1	Linear	0.02	7.14	7.14
97	30	1	1	1	-1	Equipercentile	0.02	7.14	7.14
17	30	1	1	0.6	0	Tucker	0.07	5.96	6.33
17	30	1	1	0.6	0	Levine True	0.09	5.73	6.94
17	30	1	1	0.6	0	Braun	0.09	5.96	6.33
17	30	1	1	0.6	0	FEEE	0.11	5.93	6.30

Table 4.2

Cont.

Panel	Study Condition: Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	Statistic: BIAS	SEE	RMSE
17	30	1	1	0.6	0	Chain_L	0.08	5.85	6.66
17	30	1	1	0.6	0	Chain_E	0.07	5.87	6.66
17	30	1	1	0.6	0	keNEATPSE_L	0.11	5.91	6.28
17	30	1	1	0.6	0	keNEATPSE_E	0.12	5.94	6.31
17	30	1	1	0.6	0	keNEATCE_L	0.08	5.85	6.66
17	30	1	1	0.6	0	keNEATCE_E	0.09	5.88	6.69
17	30	1	1	0.6	0	Linear	-0.01	6.05	6.05
17	30	1	1	0.6	0	Equipercentile	-0.01	6.05	6.05
98	30	1	1	1	0	Tucker	0.02	7.42	8.16
98	30	1	1	1	0	Levine True	0.02	7.10	8.66
98	30	1	1	1	0	Braun	0.10	7.39	8.14
98	30	1	1	1	0	FEEE	0.14	7.36	8.11
98	30	1	1	1	0	Chain_L	0.02	7.24	8.46
98	30	1	1	1	0	Chain_E	0.06	7.28	8.46
98	30	1	1	1	0	keNEATPSE_L	0.13	7.32	8.07
98	30	1	1	1	0	keNEATPSE_E	0.14	7.37	8.12
98	30	1	1	1	0	keNEATCE_L	0.02	7.24	8.46
98	30	1	1	1	0	keNEATCE_E	0.07	7.28	8.49
98	30	1	1	1	0	Linear	0.00	7.73	7.73
98	30	1	1	1	0	Equipercentile	0.01	7.73	7.73
18	30	1	1	0.6	1	Tucker	0.25	5.98	6.25
18	30	1	1	0.6	1	Levine True	0.50	5.84	6.96
18	30	1	1	0.6	1	Braun	0.28	5.97	6.24
18	30	1	1	0.6	1	FEEE	0.32	5.94	6.21
18	30	1	1	0.6	1	Chain_L	0.39	5.92	6.61
18	30	1	1	0.6	1	Chain_E	0.31	5.89	6.54
18	30	1	1	0.6	1	keNEATPSE_L	0.31	5.92	6.19
18	30	1	1	0.6	1	keNEATPSE_E	0.32	5.96	6.23
18	30	1	1	0.6	1	keNEATCE_L	0.39	5.92	6.61
18	30	1	1	0.6	1	keNEATCE_E	0.36	5.91	6.59
18	30	1	1	0.6	1	Linear	0.02	5.99	5.99
18	30	1	1	0.6	1	Equipercentile	0.02	5.98	5.98
99	30	1	1	1	1	Tucker	0.33	7.08	7.61

Table 4.2

Cont.

	Study Condition						Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
99	30	1	1	1	1	Levine True	0.52	6.77	8.16
99	30	1	1	1	1	Braun	0.37	7.05	7.59
99	30	1	1	1	1	FEEE	0.43	7.02	7.55
99	30	1	1	1	1	Chain_L	0.44	6.93	7.92
99	30	1	1	1	1	Chain_E	0.42	6.93	7.89
99	30	1	1	1	1	keNEATPSE_L	0.42	6.97	7.52
99	30	1	1	1	1	keNEATPSE_E	0.43	7.02	7.56
99	30	1	1	1	1	keNEATCE_L	0.44	6.93	7.92
99	30	1	1	1	1	keNEATCE_E	0.44	6.93	7.92
99	30	1	1	1	1	Linear	0.01	7.25	7.24
99	30	1	1	1	1	Equipercentile	0.01	7.24	7.24

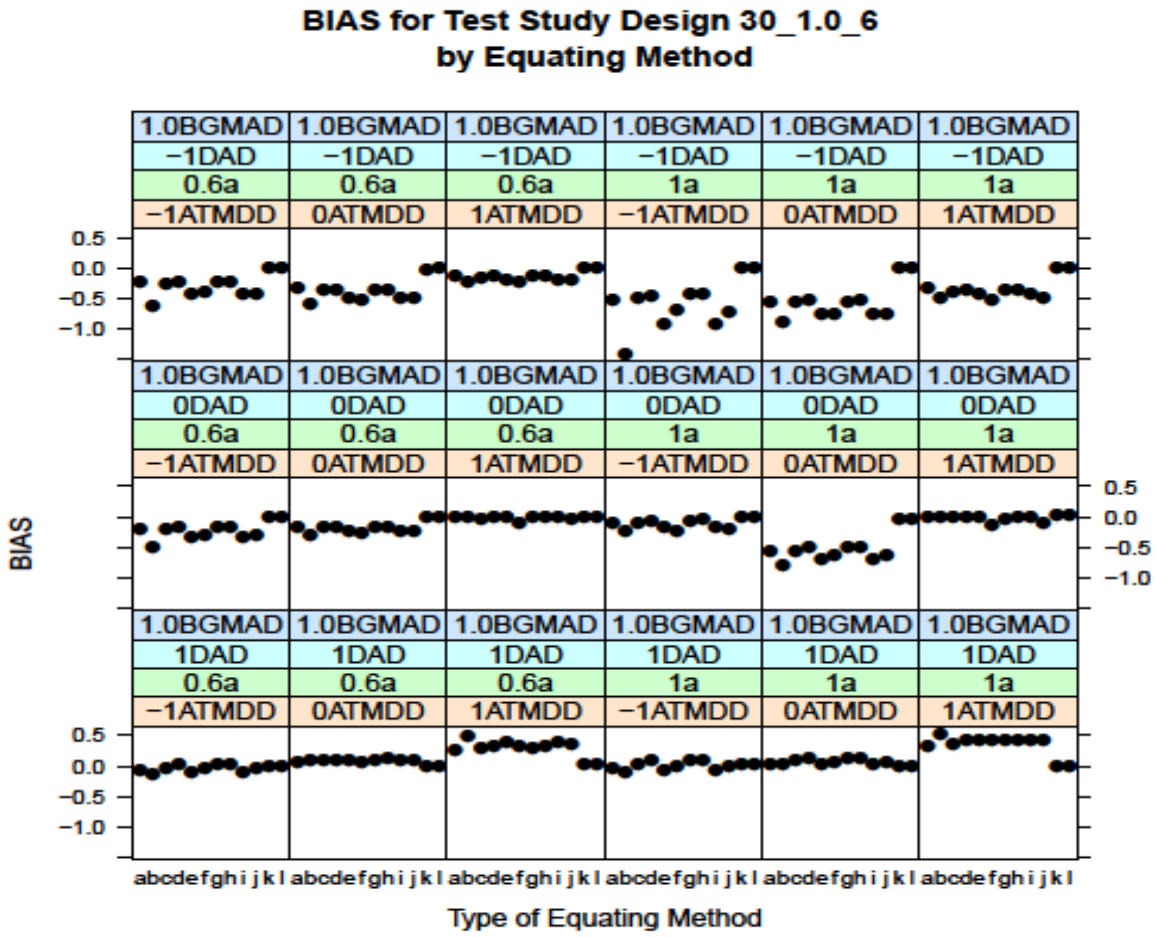

Figure 4.3. Bias for Test Study Design 30_1.0_6 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

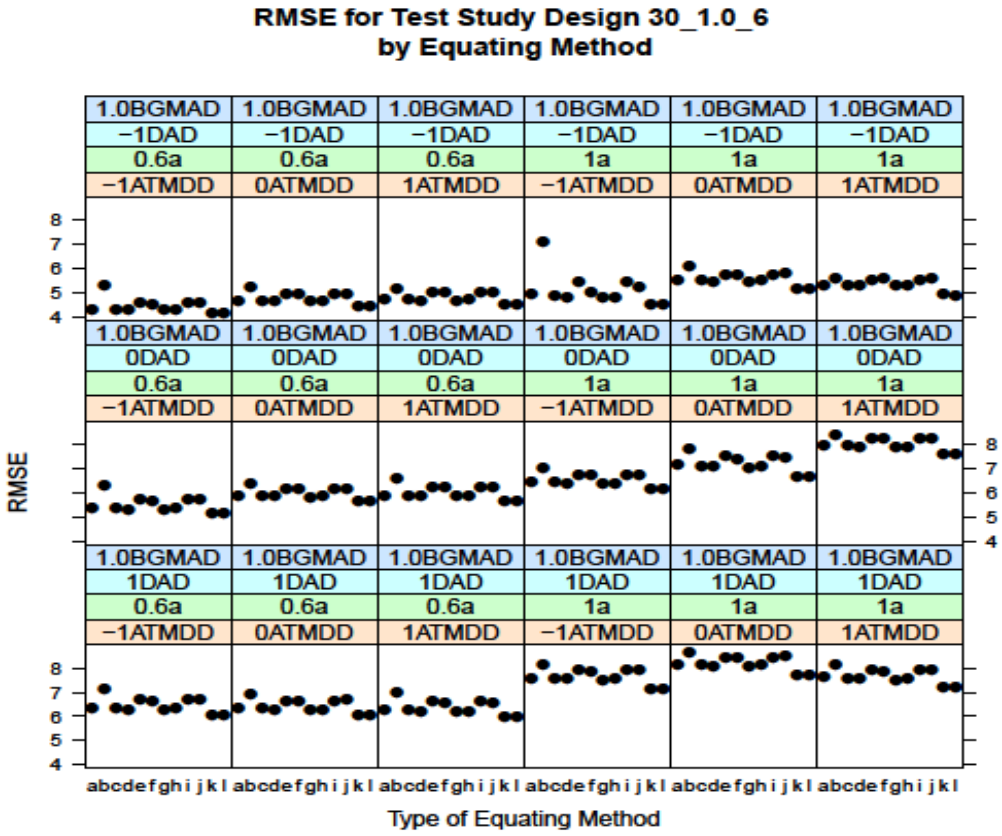

Figure 4.4. Root Mean Square Error (RMSE) for Test Study Design 30_1.0_6 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.2.3 30_1.5_6 Test Study Design

This subsection presents the results for test study design 30_1.5_6. This design had 30 total items and 6 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of group separation (or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large

respectively. The previous two subsections discussed small and medium BGMAD. This subsection mainly focuses on large BGMAD, which means 30 total items and 6 anchor test items were held constant and the other three study conditions were varied. Tables A.19-A.27 in Appendix A display average descriptive statistics and Figure B.3 in Appendix B shows the standard error of equating (SEE) for test study design 30_1.5_6. Table 4.3 represents bias, SEE, and RMSE for this test study design. Figure 4.5 demonstrates that there is negative bias for all conditions under all equating methods for large (1.5) between-grade mean ability difference (BGMAD) except for positive bias when BGMAD is large (1.5) and DAD is above average, item discrimination is high (1) and ATMDD is above average (1). Both negative and positive values of bias are very close to zero for a few study conditions. Noticeable in this regard is negative bias which is almost zero when BGMDD is large (1.5) while DAD is above average (1) and item discrimination is moderate (0.6) and ATMDD varied across its three levels. However, when the same conditions are repeated under high (1) item discrimination, only conditions under above average (1) ATMDD register positive bias which is very close to zero.

Figure 4.6 shows test study design 30_1.5_6, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (30), large (1.5) BGMAD, moderate (0.6) and high (1) item discriminations are invariant. The RMSE values fall between 5 and 10. When all conditions are held constant and manipulate item discrimination (moderate versus high), RMSE displays somewhat clear consistency for both moderate (0.6) and high (1) conditions. Also, the RMSE values for

both moderate (0.6) and high (1) discrimination when other conditions are varied performed similarly. Comparatively, conditions manipulated under moderate (0.6) item discrimination produced smaller (more accurate) RMSE values than its counterpart, high (1) item discrimination.

Addressing Research Question 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.5 revealed that difference between anchor test difficulty and the other four study conditions is smallest when BGMAD is large (1.5) for above average (1) DAD, moderate (0.6) item discrimination for below average (-1), average (0), and above average (1) ATMDD. Other study conditions produced worst results. There is sufficient evidence to believe that a large (1.5) BGMAD coupled with a short test (30 items) yield disparate bias results across all study conditions under all equating methods. This assertion is gleaned from the fact that when holding BGMAD large (1.5), item discrimination high (1), and ATMDD below average (-1) constant and vary DAD across its three levels, then the equating methods produce the largest bias compared with other study conditions.

Overall the equating results in Figure 4.6 show that there was a slight difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods. That is, all equating methods seemed to have a slightly distinguishable performance without any recognizable pattern. The worst performance was under large (1.5) BGMAD, above average (1) DAD, high (1) item discrimination where ATMDD is manipulated form below average (-1), average (0), and

above average (1) conditions. This means that other study conditions produced almost similar RMSE values under various equating methods.

At this juncture, it is worthwhile to note that the first three test study designs discussed thus far—30_0.5_6, 30_1.0_6, and 30_1.5_6—have the same number of total test items (30 items in total) and anchor test items (6 items) under all study conditions with variability in magnitude of the group separation (or BGMAD) across small (0.5), medium (1.0), and large (1.5). Contrasting these three test study designs—on the basis of magnitude of the group separation—(Figures 4.2, 4.4, and 4.6) in terms of RMSE values leads to the conclusion that the overall accuracy of the results is considerably affected by the degree of group effect (or mean ability difference between adjacent grades/BGMAD). Small (0.5) BGMAD produced more accurate results than medium (1.0) BGMAD and large (1.5) BGMAD; large (1.5) BGMAD has the largest RMSE values compared to the other two test study designs. Interestingly, small (0.5) BGMAD under all study conditions also produced the smallest bias and large (1.5) BGMAD the largest bias values.

Table 4.3

BIAS, SEE, and RMSE Statistics for Test Study Design 30_1.5_6 by Equating Method Under All Conditions

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
19	30	1.5	-1	0.6	-1	Tucker	-0.61	3.67	3.88
19	30	1.5	-1	0.6	-1	Levine True	-1.75	5.02	6.12
19	30	1.5	-1	0.6	-1	Braun	-0.61	3.73	3.93

Table 4.3

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
19	30	1.5	-1	0.6	-1	FEEE	-0.60	3.68	3.88
19	30	1.5	-1	0.6	-1	Chain_L	-1.13	3.96	4.55
19	30	1.5	-1	0.6	-1	Chain_E	-0.98	3.79	4.28
19	30	1.5	-1	0.6	-1	keNEATPSE_L	-0.59	3.66	3.86
19	30	1.5	-1	0.6	-1	keNEATPSE_E	-0.59	3.70	3.90
19	30	1.5	-1	0.6	-1	keNEATCE_L	-1.13	3.96	4.55
19	30	1.5	-1	0.6	-1	keNEATCE_E	-1.06	3.90	4.51
19	30	1.5	-1	0.6	-1	Linear	0.02	3.49	3.49
19	30	1.5	-1	0.6	-1	Equipercentile	0.02	3.49	3.49
100	30	1.5	-1	1	-1	Tucker	-0.89	3.93	4.19
100	30	1.5	-1	1	-1	Levine True	-2.40	6.02	7.04
100	30	1.5	-1	1	-1	Braun	-0.87	3.99	4.23
100	30	1.5	-1	1	-1	FEEE	-0.73	3.94	4.20
100	30	1.5	-1	1	-1	Chain_L	-1.55	4.53	5.12
100	30	1.5	-1	1	-1	Chain_E	-1.24	4.18	4.66
100	30	1.5	-1	1	-1	keNEATPSE_L	-0.82	3.85	4.09
100	30	1.5	-1	1	-1	keNEATPSE_E	-0.77	3.97	4.21
100	30	1.5	-1	1	-1	keNEATCE_L	-1.55	4.53	5.11
100	30	1.5	-1	1	-1	keNEATCE_E	-2.24	5.64	6.62
100	30	1.5	-1	1	-1	Linear	0.01	3.54	3.53
100	30	1.5	-1	1	-1	Equipercentile	0.01	3.53	3.53
20	30	1.5	-1	0.6	0	Tucker	-0.60	3.91	4.25
20	30	1.5	-1	0.6	0	Levine True	-1.34	4.46	5.71
20	30	1.5	-1	0.6	0	Braun	-0.63	4.00	4.32
20	30	1.5	-1	0.6	0	FEEE	-0.62	3.96	4.28
20	30	1.5	-1	0.6	0	Chain_L	-1.01	4.03	4.84
20	30	1.5	-1	0.6	0	Chain_E	-0.98	4.00	4.75
20	30	1.5	-1	0.6	0	keNEATPSE_L	-0.61	3.93	4.25
20	30	1.5	-1	0.6	0	keNEATPSE_E	-0.61	3.97	4.30
20	30	1.5	-1	0.6	0	keNEATCE_L	-1.01	4.03	4.84
20	30	1.5	-1	0.6	0	keNEATCE_E	-0.98	4.00	4.82

Table 4.3

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
20	30	1.5	-1	0.6	0	Linear	-0.01	3.82	3.82
20	30	1.5	-1	0.6	0	Equipercentile	-0.01	3.82	3.82
101	30	1.5	-1	1	0	Tucker	-0.97	3.79	4.36
101	30	1.5	-1	1	0	Levine True	-1.58	4.13	5.38
101	30	1.5	-1	1	0	Braun	-1.01	3.94	4.48
101	30	1.5	-1	1	0	FEEE	-0.97	3.90	4.44
101	30	1.5	-1	1	0	Chain_L	-1.33	3.86	4.83
101	30	1.5	-1	1	0	Chain_E	-1.35	3.96	4.86
101	30	1.5	-1	1	0	keNEATPSE_L	-0.97	3.90	4.44
101	30	1.5	-1	1	0	keNEATPSE_E	-0.96	3.93	4.47
101	30	1.5	-1	1	0	keNEATCE_L	-1.34	3.86	4.83
101	30	1.5	-1	1	0	keNEATCE_E	-1.39	4.00	5.00
101	30	1.5	-1	1	0	Linear	0.02	3.69	3.69
101	30	1.5	-1	1	0	Equipercentile	0.02	3.69	3.69
21	30	1.5	-1	0.6	1	Tucker	-0.28	4.07	4.51
21	30	1.5	-1	0.6	1	Levine True	-0.54	3.91	5.44
21	30	1.5	-1	0.6	1	Braun	-0.35	4.11	4.55
21	30	1.5	-1	0.6	1	FEEE	-0.32	4.06	4.49
21	30	1.5	-1	0.6	1	Chain_L	-0.44	4.02	5.05
21	30	1.5	-1	0.6	1	Chain_E	-0.54	4.07	5.05
21	30	1.5	-1	0.6	1	keNEATPSE_L	-0.32	4.03	4.47
21	30	1.5	-1	0.6	1	keNEATPSE_E	-0.31	4.07	4.51
21	30	1.5	-1	0.6	1	keNEATCE_L	-0.44	4.02	5.05
21	30	1.5	-1	0.6	1	keNEATCE_E	-0.51	4.05	5.09
21	30	1.5	-1	0.6	1	Linear	0.01	4.12	4.12
21	30	1.5	-1	0.6	1	Equipercentile	0.01	4.11	4.11
102	30	1.5	-1	1	1	Tucker	-0.22	3.62	4.18
102	30	1.5	-1	1	1	Levine True	-0.36	3.57	4.99
102	30	1.5	-1	1	1	Braun	-0.36	3.75	4.32
102	30	1.5	-1	1	1	FEEE	-0.27	3.69	4.30
102	30	1.5	-1	1	1	Chain_L	-0.31	3.61	4.69

Table 4.3

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
102	30	1.5	-1	1	1	Chain_E	-0.54	3.80	4.84
102	30	1.5	-1	1	1	keNEATPSE_L	-0.31	3.70	4.27
102	30	1.5	-1	1	1	keNEATPSE_E	-0.30	3.72	4.29
102	30	1.5	-1	1	1	keNEATCE_L	-0.31	3.61	4.69
102	30	1.5	-1	1	1	keNEATCE_E	-0.50	3.79	4.88
102	30	1.5	-1	1	1	Linear	0.01	3.54	3.54
102	30	1.5	-1	1	1	Equipercentile	0.01	3.55	3.55
22	30	1.5	0	0.6	-1	Tucker	-0.40	5.00	5.33
22	30	1.5	0	0.6	-1	Levine True	-0.92	5.35	7.13
22	30	1.5	0	0.6	-1	Braun	-0.43	5.02	5.34
22	30	1.5	0	0.6	-1	FEEE	-0.38	4.96	5.27
22	30	1.5	0	0.6	-1	Chain_L	-0.65	5.00	5.96
22	30	1.5	0	0.6	-1	Chain_E	-0.64	4.96	5.84
22	30	1.5	0	0.6	-1	keNEATPSE_L	-0.37	4.91	5.23
22	30	1.5	0	0.6	-1	keNEATPSE_E	-0.37	4.98	5.29
22	30	1.5	0	0.6	-1	keNEATCE_L	-0.65	5.00	5.96
22	30	1.5	0	0.6	-1	keNEATCE_E	-0.63	4.95	5.92
22	30	1.5	0	0.6	-1	Linear	-0.01	5.01	5.00
22	30	1.5	0	0.6	-1	Equipercentile	-0.01	5.00	5.00
103	30	1.5	0	1	-1	Tucker	-0.73	5.04	5.65
103	30	1.5	0	1	-1	Levine True	-1.39	5.51	7.35
103	30	1.5	0	1	-1	Braun	-0.82	5.14	5.75
103	30	1.5	0	1	-1	FEEE	-0.72	5.08	5.68
103	30	1.5	0	1	-1	Chain_L	-1.08	5.09	6.33
103	30	1.5	0	1	-1	Chain_E	-1.11	5.11	6.25
103	30	1.5	0	1	-1	keNEATPSE_L	-0.71	5.02	5.64
103	30	1.5	0	1	-1	keNEATPSE_E	-0.71	5.10	5.70
103	30	1.5	0	1	-1	keNEATCE_L	-1.08	5.09	6.33
103	30	1.5	0	1	-1	keNEATCE_E	-1.12	5.10	6.37
103	30	1.5	0	1	-1	Linear	0.00	5.02	5.02
103	30	1.5	0	1	-1	Equipercentile	0.00	5.02	5.02

Table 4.3

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
23	30	1.5	0	0.6	0	Tucker	-0.68	4.94	5.54
23	30	1.5	0	0.6	0	Levine True	-1.25	5.10	6.94
23	30	1.5	0	0.6	0	Braun	-0.71	4.99	5.58
23	30	1.5	0	0.6	0	FEEE	-0.69	4.95	5.53
23	30	1.5	0	0.6	0	Chain_L	-1.01	4.86	6.16
23	30	1.5	0	0.6	0	Chain_E	-1.00	4.90	6.14
23	30	1.5	0	0.6	0	keNEATPSE_L	-0.68	4.91	5.50
23	30	1.5	0	0.6	0	keNEATPSE_E	-0.68	4.96	5.54
23	30	1.5	0	0.6	0	keNEATCE_L	-1.01	4.86	6.16
23	30	1.5	0	0.6	0	keNEATCE_E	-0.98	4.88	6.16
23	30	1.5	0	0.6	0	Linear	-0.01	5.03	5.03
23	30	1.5	0	0.6	0	Equipercentile	-0.01	5.03	5.03
104	30	1.5	0	1	0	Tucker	-1.00	5.10	6.07
104	30	1.5	0	1	0	Levine True	-1.50	5.15	7.09
104	30	1.5	0	1	0	Braun	-1.05	5.20	6.16
104	30	1.5	0	1	0	FEEE	-1.00	5.16	6.11
104	30	1.5	0	1	0	Chain_L	-1.32	5.00	6.58
104	30	1.5	0	1	0	Chain_E	-1.35	5.11	6.61
104	30	1.5	0	1	0	keNEATPSE_L	-0.99	5.11	6.08
104	30	1.5	0	1	0	keNEATPSE_E	-0.99	5.16	6.12
104	30	1.5	0	1	0	keNEATCE_L	-1.32	5.00	6.58
104	30	1.5	0	1	0	keNEATCE_E	-1.34	5.10	6.66
104	30	1.5	0	1	0	Linear	0.00	5.28	5.28
104	30	1.5	0	1	0	Equipercentile	0.00	5.27	5.27
24	30	1.5	0	0.6	1	Tucker	-0.32	5.17	5.76
24	30	1.5	0	0.6	1	Levine True	-0.57	4.79	6.77
24	30	1.5	0	0.6	1	Braun	-0.36	5.17	5.77
24	30	1.5	0	0.6	1	FEEE	-0.32	5.11	5.70
24	30	1.5	0	0.6	1	Chain_L	-0.47	5.00	6.34
24	30	1.5	0	0.6	1	Chain_E	-0.53	5.03	6.33
24	30	1.5	0	0.6	1	keNEATPSE_L	-0.32	5.06	5.67

Table 4.3

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
24	30	1.5	0	0.6	1	keNEATPSE_E	-0.32	5.12	5.72
24	30	1.5	0	0.6	1	keNEATCE_L	-0.47	5.00	6.34
24	30	1.5	0	0.6	1	keNEATCE_E	-0.49	5.02	6.37
24	30	1.5	0	0.6	1	Linear	-0.01	5.30	5.30
24	30	1.5	0	0.6	1	Equipercentile	0.00	5.30	5.30
105	30	1.5	0	1	1	Tucker	-0.62	5.57	6.77
105	30	1.5	0	1	1	Levine True	-0.88	5.23	7.51
105	30	1.5	0	1	1	Braun	-0.67	5.61	6.79
105	30	1.5	0	1	1	FEEE	-0.62	5.55	6.74
105	30	1.5	0	1	1	Chain_L	-0.79	5.38	7.26
105	30	1.5	0	1	1	Chain_E	-0.88	5.50	7.32
105	30	1.5	0	1	1	keNEATPSE_L	-0.62	5.52	6.72
105	30	1.5	0	1	1	keNEATPSE_E	-0.61	5.56	6.75
105	30	1.5	0	1	1	keNEATCE_L	-0.79	5.38	7.26
105	30	1.5	0	1	1	keNEATCE_E	-0.83	5.46	7.34
105	30	1.5	0	1	1	Linear	0.01	5.94	5.94
105	30	1.5	0	1	1	Equipercentile	0.01	5.94	5.94
25	30	1.5	1	0.6	-1	Tucker	-0.15	5.57	5.98
25	30	1.5	1	0.6	-1	Levine True	-0.52	5.23	7.55
25	30	1.5	1	0.6	-1	Braun	-0.15	5.54	5.95
25	30	1.5	1	0.6	-1	FEEE	-0.06	5.48	5.88
25	30	1.5	1	0.6	-1	Chain_L	-0.33	5.48	6.65
25	30	1.5	1	0.6	-1	Chain_E	-0.26	5.41	6.51
25	30	1.5	1	0.6	-1	keNEATPSE_L	-0.05	5.43	5.83
25	30	1.5	1	0.6	-1	keNEATPSE_E	-0.06	5.49	5.89
25	30	1.5	1	0.6	-1	keNEATCE_L	-0.33	5.48	6.65
25	30	1.5	1	0.6	-1	keNEATCE_E	-0.26	5.43	6.59
25	30	1.5	1	0.6	-1	Linear	0.03	5.70	5.69
25	30	1.5	1	0.6	-1	Equipercentile	0.03	5.69	5.69
106	30	1.5	1	1	-1	Tucker	-1.45	6.45	7.48
106	30	1.5	1	1	-1	Levine True	-2.36	6.94	9.40

Table 4.3

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
106	30	1.5	1	1	-1	Braun	-1.33	6.32	7.35
106	30	1.5	1	1	-1	FEEE	-1.20	6.25	7.26
106	30	1.5	1	1	-1	Chain_L	-1.96	6.40	8.24
106	30	1.5	1	1	-1	Chain_E	-1.56	6.10	7.79
106	30	1.5	1	1	-1	keNEATPSE_L	-1.17	6.14	7.17
106	30	1.5	1	1	-1	keNEATPSE_E	-1.20	6.27	7.27
106	30	1.5	1	1	-1	keNEATCE_L	-1.96	6.40	8.24
106	30	1.5	1	1	-1	keNEATCE_E	-1.63	6.14	7.88
106	30	1.5	1	1	-1	Linear	-0.01	6.51	6.51
106	30	1.5	1	1	-1	Equipercentile	-0.01	6.51	6.51
26	30	1.5	1	0.6	0	Tucker	-0.13	5.64	6.18
26	30	1.5	1	0.6	0	Levine True	-0.28	5.20	7.50
26	30	1.5	1	0.6	0	Braun	-0.19	5.63	6.18
26	30	1.5	1	0.6	0	FEEE	-0.10	5.57	6.11
26	30	1.5	1	0.6	0	Chain_L	-0.21	5.46	6.83
26	30	1.5	1	0.6	0	Chain_E	-0.31	5.50	6.80
26	30	1.5	1	0.6	0	keNEATPSE_L	-0.10	5.51	6.06
26	30	1.5	1	0.6	0	keNEATPSE_E	-0.10	5.58	6.12
26	30	1.5	1	0.6	0	keNEATCE_L	-0.21	5.46	6.83
26	30	1.5	1	0.6	0	keNEATCE_E	-0.25	5.50	6.86
26	30	1.5	1	0.6	0	Linear	-0.01	5.77	5.77
26	30	1.5	1	0.6	0	Equipercentile	-0.01	5.77	5.76
107	30	1.5	1	1	0	Tucker	-0.33	6.73	8.01
107	30	1.5	1	1	0	Levine True	-0.56	6.24	9.05
107	30	1.5	1	1	0	Braun	-0.28	6.66	7.96
107	30	1.5	1	1	0	FEEE	-0.18	6.61	7.89
107	30	1.5	1	1	0	Chain_L	-0.47	6.48	8.63
107	30	1.5	1	1	0	Chain_E	-0.43	6.47	8.55
107	30	1.5	1	1	0	keNEATPSE_L	-0.19	6.50	7.81
107	30	1.5	1	1	0	keNEATPSE_E	-0.18	6.62	7.90
107	30	1.5	1	1	0	keNEATCE_L	-0.47	6.48	8.63

Table 4.3

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
107	30	1.5	1	1	0	keNEATCE_E	-0.41	6.50	8.65
107	30	1.5	1	1	0	Linear	0.00	7.06	7.06
107	30	1.5	1	1	0	Equipercentile	0.00	7.06	7.06
27	30	1.5	1	0.6	1	Tucker	-0.17	5.76	6.30
27	30	1.5	1	0.6	1	Levine True	-0.41	5.26	7.47
27	30	1.5	1	0.6	1	Braun	-0.21	5.74	6.29
27	30	1.5	1	0.6	1	FEEE	-0.12	5.68	6.22
27	30	1.5	1	0.6	1	Chain_L	-0.30	5.55	6.90
27	30	1.5	1	0.6	1	Chain_E	-0.31	5.57	6.88
27	30	1.5	1	0.6	1	keNEATPSE_L	-0.11	5.62	6.17
27	30	1.5	1	0.6	1	keNEATPSE_E	-0.12	5.69	6.23
27	30	1.5	1	0.6	1	keNEATCE_L	-0.30	5.55	6.90
27	30	1.5	1	0.6	1	keNEATCE_E	-0.29	5.57	6.92
27	30	1.5	1	0.6	1	Linear	0.00	5.93	5.93
27	30	1.5	1	0.6	1	Equipercentile	0.00	5.92	5.92
108	30	1.5	1	1	1	Tucker	0.24	6.56	7.74
108	30	1.5	1	1	1	Levine True	0.29	5.98	8.67
108	30	1.5	1	1	1	Braun	0.23	6.52	7.70
108	30	1.5	1	1	1	FEEE	0.33	6.44	7.62
108	30	1.5	1	1	1	Chain_L	0.27	6.25	8.30
108	30	1.5	1	1	1	Chain_E	0.11	6.34	8.31
108	30	1.5	1	1	1	keNEATPSE_L	0.31	6.39	7.59
108	30	1.5	1	1	1	keNEATPSE_E	0.33	6.45	7.63
108	30	1.5	1	1	1	keNEATCE_L	0.27	6.25	8.30
108	30	1.5	1	1	1	keNEATCE_E	0.20	6.31	8.34
108	30	1.5	1	1	1	Linear	-0.01	6.98	6.98
108	30	1.5	1	1	1	Equipercentile	-0.01	6.98	6.98

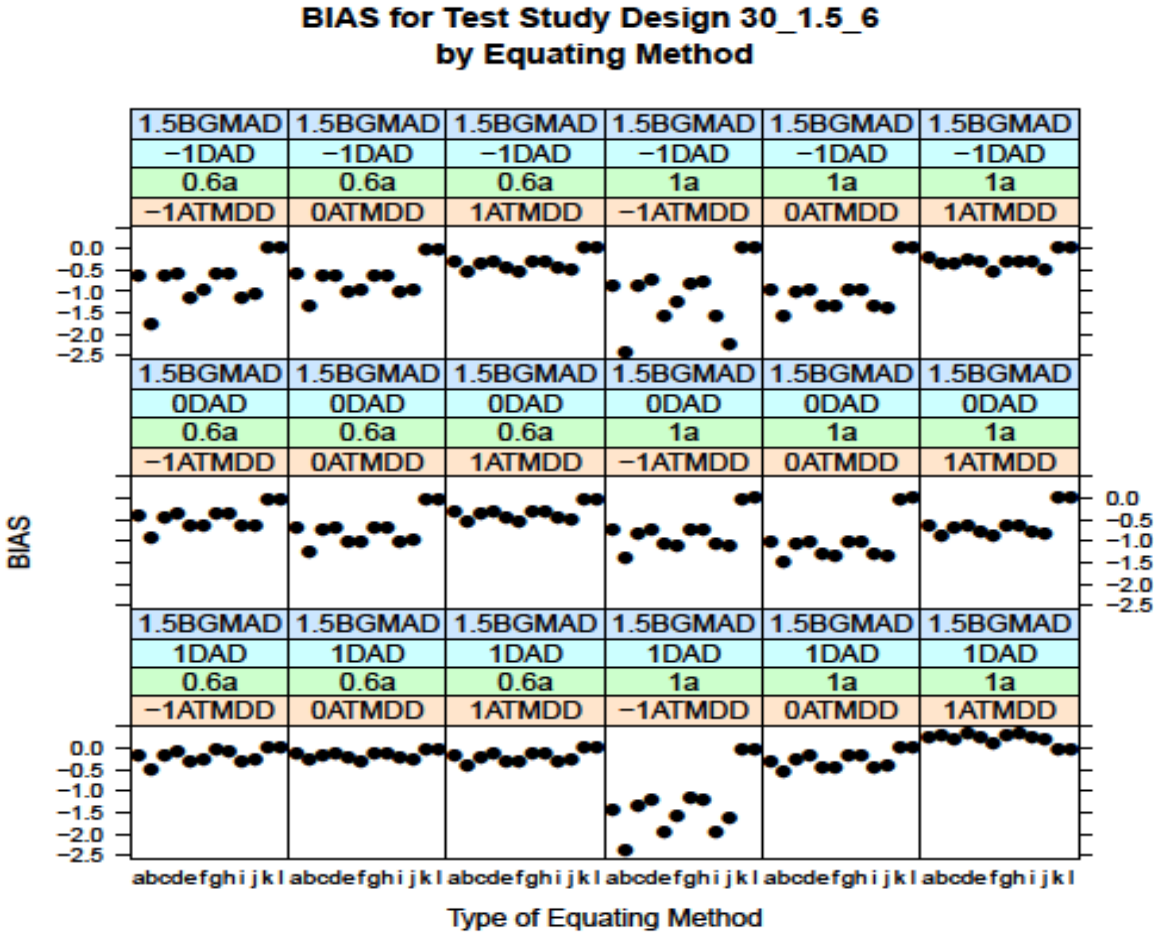

Figure 4.5. Bias for Test Study Design 30_1.5_6 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

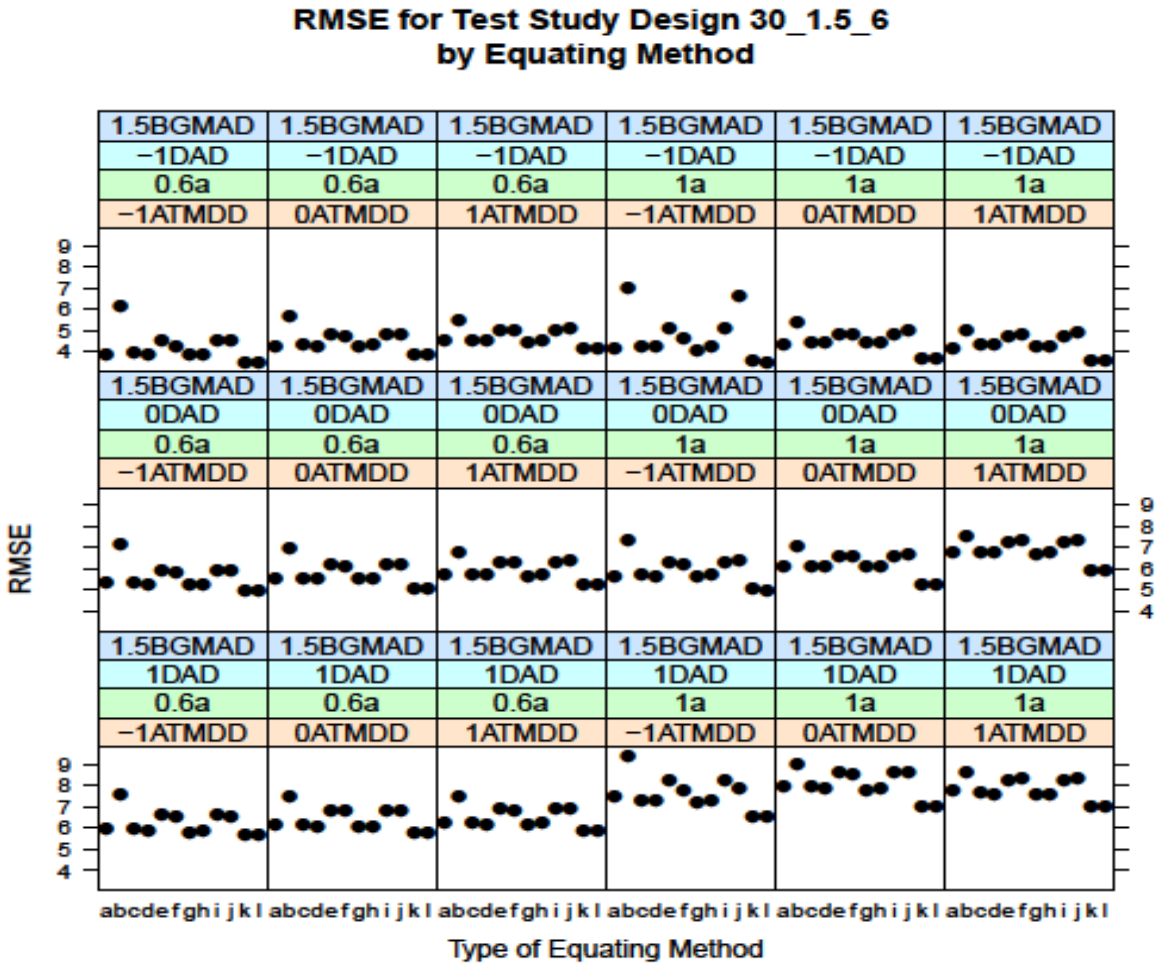

Figure 4.6. Root Mean Square Error (RMSE) for Test Study Design 30_1.5_6 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.2.4 60_0.5_12 Test Study Design

In this subsection, the results for test study design 60_0.5_12 are presented. This design had 60 total items and 12 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of

group separation (or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large respectively. Medium and large BGMAD will be discussed in the subsequent subsections. This subsection mainly focuses on small BGMAD, which means 60 total items and 12 anchor test items were held constant and the other three study conditions were varied. Tables A28-A36 in Appendix A display average descriptive statistics and Figure B.4 in Appendix B shows the standard error of equating (SEE) for test study design 60_0.5_12. Table 4.4 represents bias, SEE, and RMSE for this test study design. Figure 4.7 demonstrates that there is almost zero bias for all conditions under all equating methods for small (0.5) between-grade mean ability difference (BGMAD). Both negative and positive values of bias are almost close to zero except for a few study conditions where results are rather stable. For instance, when distribution of ability difference (DAD) is below (-1) average and average (0) and anchor test mean difficulty difference (ATMDD) is average (0) and above (1) average and item discrimination is moderate (0.6) the equating methods show consistency. This pattern of consistency is also repeated when item discrimination is high (1) where DAD is average (0) and above average (1) and especially where ATMDD is average (0). Majority of the study conditions produced inconsistent results and more so when item discrimination is high (1).

Figure 4.8 shows test study design 60_0.5_12, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (60), small (0.5) BGMAD, moderate (0.6) and high (1) item discriminations are held constant. The RMSE values fall between 9 and 15. Interestingly, when all conditions are held

constant and manipulate item discrimination (moderate versus high), RMSE displays a clear consistency when item discrimination is moderate (0.6). Also, the RMSE values for moderate (0.6) discrimination when other conditions are varied are lower (more accurate) than RMSE values when item discrimination is high (1). One trend that stood out in this design was that all equating methods consistently produced almost similar values of RMSE when magnitude of group separation or BGMAD was considerably small (0.5) and *b*-item parameter for a grade was the same as the mean ability for that grade [or DAD was average (0)].

Addressing the research question number 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.7 revealed three results. First, the bias was consistent and very close to zero for all equating methods when small (0.5) BGMAD and moderate (0.6) item discrimination were held constant and DAD varied across below (-1), average (0), and above average (1) and when ATMDD was average (0) and above average (1). Second, the equating results were inaccurate and underestimated accuracy for all equating methods, as evidenced by negative bias, under small (0.5) BGMAD where DAD was below average (-1) and average (0) for both moderate (0.6) and high (1) item discrimination and ATMDD was below average (-1) and average (0). Third, when small (0.5) BGMAD, high (1) item discrimination, and above average (1) ATMDD were held constant and manipulated DAD from below average (-1), average (0), and above average (1) the bias overestimated, suggested by positive bias values, the accuracy of the equating results for all equating methods.

Overall the equating results in Figure 4.8 show that there was a slight significant difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods. That is, all equating methods seemed to have a slightly indistinguishable performance without any discernible pattern apart from slight differences where item discrimination was moderate (0.6) and high (1) for all conditions. Comparatively, though, moderate (0.6) item discrimination produced rather more accurate overall results than high (1) item discrimination under all conditions.

Table 4.4

BIAS, SEE, and RMSE Statistics for Test Study Design 60_0.5_12 by Equating Method Under All Conditions

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
28	60	0.5	-1	0.6	-1	Tucker	-0.18	9.35	9.51
28	60	0.5	-1	0.6	-1	Levine True	-0.35	9.35	9.85
28	60	0.5	-1	0.6	-1	Braun	-0.17	9.35	9.50
28	60	0.5	-1	0.6	-1	FEEE	-0.17	9.34	9.50
28	60	0.5	-1	0.6	-1	Chain_L	-0.27	9.37	9.69
28	60	0.5	-1	0.6	-1	Chain_E	-0.25	9.33	9.64
28	60	0.5	-1	0.6	-1	keNEATPSE_L	-0.17	9.32	9.48
28	60	0.5	-1	0.6	-1	keNEATPSE_E	-0.17	9.35	9.51
28	60	0.5	-1	0.6	-1	keNEATCE_L	-0.27	9.37	9.69
28	60	0.5	-1	0.6	-1	keNEATCE_E	-0.24	9.34	9.65
28	60	0.5	-1	0.6	-1	Linear	0.02	9.28	9.28
28	60	0.5	-1	0.6	-1	Equipercentile	0.02	9.28	9.28
109	60	0.5	-1	1	-1	Tucker	-0.45	9.79	10.00
109	60	0.5	-1	1	-1	Levine True	-0.64	9.84	10.29
109	60	0.5	-1	1	-1	Braun	-0.43	9.77	9.97
109	60	0.5	-1	1	-1	FEEE	-0.42	9.76	9.97

Table 4.4

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
109	60	0.5	-1	1	-1	Chain_L	-0.55	9.83	10.16
109	60	0.5	-1	1	-1	Chain_E	-0.50	9.76	10.07
109	60	0.5	-1	1	-1	keNEATPSE_L	-0.42	9.74	9.95
109	60	0.5	-1	1	-1	keNEATPSE_E	-0.42	9.77	9.98
109	60	0.5	-1	1	-1	keNEATCE_L	-0.55	9.83	10.16
109	60	0.5	-1	1	-1	keNEATCE_E	-0.50	9.78	10.10
109	60	0.5	-1	1	-1	Linear	0.00	9.81	9.81
109	60	0.5	-1	1	-1	Equipercentile	0.00	9.80	9.80
29	60	0.5	-1	0.6	0	Tucker	-0.01	10.07	10.32
29	60	0.5	-1	0.6	0	Levine True	-0.07	10.02	10.59
29	60	0.5	-1	0.6	0	Braun	-0.02	10.07	10.32
29	60	0.5	-1	0.6	0	FEEE	-0.01	10.06	10.31
29	60	0.5	-1	0.6	0	Chain_L	-0.04	10.05	10.47
29	60	0.5	-1	0.6	0	Chain_E	-0.06	10.05	10.46
29	60	0.5	-1	0.6	0	keNEATPSE_L	-0.02	10.06	10.30
29	60	0.5	-1	0.6	0	keNEATPSE_E	-0.01	10.08	10.32
29	60	0.5	-1	0.6	0	keNEATCE_L	-0.04	10.05	10.47
29	60	0.5	-1	0.6	0	keNEATCE_E	-0.04	10.06	10.48
29	60	0.5	-1	0.6	0	Linear	-0.01	10.02	10.02
29	60	0.5	-1	0.6	0	Equipercentile	-0.01	10.02	10.02
110	60	0.5	-1	1	0	Tucker	-0.20	11.74	12.11
110	60	0.5	-1	1	0	Levine True	-0.28	11.70	12.34
110	60	0.5	-1	1	0	Braun	-0.23	11.76	12.12
110	60	0.5	-1	1	0	FEEE	-0.22	11.75	12.11
110	60	0.5	-1	1	0	Chain_L	-0.25	11.72	12.24
110	60	0.5	-1	1	0	Chain_E	-0.28	11.75	12.25
110	60	0.5	-1	1	0	keNEATPSE_L	-0.22	11.73	12.09
110	60	0.5	-1	1	0	keNEATPSE_E	-0.22	11.75	12.12
110	60	0.5	-1	1	0	keNEATCE_L	-0.25	11.72	12.24
110	60	0.5	-1	1	0	keNEATCE_E	-0.27	11.75	12.27
110	60	0.5	-1	1	0	Linear	0.00	11.75	11.74
110	60	0.5	-1	1	0	Equipercentile	0.00	11.74	11.74
30	60	0.5	-1	0.6	1	Tucker	-0.02	10.49	10.77

Table 4.4

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
30	60	0.5	-1	0.6	1	Levine True	-0.04	10.43	11.02
30	60	0.5	-1	0.6	1	Braun	-0.04	10.49	10.77
30	60	0.5	-1	0.6	1	FEEE	-0.04	10.48	10.76
30	60	0.5	-1	0.6	1	Chain_L	-0.04	10.46	10.92
30	60	0.5	-1	0.6	1	Chain_E	-0.08	10.48	10.93
30	60	0.5	-1	0.6	1	keNEATPSE_L	-0.04	10.47	10.75
30	60	0.5	-1	0.6	1	keNEATPSE_E	-0.04	10.49	10.77
30	60	0.5	-1	0.6	1	keNEATCE_L	-0.04	10.46	10.92
30	60	0.5	-1	0.6	1	keNEATCE_E	-0.06	10.48	10.95
30	60	0.5	-1	0.6	1	Linear	-0.02	10.57	10.56
30	60	0.5	-1	0.6	1	Equipercentile	-0.02	10.56	10.56
111	60	0.5	-1	1	1	Tucker	-0.03	11.93	12.35
111	60	0.5	-1	1	1	Levine True	-0.06	11.88	12.57
111	60	0.5	-1	1	1	Braun	-0.10	11.94	12.35
111	60	0.5	-1	1	1	FEEE	-0.09	11.93	12.35
111	60	0.5	-1	1	1	Chain_L	-0.05	11.90	12.49
111	60	0.5	-1	1	1	Chain_E	-0.14	11.93	12.50
111	60	0.5	-1	1	1	keNEATPSE_L	-0.10	11.90	12.32
111	60	0.5	-1	1	1	keNEATPSE_E	-0.08	11.96	12.38
111	60	0.5	-1	1	1	keNEATCE_L	-0.05	11.90	12.49
111	60	0.5	-1	1	1	keNEATCE_E	-0.12	11.96	12.54
111	60	0.5	-1	1	1	Linear	-0.01	12.05	12.05
111	60	0.5	-1	1	1	Equipercentile	-0.01	12.05	12.05
31	60	0.5	0	0.6	-1	Tucker	-0.13	11.58	11.79
31	60	0.5	0	0.6	-1	Levine True	-0.28	11.54	12.17
31	60	0.5	0	0.6	-1	Braun	-0.12	11.55	11.76
31	60	0.5	0	0.6	-1	FEEE	-0.10	11.54	11.75
31	60	0.5	0	0.6	-1	Chain_L	-0.21	11.57	11.98
31	60	0.5	0	0.6	-1	Chain_E	-0.15	11.52	11.93
31	60	0.5	0	0.6	-1	keNEATPSE_L	-0.10	11.52	11.73
31	60	0.5	0	0.6	-1	keNEATPSE_E	-0.10	11.55	11.76
31	60	0.5	0	0.6	-1	keNEATCE_L	-0.21	11.57	11.98
31	60	0.5	0	0.6	-1	keNEATCE_E	-0.14	11.53	11.94

Table 4.4

Cont.

Panel	Study Condition: Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	Statistic: BIAS	SEE	RMSE
31	60	0.5	0	0.6	-1	Linear	-0.01	11.62	11.62
31	60	0.5	0	0.6	-1	Equipercentile	-0.01	11.62	11.62
112	60	0.5	0	1	-1	Tucker	-0.15	13.34	13.73
112	60	0.5	0	1	-1	Levine True	-0.29	13.29	14.04
112	60	0.5	0	1	-1	Braun	-0.06	13.26	13.67
112	60	0.5	0	1	-1	FEEE	-0.05	13.25	13.66
112	60	0.5	0	1	-1	Chain_L	-0.23	13.32	13.90
112	60	0.5	0	1	-1	Chain_E	-0.08	13.21	13.79
112	60	0.5	0	1	-1	keNEATPSE_L	-0.05	13.23	13.64
112	60	0.5	0	1	-1	keNEATPSE_E	-0.05	13.26	13.67
112	60	0.5	0	1	-1	keNEATCE_L	-0.23	13.32	13.90
112	60	0.5	0	1	-1	keNEATCE_E	-0.08	13.22	13.81
112	60	0.5	0	1	-1	Linear	0.02	13.37	13.37
112	60	0.5	0	1	-1	Equipercentile	0.02	13.37	13.37
32	60	0.5	0	0.6	0	Tucker	-0.03	11.74	12.04
32	60	0.5	0	0.6	0	Levine True	-0.04	11.65	12.32
32	60	0.5	0	0.6	0	Braun	-0.02	11.73	12.03
32	60	0.5	0	0.6	0	FEEE	-0.01	11.72	12.02
32	60	0.5	0	0.6	0	Chain_L	-0.03	11.69	12.20
32	60	0.5	0	0.6	0	Chain_E	-0.03	11.69	12.19
32	60	0.5	0	0.6	0	keNEATPSE_L	-0.01	11.70	12.00
32	60	0.5	0	0.6	0	keNEATPSE_E	-0.01	11.72	12.03
32	60	0.5	0	0.6	0	keNEATCE_L	-0.03	11.69	12.20
32	60	0.5	0	0.6	0	keNEATCE_E	-0.02	11.70	12.21
32	60	0.5	0	0.6	0	Linear	0.00	11.78	11.78
32	60	0.5	0	0.6	0	Equipercentile	0.00	11.77	11.77
113	60	0.5	0	1	0	Tucker	-0.10	13.65	14.08
113	60	0.5	0	1	0	Levine True	-0.13	13.58	14.35
113	60	0.5	0	1	0	Braun	-0.13	13.65	14.07
113	60	0.5	0	1	0	FEEE	-0.12	13.63	14.06
113	60	0.5	0	1	0	Chain_L	-0.12	13.61	14.23
113	60	0.5	0	1	0	Chain_E	-0.18	13.63	14.23
113	60	0.5	0	1	0	keNEATPSE_L	-0.12	13.61	14.04

Table 4.4

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
113	60	0.5	0	1	0	keNEATPSE_E	-0.12	13.64	14.07
113	60	0.5	0	1	0	keNEATCE_L	-0.12	13.61	14.23
113	60	0.5	0	1	0	keNEATCE_E	-0.16	13.64	14.26
113	60	0.5	0	1	0	Linear	0.00	13.67	13.67
113	60	0.5	0	1	0	Equipercentile	0.00	13.67	13.66
33	60	0.5	0	0.6	1	Tucker	-0.03	11.40	11.65
33	60	0.5	0	0.6	1	Levine True	0.01	11.34	11.95
33	60	0.5	0	0.6	1	Braun	-0.04	11.40	11.65
33	60	0.5	0	0.6	1	FEEE	-0.03	11.39	11.63
33	60	0.5	0	0.6	1	Chain_L	-0.01	11.37	11.82
33	60	0.5	0	0.6	1	Chain_E	-0.06	11.36	11.80
33	60	0.5	0	0.6	1	keNEATPSE_L	-0.03	11.37	11.62
33	60	0.5	0	0.6	1	keNEATPSE_E	-0.03	11.40	11.64
33	60	0.5	0	0.6	1	keNEATCE_L	-0.01	11.37	11.82
33	60	0.5	0	0.6	1	keNEATCE_E	-0.04	11.38	11.82
33	60	0.5	0	0.6	1	Linear	-0.01	11.38	11.37
33	60	0.5	0	0.6	1	Equipercentile	-0.01	11.37	11.37
114	60	0.5	0	1	1	Tucker	0.23	13.02	13.39
114	60	0.5	0	1	1	Levine True	0.27	12.92	13.67
114	60	0.5	0	1	1	Braun	0.20	13.00	13.37
114	60	0.5	0	1	1	FEEE	0.22	12.98	13.35
114	60	0.5	0	1	1	Chain_L	0.25	12.97	13.54
114	60	0.5	0	1	1	Chain_E	0.17	12.95	13.51
114	60	0.5	0	1	1	keNEATPSE_L	0.21	12.96	13.33
114	60	0.5	0	1	1	keNEATPSE_E	0.22	13.00	13.37
114	60	0.5	0	1	1	keNEATCE_L	0.25	12.97	13.54
114	60	0.5	0	1	1	keNEATCE_E	0.19	12.97	13.54
114	60	0.5	0	1	1	Linear	0.03	13.10	13.10
114	60	0.5	0	1	1	Equipercentile	0.03	13.10	13.09
34	60	0.5	1	0.6	-1	Tucker	-0.17	11.77	11.95
34	60	0.5	1	0.6	-1	Levine True	-0.24	11.73	12.37
34	60	0.5	1	0.6	-1	Braun	-0.15	11.75	11.93
34	60	0.5	1	0.6	-1	FEEE	-0.12	11.74	11.92

Table 4.4

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
34	60	0.5	1	0.6	-1	Chain_L	-0.21	11.75	12.14
34	60	0.5	1	0.6	-1	Chain_E	-0.13	11.72	12.10
34	60	0.5	1	0.6	-1	keNEATPSE_L	-0.11	11.71	11.89
34	60	0.5	1	0.6	-1	keNEATPSE_E	-0.12	11.74	11.92
34	60	0.5	1	0.6	-1	keNEATCE_L	-0.21	11.75	12.14
34	60	0.5	1	0.6	-1	keNEATCE_E	-0.14	11.75	12.13
34	60	0.5	1	0.6	-1	Linear	-0.03	11.77	11.77
34	60	0.5	1	0.6	-1	Equipercentile	-0.03	11.77	11.77
115	60	0.5	1	1	-1	Tucker	0.06	13.61	14.02
115	60	0.5	1	1	-1	Levine True	-0.02	13.54	14.36
115	60	0.5	1	1	-1	Braun	0.16	13.56	13.97
115	60	0.5	1	1	-1	FEEE	0.18	13.54	13.96
115	60	0.5	1	1	-1	Chain_L	0.01	13.58	14.20
115	60	0.5	1	1	-1	Chain_E	0.17	13.51	14.13
115	60	0.5	1	1	-1	keNEATPSE_L	0.18	13.51	13.93
115	60	0.5	1	1	-1	keNEATPSE_E	0.18	13.56	13.98
115	60	0.5	1	1	-1	keNEATCE_L	0.01	13.58	14.20
115	60	0.5	1	1	-1	keNEATCE_E	0.16	13.54	14.16
115	60	0.5	1	1	-1	Linear	-0.01	13.70	13.70
115	60	0.5	1	1	-1	Equipercentile	-0.01	13.70	13.70
35	60	0.5	1	0.6	0	Tucker	0.10	12.11	12.40
35	60	0.5	1	0.6	0	Levine True	0.11	12.04	12.74
35	60	0.5	1	0.6	0	Braun	0.13	12.09	12.39
35	60	0.5	1	0.6	0	FEEE	0.14	12.08	12.38
35	60	0.5	1	0.6	0	Chain_L	0.11	12.07	12.59
35	60	0.5	1	0.6	0	Chain_E	0.12	12.06	12.57
35	60	0.5	1	0.6	0	keNEATPSE_L	0.14	12.06	12.36
35	60	0.5	1	0.6	0	keNEATPSE_E	0.14	12.09	12.38
35	60	0.5	1	0.6	0	keNEATCE_L	0.11	12.07	12.59
35	60	0.5	1	0.6	0	keNEATCE_E	0.12	12.08	12.60
35	60	0.5	1	0.6	0	Linear	0.01	12.14	12.14
35	60	0.5	1	0.6	0	Equipercentile	0.01	12.13	12.13
116	60	0.5	1	1	0	Tucker	-0.09	12.23	12.51

Table 4.4

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
116	60	0.5	1	1	0	Levine True	-0.06	12.18	12.85
116	60	0.5	1	1	0	Braun	-0.05	12.22	12.50
116	60	0.5	1	1	0	FEEE	-0.02	12.21	12.49
116	60	0.5	1	1	0	Chain_L	-0.07	12.20	12.69
116	60	0.5	1	1	0	Chain_E	-0.03	12.19	12.67
116	60	0.5	1	1	0	keNEATPSE_L	-0.03	12.19	12.47
116	60	0.5	1	1	0	keNEATPSE_E	-0.02	12.22	12.50
116	60	0.5	1	1	0	keNEATCE_L	-0.07	12.20	12.69
116	60	0.5	1	1	0	keNEATCE_E	-0.02	12.21	12.69
116	60	0.5	1	1	0	Linear	0.02	12.26	12.26
116	60	0.5	1	1	0	Equipercentile	0.02	12.26	12.26
36	60	0.5	1	0.6	1	Tucker	-0.03	11.65	11.84
36	60	0.5	1	0.6	1	Levine True	0.08	11.60	12.20
36	60	0.5	1	0.6	1	Braun	-0.01	11.64	11.83
36	60	0.5	1	0.6	1	FEEE	0.01	11.64	11.83
36	60	0.5	1	0.6	1	Chain_L	0.03	11.63	12.02
36	60	0.5	1	0.6	1	Chain_E	0.02	11.62	12.00
36	60	0.5	1	0.6	1	keNEATPSE_L	0.00	11.62	11.81
36	60	0.5	1	0.6	1	keNEATPSE_E	0.01	11.65	11.83
36	60	0.5	1	0.6	1	keNEATCE_L	0.03	11.63	12.02
36	60	0.5	1	0.6	1	keNEATCE_E	0.03	11.64	12.02
36	60	0.5	1	0.6	1	Linear	-0.02	11.64	11.64
36	60	0.5	1	0.6	1	Equipercentile	-0.02	11.64	11.64
117	60	0.5	1	1	1	Tucker	0.27	13.75	14.00
117	60	0.5	1	1	1	Levine True	0.44	13.78	14.49
117	60	0.5	1	1	1	Braun	0.31	13.75	14.00
117	60	0.5	1	1	1	FEEE	0.36	13.74	13.99
117	60	0.5	1	1	1	Chain_L	0.36	13.77	14.23
117	60	0.5	1	1	1	Chain_E	0.37	13.71	14.17
117	60	0.5	1	1	1	keNEATPSE_L	0.35	13.68	13.93
117	60	0.5	1	1	1	keNEATPSE_E	0.36	13.75	14.00
117	60	0.5	1	1	1	keNEATCE_L	0.36	13.77	14.23
117	60	0.5	1	1	1	keNEATCE_E	0.37	13.76	14.22

Table 4.4

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
117	60	0.5	1	1	1	Linear	0.00	13.70	13.69
117	60	0.5	1	1	1	Equipercentile	0.00	13.69	13.69

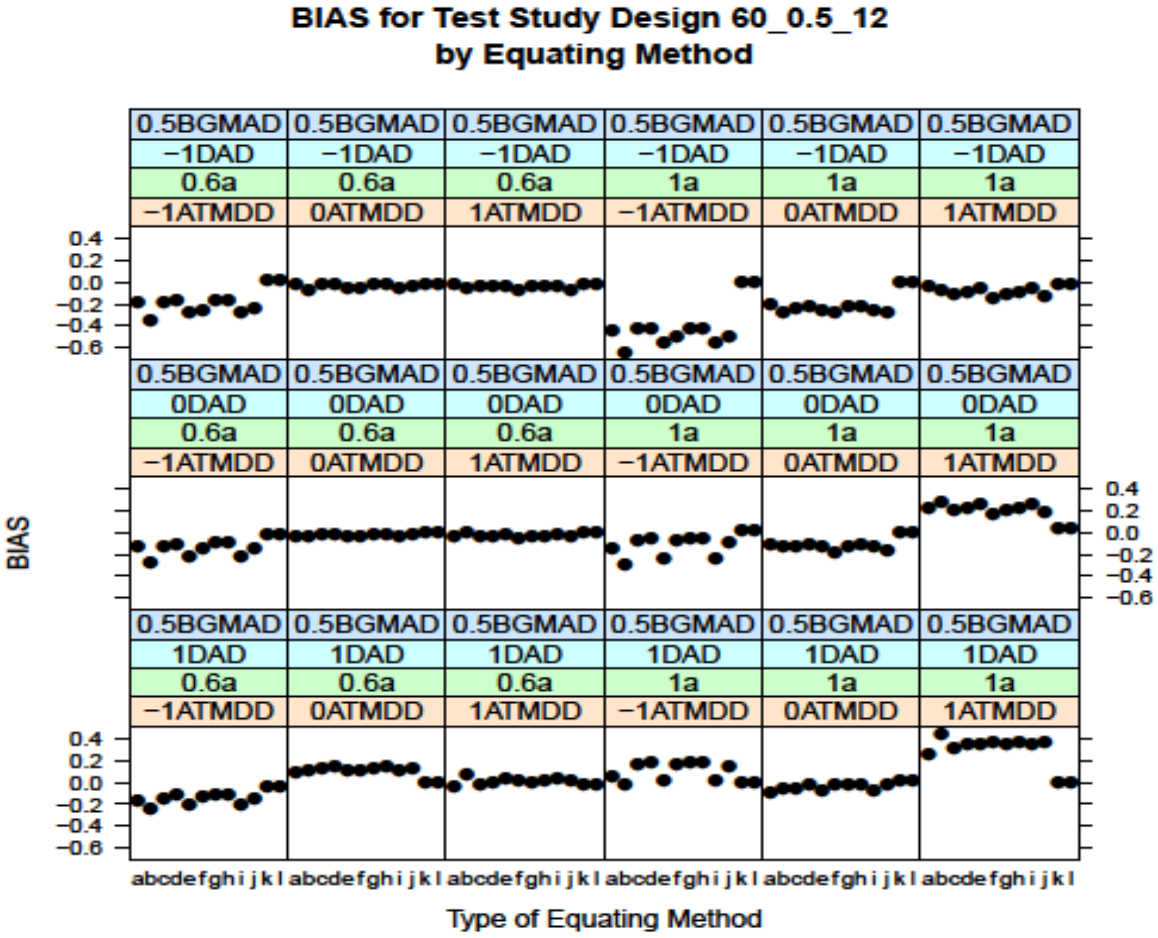

Figure 4.7. Bias for Test Study Design 60_0.5_12 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

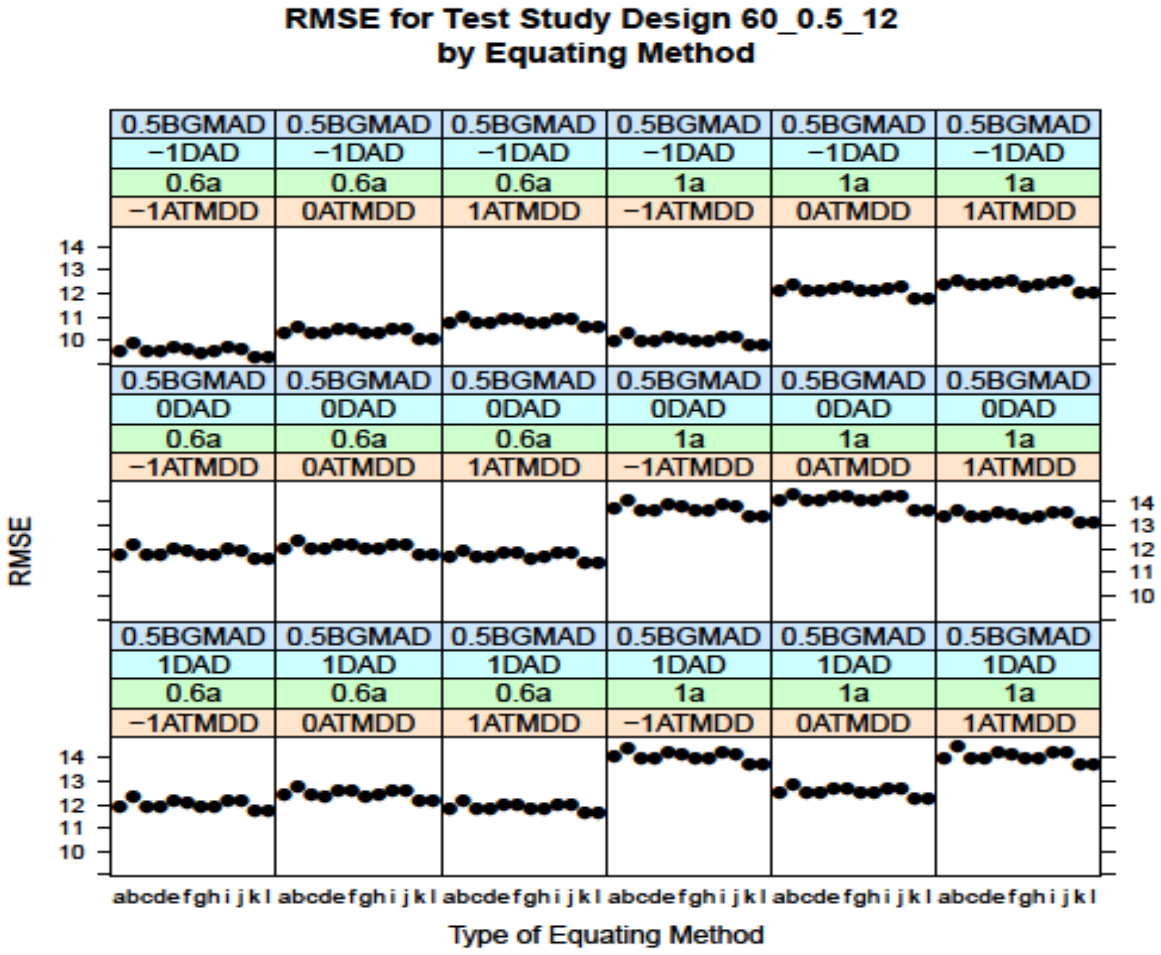

Figure 4.8. Root Mean Square Error (RMSE) for Test Study Design 60_0.5_12 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.2.5 60_1.0_12 Test Study Design

In this subsection, the results for test study design 60_1.0_12 are presented. This design had 60 total items and 12 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of

group separation (or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large respectively. Large (1.5) BGMAD will be discussed in the subsequent subsections. This subsection mainly focuses on medium (1.0) BGMAD, which means 60 total items and 12 anchor test items were held constant and the other three study conditions were varied. Tables A37-A45 in Appendix A display average descriptive statistics and Figure B.5 in Appendix B shows the standard error of equating (SEE) for test study design 60_1.0_12. Table 4.5 represents bias, SEE, and RMSE for this test study design. Figure 4.9 demonstrates that there is zero bias for all conditions under linear and equipercentile equating methods for medium (1.0) between-grade mean ability difference (BGMAD). Other equating methods show both negative and positive values of bias which are very close to zero apart from a few study conditions where results are inconsistent. For example, when distribution of ability difference (DAD) and anchor test mean difficulty difference (ATMDD) are below average (-1), average (1.0), and above average (1) and item discrimination is moderate (0.6) the equating methods show inconsistency. This pattern of inconsistency is also repeated when item discrimination is high (1) where DAD is below average (-1), average (0), and above average (1). However, where BGMDD is medium (1.0) and DAD is average (0), the equating methods perform similarly with bias about zero except when ATMDD is below average (-1) and item discrimination is high (1) resulting to negative values. Similarly, where BGMDD is medium (1.0) and DAD is average (0) and above average (1), the equating methods yield similar bias results, which show positive bias values.

Figure 4.10 shows test study design 60_1.0_12, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (60), medium (1.0) BGMAD, moderate (0.6) and high (1) item discriminations are invariant. The RMSE values fall between 10 and 17. When all conditions are held constant and manipulate item discrimination (moderate versus high), RMSE displays a clear consistency where moderate (0.6) item discrimination across the other four study conditions results in smaller (more accurate) RMSE values than RMSE values for high (1) item discrimination varied over the other four study conditions. Therefore, two patterns are clear in this design based on conditions varied under either moderate item discrimination or high item discrimination with the former performing better than the latter in terms of accuracy.

Addressing the research question number 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.9 revealed the following results. First, the bias was consistent and very close to zero for all equating methods when medium (1.0) BGMAD for moderate (0.6) item discrimination and DAD was average (0) and above (1) average, and when ATMDD was average (0) and above average (1). Second, the equating methods performed similarly when BGMAD was medium (1.0) under above (1) average DAD with high (1) item discrimination for below (-1) average and average (0) ATMDD. The rest of the results for other study conditions under this design can be deemed inaccurate and perhaps underestimated or overestimated accuracy for all equating methods, because of negative and positive bias values.

Overall the equating results in Figure 4.10 show that the smallest difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods was under below average (-1), average (0), and above average (1) DAD when item discrimination was moderate (0.6) and ATMDD below average (-1) conditions. However, largest difference between anchor test mean difficulty and the other four study conditions when DAD varied across its three levels with a high (1) item discrimination and ATMDD was average (0). Moderate (0.6) item discrimination produced rather more accurate overall results than high (1) item discrimination under all conditions across all equating methods.

Table 4.5

BIAS, SEE, and RMSE Statistics for Test Study Design 60_1.0_12 by Equating Method Under All Conditions

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
37	60	1	-1	0.6	-1	Tucker	-0.65	8.21	8.66
37	60	1	-1	0.6	-1	Levine True	-1.28	8.50	9.99
37	60	1	-1	0.6	-1	Braun	-0.68	8.27	8.71
37	60	1	-1	0.6	-1	FEEE	-0.66	8.25	8.69
37	60	1	-1	0.6	-1	Chain_L	-0.98	8.28	9.22
37	60	1	-1	0.6	-1	Chain_E	-0.93	8.24	9.13
37	60	1	-1	0.6	-1	keNEATPSE_L	-0.66	8.20	8.64
37	60	1	-1	0.6	-1	keNEATPSE_E	-0.66	8.26	8.70
37	60	1	-1	0.6	-1	keNEATCE_L	-0.98	8.28	9.22
37	60	1	-1	0.6	-1	keNEATCE_E	-0.91	8.25	9.17
37	60	1	-1	0.6	-1	Linear	-0.03	8.08	8.07
37	60	1	-1	0.6	-1	Equipercentile	-0.02	8.07	8.07
118	60	1	-1	1	-1	Tucker	-1.07	9.23	9.84
118	60	1	-1	1	-1	Levine True	-1.78	9.92	11.51
118	60	1	-1	1	-1	Braun	-1.09	9.20	9.80

Table 4.5

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
118	60	1	-1	1	-1	FEEE	-1.05	9.19	9.80
118	60	1	-1	1	-1	Chain_L	-1.43	9.38	10.47
118	60	1	-1	1	-1	Chain_E	-1.32	9.20	10.23
118	60	1	-1	1	-1	keNEATPSE_L	-1.04	9.07	9.68
118	60	1	-1	1	-1	keNEATPSE_E	-1.05	9.20	9.80
118	60	1	-1	1	-1	keNEATCE_L	-1.43	9.38	10.47
118	60	1	-1	1	-1	keNEATCE_E	-1.35	9.21	10.32
118	60	1	-1	1	-1	Linear	0.01	8.98	8.97
118	60	1	-1	1	-1	Equipercentile	0.01	8.98	8.97
38	60	1	-1	0.6	0	Tucker	-0.45	8.56	9.11
38	60	1	-1	0.6	0	Levine True	-0.78	8.55	10.18
38	60	1	-1	0.6	0	Braun	-0.53	8.61	9.15
38	60	1	-1	0.6	0	FEEE	-0.50	8.58	9.13
38	60	1	-1	0.6	0	Chain_L	-0.63	8.58	9.67
38	60	1	-1	0.6	0	Chain_E	-0.70	8.58	9.64
38	60	1	-1	0.6	0	keNEATPSE_L	-0.51	8.51	9.06
38	60	1	-1	0.6	0	keNEATPSE_E	-0.50	8.60	9.15
38	60	1	-1	0.6	0	keNEATCE_L	-0.63	8.58	9.67
38	60	1	-1	0.6	0	keNEATCE_E	-0.66	8.60	9.70
38	60	1	-1	0.6	0	Linear	-0.01	8.54	8.54
38	60	1	-1	0.6	0	Equipercentile	-0.01	8.54	8.54
119	60	1	-1	1	0	Tucker	-0.90	9.31	10.21
119	60	1	-1	1	0	Levine True	-1.26	9.34	11.12
119	60	1	-1	1	0	Braun	-1.03	9.35	10.23
119	60	1	-1	1	0	FEEE	-1.00	9.34	10.21
119	60	1	-1	1	0	Chain_L	-1.11	9.35	10.72
119	60	1	-1	1	0	Chain_E	-1.24	9.35	10.66
119	60	1	-1	1	0	keNEATPSE_L	-1.02	9.25	10.14
119	60	1	-1	1	0	keNEATPSE_E	-1.00	9.34	10.22
119	60	1	-1	1	0	keNEATCE_L	-1.11	9.35	10.72
119	60	1	-1	1	0	keNEATCE_E	-1.22	9.34	10.71
119	60	1	-1	1	0	Linear	0.01	9.23	9.22
119	60	1	-1	1	0	Equipercentile	0.01	9.22	9.22

Table 4.5

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
39	60	1	-1	0.6	1	Tucker	-0.62	8.97	9.80
39	60	1	-1	0.6	1	Levine True	-0.87	8.85	10.60
39	60	1	-1	0.6	1	Braun	-0.71	9.02	9.84
39	60	1	-1	0.6	1	FEEE	-0.68	9.00	9.82
39	60	1	-1	0.6	1	Chain_L	-0.78	8.92	10.29
39	60	1	-1	0.6	1	Chain_E	-0.86	8.98	10.33
39	60	1	-1	0.6	1	keNEATPSE_L	-0.68	8.96	9.79
39	60	1	-1	0.6	1	keNEATPSE_E	-0.68	9.01	9.83
39	60	1	-1	0.6	1	keNEATCE_L	-0.78	8.92	10.29
39	60	1	-1	0.6	1	keNEATCE_E	-0.81	8.98	10.35
39	60	1	-1	0.6	1	Linear	0.00	9.06	9.06
39	60	1	-1	0.6	1	Equipercentile	0.00	9.06	9.06
120	60	1	-1	1	1	Tucker	-0.99	10.17	11.61
120	60	1	-1	1	1	Levine True	-1.20	10.01	12.09
120	60	1	-1	1	1	Braun	-1.08	10.25	11.67
120	60	1	-1	1	1	FEEE	-1.06	10.24	11.66
120	60	1	-1	1	1	Chain_L	-1.13	10.07	11.94
120	60	1	-1	1	1	Chain_E	-1.26	10.25	12.07
120	60	1	-1	1	1	keNEATPSE_L	-1.07	10.22	11.64
120	60	1	-1	1	1	keNEATPSE_E	-1.07	10.25	11.67
120	60	1	-1	1	1	keNEATCE_L	-1.13	10.07	11.94
120	60	1	-1	1	1	keNEATCE_E	-1.22	10.22	12.07
120	60	1	-1	1	1	Linear	-0.01	10.37	10.36
120	60	1	-1	1	1	Equipercentile	-0.01	10.36	10.36
40	60	1	0	0.6	-1	Tucker	-0.71	11.07	11.77
40	60	1	0	0.6	-1	Levine True	-1.27	11.01	13.18
40	60	1	0	0.6	-1	Braun	-0.70	11.00	11.70
40	60	1	0	0.6	-1	FEEE	-0.64	10.97	11.67
40	60	1	0	0.6	-1	Chain_L	-1.01	11.09	12.47
40	60	1	0	0.6	-1	Chain_E	-0.83	10.89	12.22
40	60	1	0	0.6	-1	keNEATPSE_L	-0.63	10.85	11.56
40	60	1	0	0.6	-1	keNEATPSE_E	-0.64	10.98	11.68
40	60	1	0	0.6	-1	keNEATCE_L	-1.01	11.09	12.47

Table 4.5

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
40	60	1	0	0.6	-1	keNEATCE_E	-0.82	10.94	12.30
40	60	1	0	0.6	-1	Linear	0.00	11.02	11.02
40	60	1	0	0.6	-1	Equipercentile	0.00	11.02	11.02
121	60	1	0	1	-1	Tucker	-0.52	11.13	12.06
121	60	1	0	1	-1	Levine True	-0.84	11.14	13.40
121	60	1	0	1	-1	Braun	-0.55	11.09	12.01
121	60	1	0	1	-1	FEEE	-0.48	11.06	11.98
121	60	1	0	1	-1	Chain_L	-0.69	11.16	12.73
121	60	1	0	1	-1	Chain_E	-0.67	11.05	12.58
121	60	1	0	1	-1	keNEATPSE_L	-0.48	10.95	11.88
121	60	1	0	1	-1	keNEATPSE_E	-0.47	11.08	12.00
121	60	1	0	1	-1	keNEATCE_L	-0.69	11.16	12.73
121	60	1	0	1	-1	keNEATCE_E	-0.63	11.07	12.65
121	60	1	0	1	-1	Linear	0.02	11.01	11.01
121	60	1	0	1	-1	Equipercentile	0.02	11.00	11.00
41	60	1	0	0.6	0	Tucker	-0.21	11.34	12.49
41	60	1	0	0.6	0	Levine True	-0.39	11.07	13.47
41	60	1	0	0.6	0	Braun	-0.22	11.31	12.45
41	60	1	0	0.6	0	FEEE	-0.18	11.28	12.42
41	60	1	0	0.6	0	Chain_L	-0.32	11.21	13.05
41	60	1	0	0.6	0	Chain_E	-0.35	11.23	13.04
41	60	1	0	0.6	0	keNEATPSE_L	-0.19	11.21	12.36
41	60	1	0	0.6	0	keNEATPSE_E	-0.18	11.29	12.43
41	60	1	0	0.6	0	keNEATCE_L	-0.32	11.21	13.05
41	60	1	0	0.6	0	keNEATCE_E	-0.31	11.25	13.09
41	60	1	0	0.6	0	Linear	0.02	11.45	11.45
41	60	1	0	0.6	0	Equipercentile	0.02	11.45	11.45
122	60	1	0	1	0	Tucker	-0.89	12.53	14.19
122	60	1	0	1	0	Levine True	-1.18	12.38	15.00
122	60	1	0	1	0	Braun	-0.88	12.47	14.13
122	60	1	0	1	0	FEEE	-0.85	12.46	14.11
122	60	1	0	1	0	Chain_L	-1.06	12.45	14.68
122	60	1	0	1	0	Chain_E	-1.02	12.40	14.61

Table 4.5

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
122	60	1	0	1	0	keNEATPSE_L	-0.88	12.39	14.05
122	60	1	0	1	0	keNEATPSE_E	-0.86	12.48	14.13
122	60	1	0	1	0	keNEATCE_L	-1.06	12.45	14.68
122	60	1	0	1	0	keNEATCE_E	-0.98	12.43	14.67
122	60	1	0	1	0	Linear	0.03	12.64	12.64
122	60	1	0	1	0	Equipercentile	0.03	12.64	12.64
42	60	1	0	0.6	1	Tucker	0.09	11.00	12.05
42	60	1	0	0.6	1	Levine True	0.07	10.73	13.01
42	60	1	0	0.6	1	Braun	0.05	10.99	12.03
42	60	1	0	0.6	1	FEEE	0.08	10.95	12.00
42	60	1	0	0.6	1	Chain_L	0.08	10.85	12.61
42	60	1	0	0.6	1	Chain_E	-0.07	10.90	12.62
42	60	1	0	0.6	1	keNEATPSE_L	0.07	10.88	11.93
42	60	1	0	0.6	1	keNEATPSE_E	0.08	10.96	12.01
42	60	1	0	0.6	1	keNEATCE_L	0.08	10.85	12.61
42	60	1	0	0.6	1	keNEATCE_E	-0.01	10.91	12.66
42	60	1	0	0.6	1	Linear	0.01	11.19	11.19
42	60	1	0	0.6	1	Equipercentile	0.01	11.19	11.19
123	60	1	0	1	1	Tucker	0.37	12.17	13.59
123	60	1	0	1	1	Levine True	0.46	11.90	14.51
123	60	1	0	1	1	Braun	0.24	12.10	13.51
123	60	1	0	1	1	FEEE	0.28	12.06	13.47
123	60	1	0	1	1	Chain_L	0.42	12.02	14.13
123	60	1	0	1	1	Chain_E	0.05	12.02	14.06
123	60	1	0	1	1	keNEATPSE_L	0.25	11.98	13.40
123	60	1	0	1	1	keNEATPSE_E	0.27	12.07	13.48
123	60	1	0	1	1	keNEATCE_L	0.42	12.02	14.13
123	60	1	0	1	1	keNEATCE_E	0.13	12.01	14.08
123	60	1	0	1	1	Linear	0.00	12.41	12.41
123	60	1	0	1	1	Equipercentile	0.00	12.41	12.41
43	60	1	1	0.6	-1	Tucker	-0.63	12.22	13.26
43	60	1	1	0.6	-1	Levine True	-0.97	11.97	14.51
43	60	1	1	0.6	-1	Braun	-0.50	12.13	13.18

Table 4.5

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
43	60	1	1	0.6	-1	FEEE	-0.44	12.09	13.15
43	60	1	1	0.6	-1	Chain_L	-0.82	12.11	13.92
43	60	1	1	0.6	-1	Chain_E	-0.54	12.00	13.80
43	60	1	1	0.6	-1	keNEATPSE_L	-0.44	12.00	13.06
43	60	1	1	0.6	-1	keNEATPSE_E	-0.44	12.11	13.16
43	60	1	1	0.6	-1	keNEATCE_L	-0.82	12.11	13.92
43	60	1	1	0.6	-1	keNEATCE_E	-0.55	12.06	13.87
43	60	1	1	0.6	-1	Linear	-0.01	12.29	12.29
43	60	1	1	0.6	-1	Equipercentile	-0.01	12.29	12.29
124	60	1	1	1	-1	Tucker	-0.34	13.24	14.66
124	60	1	1	1	-1	Levine True	-0.46	12.96	15.84
124	60	1	1	1	-1	Braun	-0.14	13.07	14.50
124	60	1	1	1	-1	FEEE	-0.07	13.04	14.47
124	60	1	1	1	-1	Chain_L	-0.41	13.10	15.29
124	60	1	1	1	-1	Chain_E	-0.16	12.93	15.08
124	60	1	1	1	-1	keNEATPSE_L	-0.08	12.90	14.33
124	60	1	1	1	-1	keNEATPSE_E	-0.06	13.08	14.50
124	60	1	1	1	-1	keNEATCE_L	-0.41	13.10	15.29
124	60	1	1	1	-1	keNEATCE_E	-0.15	13.03	15.21
124	60	1	1	1	-1	Linear	0.03	13.55	13.55
124	60	1	1	1	-1	Equipercentile	0.03	13.55	13.55
44	60	1	1	0.6	0	Tucker	0.02	11.23	12.03
44	60	1	1	0.6	0	Levine True	0.07	11.04	13.32
44	60	1	1	0.6	0	Braun	0.08	11.19	11.99
44	60	1	1	0.6	0	FEEE	0.13	11.17	11.97
44	60	1	1	0.6	0	Chain_L	0.05	11.14	12.69
44	60	1	1	0.6	0	Chain_E	0.05	11.11	12.64
44	60	1	1	0.6	0	keNEATPSE_L	0.12	11.10	11.90
44	60	1	1	0.6	0	keNEATPSE_E	0.13	11.19	11.99
44	60	1	1	0.6	0	keNEATCE_L	0.05	11.14	12.69
44	60	1	1	0.6	0	keNEATCE_E	0.08	11.16	12.72
44	60	1	1	0.6	0	Linear	0.00	11.29	11.29
44	60	1	1	0.6	0	Equipercentile	0.00	11.29	11.29

Table 4.5

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
125	60	1	1	1	0	Tucker	-0.08	14.18	16.18
125	60	1	1	1	0	Levine True	-0.11	13.81	16.96
125	60	1	1	1	0	Braun	0.19	14.15	16.18
125	60	1	1	1	0	FEEE	0.22	14.12	16.15
125	60	1	1	1	0	Chain_L	-0.10	13.96	16.65
125	60	1	1	1	0	Chain_E	0.12	14.06	16.74
125	60	1	1	1	0	keNEATPSE_L	0.21	14.05	16.09
125	60	1	1	1	0	keNEATPSE_E	0.22	14.13	16.15
125	60	1	1	1	0	keNEATCE_L	-0.10	13.96	16.65
125	60	1	1	1	0	keNEATCE_E	0.13	14.07	16.77
125	60	1	1	1	0	Linear	0.01	14.89	14.89
125	60	1	1	1	0	Equipercentile	0.01	14.89	14.89
45	60	1	1	0.6	1	Tucker	0.29	12.03	12.85
45	60	1	1	0.6	1	Levine True	0.59	11.81	14.18
45	60	1	1	0.6	1	Braun	0.34	11.99	12.81
45	60	1	1	0.6	1	FEEE	0.39	11.95	12.77
45	60	1	1	0.6	1	Chain_L	0.46	11.93	13.53
45	60	1	1	0.6	1	Chain_E	0.33	11.86	13.42
45	60	1	1	0.6	1	keNEATPSE_L	0.37	11.84	12.67
45	60	1	1	0.6	1	keNEATPSE_E	0.39	11.97	12.79
45	60	1	1	0.6	1	keNEATCE_L	0.46	11.93	13.53
45	60	1	1	0.6	1	keNEATCE_E	0.39	11.92	13.51
45	60	1	1	0.6	1	Linear	0.00	12.20	12.20
45	60	1	1	0.6	1	Equipercentile	0.00	12.20	12.19
126	60	1	1	1	1	Tucker	0.70	13.65	15.00
126	60	1	1	1	1	Levine True	1.13	13.35	16.14
126	60	1	1	1	1	Braun	0.74	13.53	14.89
126	60	1	1	1	1	FEEE	0.81	13.48	14.83
126	60	1	1	1	1	Chain_L	0.96	13.50	15.64
126	60	1	1	1	1	Chain_E	0.61	13.37	15.43
126	60	1	1	1	1	keNEATPSE_L	0.76	13.37	14.74
126	60	1	1	1	1	keNEATPSE_E	0.80	13.50	14.85
126	60	1	1	1	1	keNEATCE_L	0.96	13.50	15.64

Table 4.5

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
126	60	1	1	1	1	keNEATCE_E	0.72	13.41	15.51
126	60	1	1	1	1	Linear	0.00	13.93	13.93
126	60	1	1	1	1	Equipercentile	0.00	13.93	13.93

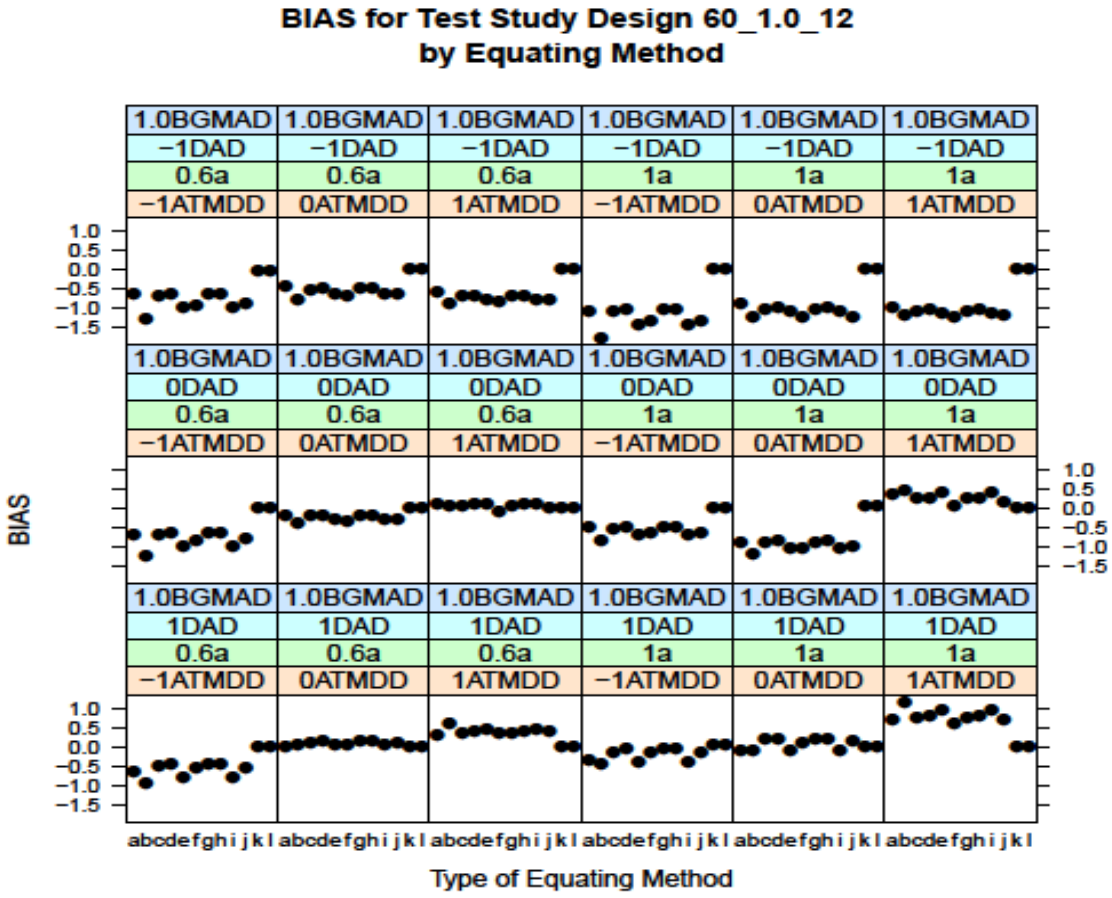

Figure 4.9. Bias for Test Study Design 60_1.0_12 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

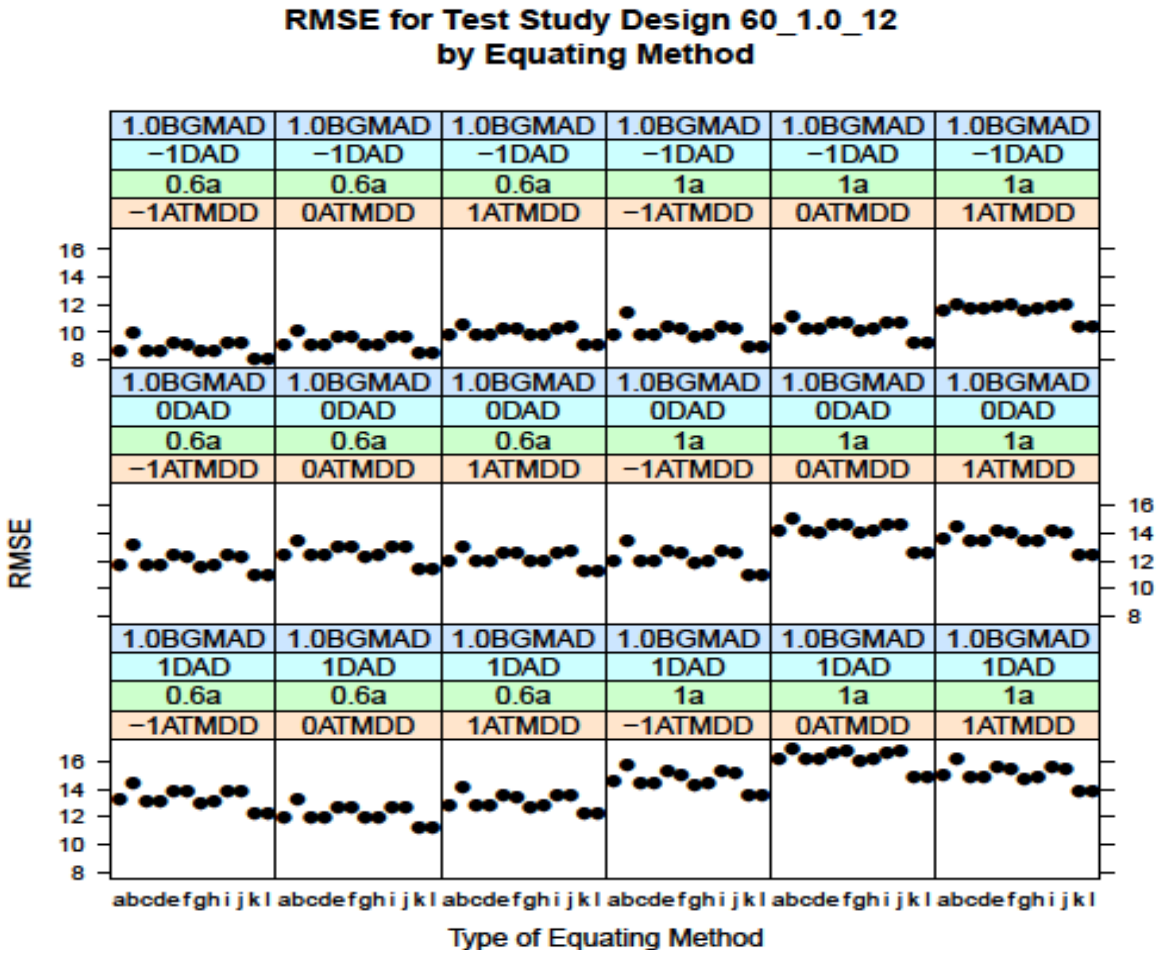

Figure 4.10. Root Mean Square Error (RMSE) for Test Study Design 60_1.0_12 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.2.6 60_1.5_12 Test Study Design

This subsection presents the results for test study design 60_1.5_12. This design had 60 total items and 12 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of group separation

(or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large respectively. The previous two subsections discussed small and medium BGMAD. This subsection mainly focuses on large BGMAD, which means 60 total items and 12 anchor test items were held constant and the other three study conditions were varied. Tables A46-A54 in Appendix A display average descriptive statistics and Figure B.6 in Appendix B shows the standard error of equating (SEE) for test study design 60_1.5_12. Table 4.6 represents bias, SEE, and RMSE for this test study design. Figure 4.11 demonstrates that there is negative bias for all conditions under all equating methods for large (1.5) between-grade mean ability difference (BGMAD) except for positive bias when BGMAD is large (1.5) and DAD is average (0) and above average (1), item discrimination is moderate (0.6) and high (1) and ATMDD is above average (1). Both negative and positive values of bias are very close to zero for a few study conditions. Noticeable in this regard is negative bias which is almost zero when BGMDD is large (1.5) while DAD is above average (1) and item discrimination is moderate (0.6) and ATMDD varied across its three levels. However, when the same conditions are repeated under high (1) item discrimination (and to some extent under moderate item discrimination), only linear and equipercentile equating methods produce zero bias.

Figure 4.12 shows test study design 60_1.5_12, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (60), large (1.5) BGMAD, moderate (0.6) and high (1) item discriminations are unchanging factors. The RMSE values fall between 5 and 19. When all conditions are held constant and manipulate item discrimination (moderate versus high), RMSE displays some

consistency for both moderate (0.6) and high (1) conditions. Also, the RMSE values for both moderate (0.6) and high (1) discrimination when other conditions are varied performed similarly. Comparatively, conditions manipulated under moderate (0.6) item discrimination produced smaller (more accurate) RMSE values than its counterpart, high (1) item discrimination.

Addressing the research question number 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.11 revealed that difference between anchor test difficulty and the other four study conditions is smallest when large (1.5) BGMAD, moderate (0.6) item discrimination, and above average (1) ATMDD are held constant and then grade-to-grade ability variability (DAD) is manipulated –that is, across its three levels (below average, average, and above average). Similarly, under the above conditions, average (0) ATMDD also produced bias values too close to zero. This means that a medium (60) test length, a large (1.5) BGMAD with average (0) and above average (1) ATMDD conditioned on different ability distribution (DAD) within a grade has the smallest bias (or best results) compared to other study conditions in this design. Other study conditions produced worst results. Therefore, there is sufficient evidence to believe that a large (1.5) BGMAD together with a medium test (60 items) produce heterogeneous bias results across all study conditions under all equating methods. This assertion is supported by the fact that when holding large (1.5) BGMAD, high (1) item discrimination, and below average (-1) ATMDD constant and vary DAD across its three levels, then the equating methods produce the largest bias (or worst results) compared with other study conditions.

Overall the equating results in Figure 4.12 reveal that there was a slight difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods. In different words, all equating methods seemed to have performed differently without any particular pattern. The worst scenario was for large (1.5) BGMAD, above average (1) DAD, high (1) item discrimination where ATMDD is varied as below average (-1), average (0), and above average (1) conditions. This means that other study conditions produced almost close RMSE values under various equating methods.

At this point, it is important to reflect on the second set of the three test study designs (the first set of the three test study design included 30_0.5_6, 30_1.0_6, and 30_1.5_6 as outlined previously) discussed so far—60_0.5_12, 60_1.0_12, and 60_1.5_12—have the same number of total test items or medium test (60 items in total) and anchor test items (12 items) under all study conditions with variability in magnitude of the group separation [or BGMAD across small (0.5), medium (1.0), and large (1.5)]. Juxtaposing these three test study designs—on the basis of magnitude of the group separation—(Figures 4.8, 4.10, and 4.12) in terms of RMSE values leads to the conclusion that the overall accuracy or stability of the results is considerably affected by the magnitude of group separation/group effect (or mean ability difference between adjacent grades/BGMAD). Degree of accuracy of the results decreased from small (0.5) BGMAD to large (1.5) BGMAD under all conditions with large (1.5) BGMAD producing the largest RMSE values compared to the other two test study designs.

Remarkably, small (0.5) BGMAD under all study conditions also had the smallest bias while large (1.5) BGMAD had the largest bias values.

Table 4.6

BIAS, SEE, and RMSE Statistics for Test Study Design 60_1.5_12 by Equating Method Under All Conditions

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
46	60	1.5	-1	0.6	-1	Tucker	-1.42	6.96	7.63
46	60	1.5	-1	0.6	-1	Levine True	-3.03	8.64	10.85
46	60	1.5	-1	0.6	-1	Braun	-1.39	6.99	7.62
46	60	1.5	-1	0.6	-1	FEEE	-1.39	7.00	7.64
46	60	1.5	-1	0.6	-1	Chain_L	-2.22	7.40	8.82
46	60	1.5	-1	0.6	-1	Chain_E	-1.97	7.04	8.33
46	60	1.5	-1	0.6	-1	keNEATPSE_L	-1.06	6.85	7.67
46	60	1.5	-1	0.6	-1	keNEATPSE_E	-1.13	6.94	7.71
46	60	1.5	-1	0.6	-1	keNEATCE_L	-1.97	7.42	9.01
46	60	1.5	-1	0.6	-1	keNEATCE_E	-1.86	7.18	8.65
46	60	1.5	-1	0.6	-1	Linear	0.00	6.55	6.55
46	60	1.5	-1	0.6	-1	Equipercentile	0.00	6.54	6.54
127	60	1.5	-1	1	-1	Tucker	-1.54	6.60	7.57
127	60	1.5	-1	1	-1	Levine True	-2.58	7.35	9.53
127	60	1.5	-1	1	-1	Braun	-1.62	6.80	7.73
127	60	1.5	-1	1	-1	FEEE	-1.61	6.83	7.78
127	60	1.5	-1	1	-1	Chain_L	-2.11	6.77	8.40
127	60	1.5	-1	1	-1	Chain_E	-2.13	6.86	8.38
127	60	1.5	-1	1	-1	keNEATPSE_L	-1.57	6.76	7.70
127	60	1.5	-1	1	-1	keNEATPSE_E	-1.57	6.84	7.76
127	60	1.5	-1	1	-1	keNEATCE_L	-2.10	6.78	8.40
127	60	1.5	-1	1	-1	keNEATCE_E	-2.23	7.07	8.69
127	60	1.5	-1	1	-1	Linear	0.00	6.30	6.30
127	60	1.5	-1	1	-1	Equipercentile	0.00	6.32	6.32
47	60	1.5	-1	0.6	0	Tucker	-1.00	7.48	8.50
47	60	1.5	-1	0.6	0	Levine True	-1.77	7.70	10.49
47	60	1.5	-1	0.6	0	Braun	-1.12	7.60	8.60

Table 4.6

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
47	60	1.5	-1	0.6	0	FEEE	-1.11	7.59	8.59
47	60	1.5	-1	0.6	0	Chain_L	-1.43	7.51	9.48
47	60	1.5	-1	0.6	0	Chain_E	-1.58	7.61	9.51
47	60	1.5	-1	0.6	0	keNEATPSE_L	-1.10	7.48	8.50
47	60	1.5	-1	0.6	0	keNEATPSE_E	-1.09	7.60	8.60
47	60	1.5	-1	0.6	0	keNEATCE_L	-1.43	7.51	9.48
47	60	1.5	-1	0.6	0	keNEATCE_E	-1.53	7.67	9.64
47	60	1.5	-1	0.6	0	Linear	0.03	7.49	7.48
47	60	1.5	-1	0.6	0	Equipercentile	0.03	7.49	7.48
128	60	1.5	-1	1	0	Tucker	-2.20	7.54	9.07
128	60	1.5	-1	1	0	Levine True	-3.05	7.99	10.56
128	60	1.5	-1	1	0	Braun	-2.36	7.78	9.25
128	60	1.5	-1	1	0	FEEE	-2.40	7.85	9.34
128	60	1.5	-1	1	0	Chain_L	-2.72	7.67	9.84
128	60	1.5	-1	1	0	Chain_E	-2.86	7.85	9.93
128	60	1.5	-1	1	0	keNEATPSE_L	-2.31	7.78	9.33
128	60	1.5	-1	1	0	keNEATPSE_E	-2.31	7.87	9.39
128	60	1.5	-1	1	0	keNEATCE_L	-2.69	7.67	9.85
128	60	1.5	-1	1	0	keNEATCE_E	-2.85	7.98	10.18
128	60	1.5	-1	1	0	Linear	0.01	7.17	7.17
128	60	1.5	-1	1	0	Equipercentile	0.01	7.18	7.18
48	60	1.5	-1	0.6	1	Tucker	-0.79	7.38	8.73
48	60	1.5	-1	0.6	1	Levine True	-1.26	7.18	10.14
48	60	1.5	-1	0.6	1	Braun	-1.03	7.56	8.88
48	60	1.5	-1	0.6	1	FEEE	-1.03	7.53	8.87
48	60	1.5	-1	0.6	1	Chain_L	-1.10	7.29	9.62
48	60	1.5	-1	0.6	1	Chain_E	-1.45	7.58	9.86
48	60	1.5	-1	0.6	1	keNEATPSE_L	-1.03	7.49	8.83
48	60	1.5	-1	0.6	1	keNEATPSE_E	-1.02	7.55	8.88
48	60	1.5	-1	0.6	1	keNEATCE_L	-1.10	7.29	9.62
48	60	1.5	-1	0.6	1	keNEATCE_E	-1.36	7.55	9.88
48	60	1.5	-1	0.6	1	Linear	-0.04	7.39	7.39
48	60	1.5	-1	0.6	1	Equipercentile	-0.04	7.39	7.39

Table 4.6

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
129	60	1.5	-1	1	1	Tucker	-1.58	8.09	10.28
129	60	1.5	-1	1	1	Levine True	-2.00	7.82	11.09
129	60	1.5	-1	1	1	Braun	-1.90	8.42	10.56
129	60	1.5	-1	1	1	FEEE	-1.90	8.43	10.59
129	60	1.5	-1	1	1	Chain_L	-1.88	7.93	10.86
129	60	1.5	-1	1	1	Chain_E	-2.29	8.48	11.34
129	60	1.5	-1	1	1	keNEATPSE_L	-1.90	8.44	10.63
129	60	1.5	-1	1	1	keNEATPSE_E	-1.90	8.47	10.63
129	60	1.5	-1	1	1	keNEATCE_L	-1.88	7.93	10.86
129	60	1.5	-1	1	1	keNEATCE_E	-2.29	8.50	11.40
129	60	1.5	-1	1	1	Linear	-0.01	8.52	8.52
129	60	1.5	-1	1	1	Equipercentile	-0.01	8.53	8.52
49	60	1.5	0	0.6	-1	Tucker	-2.29	10.36	11.89
49	60	1.5	0	0.6	-1	Levine True	-4.04	11.66	15.38
49	60	1.5	0	0.6	-1	Braun	-2.14	10.18	11.67
49	60	1.5	0	0.6	-1	FEEE	-2.08	10.16	11.66
49	60	1.5	0	0.6	-1	Chain_L	-3.21	10.60	13.30
49	60	1.5	0	0.6	-1	Chain_E	-2.62	9.99	12.53
49	60	1.5	0	0.6	-1	keNEATPSE_L	-2.04	9.97	11.49
49	60	1.5	0	0.6	-1	keNEATPSE_E	-2.07	10.18	11.67
49	60	1.5	0	0.6	-1	keNEATCE_L	-3.21	10.60	13.30
49	60	1.5	0	0.6	-1	keNEATCE_E	-2.66	10.07	12.66
49	60	1.5	0	0.6	-1	Linear	0.01	9.96	9.96
49	60	1.5	0	0.6	-1	Equipercentile	0.01	9.96	9.96
130	60	1.5	0	1	-1	Tucker	-3.89	11.63	13.63
130	60	1.5	0	1	-1	Levine True	-6.04	13.30	17.20
130	60	1.5	0	1	-1	Braun	-3.47	11.07	13.00
130	60	1.5	0	1	-1	FEEE	-3.42	11.11	13.04
130	60	1.5	0	1	-1	Chain_L	-5.02	12.04	15.05
130	60	1.5	0	1	-1	Chain_E	-4.01	10.90	13.69
130	60	1.5	0	1	-1	keNEATPSE_L	-3.36	10.82	12.80
130	60	1.5	0	1	-1	keNEATPSE_E	-3.41	11.11	13.05
130	60	1.5	0	1	-1	keNEATCE_L	-5.02	12.04	15.05

Table 4.6

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
130	60	1.5	0	1	-1	keNEATCE_E	-4.11	10.94	13.79
130	60	1.5	0	1	-1	Linear	0.00	10.87	10.87
130	60	1.5	0	1	-1	Equipercentile	0.00	10.87	10.87
50	60	1.5	0	0.6	0	Tucker	-0.95	9.19	10.85
50	60	1.5	0	0.6	0	Levine True	-1.44	8.75	12.66
50	60	1.5	0	0.6	0	Braun	-1.01	9.23	10.88
50	60	1.5	0	0.6	0	FEEE	-0.97	9.20	10.86
50	60	1.5	0	0.6	0	Chain_L	-1.25	8.99	11.88
50	60	1.5	0	0.6	0	Chain_E	-1.33	9.11	11.96
50	60	1.5	0	0.6	0	keNEATPSE_L	-0.97	9.11	10.79
50	60	1.5	0	0.6	0	keNEATPSE_E	-0.97	9.21	10.87
50	60	1.5	0	0.6	0	keNEATCE_L	-1.25	8.99	11.88
50	60	1.5	0	0.6	0	keNEATCE_E	-1.24	9.10	11.99
50	60	1.5	0	0.6	0	Linear	-0.02	9.46	9.46
50	60	1.5	0	0.6	0	Equipercentile	-0.02	9.46	9.46
131	60	1.5	0	1	0	Tucker	-1.93	11.34	14.15
131	60	1.5	0	1	0	Levine True	-2.53	10.90	15.55
131	60	1.5	0	1	0	Braun	-2.03	11.30	14.06
131	60	1.5	0	1	0	FEEE	-1.98	11.28	14.08
131	60	1.5	0	1	0	Chain_L	-2.29	11.12	14.98
131	60	1.5	0	1	0	Chain_E	-2.42	11.19	14.98
131	60	1.5	0	1	0	keNEATPSE_L	-2.00	11.10	13.93
131	60	1.5	0	1	0	keNEATPSE_E	-1.96	11.29	14.07
131	60	1.5	0	1	0	keNEATCE_L	-2.29	11.12	14.98
131	60	1.5	0	1	0	keNEATCE_E	-2.40	11.27	15.15
131	60	1.5	0	1	0	Linear	0.01	11.85	11.84
131	60	1.5	0	1	0	Equipercentile	0.02	11.85	11.85
51	60	1.5	0	0.6	1	Tucker	-0.19	9.62	11.51
51	60	1.5	0	0.6	1	Levine True	-0.39	9.10	13.20
51	60	1.5	0	0.6	1	Braun	-0.40	9.66	11.53
51	60	1.5	0	0.6	1	FEEE	-0.35	9.60	11.48
51	60	1.5	0	0.6	1	Chain_L	-0.31	9.34	12.52
51	60	1.5	0	0.6	1	Chain_E	-0.75	9.54	12.64

Table 4.6

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
51	60	1.5	0	0.6	1	keNEATPSE_L	-0.38	9.48	11.38
51	60	1.5	0	0.6	1	keNEATPSE_E	-0.36	9.61	11.49
51	60	1.5	0	0.6	1	keNEATCE_L	-0.31	9.34	12.52
51	60	1.5	0	0.6	1	keNEATCE_E	-0.60	9.52	12.68
51	60	1.5	0	0.6	1	Linear	0.00	9.98	9.98
51	60	1.5	0	0.6	1	Equipercentile	0.00	9.98	9.98
132	60	1.5	0	1	1	Tucker	-0.37	11.71	14.83
132	60	1.5	0	1	1	Levine True	-0.48	11.09	16.41
132	60	1.5	0	1	1	Braun	-1.20	11.78	14.87
132	60	1.5	0	1	1	FEEE	-1.14	11.73	14.81
132	60	1.5	0	1	1	Chain_L	-0.44	11.36	15.79
132	60	1.5	0	1	1	Chain_E	-1.79	11.71	15.88
132	60	1.5	0	1	1	keNEATPSE_L	-1.24	11.52	14.64
132	60	1.5	0	1	1	keNEATPSE_E	-1.14	11.75	14.82
132	60	1.5	0	1	1	keNEATCE_L	-0.44	11.36	15.79
132	60	1.5	0	1	1	keNEATCE_E	-1.69	11.70	15.91
132	60	1.5	0	1	1	Linear	0.03	12.60	12.60
132	60	1.5	0	1	1	Equipercentile	0.03	12.60	12.60
52	60	1.5	1	0.6	-1	Tucker	-1.60	11.17	13.30
52	60	1.5	1	0.6	-1	Levine True	-2.50	10.90	15.65
52	60	1.5	1	0.6	-1	Braun	-1.34	10.98	13.10
52	60	1.5	1	0.6	-1	FEEE	-1.22	10.94	13.06
52	60	1.5	1	0.6	-1	Chain_L	-2.11	10.98	14.52
52	60	1.5	1	0.6	-1	Chain_E	-1.50	10.72	14.19
52	60	1.5	1	0.6	-1	keNEATPSE_L	-1.21	10.77	12.92
52	60	1.5	1	0.6	-1	keNEATPSE_E	-1.22	10.95	13.07
52	60	1.5	1	0.6	-1	keNEATCE_L	-2.11	10.98	14.52
52	60	1.5	1	0.6	-1	keNEATCE_E	-1.52	10.76	14.25
52	60	1.5	1	0.6	-1	Linear	0.02	11.42	11.42
52	60	1.5	1	0.6	-1	Equipercentile	0.02	11.42	11.41
133	60	1.5	1	1	-1	Tucker	-2.12	12.05	15.06
133	60	1.5	1	1	-1	Levine True	-2.78	11.46	16.63
133	60	1.5	1	1	-1	Braun	-1.61	11.87	14.90

Table 4.6

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
133	60	1.5	1	1	-1	FEEE	-1.52	11.81	14.87
133	60	1.5	1	1	-1	Chain_L	-2.51	11.75	15.98
133	60	1.5	1	1	-1	Chain_E	-1.76	11.65	15.82
133	60	1.5	1	1	-1	keNEATPSE_L	-1.50	11.69	14.79
133	60	1.5	1	1	-1	keNEATPSE_E	-1.52	11.82	14.88
133	60	1.5	1	1	-1	keNEATCE_L	-2.51	11.75	15.98
133	60	1.5	1	1	-1	keNEATCE_E	-1.80	11.63	15.83
133	60	1.5	1	1	-1	Linear	-0.04	12.66	12.66
133	60	1.5	1	1	-1	Equipercentile	-0.04	12.66	12.66
53	60	1.5	1	0.6	0	Tucker	-0.78	11.10	13.23
53	60	1.5	1	0.6	0	Levine True	-1.19	10.52	15.19
53	60	1.5	1	0.6	0	Braun	-0.68	11.01	13.14
53	60	1.5	1	0.6	0	FEEE	-0.58	10.95	13.09
53	60	1.5	1	0.6	0	Chain_L	-1.02	10.81	14.35
53	60	1.5	1	0.6	0	Chain_E	-0.80	10.79	14.29
53	60	1.5	1	0.6	0	keNEATPSE_L	-0.57	10.81	12.97
53	60	1.5	1	0.6	0	keNEATPSE_E	-0.58	10.97	13.10
53	60	1.5	1	0.6	0	keNEATCE_L	-1.02	10.81	14.35
53	60	1.5	1	0.6	0	keNEATCE_E	-0.78	10.83	14.38
53	60	1.5	1	0.6	0	Linear	0.00	11.47	11.47
53	60	1.5	1	0.6	0	Equipercentile	0.00	11.47	11.47
134	60	1.5	1	1	0	Tucker	-2.00	13.60	16.95
134	60	1.5	1	1	0	Levine True	-2.69	12.95	18.73
134	60	1.5	1	1	0	Braun	-1.59	13.26	16.60
134	60	1.5	1	1	0	FEEE	-1.43	13.18	16.56
134	60	1.5	1	1	0	Chain_L	-2.41	13.27	17.98
134	60	1.5	1	1	0	Chain_E	-1.62	12.97	17.61
134	60	1.5	1	1	0	keNEATPSE_L	-1.41	12.97	16.40
134	60	1.5	1	1	0	keNEATPSE_E	-1.42	13.19	16.57
134	60	1.5	1	1	0	keNEATCE_L	-2.41	13.27	17.98
134	60	1.5	1	1	0	keNEATCE_E	-1.65	12.99	17.68
134	60	1.5	1	1	0	Linear	-0.01	14.25	14.25
134	60	1.5	1	1	0	Equipercentile	0.00	14.25	14.25

Table 4.6

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
54	60	1.5	1	0.6	1	Tucker	-0.25	11.25	13.36
54	60	1.5	1	0.6	1	Levine True	-0.38	10.58	15.39
54	60	1.5	1	0.6	1	Braun	-0.30	11.17	13.27
54	60	1.5	1	0.6	1	FEEE	-0.19	11.10	13.21
54	60	1.5	1	0.6	1	Chain_L	-0.33	10.91	14.49
54	60	1.5	1	0.6	1	Chain_E	-0.46	10.94	14.47
54	60	1.5	1	0.6	1	keNEATPSE_L	-0.20	10.93	13.07
54	60	1.5	1	0.6	1	keNEATPSE_E	-0.19	11.12	13.22
54	60	1.5	1	0.6	1	keNEATCE_L	-0.33	10.91	14.49
54	60	1.5	1	0.6	1	keNEATCE_E	-0.35	10.97	14.55
54	60	1.5	1	0.6	1	Linear	-0.03	11.77	11.76
54	60	1.5	1	0.6	1	Equipercentile	-0.03	11.76	11.76
135	60	1.5	1	1	1	Tucker	-0.13	12.93	15.82
135	60	1.5	1	1	1	Levine True	-0.16	12.40	17.97
135	60	1.5	1	1	1	Braun	-0.23	12.68	15.55
135	60	1.5	1	1	1	FEEE	-0.08	12.59	15.45
135	60	1.5	1	1	1	Chain_L	-0.15	12.65	17.02
135	60	1.5	1	1	1	Chain_E	-0.34	12.42	16.68
135	60	1.5	1	1	1	keNEATPSE_L	-0.13	12.31	15.21
135	60	1.5	1	1	1	keNEATPSE_E	-0.09	12.62	15.46
135	60	1.5	1	1	1	keNEATCE_L	-0.15	12.65	17.02
135	60	1.5	1	1	1	keNEATCE_E	-0.25	12.48	16.80
135	60	1.5	1	1	1	Linear	-0.01	13.49	13.49
135	60	1.5	1	1	1	Equipercentile	-0.01	13.49	13.49

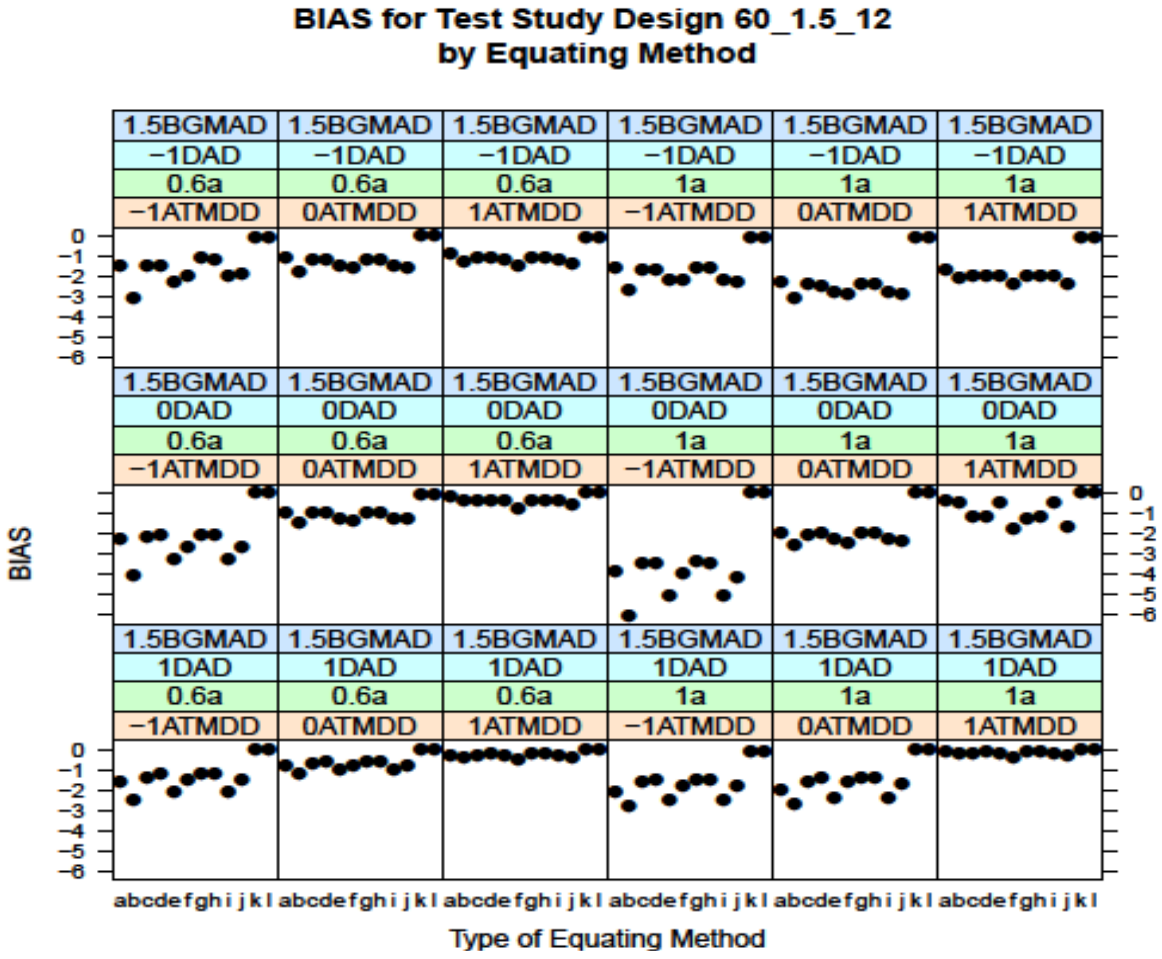

Figure 4.11. Bias for Test Study Design 60_1.5_12 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

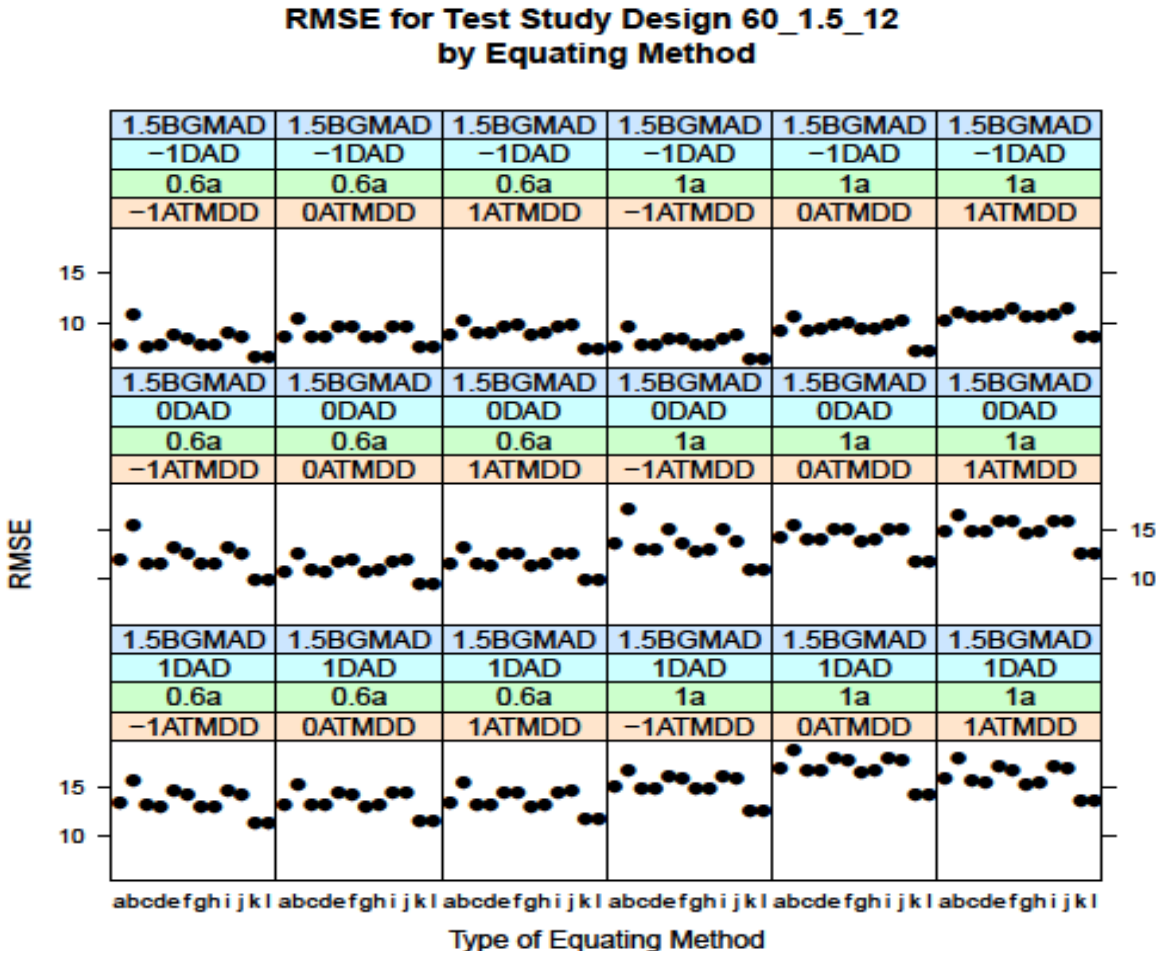

Figure 4.12. Root Mean Square Error (RMSE) for Test Study Design 60_1.5_12 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.2.7 120_0.5_24 Test Study Design

This subsection presents the results for test study design 120_0.5_24. This design had 120 total items and 24 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of group separation

(or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large respectively. Medium and large BGMAD will be discussed in the subsequent subsections. This subsection mainly focuses on small BGMAD, which means 120 total items and 24 anchor test items were held constant and the other three study conditions were varied. Tables A55-A63 in Appendix A display average descriptive statistics and Figure B.7 in Appendix B shows the standard error of equating (SEE) for test study design 120_0.5_24. Table 4.7 represents bias, SEE, and RMSE for this test study design. Figure 4.13 demonstrates that there is almost zero bias for all conditions under all equating methods for small (0.5) between-grade mean ability difference (BGMAD). Both negative and positive values of bias are almost close to zero except for a few study conditions where results are rather stable. For example, when distribution of ability difference (DAD) is below (-1) average, average (0), and above (1) average and anchor test mean difficulty difference (ATMDD) is above (1) average and item discrimination is moderate (0.6), the equating methods show consistency. This pattern of consistency is also repeated when item discrimination is high (1) where DAD is below (-1) average and average (0) and especially where ATMDD is above average (1). The rest of the study conditions particularly when item discrimination was high (1) produced inconsistent results.

Figure 4.14 shows test study design 120_0.5_24, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (120), small (0.5) BGMAD, moderate (0.6) and high (1) item discriminations are held constant. The RMSE values fall between 18 and 23. Strikingly, when all conditions are

held constant and manipulate item discrimination (moderate versus high), RMSE displays a clear consistency and stability when item discrimination is moderate (0.6). Also, the RMSE values for moderate (0.6) discrimination when other conditions are manipulated are more accurate (or lower) than RMSE values when item discrimination is high (1). One noticeable trend in this design was that all equating methods consistently produced almost similar values of RMSE when magnitude of group separation or BGMAD was considerably small (0.5) and item discrimination was moderate (0.6) and regular test (RT) *b*-item parameter for a grade was equal to the mean ability for that grade [or DAD was average (0)] and DAD was above (1) average.

Addressing the research question number 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.13 revealed the following results. First, the bias was consistent (or stable) and very close to zero for all equating methods when small (0.5) BGMAD and moderate (0.6) item discrimination were held constant and DAD varied across below (-1), average (0), and above average (1) and when ATMDD was average (0) and above average (1). Second, the equating results were inaccurate and underestimated accuracy for all equating methods, as evidenced by negative bias, under small (0.5) BGMAD where DAD was below average (-1) and average (0) for both moderate (0.6) and high (1) item discrimination and ATMDD was below (-1) average and average (0). Third, when small (0.5) BGMAD, high (1) item discrimination, and above average (1) DAD were held constant and manipulated ATMDD from average (0) to above (1)

average the bias overestimated, suggested by positive bias values, the accuracy of the equating results for all equating methods.

Overall the equating results in Figure 4.14 show that there was an insignificant difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods. Stated differently, all equating methods seemed to have a similar performance without any discernible pattern save for slight differences where item discrimination was moderate (0.6) and high (1) for all conditions. Relatively, high (1) item discrimination produced less accurate overall results than moderate (0.6) item discrimination under all conditions.

Table 4.7

BIAS, SEE, and RMSE Statistics for Test Study Design 120_0.5_24 by Equating Method Under All Conditions

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
55	120	0.5	-1	0.6	-1	Tucker	-0.37	19.21	19.72
55	120	0.5	-1	0.6	-1	Levine True	-0.54	19.17	20.19
55	120	0.5	-1	0.6	-1	Braun	-0.31	19.13	19.64
55	120	0.5	-1	0.6	-1	FEEE	-0.32	19.14	19.67
55	120	0.5	-1	0.6	-1	Chain_L	-0.47	19.20	19.98
55	120	0.5	-1	0.6	-1	Chain_E	-0.37	19.10	19.89
55	120	0.5	-1	0.6	-1	keNEATPSE_L	-0.32	19.09	19.62
55	120	0.5	-1	0.6	-1	keNEATPSE_E	-0.32	19.15	19.67
55	120	0.5	-1	0.6	-1	keNEATCE_L	-0.47	19.20	19.98
55	120	0.5	-1	0.6	-1	keNEATCE_E	-0.35	19.13	19.92
55	120	0.5	-1	0.6	-1	Linear	0.04	19.36	19.35
55	120	0.5	-1	0.6	-1	Equipercentile	0.04	19.36	19.35
136	120	0.5	-1	1	-1	Tucker	-1.04	21.12	21.74
136	120	0.5	-1	1	-1	Levine True	-1.33	21.21	22.24
136	120	0.5	-1	1	-1	Braun	-0.90	20.90	21.55

Table 4.7

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
136	120	0.5	-1	1	-1	FEEE	-0.93	20.94	21.60
136	120	0.5	-1	1	-1	Chain_L	-1.20	21.18	22.02
136	120	0.5	-1	1	-1	Chain_E	-1.01	20.90	21.78
136	120	0.5	-1	1	-1	keNEATPSE_L	-0.92	20.90	21.56
136	120	0.5	-1	1	-1	keNEATPSE_E	-0.92	20.94	21.60
136	120	0.5	-1	1	-1	keNEATCE_L	-1.20	21.18	22.02
136	120	0.5	-1	1	-1	keNEATCE_E	-1.00	20.91	21.79
136	120	0.5	-1	1	-1	Linear	0.01	20.58	20.58
136	120	0.5	-1	1	-1	Equipercentile	0.01	20.57	20.57
56	120	0.5	-1	0.6	0	Tucker	-0.54	18.33	18.89
56	120	0.5	-1	0.6	0	Levine True	-0.67	18.31	19.32
56	120	0.5	-1	0.6	0	Braun	-0.57	18.33	18.88
56	120	0.5	-1	0.6	0	FEEE	-0.57	18.34	18.90
56	120	0.5	-1	0.6	0	Chain_L	-0.62	18.32	19.14
56	120	0.5	-1	0.6	0	Chain_E	-0.65	18.33	19.13
56	120	0.5	-1	0.6	0	keNEATPSE_L	-0.57	18.31	18.86
56	120	0.5	-1	0.6	0	keNEATPSE_E	-0.57	18.35	18.90
56	120	0.5	-1	0.6	0	keNEATCE_L	-0.62	18.32	19.14
56	120	0.5	-1	0.6	0	keNEATCE_E	-0.62	18.36	19.17
56	120	0.5	-1	0.6	0	Linear	0.00	18.31	18.30
56	120	0.5	-1	0.6	0	Equipercentile	0.00	18.30	18.30
137	120	0.5	-1	1	0	Tucker	-0.44	21.59	22.44
137	120	0.5	-1	1	0	Levine True	-0.53	21.54	22.76
137	120	0.5	-1	1	0	Braun	-0.50	21.59	22.42
137	120	0.5	-1	1	0	FEEE	-0.50	21.60	22.43
137	120	0.5	-1	1	0	Chain_L	-0.50	21.57	22.63
137	120	0.5	-1	1	0	Chain_E	-0.57	21.59	22.63
137	120	0.5	-1	1	0	keNEATPSE_L	-0.52	21.55	22.39
137	120	0.5	-1	1	0	keNEATPSE_E	-0.50	21.63	22.46
137	120	0.5	-1	1	0	keNEATCE_L	-0.50	21.57	22.63
137	120	0.5	-1	1	0	keNEATCE_E	-0.54	21.63	22.68
137	120	0.5	-1	1	0	Linear	-0.04	21.45	21.45
137	120	0.5	-1	1	0	Equipercentile	-0.04	21.45	21.44

Table 4.7

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
57	120	0.5	-1	0.6	1	Tucker	-0.19	20.98	21.78
57	120	0.5	-1	0.6	1	Levine True	-0.20	20.90	22.12
57	120	0.5	-1	0.6	1	Braun	-0.26	20.99	21.77
57	120	0.5	-1	0.6	1	FEEE	-0.26	20.98	21.77
57	120	0.5	-1	0.6	1	Chain_L	-0.20	20.93	21.99
57	120	0.5	-1	0.6	1	Chain_E	-0.33	20.98	22.02
57	120	0.5	-1	0.6	1	keNEATPSE_L	-0.27	20.96	21.75
57	120	0.5	-1	0.6	1	keNEATPSE_E	-0.26	20.99	21.78
57	120	0.5	-1	0.6	1	keNEATCE_L	-0.20	20.93	21.99
57	120	0.5	-1	0.6	1	keNEATCE_E	-0.30	20.99	22.04
57	120	0.5	-1	0.6	1	Linear	0.07	21.14	21.14
57	120	0.5	-1	0.6	1	Equipercentile	0.07	21.15	21.14
138	120	0.5	-1	1	1	Tucker	0.23	21.69	22.60
138	120	0.5	-1	1	1	Levine True	0.21	21.66	23.00
138	120	0.5	-1	1	1	Braun	0.01	21.72	22.61
138	120	0.5	-1	1	1	FEEE	0.00	21.73	22.63
138	120	0.5	-1	1	1	Chain_L	0.22	21.68	22.84
138	120	0.5	-1	1	1	Chain_E	-0.10	21.75	22.86
138	120	0.5	-1	1	1	keNEATPSE_L	-0.01	21.69	22.58
138	120	0.5	-1	1	1	keNEATPSE_E	0.01	21.76	22.66
138	120	0.5	-1	1	1	keNEATCE_L	0.22	21.68	22.84
138	120	0.5	-1	1	1	keNEATCE_E	-0.06	21.79	22.90
138	120	0.5	-1	1	1	Linear	0.05	21.74	21.74
138	120	0.5	-1	1	1	Equipercentile	0.05	21.74	21.74
58	120	0.5	0	0.6	-1	Tucker	-0.43	22.62	23.38
58	120	0.5	0	0.6	-1	Levine True	-0.62	22.54	23.84
58	120	0.5	0	0.6	-1	Braun	-0.29	22.52	23.29
58	120	0.5	0	0.6	-1	FEEE	-0.30	22.53	23.31
58	120	0.5	0	0.6	-1	Chain_L	-0.54	22.58	23.64
58	120	0.5	0	0.6	-1	Chain_E	-0.33	22.48	23.55
58	120	0.5	0	0.6	-1	keNEATPSE_L	-0.30	22.50	23.28
58	120	0.5	0	0.6	-1	keNEATPSE_E	-0.30	22.54	23.32
58	120	0.5	0	0.6	-1	keNEATCE_L	-0.54	22.58	23.64

Table 4.7

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
58	120	0.5	0	0.6	-1	keNEATCE_E	-0.33	22.50	23.57
58	120	0.5	0	0.6	-1	Linear	0.03	22.80	22.80
58	120	0.5	0	0.6	-1	Equipercentile	0.03	22.80	22.79
139	120	0.5	0	1	-1	Tucker	-0.82	26.15	27.08
139	120	0.5	0	1	-1	Levine True	-1.04	26.12	27.57
139	120	0.5	0	1	-1	Braun	-0.60	25.90	26.89
139	120	0.5	0	1	-1	FEEE	-0.59	25.89	26.90
139	120	0.5	0	1	-1	Chain_L	-0.94	26.14	27.36
139	120	0.5	0	1	-1	Chain_E	-0.61	25.82	27.10
139	120	0.5	0	1	-1	keNEATPSE_L	-0.58	25.86	26.87
139	120	0.5	0	1	-1	keNEATPSE_E	-0.58	25.91	26.91
139	120	0.5	0	1	-1	keNEATCE_L	-0.94	26.14	27.36
139	120	0.5	0	1	-1	keNEATCE_E	-0.60	25.85	27.12
139	120	0.5	0	1	-1	Linear	0.00	26.11	26.10
139	120	0.5	0	1	-1	Equipercentile	0.00	26.11	26.11
59	120	0.5	0	0.6	0	Tucker	-0.31	22.47	23.18
59	120	0.5	0	0.6	0	Levine True	-0.39	22.42	23.68
59	120	0.5	0	0.6	0	Braun	-0.25	22.42	23.14
59	120	0.5	0	0.6	0	FEEE	-0.25	22.42	23.14
59	120	0.5	0	0.6	0	Chain_L	-0.35	22.44	23.46
59	120	0.5	0	0.6	0	Chain_E	-0.29	22.42	23.43
59	120	0.5	0	0.6	0	keNEATPSE_L	-0.25	22.38	23.09
59	120	0.5	0	0.6	0	keNEATPSE_E	-0.24	22.43	23.15
59	120	0.5	0	0.6	0	keNEATCE_L	-0.35	22.44	23.46
59	120	0.5	0	0.6	0	keNEATCE_E	-0.27	22.45	23.47
59	120	0.5	0	0.6	0	Linear	0.01	22.43	22.43
59	120	0.5	0	0.6	0	Equipercentile	0.01	22.43	22.43
140	120	0.5	0	1	0	Tucker	-0.56	26.46	27.59
140	120	0.5	0	1	0	Levine True	-0.63	26.39	27.92
140	120	0.5	0	1	0	Braun	-0.47	26.42	27.55
140	120	0.5	0	1	0	FEEE	-0.47	26.41	27.55
140	120	0.5	0	1	0	Chain_L	-0.60	26.42	27.79
140	120	0.5	0	1	0	Chain_E	-0.51	26.39	27.77

Table 4.7

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
140	120	0.5	0	1	0	keNEATPSE_L	-0.47	26.38	27.51
140	120	0.5	0	1	0	keNEATPSE_E	-0.47	26.43	27.56
140	120	0.5	0	1	0	keNEATCE_L	-0.60	26.42	27.79
140	120	0.5	0	1	0	keNEATCE_E	-0.49	26.42	27.80
140	120	0.5	0	1	0	Linear	-0.05	26.71	26.71
140	120	0.5	0	1	0	Equipercentile	-0.06	26.70	26.70
60	120	0.5	0	0.6	1	Tucker	0.14	23.58	24.33
60	120	0.5	0	0.6	1	Levine True	0.20	23.51	24.85
60	120	0.5	0	0.6	1	Braun	0.09	23.53	24.28
60	120	0.5	0	0.6	1	FEEE	0.10	23.53	24.28
60	120	0.5	0	0.6	1	Chain_L	0.17	23.54	24.63
60	120	0.5	0	0.6	1	Chain_E	0.01	23.52	24.59
60	120	0.5	0	0.6	1	keNEATPSE_L	0.09	23.48	24.23
60	120	0.5	0	0.6	1	keNEATPSE_E	0.10	23.54	24.29
60	120	0.5	0	0.6	1	keNEATCE_L	0.17	23.54	24.63
60	120	0.5	0	0.6	1	keNEATCE_E	0.05	23.55	24.63
60	120	0.5	0	0.6	1	Linear	0.07	23.62	23.61
60	120	0.5	0	0.6	1	Equipercentile	0.07	23.61	23.61
141	120	0.5	0	1	1	Tucker	0.16	26.86	27.87
141	120	0.5	0	1	1	Levine True	0.23	26.78	28.32
141	120	0.5	0	1	1	Braun	0.04	26.77	27.78
141	120	0.5	0	1	1	FEEE	0.04	26.76	27.78
141	120	0.5	0	1	1	Chain_L	0.20	26.82	28.13
141	120	0.5	0	1	1	Chain_E	-0.08	26.75	28.05
141	120	0.5	0	1	1	keNEATPSE_L	0.02	26.72	27.73
141	120	0.5	0	1	1	keNEATPSE_E	0.04	26.79	27.81
141	120	0.5	0	1	1	keNEATCE_L	0.20	26.82	28.13
141	120	0.5	0	1	1	keNEATCE_E	-0.04	26.79	28.09
141	120	0.5	0	1	1	Linear	0.00	26.94	26.94
141	120	0.5	0	1	1	Equipercentile	0.00	26.94	26.93
61	120	0.5	1	0.6	-1	Tucker	-0.29	22.37	23.14
61	120	0.5	1	0.6	-1	Levine True	-0.39	22.28	23.68
61	120	0.5	1	0.6	-1	Braun	-0.10	22.32	23.10

Table 4.7

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
61	120	0.5	1	0.6	-1	FEEE	-0.08	22.30	23.09
61	120	0.5	1	0.6	-1	Chain_L	-0.35	22.32	23.44
61	120	0.5	1	0.6	-1	Chain_E	-0.08	22.28	23.40
61	120	0.5	1	0.6	-1	keNEATPSE_L	-0.08	22.28	23.07
61	120	0.5	1	0.6	-1	keNEATPSE_E	-0.08	22.31	23.10
61	120	0.5	1	0.6	-1	keNEATCE_L	-0.35	22.32	23.44
61	120	0.5	1	0.6	-1	keNEATCE_E	-0.09	22.31	23.42
61	120	0.5	1	0.6	-1	Linear	-0.05	22.67	22.66
61	120	0.5	1	0.6	-1	Equipercentile	-0.05	22.67	22.66
142	120	0.5	1	1	-1	Tucker	-0.17	27.02	28.16
142	120	0.5	1	1	-1	Levine True	-0.27	26.91	28.52
142	120	0.5	1	1	-1	Braun	0.20	26.95	28.15
142	120	0.5	1	1	-1	FEEE	0.19	26.92	28.13
142	120	0.5	1	1	-1	Chain_L	-0.23	26.95	28.38
142	120	0.5	1	1	-1	Chain_E	0.20	26.91	28.34
142	120	0.5	1	1	-1	keNEATPSE_L	0.19	26.92	28.13
142	120	0.5	1	1	-1	keNEATPSE_E	0.19	26.93	28.13
142	120	0.5	1	1	-1	keNEATCE_L	-0.23	26.95	28.38
142	120	0.5	1	1	-1	keNEATCE_E	0.19	26.90	28.34
142	120	0.5	1	1	-1	Linear	0.00	26.88	26.88
142	120	0.5	1	1	-1	Equipercentile	0.00	26.88	26.87
62	120	0.5	1	0.6	0	Tucker	0.04	22.87	23.69
62	120	0.5	1	0.6	0	Levine True	0.03	22.77	24.16
62	120	0.5	1	0.6	0	Braun	0.17	22.86	23.68
62	120	0.5	1	0.6	0	FEEE	0.18	22.86	23.68
62	120	0.5	1	0.6	0	Chain_L	0.04	22.82	23.96
62	120	0.5	1	0.6	0	Chain_E	0.15	22.83	23.97
62	120	0.5	1	0.6	0	keNEATPSE_L	0.18	22.83	23.65
62	120	0.5	1	0.6	0	keNEATPSE_E	0.18	22.87	23.69
62	120	0.5	1	0.6	0	keNEATCE_L	0.04	22.82	23.96
62	120	0.5	1	0.6	0	keNEATCE_E	0.15	22.86	24.00
62	120	0.5	1	0.6	0	Linear	-0.03	23.07	23.07
62	120	0.5	1	0.6	0	Equipercentile	-0.03	23.07	23.07

Table 4.7

Cont.

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
143	120	0.5	1	1	0	Tucker	0.68	25.75	26.76
143	120	0.5	1	1	0	Levine True	0.72	25.67	27.21
143	120	0.5	1	1	0	Braun	0.81	25.73	26.73
143	120	0.5	1	1	0	FEEE	0.84	25.74	26.74
143	120	0.5	1	1	0	Chain_L	0.70	25.71	27.02
143	120	0.5	1	1	0	Chain_E	0.79	25.70	27.01
143	120	0.5	1	1	0	keNEATPSE_L	0.84	25.69	26.69
143	120	0.5	1	1	0	keNEATPSE_E	0.84	25.75	26.75
143	120	0.5	1	1	0	keNEATCE_L	0.70	25.71	27.02
143	120	0.5	1	1	0	keNEATCE_E	0.80	25.73	27.04
143	120	0.5	1	1	0	Linear	0.01	25.79	25.79
143	120	0.5	1	1	0	Equipercentile	0.00	25.79	25.79
63	120	0.5	1	0.6	1	Tucker	0.22	22.14	22.73
63	120	0.5	1	0.6	1	Levine True	0.40	22.07	23.31
63	120	0.5	1	0.6	1	Braun	0.27	22.12	22.70
63	120	0.5	1	0.6	1	FEEE	0.30	22.12	22.70
63	120	0.5	1	0.6	1	Chain_L	0.32	22.10	23.03
63	120	0.5	1	0.6	1	Chain_E	0.28	22.08	23.00
63	120	0.5	1	0.6	1	keNEATPSE_L	0.29	22.07	22.66
63	120	0.5	1	0.6	1	keNEATPSE_E	0.30	22.13	22.72
63	120	0.5	1	0.6	1	keNEATCE_L	0.32	22.10	23.03
63	120	0.5	1	0.6	1	keNEATCE_E	0.29	22.12	23.05
63	120	0.5	1	0.6	1	Linear	-0.04	22.23	22.23
63	120	0.5	1	0.6	1	Equipercentile	-0.04	22.23	22.22
144	120	0.5	1	1	1	Tucker	0.75	26.48	27.20
144	120	0.5	1	1	1	Levine True	1.11	26.50	27.87
144	120	0.5	1	1	1	Braun	0.74	26.35	27.07
144	120	0.5	1	1	1	FEEE	0.79	26.35	27.07
144	120	0.5	1	1	1	Chain_L	0.96	26.50	27.57
144	120	0.5	1	1	1	Chain_E	0.71	26.27	27.33
144	120	0.5	1	1	1	keNEATPSE_L	0.76	26.28	27.01
144	120	0.5	1	1	1	keNEATPSE_E	0.79	26.36	27.09
144	120	0.5	1	1	1	keNEATCE_L	0.96	26.50	27.57

Table 4.7

Cont.

	Study Condition							Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
144	120	0.5	1	1	1	keNEATCE_E	0.75	26.33	27.40
144	120	0.5	1	1	1	Linear	0.00	26.43	26.43
144	120	0.5	1	1	1	Equipercentile	0.00	26.43	26.43

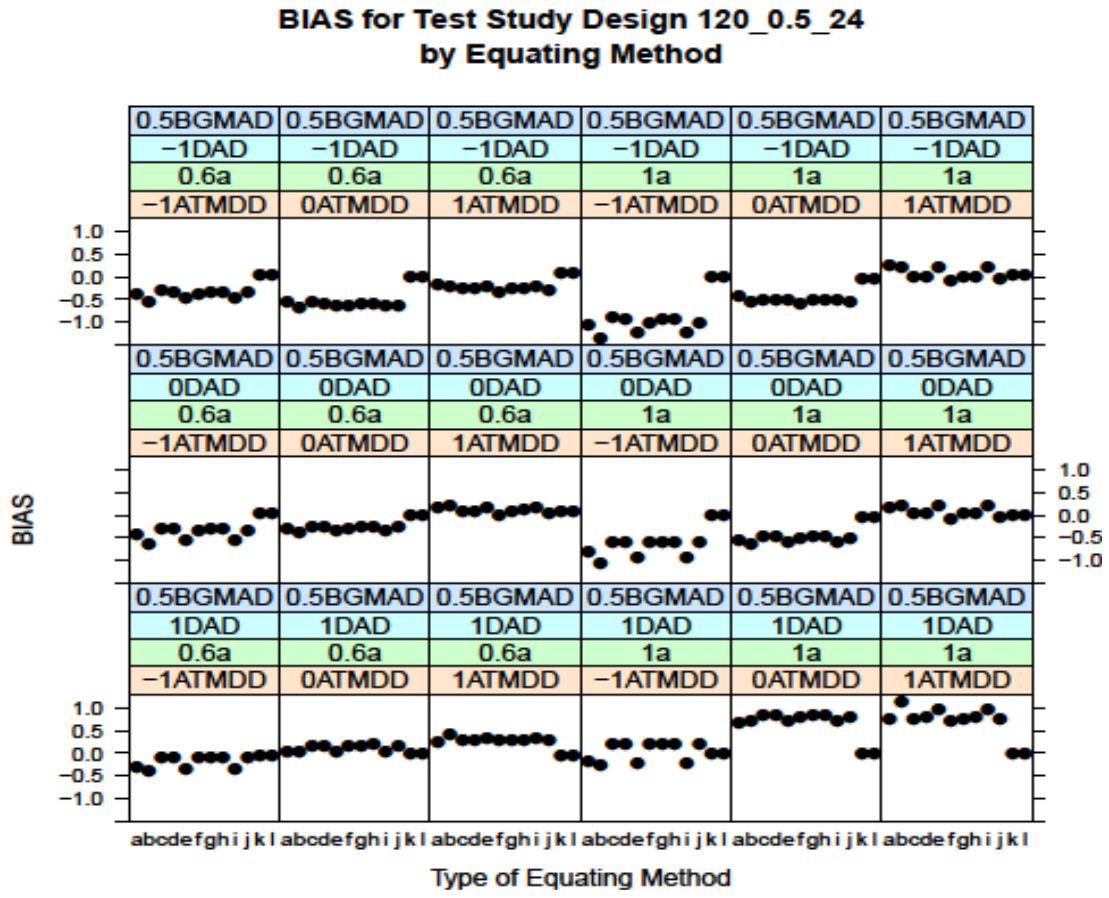

Figure 4.13. Bias for Test Study Design 120_0.5_24 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

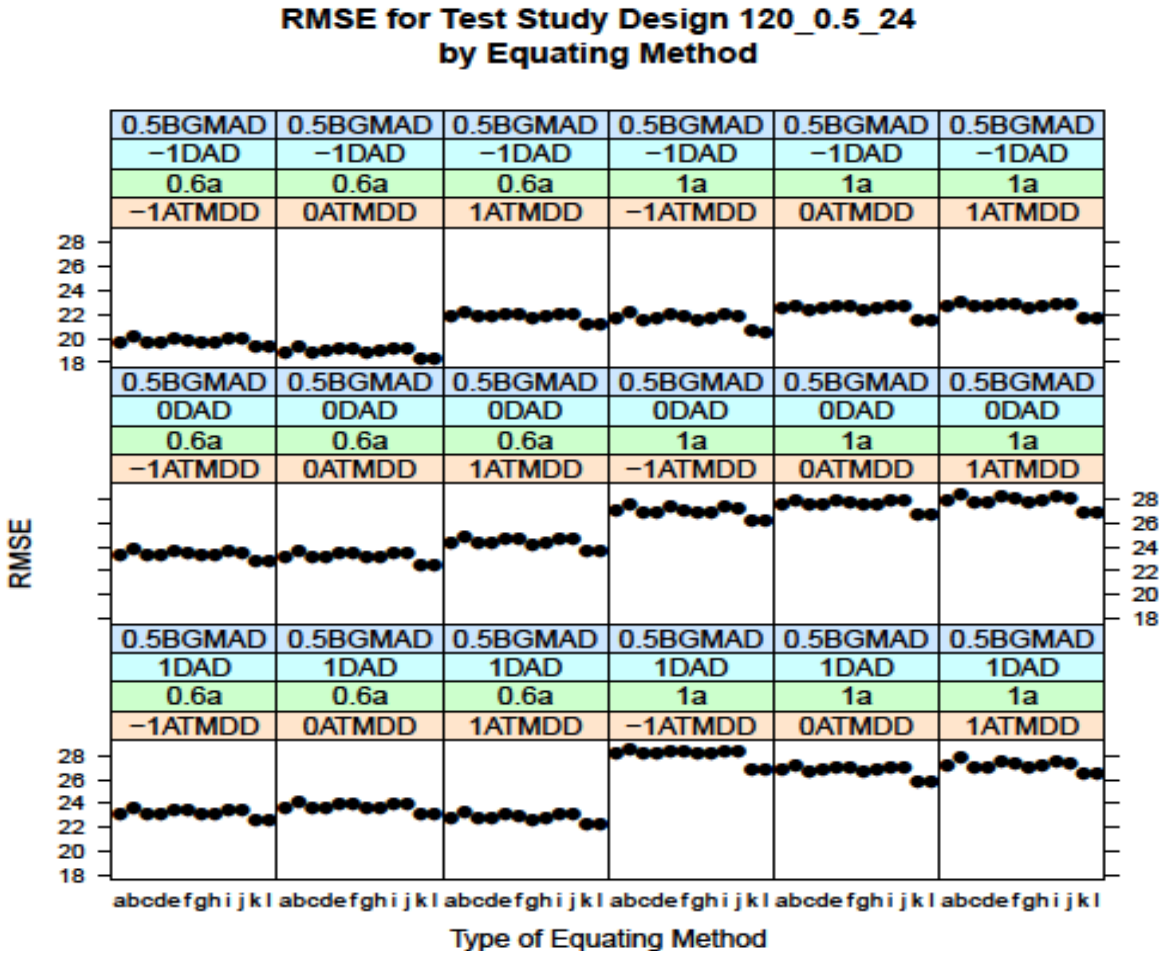

Figure 4.14. Root Mean Square Error (RMSE) for Test Study Design 120_0.5_24 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.2.8 120_1.0_24 Test Study Design

In this subsection, the results for test study design 120_1.0_24 are presented. This design had 120 total items and 24 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of

group separation (or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large respectively. Large (1.5) BGMAD will be discussed in the subsequent subsection. This subsection mainly focuses on medium (1.0) BGMAD, which means 120 total items and 24 anchor test items were held constant and the other three study conditions were varied. Tables A64-A72 in Appendix A display average descriptive statistics and Figure B.8 in Appendix B shows the standard error of equating (SEE) for test study design 120_1.0_24. Table 4.8 represents bias, SEE, and RMSE for this test study design. Figure 4.15 demonstrates that there is zero bias for all conditions under linear and equipercentile equating methods for medium (1.0) between-grade mean ability difference (BGMAD). Other equating methods show both negative and positive values of bias which are very close to zero apart from a few study conditions where results are inconsistent. For example, when distribution of ability difference (DAD) and anchor test mean difficulty difference (ATMDD) are below average (-1), average (0), and above average (1) and item discrimination is moderate (0.6) the equating methods show inconsistency. This pattern of inconsistency is also repeated when item discrimination is high (1) where DAD is below average (-1), average (0), and above average (1). However, where BGMDD is medium (1), DAD is above average (1), item discrimination is high (1), and ATMDD is average (0) the equating methods perform similarly with bias about zero with other conditions resulting to negative values. Similarly, where BGMDD is medium (1), DAD is above average (1), item discrimination is high (1), and ATMDD is above average (1), the equating methods yield similar bias results, which show positive bias values.

Figure 4.16 shows test study design 120_1.0_24, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (120), medium (1.0) BGMAD, moderate (0.6) and high (1) item discriminations are invariant. The RMSE values fall between 15 and 34. When all conditions are held constant and manipulate item discrimination (moderate versus high), RMSE displays a clear consistency where moderate (0.6) item discrimination across the other four study conditions results in smaller (more accurate) RMSE values than RMSE values for high (1) item discrimination varied over the other four study conditions. There are two patterns in this design based on conditions varied under either moderate item discrimination or high item discrimination with the former performing better than the latter in terms of accuracy.

Addressing the research question number 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.15 revealed the following results. First, the bias was consistent and very close to zero for all equating methods when medium (1.0) BGMAD for moderate (0.6) item discrimination and DAD was below average (-1) and average (0), and when ATMDD was average (0) and item discrimination was high (1). Second, the equating methods performed similarly when BGMAD was medium (1.0) for below (-1) average and average (0) DAD with moderate (0.6) item discrimination for above (1) average ATMDD. The results for the rest of the remaining study conditions under this design can be considered inaccurate and perhaps underestimated or overestimated accuracy for all equating methods, because of negative and positive bias values.

Overall the equating results in Figure 4.16 show that the smallest difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods was under below average (-1), average (0), and above average (1) DAD when item discrimination was moderate (0.6) and ATMDD below average (-1) conditions. However, largest difference between anchor test mean difficulty and the other four study conditions was when DAD varied across its three levels with a high (1) item discrimination and ATMDD was below average (-1). Moderate (0.6) item discrimination produced rather more accurate overall results, when considering RMSE values, than high (1) item discrimination under all conditions across all equating methods.

Table 4.8

BIAS, SEE, and RMSE Statistics for Test Study Design 120_1.0_24 by Equating Method Under All Conditions

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
64	120	1	-1	0.6	-1	Tucker	-1.63	16.13	17.47
64	120	1	-1	0.6	-1	Levine True	-2.57	16.40	19.35
64	120	1	-1	0.6	-1	Braun	-1.49	15.95	17.24
64	120	1	-1	0.6	-1	FEEE	-1.51	16.00	17.33
64	120	1	-1	0.6	-1	Chain_L	-2.14	16.34	18.47
64	120	1	-1	0.6	-1	Chain_E	-1.83	15.94	18.05
64	120	1	-1	0.6	-1	keNEATPSE_L	-1.48	15.86	17.18
64	120	1	-1	0.6	-1	keNEATPSE_E	-1.48	15.98	17.29
64	120	1	-1	0.6	-1	keNEATCE_L	-2.14	16.34	18.47
64	120	1	-1	0.6	-1	keNEATCE_E	-1.80	16.04	18.16
64	120	1	-1	0.6	-1	Linear	0.02	15.57	15.57
64	120	1	-1	0.6	-1	Equipercentile	0.02	15.58	15.58
145	120	1	-1	1	-1	Tucker	-2.95	17.20	18.80

Table 4.8

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
145	120	1	-1	1	-1	Levine True	-3.94	18.12	20.81
145	120	1	-1	1	-1	Braun	-2.60	16.59	18.19
145	120	1	-1	1	-1	FEEE	-2.70	16.76	18.41
145	120	1	-1	1	-1	Chain_L	-3.49	17.44	19.63
145	120	1	-1	1	-1	Chain_E	-3.04	16.64	18.86
145	120	1	-1	1	-1	keNEATPSE_L	-2.68	16.59	18.25
145	120	1	-1	1	-1	keNEATPSE_E	-2.68	16.77	18.41
145	120	1	-1	1	-1	keNEATCE_L	-3.49	17.44	19.62
145	120	1	-1	1	-1	keNEATCE_E	-3.03	16.73	18.97
145	120	1	-1	1	-1	Linear	0.01	16.36	16.35
145	120	1	-1	1	-1	Equipercentile	0.02	16.36	16.35
65	120	1	-1	0.6	0	Tucker	-1.29	16.80	18.83
65	120	1	-1	0.6	0	Levine True	-1.73	16.73	20.11
65	120	1	-1	0.6	0	Braun	-1.46	16.89	18.84
65	120	1	-1	0.6	0	FEEE	-1.52	16.96	18.96
65	120	1	-1	0.6	0	Chain_L	-1.56	16.78	19.59
65	120	1	-1	0.6	0	Chain_E	-1.77	16.96	19.72
65	120	1	-1	0.6	0	keNEATPSE_L	-1.54	16.89	18.89
65	120	1	-1	0.6	0	keNEATPSE_E	-1.53	17.00	18.99
65	120	1	-1	0.6	0	keNEATCE_L	-1.56	16.78	19.59
65	120	1	-1	0.6	0	keNEATCE_E	-1.72	17.10	19.90
65	120	1	-1	0.6	0	Linear	0.01	16.69	16.69
65	120	1	-1	0.6	0	Equipercentile	0.01	16.69	16.69
146	120	1	-1	1	0	Tucker	-2.63	18.72	21.42
146	120	1	-1	1	0	Levine True	-3.05	18.61	22.20
146	120	1	-1	1	0	Braun	-2.63	18.78	21.42
146	120	1	-1	1	0	FEEE	-2.69	18.85	21.51
146	120	1	-1	1	0	Chain_L	-2.89	18.67	21.91
146	120	1	-1	1	0	Chain_E	-2.87	18.82	22.05
146	120	1	-1	1	0	keNEATPSE_L	-2.74	18.86	21.54
146	120	1	-1	1	0	keNEATPSE_E	-2.73	18.91	21.58
146	120	1	-1	1	0	keNEATCE_L	-2.89	18.67	21.91
146	120	1	-1	1	0	keNEATCE_E	-2.85	18.89	22.15

Table 4.8

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
146	120	1	-1	1	0	Linear	0.02	18.80	18.80
146	120	1	-1	1	0	Equipercentile	0.02	18.80	18.79
66	120	1	-1	0.6	1	Tucker	-0.35	17.82	20.26
66	120	1	-1	0.6	1	Levine True	-0.56	17.61	21.47
66	120	1	-1	0.6	1	Braun	-0.67	17.92	20.32
66	120	1	-1	0.6	1	FEEE	-0.67	17.93	20.35
66	120	1	-1	0.6	1	Chain_L	-0.49	17.70	21.01
66	120	1	-1	0.6	1	Chain_E	-0.93	17.92	21.15
66	120	1	-1	0.6	1	keNEATPSE_L	-0.70	17.84	20.27
66	120	1	-1	0.6	1	keNEATPSE_E	-0.67	17.95	20.36
66	120	1	-1	0.6	1	keNEATCE_L	-0.49	17.70	21.01
66	120	1	-1	0.6	1	keNEATCE_E	-0.85	17.98	21.24
66	120	1	-1	0.6	1	Linear	0.03	18.20	18.20
66	120	1	-1	0.6	1	Equipercentile	0.03	18.20	18.20
147	120	1	-1	1	1	Tucker	-1.21	18.37	21.09
147	120	1	-1	1	1	Levine True	-1.48	18.32	22.20
147	120	1	-1	1	1	Braun	-1.71	18.41	21.03
147	120	1	-1	1	1	FEEE	-1.77	18.49	21.14
147	120	1	-1	1	1	Chain_L	-1.38	18.34	21.78
147	120	1	-1	1	1	Chain_E	-2.02	18.48	21.79
147	120	1	-1	1	1	keNEATPSE_L	-1.85	18.40	21.06
147	120	1	-1	1	1	keNEATPSE_E	-1.79	18.57	21.22
147	120	1	-1	1	1	keNEATCE_L	-1.38	18.35	21.78
147	120	1	-1	1	1	keNEATCE_E	-1.93	18.55	21.88
147	120	1	-1	1	1	Linear	0.02	18.49	18.49
147	120	1	-1	1	1	Equipercentile	0.02	18.49	18.49
67	120	1	0	0.6	-1	Tucker	-1.67	20.99	23.30
67	120	1	0	0.6	-1	Levine True	-2.40	20.88	25.14
67	120	1	0	0.6	-1	Braun	-1.37	20.67	22.95
67	120	1	0	0.6	-1	FEEE	-1.42	20.72	23.04
67	120	1	0	0.6	-1	Chain_L	-2.08	20.96	24.31
67	120	1	0	0.6	-1	Chain_E	-1.58	20.57	23.92
67	120	1	0	0.6	-1	keNEATPSE_L	-1.41	20.52	22.85

Table 4.8

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
67	120	1	0	0.6	-1	keNEATPSE_E	-1.40	20.73	23.04
67	120	1	0	0.6	-1	keNEATCE_L	-2.08	20.96	24.31
67	120	1	0	0.6	-1	keNEATCE_E	-1.55	20.67	24.04
67	120	1	0	0.6	-1	Linear	0.03	20.91	20.90
67	120	1	0	0.6	-1	Equipercentile	0.03	20.91	20.90
148	120	1	0	1	-1	Tucker	-2.03	24.13	27.53
148	120	1	0	1	-1	Levine True	-2.48	23.91	29.08
148	120	1	0	1	-1	Braun	-1.68	23.71	27.09
148	120	1	0	1	-1	FEEE	-1.75	23.76	27.17
148	120	1	0	1	-1	Chain_L	-2.28	24.03	28.38
148	120	1	0	1	-1	Chain_E	-1.90	23.64	27.99
148	120	1	0	1	-1	keNEATPSE_L	-1.79	23.53	26.96
148	120	1	0	1	-1	keNEATPSE_E	-1.75	23.78	27.18
148	120	1	0	1	-1	keNEATCE_L	-2.28	24.03	28.38
148	120	1	0	1	-1	keNEATCE_E	-1.87	23.73	28.12
148	120	1	0	1	-1	Linear	-0.04	24.42	24.42
148	120	1	0	1	-1	Equipercentile	-0.04	24.41	24.41
68	120	1	0	0.6	0	Tucker	-1.17	20.31	22.72
68	120	1	0	0.6	0	Levine True	-1.61	20.15	24.48
68	120	1	0	0.6	0	Braun	-1.04	20.15	22.53
68	120	1	0	0.6	0	FEEE	-1.05	20.16	22.57
68	120	1	0	0.6	0	Chain_L	-1.42	20.24	23.71
68	120	1	0	0.6	0	Chain_E	-1.19	20.07	23.53
68	120	1	0	0.6	0	keNEATPSE_L	-1.05	19.98	22.40
68	120	1	0	0.6	0	keNEATPSE_E	-1.04	20.18	22.59
68	120	1	0	0.6	0	keNEATCE_L	-1.42	20.24	23.71
68	120	1	0	0.6	0	keNEATCE_E	-1.15	20.16	23.64
68	120	1	0	0.6	0	Linear	-0.04	20.54	20.54
68	120	1	0	0.6	0	Equipercentile	-0.05	20.54	20.54
149	120	1	0	1	0	Tucker	-1.55	24.84	28.83
149	120	1	0	1	0	Levine True	-1.75	24.49	29.85
149	120	1	0	1	0	Braun	-1.64	24.76	28.72
149	120	1	0	1	0	FEEE	-1.67	24.76	28.74

Table 4.8

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
149	120	1	0	1	0	Chain_L	-1.67	24.63	29.45
149	120	1	0	1	0	Chain_E	-1.93	24.73	29.49
149	120	1	0	1	0	keNEATPSE_L	-1.72	24.63	28.62
149	120	1	0	1	0	keNEATPSE_E	-1.67	24.79	28.76
149	120	1	0	1	0	keNEATCE_L	-1.67	24.63	29.45
149	120	1	0	1	0	keNEATCE_E	-1.86	24.82	29.62
149	120	1	0	1	0	Linear	-0.05	25.87	25.87
149	120	1	0	1	0	Equipercentile	-0.05	25.87	25.87
69	120	1	0	0.6	1	Tucker	-0.68	21.96	24.62
69	120	1	0	0.6	1	Levine True	-0.88	21.66	26.23
69	120	1	0	0.6	1	Braun	-0.81	21.85	24.48
69	120	1	0	0.6	1	FEEE	-0.79	21.84	24.49
69	120	1	0	0.6	1	Chain_L	-0.80	21.80	25.54
69	120	1	0	0.6	1	Chain_E	-0.97	21.75	25.46
69	120	1	0	0.6	1	keNEATPSE_L	-0.81	21.68	24.34
69	120	1	0	0.6	1	keNEATPSE_E	-0.79	21.87	24.51
69	120	1	0	0.6	1	keNEATCE_L	-0.80	21.80	25.54
69	120	1	0	0.6	1	keNEATCE_E	-0.90	21.82	25.56
69	120	1	0	0.6	1	Linear	0.02	22.27	22.27
69	120	1	0	0.6	1	Equipercentile	0.02	22.27	22.27
150	120	1	0	1	1	Tucker	-0.18	25.79	29.77
150	120	1	0	1	1	Levine True	-0.07	25.46	31.19
150	120	1	0	1	1	Braun	-1.00	25.56	29.49
150	120	1	0	1	1	FEEE	-1.01	25.54	29.48
150	120	1	0	1	1	Chain_L	-0.11	25.60	30.62
150	120	1	0	1	1	Chain_E	-1.41	25.50	30.35
150	120	1	0	1	1	keNEATPSE_L	-1.07	25.38	29.33
150	120	1	0	1	1	keNEATPSE_E	-1.01	25.58	29.51
150	120	1	0	1	1	keNEATCE_L	-0.11	25.60	30.62
150	120	1	0	1	1	keNEATCE_E	-1.30	25.54	30.40
150	120	1	0	1	1	Linear	0.01	26.44	26.44
150	120	1	0	1	1	Equipercentile	0.01	26.44	26.43
70	120	1	1	0.6	-1	Tucker	-0.70	22.42	25.26

Table 4.8

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
70	120	1	1	0.6	-1	Levine True	-1.02	22.12	27.05
70	120	1	1	0.6	-1	Braun	-0.18	22.19	25.06
70	120	1	1	0.6	-1	FEEE	-0.14	22.16	25.05
70	120	1	1	0.6	-1	Chain_L	-0.89	22.27	26.26
70	120	1	1	0.6	-1	Chain_E	-0.15	22.05	26.04
70	120	1	1	0.6	-1	keNEATPSE_L	-0.14	22.00	24.90
70	120	1	1	0.6	-1	keNEATPSE_E	-0.14	22.19	25.07
70	120	1	1	0.6	-1	keNEATCE_L	-0.89	22.27	26.26
70	120	1	1	0.6	-1	keNEATCE_E	-0.16	22.14	26.14
70	120	1	1	0.6	-1	Linear	0.00	22.75	22.74
70	120	1	1	0.6	-1	Equipercentile	0.00	22.74	22.74
151	120	1	1	1	-1	Tucker	-1.74	26.78	31.11
151	120	1	1	1	-1	Levine True	-2.05	26.36	32.38
151	120	1	1	1	-1	Braun	-0.70	26.41	30.85
151	120	1	1	1	-1	FEEE	-0.72	26.39	30.85
151	120	1	1	1	-1	Chain_L	-1.93	26.54	31.86
151	120	1	1	1	-1	Chain_E	-0.74	26.31	31.67
151	120	1	1	1	-1	keNEATPSE_L	-0.71	26.26	30.74
151	120	1	1	1	-1	keNEATPSE_E	-0.71	26.43	30.88
151	120	1	1	1	-1	keNEATCE_L	-1.93	26.54	31.86
151	120	1	1	1	-1	keNEATCE_E	-0.74	26.38	31.75
151	120	1	1	1	-1	Linear	-0.02	27.92	27.92
151	120	1	1	1	-1	Equipercentile	-0.03	27.92	27.91
71	120	1	1	0.6	0	Tucker	0.29	23.09	26.30
71	120	1	1	0.6	0	Levine True	0.30	22.64	27.73
71	120	1	1	0.6	0	Braun	0.45	23.01	26.21
71	120	1	1	0.6	0	FEEE	0.46	22.99	26.20
71	120	1	1	0.6	0	Chain_L	0.30	22.84	27.14
71	120	1	1	0.6	0	Chain_E	0.32	22.88	27.17
71	120	1	1	0.6	0	keNEATPSE_L	0.45	22.87	26.09
71	120	1	1	0.6	0	keNEATPSE_E	0.46	23.01	26.22
71	120	1	1	0.6	0	keNEATCE_L	0.30	22.84	27.14
71	120	1	1	0.6	0	keNEATCE_E	0.37	22.95	27.26

Table 4.8

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
71	120	1	1	0.6	0	Linear	-0.05	23.79	23.78
71	120	1	1	0.6	0	Equipercentile	-0.05	23.78	23.78
152	120	1	1	1	0	Tucker	0.35	26.31	30.54
152	120	1	1	1	0	Levine True	0.42	25.90	31.75
152	120	1	1	1	0	Braun	0.45	26.21	30.44
152	120	1	1	1	0	FEEE	0.44	26.18	30.42
152	120	1	1	1	0	Chain_L	0.39	26.07	31.27
152	120	1	1	1	0	Chain_E	0.18	26.10	31.28
152	120	1	1	1	0	keNEATPSE_L	0.42	26.04	30.29
152	120	1	1	1	0	keNEATPSE_E	0.44	26.20	30.43
152	120	1	1	1	0	keNEATCE_L	0.39	26.07	31.27
152	120	1	1	1	0	keNEATCE_E	0.24	26.18	31.38
152	120	1	1	1	0	Linear	0.04	27.37	27.37
152	120	1	1	1	0	Equipercentile	0.04	27.37	27.36
72	120	1	1	0.6	1	Tucker	1.32	23.04	25.79
72	120	1	1	0.6	1	Levine True	1.73	22.70	27.61
72	120	1	1	0.6	1	Braun	1.36	22.87	25.61
72	120	1	1	0.6	1	FEEE	1.44	22.86	25.62
72	120	1	1	0.6	1	Chain_L	1.56	22.86	26.83
72	120	1	1	0.6	1	Chain_E	1.24	22.72	26.65
72	120	1	1	0.6	1	keNEATPSE_L	1.41	22.68	25.46
72	120	1	1	0.6	1	keNEATPSE_E	1.44	22.89	25.64
72	120	1	1	0.6	1	keNEATCE_L	1.56	22.86	26.83
72	120	1	1	0.6	1	keNEATCE_E	1.33	22.82	26.78
72	120	1	1	0.6	1	Linear	0.02	23.43	23.43
72	120	1	1	0.6	1	Equipercentile	0.02	23.43	23.43
153	120	1	1	1	1	Tucker	1.06	26.93	30.68
153	120	1	1	1	1	Levine True	1.53	26.52	32.13
153	120	1	1	1	1	Braun	0.78	26.57	30.31
153	120	1	1	1	1	FEEE	0.85	26.53	30.27
153	120	1	1	1	1	Chain_L	1.34	26.70	31.54
153	120	1	1	1	1	Chain_E	0.45	26.39	31.15
153	120	1	1	1	1	keNEATPSE_L	0.80	26.37	30.14

Table 4.8

Cont.

Panel	Study Condition: Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	Statistic: SEE	RMSE
153	120	1	1	1	1	keNEATPSE_E	0.85	26.56	30.29
153	120	1	1	1	1	keNEATCE_L	1.34	26.70	31.54
153	120	1	1	1	1	keNEATCE_E	0.58	26.48	31.25
153	120	1	1	1	1	Linear	-0.04	27.70	27.69
153	120	1	1	1	1	Equipercentile	-0.05	27.69	27.68

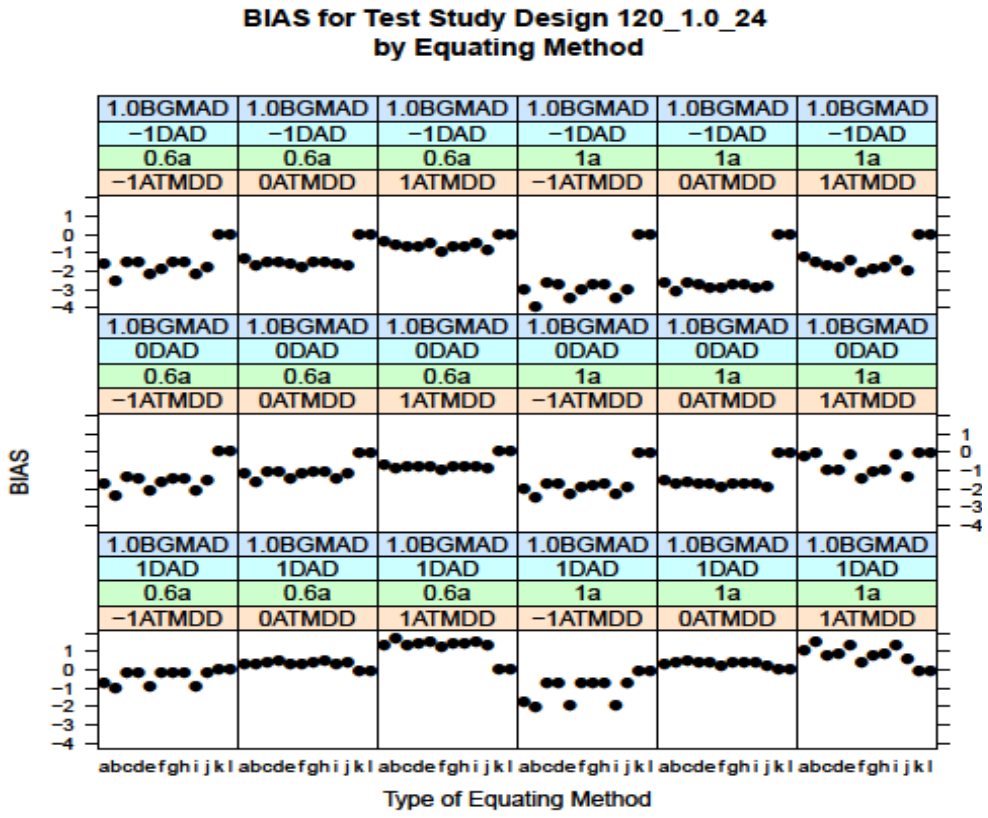

Figure 4.15. Bias for Test Study Design 120_1.0_24 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

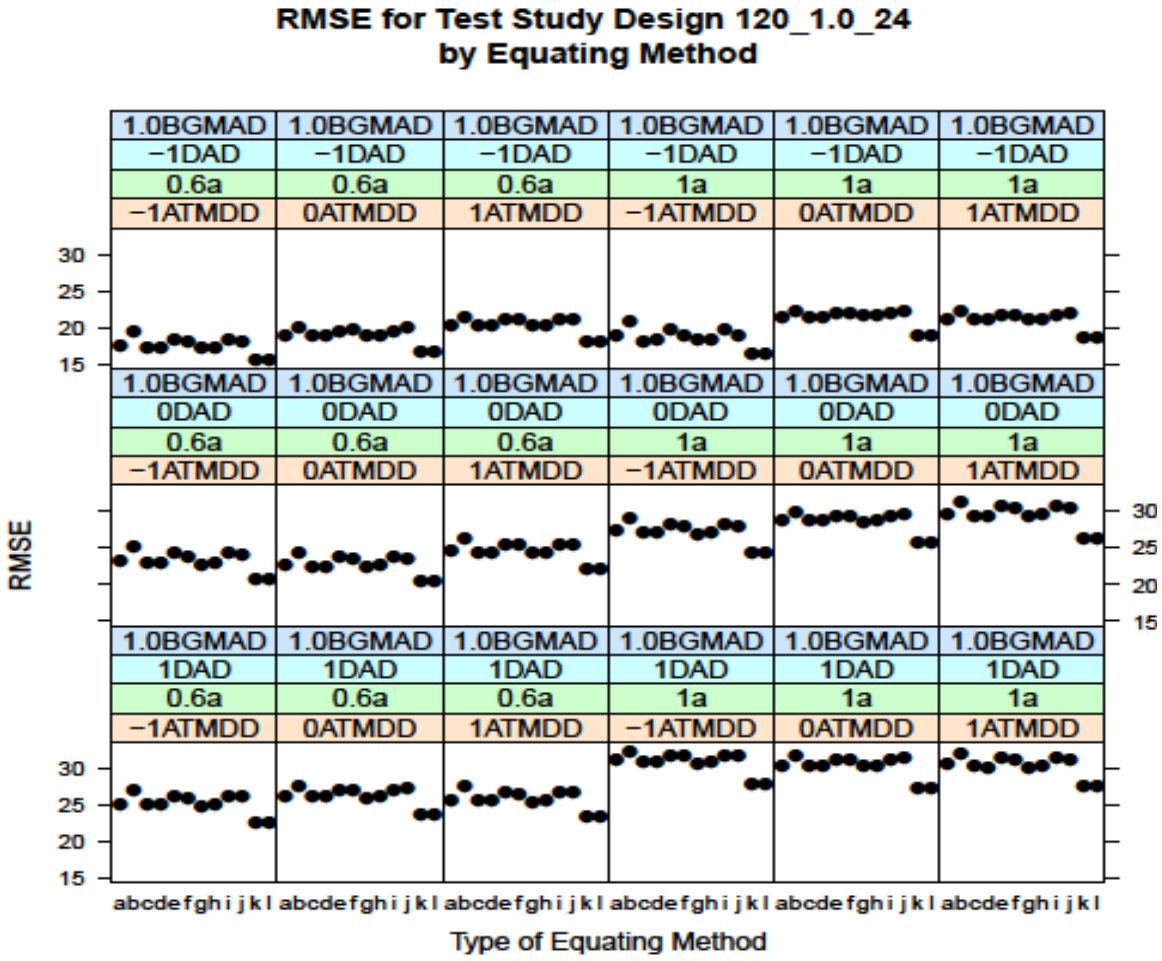

Figure 4.16. Root Mean Square Error (RMSE) for Test Study Design 120_1.0_24 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.2.9 120_1.5_24 Test Study Design

This subsection presents the results for the last test study design 120_1.5_24. This design had 120 total items and 24 anchor test items. A proportion of 20% of the total test items was used to obtain the total anchor test items. As previously discussed in subsection 3.4 (c), between-grade mean ability difference (BGMAD) or magnitude of

group separation (or group effect) had three levels—0.5, 1.0, and 1.5 labeled as small, medium, and large respectively. The previous two subsections discussed small and medium BGMAD. This subsection mainly zero in on large BGMAD, which means 120 total items and 24 anchor test items were held constant and the other three study conditions were varied. Tables A73-A81 in Appendix A display average descriptive statistics and Figure B.9 in Appendix B shows the standard error of equating (SEE) for test study design 120_1.5_24. Table 4.9 displays bias, SEE, and RMSE for this test study design. Figure 4.17 demonstrates that there is negative bias for all conditions under all equating methods for large (1.5) between-grade mean ability difference (BGMAD) except for positive bias when BGMAD is large (1.5) and DAD is above average (1), item discrimination is moderate (0.6) and high (1) and ATMDD is above average (1). Both negative and positive values of bias are almost zero for a few study conditions. For instance, there was negative bias which was near zero when BGMDD was large (1.5) while DAD was above average (1) and item discrimination was moderate (0.6) and ATMDD was average (0). However, when the same conditions are repeated under high (1) item discrimination (and to some extent under moderate item discrimination), only linear and equipercentile equating methods produce zero bias.

Figure 4.18 shows test study design 120_1.5_24, amount of root mean square error (RMSE) for multiple study conditions under each equating method when test length (120), large (1.5) BGMAD, moderate (0.6) and high (1) item discriminations are unchanging factors. The RMSE values fall between 14 and 38. When all conditions are held constant and manipulate item discrimination (moderate versus high), RMSE was

somewhat consistent for both moderate (0.6) and high (1) conditions. Also, the RMSE values for both moderate (0.6) and high (1) item discrimination when other conditions are varied performed similarly. In comparison, conditions manipulated under moderate (0.6) item discrimination produced much smaller (more accurate) RMSE values than its counterpart, high (1) item discrimination.

Addressing the Research Question 2 (How much difference between anchor test difficulty and the other four study conditions can be tolerated under each equating method?) and Figure 4.17 revealed that difference between anchor test difficulty and the other four study conditions is smallest when large (1.5) BGMAD, moderate (0.6) item discrimination, and above average (1) ATMDD are held constant and then grade-to-grade ability variability (DAD) is manipulated—i.e., across its three levels (below average, average, and above average). Similarly, under the above conditions, average (0) ATMDD also produced bias values close to zero. This means that a long (120) test length, a large (1.5) BGMAD with average (0) and above average (1) ATMDD conditioned on above (1) different ability distribution (DAD) within a grade has the smallest bias (or best results) compared to other study conditions in this design. Other study conditions produced worst results. Therefore, there is sufficient evidence to hold the view that a large (1.5) BGMAD together with a long test (120 items) produced heterogeneous bias results across all study conditions under all equating methods. This conclusion was supported by the fact that when holding large (1.5) BGMAD, high (1) item discrimination, and below average (-1) ATMDD constant and vary DAD across its three levels, then the equating methods

produced the largest bias (or worst results) compared with the rest of the study conditions.

Overall the equating results in Figure 4.18 reveal that there was a small difference between anchor test mean difficulty and the other four study conditions in terms of the values of RMSE for all equating methods. That is, all equating methods seemed to have performed in different ways without any clear particular pattern. The worst part (RMSE values greater than 32) was for large (1.5) BGMAD, above average (1) DAD, high (1) item discrimination where ATMDD was varied across below average (-1), average (0), and above average (1) conditions. This means that other study conditions produced RMSE values less than 32 under various equating methods.

At this point, it is important to reflect on the third and last set of the three test study designs—the first set of the three test study design included 30_0.5_6, 30_1.0_6, and 30_1.5_6 and the second set of the three study design comprised 60_0.5_12, 60_1.0_12, and 60_1.5_12 as outlined in previous subsections—discussed so far. i.e., 120_0.5_24, 120_1.0_24, and 120_1.5_24, which have the same number of total test items or large test (120 items in total) and anchor test items (24 items) under all study conditions with variability in magnitude of the group separation [or BGMAD across small (0.5), medium (1.0), and large (1.5)]. Comparing and contrasting the last set of the three test study designs—according to their magnitude of the group separation—(Figures 4.14, 4.16, and 4.18) in terms of RMSE values leads to the conclusion that the overall accuracy (or stability) of the results is substantially affected by the magnitude of group separation/group effect (or mean ability difference between adjacent grades/BGMAD). In

terms of degree of accuracy of the result, small (0.5) BGMAD produced smallest RMSE values while large (1.5) BGMAD produced largest RMSE values under all study conditions. Also, small (0.5) BGMAD under all study conditions had the smallest bias while large (1.5) BGMAD had the largest bias values.

Table 4.9

BIAS, SEE, and RMSE Statistics for Test Study Design 120_1.5_24 by Equating Method Under All Conditions

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
73	120	1.5	-1	0.6	-1	Tucker	-2.57	14.14	16.53
73	120	1.5	-1	0.6	-1	Levine True	-4.11	15.04	20.25
73	120	1.5	-1	0.6	-1	Braun	-2.68	14.16	16.45
73	120	1.5	-1	0.6	-1	FEEE	-2.73	14.25	16.60
73	120	1.5	-1	0.6	-1	Chain_L	-3.40	14.34	18.20
73	120	1.5	-1	0.6	-1	Chain_E	-3.43	14.18	17.96
73	120	1.5	-1	0.6	-1	keNEATPSE_L	-1.37	13.87	17.08
73	120	1.5	-1	0.6	-1	keNEATPSE_E	-1.63	14.01	17.00
73	120	1.5	-1	0.6	-1	keNEATCE_L	-2.62	14.30	18.70
73	120	1.5	-1	0.6	-1	keNEATCE_E	-2.80	14.39	18.66
73	120	1.5	-1	0.6	-1	Linear	0.00	13.73	13.73
73	120	1.5	-1	0.6	-1	Equipercentile	0.00	13.74	13.73
154	120	1.5	-1	1	-1	Tucker	-4.94	15.53	17.79
154	120	1.5	-1	1	-1	Levine True	-7.80	19.02	23.16
154	120	1.5	-1	1	-1	Braun	-4.10	14.06	16.17
154	120	1.5	-1	1	-1	FEEE	-3.81	14.47	17.07
154	120	1.5	-1	1	-1	Chain_L	-6.41	16.79	20.04
154	120	1.5	-1	1	-1	Chain_E	-5.23	14.64	17.73
154	120	1.5	-1	1	-1	keNEATPSE_L	-4.12	14.15	16.33
154	120	1.5	-1	1	-1	keNEATPSE_E	-4.19	14.32	16.48
154	120	1.5	-1	1	-1	keNEATCE_L	-6.39	16.79	20.04
154	120	1.5	-1	1	-1	keNEATCE_E	-5.26	14.69	17.84
154	120	1.5	-1	1	-1	Linear	0.01	12.52	12.51
154	120	1.5	-1	1	-1	Equipercentile	0.01	12.51	12.51

Table 4.9

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
74	120	1.5	-1	0.6	0	Tucker	-3.22	14.13	17.30
74	120	1.5	-1	0.6	0	Levine True	-4.48	14.39	19.83
74	120	1.5	-1	0.6	0	Braun	-3.17	14.08	17.11
74	120	1.5	-1	0.6	0	FEEE	-3.34	14.28	17.42
74	120	1.5	-1	0.6	0	Chain_L	-3.98	14.23	18.73
74	120	1.5	-1	0.6	0	Chain_E	-3.89	14.21	18.67
74	120	1.5	-1	0.6	0	keNEATPSE_L	-3.29	14.10	17.22
74	120	1.5	-1	0.6	0	keNEATPSE_E	-3.27	14.28	17.36
74	120	1.5	-1	0.6	0	keNEATCE_L	-3.98	14.23	18.72
74	120	1.5	-1	0.6	0	keNEATCE_E	-3.86	14.48	18.97
74	120	1.5	-1	0.6	0	Linear	0.02	13.87	13.87
74	120	1.5	-1	0.6	0	Equipercentile	0.02	13.88	13.87
155	120	1.5	-1	1	0	Tucker	-4.20	13.71	17.44
155	120	1.5	-1	1	0	Levine True	-5.03	13.70	18.82
155	120	1.5	-1	1	0	Braun	-4.41	14.14	17.69
155	120	1.5	-1	1	0	FEEE	-4.67	14.42	18.09
155	120	1.5	-1	1	0	Chain_L	-4.73	13.73	18.34
155	120	1.5	-1	1	0	Chain_E	-5.12	14.42	18.97
155	120	1.5	-1	1	0	keNEATPSE_L	-4.82	15.00	18.92
155	120	1.5	-1	1	0	keNEATPSE_E	-4.84	15.19	18.96
155	120	1.5	-1	1	0	keNEATCE_L	-4.78	13.71	18.29
155	120	1.5	-1	1	0	keNEATCE_E	-5.13	14.86	19.49
155	120	1.5	-1	1	0	Linear	0.02	13.62	13.62
155	120	1.5	-1	1	0	Equipercentile	0.02	13.65	13.65
75	120	1.5	-1	0.6	1	Tucker	-2.03	14.80	18.90
75	120	1.5	-1	0.6	1	Levine True	-2.59	14.39	20.80
75	120	1.5	-1	0.6	1	Braun	-2.68	15.16	19.19
75	120	1.5	-1	0.6	1	FEEE	-2.75	15.22	19.31
75	120	1.5	-1	0.6	1	Chain_L	-2.40	14.56	20.14
75	120	1.5	-1	0.6	1	Chain_E	-3.32	15.23	20.71
75	120	1.5	-1	0.6	1	keNEATPSE_L	-2.80	15.17	19.29
75	120	1.5	-1	0.6	1	keNEATPSE_E	-2.76	15.34	19.39
75	120	1.5	-1	0.6	1	keNEATCE_L	-2.40	14.56	20.14

Table 4.9

Cont.

Study Condition								Statistic	
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
75	120	1.5	-1	0.6	1	keNEATCE_E	-3.25	15.43	20.96
75	120	1.5	-1	0.6	1	Linear	0.04	15.60	15.60
75	120	1.5	-1	0.6	1	Equipercentile	0.04	15.60	15.60
156	120	1.5	-1	1	1	Tucker	-3.31	15.11	19.81
156	120	1.5	-1	1	1	Levine True	-3.85	14.85	21.01
156	120	1.5	-1	1	1	Braun	-4.16	15.90	20.53
156	120	1.5	-1	1	1	FEEE	-4.32	16.03	20.74
156	120	1.5	-1	1	1	Chain_L	-3.69	14.95	20.66
156	120	1.5	-1	1	1	Chain_E	-4.81	16.10	21.73
156	120	1.5	-1	1	1	keNEATPSE_L	-4.68	16.42	21.26
156	120	1.5	-1	1	1	keNEATPSE_E	-4.63	16.62	21.34
156	120	1.5	-1	1	1	keNEATCE_L	-3.69	14.98	20.68
156	120	1.5	-1	1	1	keNEATCE_E	-4.79	16.30	21.95
156	120	1.5	-1	1	1	Linear	0.01	15.60	15.60
156	120	1.5	-1	1	1	Equipercentile	0.01	15.63	15.63
76	120	1.5	0	0.6	-1	Tucker	-3.03	19.32	23.42
76	120	1.5	0	0.6	-1	Levine True	-4.53	19.11	26.88
76	120	1.5	0	0.6	-1	Braun	-2.96	19.07	23.07
76	120	1.5	0	0.6	-1	FEEE	-3.03	19.16	23.25
76	120	1.5	0	0.6	-1	Chain_L	-3.86	19.29	25.29
76	120	1.5	0	0.6	-1	Chain_E	-3.57	18.96	24.90
76	120	1.5	0	0.6	-1	keNEATPSE_L	-3.01	18.75	22.88
76	120	1.5	0	0.6	-1	keNEATPSE_E	-2.97	19.16	23.21
76	120	1.5	0	0.6	-1	keNEATCE_L	-3.86	19.29	25.29
76	120	1.5	0	0.6	-1	keNEATCE_E	-3.51	19.22	25.24
76	120	1.5	0	0.6	-1	Linear	0.02	19.25	19.25
76	120	1.5	0	0.6	-1	Equipercentile	0.02	19.24	19.24
157	120	1.5	0	1	-1	Tucker	-6.20	20.85	26.19
157	120	1.5	0	1	-1	Levine True	-7.73	21.91	29.51
157	120	1.5	0	1	-1	Braun	-5.21	19.89	25.10
157	120	1.5	0	1	-1	FEEE	-5.46	20.20	25.56
157	120	1.5	0	1	-1	Chain_L	-7.07	20.97	27.67
157	120	1.5	0	1	-1	Chain_E	-5.85	19.96	26.64

Table 4.9

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
157	120	1.5	0	1	-1	keNEATPSE_L	-5.41	19.91	25.32
157	120	1.5	0	1	-1	keNEATPSE_E	-5.40	20.18	25.51
157	120	1.5	0	1	-1	keNEATCE_L	-7.08	20.97	27.67
157	120	1.5	0	1	-1	keNEATCE_E	-5.86	20.14	26.85
157	120	1.5	0	1	-1	Linear	0.03	20.01	20.01
157	120	1.5	0	1	-1	Equipercentile	0.03	20.02	20.01
77	120	1.5	0	0.6	0	Tucker	-3.18	18.47	23.48
77	120	1.5	0	0.6	0	Levine True	-4.07	18.03	25.93
77	120	1.5	0	0.6	0	Braun	-3.23	18.42	23.36
77	120	1.5	0	0.6	0	FEEE	-3.27	18.48	23.49
77	120	1.5	0	0.6	0	Chain_L	-3.72	18.24	24.95
77	120	1.5	0	0.6	0	Chain_E	-3.71	18.31	25.00
77	120	1.5	0	0.6	0	keNEATPSE_L	-3.28	18.26	23.29
77	120	1.5	0	0.6	0	keNEATPSE_E	-3.24	18.51	23.48
77	120	1.5	0	0.6	0	keNEATCE_L	-3.72	18.24	24.95
77	120	1.5	0	0.6	0	keNEATCE_E	-3.63	18.50	25.24
77	120	1.5	0	0.6	0	Linear	0.03	19.25	19.25
77	120	1.5	0	0.6	0	Equipercentile	0.03	19.25	19.25
158	120	1.5	0	1	0	Tucker	-4.91	21.70	28.33
158	120	1.5	0	1	0	Levine True	-5.91	21.44	30.37
158	120	1.5	0	1	0	Braun	-4.87	21.43	27.89
158	120	1.5	0	1	0	FEEE	-5.06	21.64	28.24
158	120	1.5	0	1	0	Chain_L	-5.51	21.57	29.56
158	120	1.5	0	1	0	Chain_E	-5.50	21.49	29.42
158	120	1.5	0	1	0	keNEATPSE_L	-5.09	21.34	27.99
158	120	1.5	0	1	0	keNEATPSE_E	-5.00	21.66	28.22
158	120	1.5	0	1	0	keNEATCE_L	-5.51	21.57	29.56
158	120	1.5	0	1	0	keNEATCE_E	-5.53	21.78	29.77
158	120	1.5	0	1	0	Linear	0.02	22.06	22.06
158	120	1.5	0	1	0	Equipercentile	0.02	22.06	22.06
78	120	1.5	0	0.6	1	Tucker	-1.45	19.32	24.74
78	120	1.5	0	0.6	1	Levine True	-1.78	18.70	27.16
78	120	1.5	0	0.6	1	Braun	-2.14	19.43	24.80

Table 4.9

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
78	120	1.5	0	0.6	1	FEEE	-2.16	19.43	24.84
78	120	1.5	0	0.6	1	Chain_L	-1.66	18.96	26.23
78	120	1.5	0	0.6	1	Chain_E	-2.71	19.34	26.50
78	120	1.5	0	0.6	1	keNEATPSE_L	-2.22	19.22	24.66
78	120	1.5	0	0.6	1	keNEATPSE_E	-2.15	19.47	24.84
78	120	1.5	0	0.6	1	keNEATCE_L	-1.66	18.96	26.23
78	120	1.5	0	0.6	1	keNEATCE_E	-2.56	19.44	26.64
78	120	1.5	0	0.6	1	Linear	-0.04	20.40	20.40
78	120	1.5	0	0.6	1	Equipercentile	-0.04	20.40	20.40
159	120	1.5	0	1	1	Tucker	-2.54	20.63	27.62
159	120	1.5	0	1	1	Levine True	-2.90	19.98	29.01
159	120	1.5	0	1	1	Braun	-3.70	21.05	28.00
159	120	1.5	0	1	1	FEEE	-3.67	21.09	28.10
159	120	1.5	0	1	1	Chain_L	-2.78	20.22	28.54
159	120	1.5	0	1	1	Chain_E	-4.30	21.13	29.39
159	120	1.5	0	1	1	keNEATPSE_L	-3.74	21.01	28.04
159	120	1.5	0	1	1	keNEATPSE_E	-3.67	21.19	28.13
159	120	1.5	0	1	1	keNEATCE_L	-2.78	20.22	28.54
159	120	1.5	0	1	1	keNEATCE_E	-4.27	21.33	29.61
159	120	1.5	0	1	1	Linear	0.05	22.56	22.56
159	120	1.5	0	1	1	Equipercentile	0.05	22.56	22.55
79	120	1.5	1	0.6	-1	Tucker	-2.19	21.57	27.51
79	120	1.5	1	0.6	-1	Levine True	-2.93	20.94	30.67
79	120	1.5	1	0.6	-1	Braun	-1.50	21.10	26.99
79	120	1.5	1	0.6	-1	FEEE	-1.58	21.13	27.10
79	120	1.5	1	0.6	-1	Chain_L	-2.63	21.26	29.30
79	120	1.5	1	0.6	-1	Chain_E	-1.83	20.87	28.85
79	120	1.5	1	0.6	-1	keNEATPSE_L	-1.57	20.76	26.76
79	120	1.5	1	0.6	-1	keNEATPSE_E	-1.53	21.17	27.09
79	120	1.5	1	0.6	-1	keNEATCE_L	-2.63	21.26	29.30
79	120	1.5	1	0.6	-1	keNEATCE_E	-1.80	21.09	29.15
79	120	1.5	1	0.6	-1	Linear	-0.07	22.46	22.46
79	120	1.5	1	0.6	-1	Equipercentile	-0.07	22.46	22.45

Table 4.9

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
160	120	1.5	1	1	-1	Tucker	-5.98	26.45	33.29
160	120	1.5	1	1	-1	Levine True	-7.85	26.56	37.40
160	120	1.5	1	1	-1	Braun	-3.71	24.59	31.42
160	120	1.5	1	1	-1	FEEE	-3.71	24.64	31.60
160	120	1.5	1	1	-1	Chain_L	-7.02	26.59	35.55
160	120	1.5	1	1	-1	Chain_E	-3.65	24.28	33.12
160	120	1.5	1	1	-1	keNEATPSE_L	-3.68	24.24	31.23
160	120	1.5	1	1	-1	keNEATPSE_E	-3.67	24.66	31.60
160	120	1.5	1	1	-1	keNEATCE_L	-7.02	26.59	35.55
160	120	1.5	1	1	-1	keNEATCE_E	-3.68	24.37	33.24
160	120	1.5	1	1	-1	Linear	0.02	25.54	25.53
160	120	1.5	1	1	-1	Equipercentile	0.02	25.53	25.53
80	120	1.5	1	0.6	0	Tucker	-0.96	21.15	26.87
80	120	1.5	1	0.6	0	Levine True	-1.19	20.49	29.93
80	120	1.5	1	0.6	0	Braun	-0.78	20.94	26.65
80	120	1.5	1	0.6	0	FEEE	-0.77	20.92	26.66
80	120	1.5	1	0.6	0	Chain_L	-1.10	20.79	28.65
80	120	1.5	1	0.6	0	Chain_E	-1.05	20.71	28.52
80	120	1.5	1	0.6	0	keNEATPSE_L	-0.80	20.60	26.39
80	120	1.5	1	0.6	0	keNEATPSE_E	-0.76	20.97	26.67
80	120	1.5	1	0.6	0	keNEATCE_L	-1.10	20.79	28.65
80	120	1.5	1	0.6	0	keNEATCE_E	-0.96	20.87	28.74
80	120	1.5	1	0.6	0	Linear	-0.02	22.14	22.13
80	120	1.5	1	0.6	0	Equipercentile	-0.02	22.13	22.13
161	120	1.5	1	1	0	Tucker	-1.41	24.88	33.10
161	120	1.5	1	1	0	Levine True	-1.68	24.11	35.14
161	120	1.5	1	1	0	Braun	-1.45	24.69	32.89
161	120	1.5	1	1	0	FEEE	-1.44	24.67	32.90
161	120	1.5	1	1	0	Chain_L	-1.57	24.43	34.33
161	120	1.5	1	1	0	Chain_E	-1.97	24.54	34.40
161	120	1.5	1	1	0	keNEATPSE_L	-1.49	24.44	32.70
161	120	1.5	1	1	0	keNEATPSE_E	-1.43	24.72	32.90
161	120	1.5	1	1	0	keNEATCE_L	-1.57	24.43	34.33

Table 4.9

Cont.

Study Condition							Statistic		
Panel	Test Length	BGMAD	DAD	μ(a)	ATMDD	Equating Method	BIAS	SEE	RMSE
161	120	1.5	1	1	0	keNEATCE_E	-1.87	24.67	34.60
161	120	1.5	1	1	0	Linear	0.00	26.77	26.76
161	120	1.5	1	1	0	Equipercentile	0.00	26.76	26.76
81	120	1.5	1	0.6	1	Tucker	1.13	21.59	27.25
81	120	1.5	1	0.6	1	Levine True	1.50	20.72	30.18
81	120	1.5	1	0.6	1	Braun	0.86	21.35	26.98
81	120	1.5	1	0.6	1	FEEE	0.90	21.29	26.96
81	120	1.5	1	0.6	1	Chain_L	1.35	21.10	28.99
81	120	1.5	1	0.6	1	Chain_E	0.36	21.11	28.89
81	120	1.5	1	0.6	1	keNEATPSE_L	0.84	21.01	26.71
81	120	1.5	1	0.6	1	keNEATPSE_E	0.90	21.33	26.96
81	120	1.5	1	0.6	1	keNEATCE_L	1.35	21.10	28.99
81	120	1.5	1	0.6	1	keNEATCE_E	0.55	21.21	29.06
81	120	1.5	1	0.6	1	Linear	-0.02	22.62	22.61
81	120	1.5	1	0.6	1	Equipercentile	-0.02	22.62	22.61
162	120	1.5	1	1	1	Tucker	2.23	26.70	34.82
162	120	1.5	1	1	1	Levine True	2.73	25.70	37.71
162	120	1.5	1	1	1	Braun	0.55	25.97	33.89
162	120	1.5	1	1	1	FEEE	0.63	25.94	33.87
162	120	1.5	1	1	1	Chain_L	2.53	26.15	36.51
162	120	1.5	1	1	1	Chain_E	-0.39	25.81	35.72
162	120	1.5	1	1	1	keNEATPSE_L	0.48	25.50	33.50
162	120	1.5	1	1	1	keNEATPSE_E	0.63	26.00	33.88
162	120	1.5	1	1	1	keNEATCE_L	2.53	26.15	36.51
162	120	1.5	1	1	1	keNEATCE_E	-0.19	25.96	35.94
162	120	1.5	1	1	1	Linear	0.00	28.74	28.74
162	120	1.5	1	1	1	Equipercentile	0.00	28.74	28.74

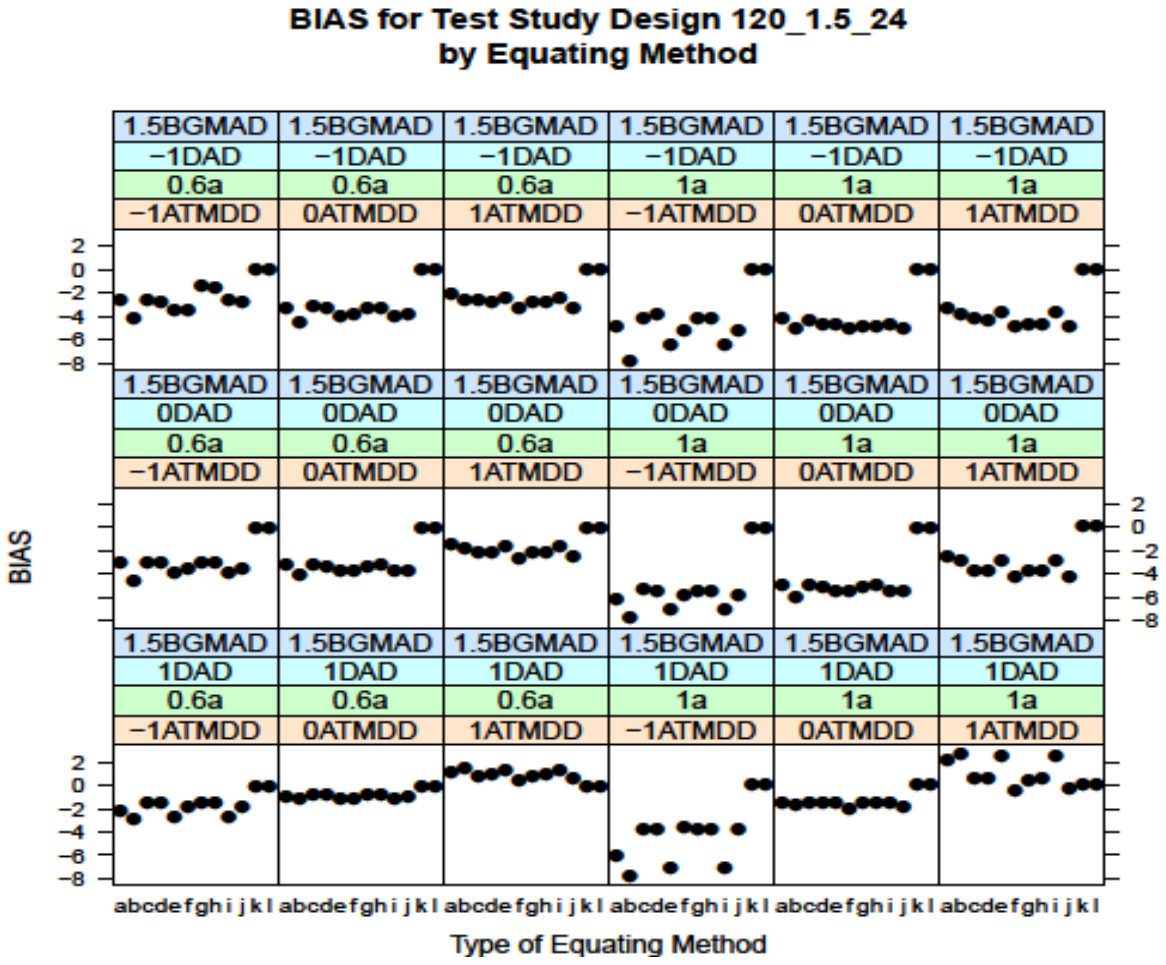

Figure 4.17. Bias for Test Study Design 120_1.5_24 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

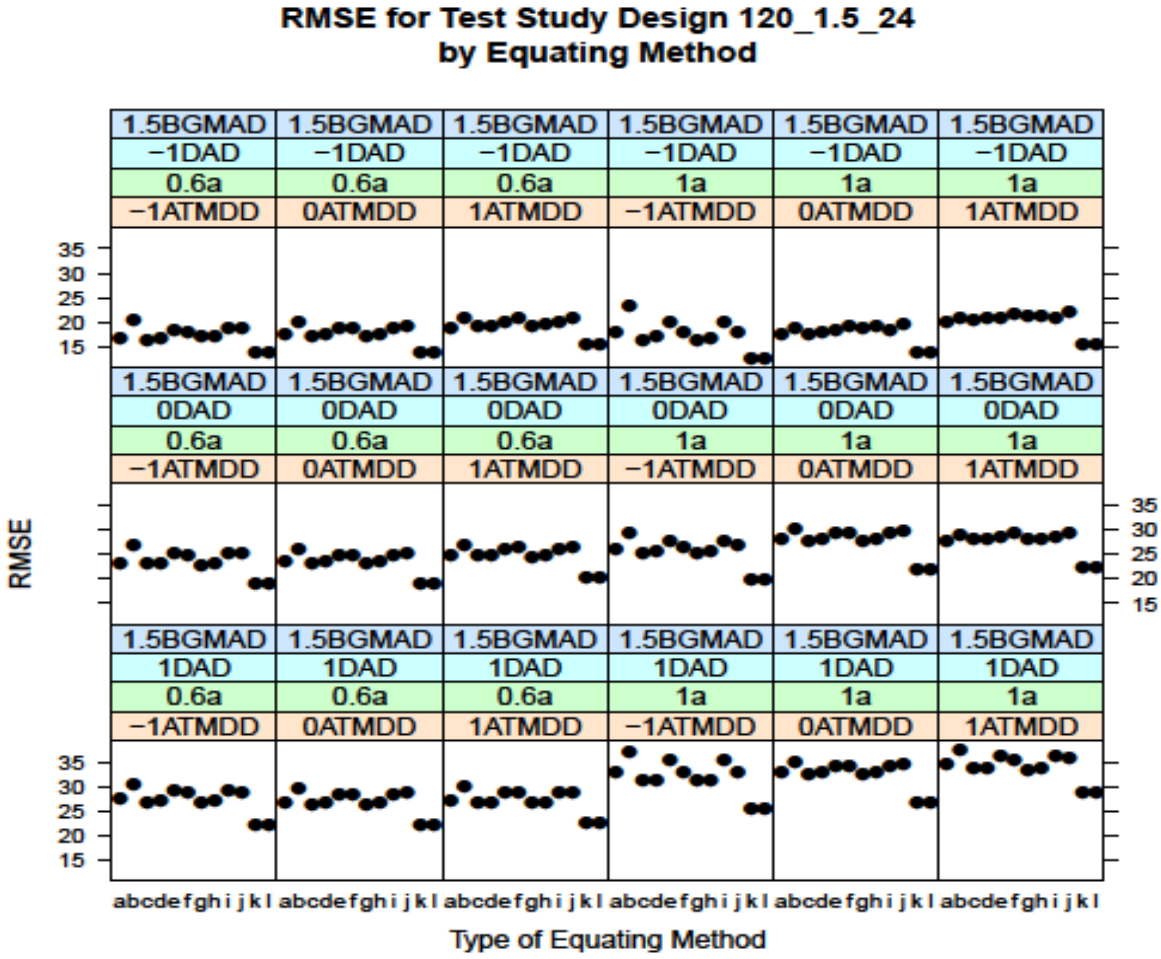

Figure 4.18. Root Mean Square Error (RMSE) for Test Study Design 120_1.5_24 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

4.3 Summary of the Nine Test Study Designs

To recapitulate, each of the nine test study designs were modeled such that test length, between-grade mean ability differences, BGMAD (or examinee proficiency on theta scale, or separation of grade ability distribution, or simply group effect), moderate (0.6) item discrimination and high (1) item discrimination were held constant. Then

followed manipulation of levels of DAD and ATMDD across levels of other factors—that is, test length, BGMAD, and item discrimination under each equating method. This process produced sets of three test study design, which were discussed as first, second, and third sets of test study designs in the preceding subsections. Broadly speaking, these nine test study designs were outlined at the beginning of this chapter. Although there were disparities in the total number of test items—30, 60, and 120—and anchor test items—6, 12, and 24—as displayed in figures 4.2, 4.8, and 4.14 (other examples include figures 4.4, 4.10, 4.16; figures 4.6, 4.12, 4.18) for the nine test study designs, equivalent results were evidenced across test study designs that exhibited similar BGMAD levels and other homologous study conditions. For instance, high (1) item discrimination items for all similar BGMAD levels produced highest values of RMSE compared to moderately (0.6) item discrimination items for similar BGMAD levels, which produced smaller RMSE values; it was also observed that small BGMAD produced smaller values of RMSE than medium and large BGMAD. This is not surprising because it is expected that when BGMAD is small and item discrimination is moderate the students in the lower adjacent grade levels are more likely to perform much better than the students in the upper adjacent grade levels. Part of the reason can be attributed to the fact that examinees in the lower adjacent grade levels have been exposed to the materials more recently unlike the examinees in the upper adjacent levels who could have forgotten the material over time. Conversely, the test items could have been much easier for the students in the upper grade levels and also their maturation (or being older) could have an influence on their performance. However, when BGMAD is large it is contemplated that the students

in the upper adjacent grade levels will outperform those in the lower adjacent grade level because some of the questions are hard and most probably have not been covered in adjacent lower class. When the number of total items and anchor items were increased in the context of vertical scaling and conditions studied in this dissertation lead to large values of RMSE. This can be associated with the fact that when sampling common items from a small BGMAD one is likely to get high quality items which reflect closely related test items in terms of difficulty compared to sampling test items from a large BGMAD, where sampling of more difficult items is likely.

In addition, test study designs with small items indicated small bias very close to zero. This means that most likely the overlapping areas of the adjacent grade—where anchor items were sampled—can produce small number of items for vertical scaling while large number of items means more above or below grade level items are more likely to be sampled; therefore, large bias results for the NEAT design. This finding is important because it can help test designers, researchers, and psychometricians or practitioners to examine and identify testing realities that lead to best or worst vertical scaling results. Besides, linear and equipercentile equating methods under RG/EG remained consistently at or near zero bias for all the nine test study designs under all the study conditions investigated in this dissertation. This could be attributed to the fact that grade 5 was considered as the base grade for vertical scaling in this study and that the RG/EG equating design results actually compared to the same grade (grade 5) with only variations in grade 5 forms; in addition to this design, no adjacent sampling of items was done.

Viewing the current results through vertical scaling lenses, all equating methods produced different results depending on whether the method is linear or nonlinear or considered under NEAT or RG/EG design paradigm and that equating error somewhat depended on satisfaction of the underlying equating assumptions that are unique to each equating method under each study condition. This leads to different discernible patterns from the performance of the multiple equating methods used in this study. Under RG/EG design the linear and equipercentile outperformed all NEAT equating methods both in terms of bias and RMSE values. However, within NEAT design equating methods Braun/Holland, Frequency estimation equipercentile equating, keNEATPSE linear, and keNEATPSE equipercentile methods performed best with very close bias and RMSE values across all study conditions. The other equating methods performed poorly—even though Chained linear, Chained equipercentile, keNEATCE linear, and keNEATCE equipercentile equating methods performed almost the same—with Levine linear methods being the poorest method overall. Last, as BGMAD or separation of grade ability distribution increases (i.e., from .5, 1, and 1.5) systematic error (bias), random error (SEE), and overall equating error (RMSE) increased under all conditions; however, random error was more impacted than its counterpart systematic error. Fundamentally, the results of overall equating error somehow lined up with those of random error.

Last, bias for test study designs 30_1.5_6 or short test with large BGMAD, 60_1.5_6 or medium test with large BGMAD, and 120_1.5_6 or long test with large BGMAD–i.e., varying total test length while holding separation of grade ability distribution (BGMAD) constant—for all levels of DAD (below average DAD or -1DAD,

average DAD or 0DAD, and above average DAD or 1DAD), high *a*-parameter or $\mu(a)$=1, and below average ATMDD (-1ATMDD) resulted in big values of bias and increasing RMSE with considerable inconsistency for all NEAT equating methods. Since this dissertation investigated through simulation study situations where equating procedures would work or fail when constructing a vertical scale, then it can be argued that large values of bias witnessed in the above conditions are clear evidence where equating procedures completely breakdown in this study.[1] Similarly, due to variability occasioned by test length and other study conditions RMSE kept on increasing. Also, it was observed that the standard deviations for total test length (TT), anchor test (AT), and regular test (RT) for high *a*-parameter or $\mu(a)$=1 are greater than the standard deviation for moderate *a*-parameter or $\mu(a)$=.6 across all study conditions. The reader is referred to Appendix A Tables A.1-A.81 for more details. Therefore, design effect (Kish, 1965) due to test length, high *a*-parameter associated with big standard deviations, and how close average *b*-parameters are to the population group are tied to the inconsistency behavior of bias values in some study conditions, which made equating procedures to collapse.

4.4 Results of the Real Data Analysis

Unlike the anchor test items, which have similar statistical and psychometric properties in both test forms, the unique items in each test form have different statistical and psychometric characteristics. For this reason, the test forms do not necessarily need to have the same level of difficulty. Specifically, Table 4.10 shows that test Form Y was harder than Form X on the basis of anchor tests performance (both anchor tests statistics

[1] Computational procedures were carefully checked and verified.

are in bold)—i.e., the examinees who took Form X performed much better than (or outperformed) the examinees who did Form Y, because test Form X was perhaps much easier than test Form Y, or the examinees who took Form X were more able than those who did test Form Y. The means for Form X anchor test and Form Y anchor test are 13.35 and 12.16 respectively; thus, examinees who took Form X were more proficient than those who took Form Y. The averages for both anchor test scores are used to compare the difficulty of these tests, because the examinees in both forms were exposed to the same anchor test. In other words, the same anchor test items in Form X were exactly the same anchor test in Form Y.

Table 4.10

Summary Descriptive Statistics for the Observed Score Equating Using an External Anchor

Test Form	Test Type	Means	SD	Min.	Max.	Skew	Kurt
X	Unique	55.24	11.84	0.00	80.00	-0.32	-0.44
	Anchor	**13.35**	**3.46**	**0.00**	**20.00**	**-0.10**	**-0.64**
Y	Unique	49.47	13.69	0.00	80.00	0.03	-0.66
	Anchor	**12.16**	**4.09**	**0.00**	**20.00**	**-0.03**	**-0.80**

Even though it is hard to construct a strictly parallel test, as evidenced in equating literature, the reliability of the scale for the test Form X and Form Y is Cronbach`s coefficient alpha 0.93 and 0.94 respectively—i.e., we can assume that the two tests measure similar underlying hypothetical construct of international language ability. This means the relative error variances are considerably small and the reliability is greater;

therefore, we can infer that the variance of the underlying latent trait (or true score) very closely estimates the variance of the observed score. This is shown in Table 4.11 by Cronbach's coefficient alpha values in parenthesis. Also, the reliability of the scale for the individual type of test is displayed in the same table besides the values in curved brackets. For example, test Form X unique test has a reliability (or Cronbach`s coefficient alpha) of 0.91 and that of test Form Y anchor test has 0.79. As expected, the probable underlying reason for the moderately low reliability for the scales under anchor tests for both test Form X and Form Y could be attributed to a small number of items—in this case, each of the anchor test has a total of 20 items compared to their counterparts, unique test items, which has a total of 80 items.

Essentially, the role of correlation in test score equating has been intensively discussed. There is consensus among the test score equating experts and practitioners that higher correlation between an anchor test and the total test oftentimes leads to better equating results. For instance, Petersen et al. (1989) and von Davier et al. (2004) have demonstrated that when the correlation between the anchor test scores and the total test scores is higher, then the anchor test would be a better candidate for equating the two test forms. Table 4.11 reports the correlation coefficients of anchor test scores to the total test scores in Form X and Form Y. There was a significant strong positive relationship between the total test scores and the anchor test scores for Form X, r (47280) = .88, $p < .05$; likewise, there was a statistically significant strong positive association between the total test scores and the anchor test scores for Form Y, r (47280) = .90, $p <$. 05.

Table 4.11

Reliability of the Scale and Anchor-Test to Total-Test Score Correlations

Test Form	Test Type	Reliability	Correlation (sig)
X	Unique	.91 (0.93)	0.88 (.000)
	Anchor	.72 (0.93)	
Y	Unique	.92 (0.94)	0.90 (.000)
	Anchor	.79 (0.94)	

Note. Significance levels for Correlation are denoted by parenthesis.

This means the high correlation between the anchor test scores and the total test scores has two important implications in equating: first, it could be used as a global measure of the efficiency of the equating (Budescu, 1985; Dorans et al., 1998); second, high quality items—i.e., in terms of number of items, content, and statistical representation—have been selected and effectively incorporated into the operational forms (Angoff, 1968). For a thorough treatment on the requirement of a good anchor test, the reader is referred to consult the literature (notably, the works of Angoff, 1971; Kolen & Brennan, 2004, 2014; Petersen et al., 1989). It is worthwhile to note from Table 4.11 that one of the equating assumptions of equal reliability between the test and the anchor has been violated (see Dorans & Holland, 2000).

The major finding in Table 4.10 and Table 4.11 is that even though Form X was easier than Form Y, both test forms demonstrated close reliability and correlation. The reliability of both the anchor test scales appeared moderate.

The equating results for the real data analysis under Kernel Equating within NEAT design are presented. To start with, Table 4.12 shows the equated scores and

standard error of equating for Form Y, when equated to the score scale of Form X for both Chained equating and Post-Stratification equating under general foundation of NEAT design. The x score represents the score scale for Form X while ey(x) denotes the scaled score equivalent of Form Y after Form Y test score were equated to Form X score scale, hence making the scores from both test forms statistically comparable because they are put into a common metric. After statistically adjusting and successfully converting raw scores onto a common scale in order to account for differences in difficulty across the two test forms, it should be a matter of indifference to a test taker as to which test form or time of the year the test taker takes—i.e., regardless of whether any test form was conceived easy or difficult. This means an examinee taking an easier test form needs to answer extra questions correctly in order to attain a specific scaled score. Besides the equated scores, that is ey(x), reported in Table 4.10 for each equating method, there are standard errors of equating at each score point. It can also be observed that out of range values are reported after equating under each method. For this study, under KENEATCE and KENEATPSE out of range scores were only 81.56 and 80.77 respectively, which is quite reasonable because is not far-fetched given the maximum score of 80 on the x score scale. Each equating method has a different scaled score that corresponds to the score on Form X. For instance, an equated score of 12 points and 1 point for KENEATCE and KENEATPSE respectively corresponds to a zero score on the x score scale. The SEE values are accuracy measures to detect the extent to which the equating method introduces random error.

Table 4.12

Equated Scores and Standard Error of Equating Under Kernel Equating for Both Chained Equating and Post-Stratification Equating for NEAT Design

Score x	KENEATCE		KENEATPSE	
	ey(x)	SEE	ey(x)	SEE
0	12.006	1.236	1.073	0.603
1	15.615	0.528	2.704	1.078
2	16.715	0.487	5.675	4.193
3	17.069	0.479	10.659	2.603
4	17.176	0.478	12.085	1.613
5	17.224	0.478	12.734	1.364
6	17.260	0.478	13.093	1.273
7	17.294	0.477	13.319	1.231
8	17.330	0.476	13.488	1.201
9	17.370	0.475	13.639	1.169
10	17.417	0.472	13.796	1.130
11	17.476	0.467	13.975	1.080
12	17.554	0.461	14.192	1.017
13	17.661	0.453	14.462	0.942
14	17.812	0.442	14.800	0.857
15	18.026	0.427	15.217	0.766
16	18.326	0.410	15.716	0.674
17	18.731	0.389	16.296	0.587
18	19.256	0.365	16.948	0.509
19	19.896	0.337	17.661	0.442
20	20.638	0.308	18.421	0.384
21	21.457	0.278	19.219	0.336
22	22.329	0.250	20.044	0.296
23	23.234	0.226	20.890	0.263
24	24.157	0.205	21.752	0.235
25	25.091	0.188	22.627	0.212
26	26.029	0.174	23.513	0.193
27	26.970	0.162	24.409	0.177
28	27.910	0.152	25.314	0.164
29	28.851	0.143	26.229	0.153
30	29.792	0.135	27.155	0.143
31	30.733	0.128	28.091	0.135

Table 4.12

Cont.

Score x	KENEATCE		KENEATPSE	
	ey(x)	SEE	ey(x)	SEE
32	31.673	0.123	29.038	0.127
33	32.614	0.118	29.996	0.121
34	33.555	0.113	30.964	0.116
35	34.496	0.109	31.943	0.111
36	35.435	0.105	32.932	0.107
37	36.371	0.102	33.928	0.103
38	37.304	0.098	34.930	0.100
39	38.231	0.095	35.936	0.096
40	39.152	0.092	36.943	0.093
41	40.064	0.089	37.950	0.090
42	40.968	0.087	38.954	0.087
43	41.862	0.085	39.953	0.085
44	42.746	0.083	40.946	0.083
45	43.620	0.081	41.931	0.081
46	44.486	0.079	42.910	0.079
47	45.344	0.077	43.881	0.077
48	46.196	0.076	44.847	0.075
49	47.045	0.074	45.809	0.074
50	47.893	0.073	46.769	0.072
51	48.743	0.072	47.731	0.071
52	49.598	0.071	48.696	0.070
53	50.462	0.070	49.668	0.069
54	51.338	0.070	50.651	0.069
55	52.229	0.070	51.646	0.068
56	53.137	0.070	52.658	0.068
57	54.067	0.070	53.688	0.068
58	55.021	0.071	54.738	0.068
59	56.000	0.071	55.812	0.069
60	57.008	0.072	56.908	0.070
61	58.045	0.073	58.030	0.070
62	59.113	0.074	59.177	0.071
63	60.213	0.076	60.349	0.072
64	61.346	0.077	61.544	0.073

Table 4.12

Cont.

Score x	KENEATCE		KENEATPSE	
	ey(x)	SEE	ey(x)	SEE
65	62.511	0.078	62.763	0.074
66	63.709	0.080	64.001	0.076
67	64.938	0.082	65.257	0.077
68	66.198	0.084	66.526	0.078
69	67.485	0.086	67.802	0.079
70	68.797	0.087	69.081	0.080
71	70.128	0.088	70.356	0.080
72	71.474	0.089	71.621	0.080
73	72.827	0.089	72.870	0.080
74	74.179	0.091	74.095	0.080
75	75.517	0.093	75.292	0.081
76	76.825	0.094	76.455	0.081
77	78.082	0.093	77.582	0.081
78	79.273	0.090	78.673	0.081
79	80.417	0.088	79.732	0.080
80	81.558	0.085	80.765	0.079

Figure 4.20 shows a linear relationship between the score scale of Form X and the equated scores from each equating method after their Kernel equating functions were computed and applied. Although the equated scores form KENEATCE and KENEATPSE were nonlinear at score points less than 20 than at any other score point, the majority of equated scores depicted a strong positive linear relationship. The solid line on Figure 4.20 is akin to identity equating, and it can be observed that the equated scores are very close to the black solid line between score range 21-81 with an exception of overlapping or close to overlapping points between score range 56-81. The small differences between the two equating procedures suggest that their choice to use one over

the other is a matter of policy or preference of the testing program, because from these results the two equating functions produced very close equated scores between score range 56-81. Overall, there are nuances of the two equating methods where KENEATCE performed better than KENEATPSE.

Figure 4.19 shows a linear relationship between the score scale of Form X and the equated scores from each equating method after their Kernel equating functions were computed and applied. Although the equated scores form KENEATCE and KENEATPSE were nonlinear at score points less than 20 than at any other score point, the majority of equated scores depicted a strong positive linear relationship. The solid line on Figure 4.20 is akin to identity equating, and it can be observed that the equated scores are very close to the black solid line between score range 21-81 with an exception of overlapping or close to overlapping points between score range 56-81. The small differences between the two equating procedures suggest that their choice to use one over the other is a matter of policy or preference of the testing program, because from these results the two equating functions produced very close equated scores between score range 56-81. Overall, there are nuances of the two equating methods where KENEATCE performed better than KENEATPSE.

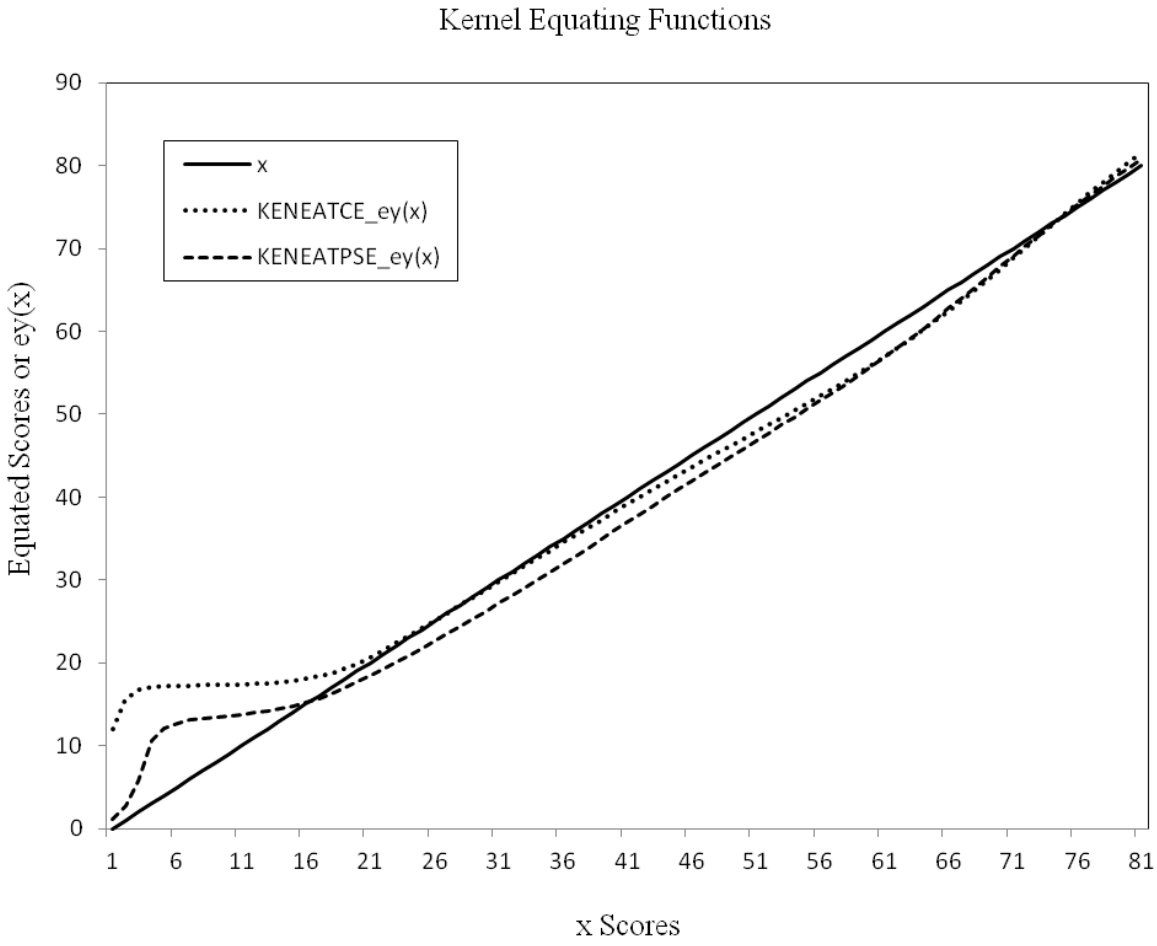

Figure 4.19. Relationship between Equated Scores and the x-score Scale under Kernel Equating for Both Chained Equating and Post-Stratification Equating for NEAT Design.

In Figure 4.20 standard error of equating across each score point is plotted for each equating method. Small SEE values that are very close to zero or at zero imply small random error introduced by the equating method. In this case, therefore, it can be deduced that score points below 26 points registered larger SEE values for the KENEATPSE than KENEATCE. Beyond point 26 there are no noticeable differences between the two curves. In fact, the two curves overlap and their SEE values stabilize across the score points at near point zero.

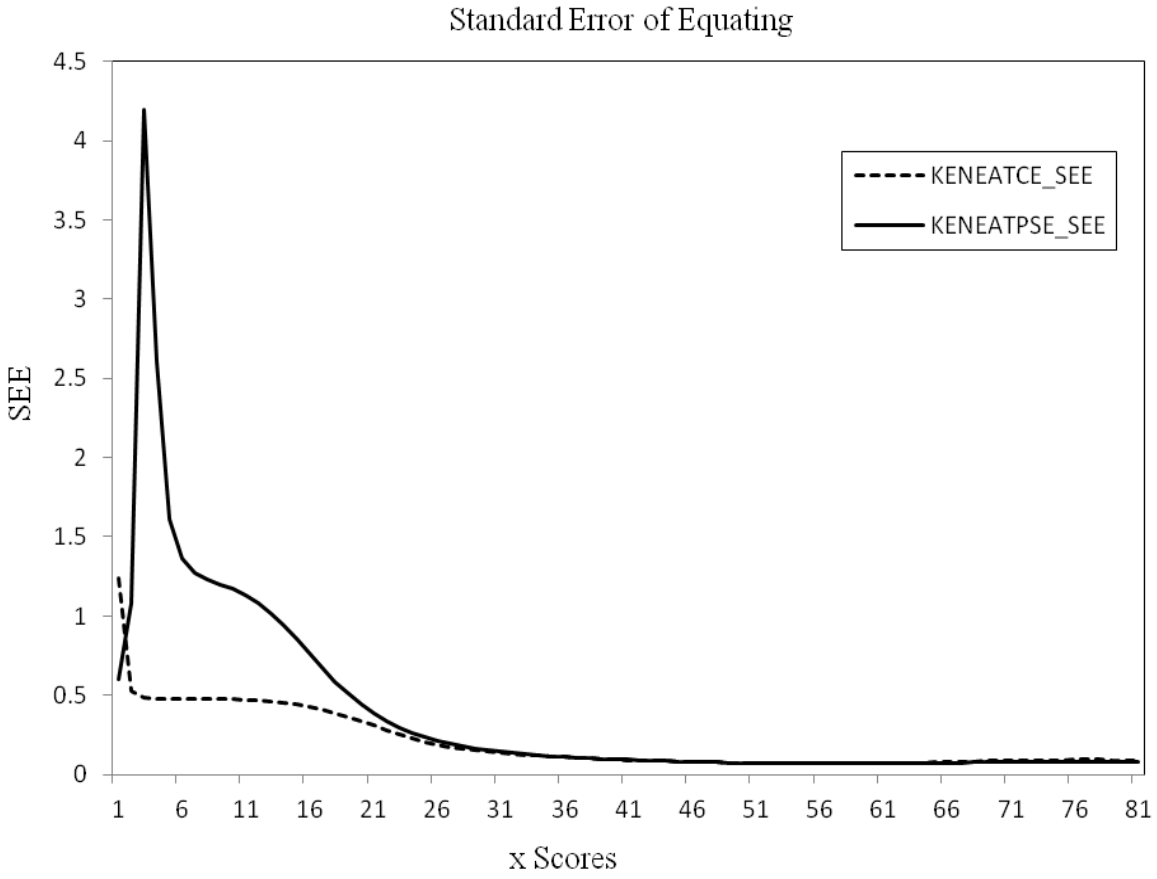

Figure 4.20. Standard Error of Equating across the x-score Scale under Kernel Equating for Both Chained Equating and Post-Stratification Equating for NEAT Design.

In summary, the equating differences for the two equating methods are plotted in Figure 4.21. This Figure displays a combination of Table 4.10, Figure 4.19, and Figure 4.20 by plotting equated scores and SEE on the same y-axis, and the x score scale on x-axis. As indicated previously, the biggest difference between the two equating methods seem to appear in the score range 0–20; however, a close scrutiny of the SEE across the score points reveals very small values, save for less than 16 points in case of KENEATPSE which has more than 1 SEE values. Also, as the linear relationship between the equated scores and x score scale increases and get stronger, the SEE

approaches zero, a clear implication that, in general terms, the two equating methods introduced random errors across the score points that can be tolerated.

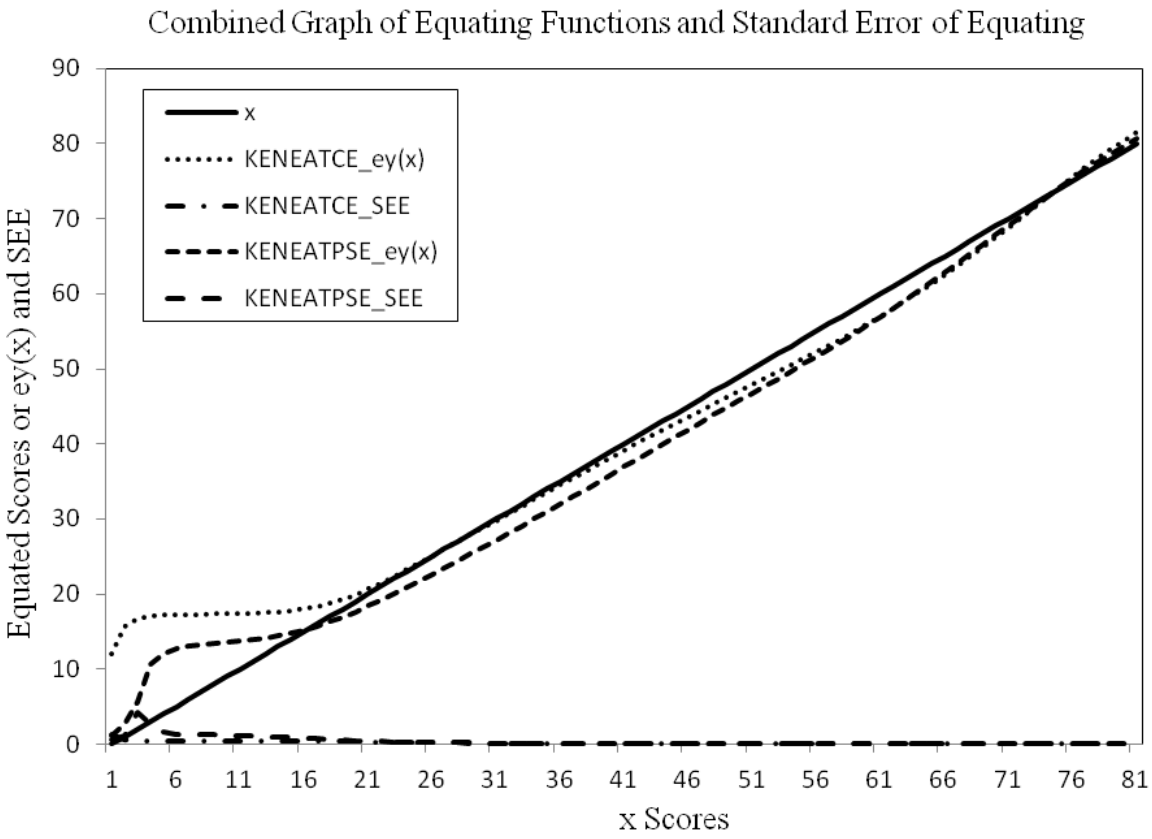

Figure 4.21. Combination of Kernel Equating Functions and Standard Error of Equating across the x-score Scale under Kernel Equating for Both Chained Equating and Post-Stratification Equating for NEAT Design.

CHAPTER V

CONCLUSION AND DISCUSSION

5.1 Overview of the Chapter

The main purpose of the current research study was to explore some of the empirical issues and complications associated with vertical scaling methods for a particular class of equating designs called the NEAT design; a comparison of the performance of different equating methods under diverse simulation conditions that mimic real world testing practices—even though generated data provided extremes that are rarely witnessed in practice—was examined. A simulation research study was undertaken using 162 conditions, and the findings of the study were outlined in the preceding chapter based on the major factors that were manipulated in the study. Results from simulated data, on one hand, indicated that small between-grade mean ability difficult when considered together with a short test length, a moderate item *a*-discrimination parameter, below average distribution of ability difference, and below average anchor test mean ability difference produced the most reasonable results. Also, the results revealed that equating error depended on the extent the underlying equating assumptions are met in relation to a particular equating method under each study condition. Results from real data, on the other hand, show a small difference between KENEATCE and KENEATPSE equating procedures, because they produced very close equated scores. Similarly, as the linear positive relationship between the equated scores

and x score scale increases, the SEE approaches zero which is an indication that the two equating methods introduce random errors across the score points that can reasonably be tolerated. In this chapter, these results are first described in connection with original research questions that guided this study. In addition, limitations and the implications of the research study are discussed. Last but not least, conclusions are made on the basis of the results of the current research study and consequently, possible suggestions for future study are made.

5.2 Summary of Key Research Findings

5.2.1 Research Question Number 1

How do variations of multiple study conditions (i.e., test length, test mean discrimination, between-grade mean ability difference, distribution of ability difference, and anchor test mean difficulty differences) affect equating errors—i.e., bias, standard error, and root mean square error—for different equating methods when constructing a vertical scale using a special NEAT design? This main question is partitioned into two sub-questions:

(i) *How does this variation affect the equating accuracy across the five study conditions?*

(ii) *How consistent are the results across the five study conditions?*

This is the main research question for the study together with its associated sub-questions, which dealt with the impact (that is, variations and consistency) of each main factor on equating error: bias, SEE, and RMSE. It is important to note that equating error depended on satisfaction of the equating assumptions that are particular to a specific

equating method under each study condition. For instance, Braun/Holland, Frequency Estimation Equating, keNEATPSE linear, and keNEATPSE equipercentile methods performed almost similarly under all study conditions; however, a closer examination of the above equating methods reveal that when the equating relationship was linear, keNEATPSE linear outperformed all linear-related equating methods considered in this dissertation. Similarly, when the equating relationship was non-linear (or curvilinear), keNEATPSE equipercentile was more accurate in terms of total error—i.e., it produced the smallest RMSE/equating total error—than all non-linear equating methods. Therefore, implementation of these equating methods is preferred within the framework of vertical scaling where NEAT design is used to collect data.

The overall equating error (RMSE) was affected by total test length and the number of anchor test items. See Chapter III for details on the meaning of short test length (30), medium test length (60), and long test length (120). As the total test length increases—which also led to increase of common items, even though the proportion of common items to the total test length remained invariant at 20%—the total equating error increased. Short test length had the smallest overall equating error and long test registered the highest total equating error, whilst medium test recorded total equating errors in-between the two.

Differences in item discrimination parameters (*a*-parameters) played an important role in the accuracy of the overall equating error. See Chapter III for details on the meaning of moderate item *a*-discrimination (.6) and high item a-discrimination (1).

Moderately discriminating items produced more accurate and consistent overall equating errors than highly discriminating items for all study conditions.

As distribution of ability difference (Pool information or grade-to-grade ability variability) differs in terms of below average, average, and above average, the total equating errors increase. See Chapter III for details on the meaning of below average (-1), average (0), and above average (1) DAD. Total equating errors were at the lowest when grade-to-grade ability variability was below average, but the results were not consistent across all test study designs. However, when grade-to-grade ability variability was average the total equating errors were consistent even though not as accurate as when DAD below average. Overall for this condition, large grade-to-grade ability variability produced the worst results in terms of the largest total equating error under all study conditions.

As between-grade mean ability differences (θ, examinee proficiency on the theta scale or the separation of grade ability distributions) increase, the RMSE values tend to increase from small BGMAD to large BGMAD under all study conditions. See Chapter III for details on the meaning of small (.5), medium (1), and large (1.5) ATMDD. Small BGMAD recorded the smallest errors in terms of total equating errors whereas large BGMAD recorded largest errors under all study conditions. The total equating errors for the medium BGMAD was somewhere between the RMSE values of small BGMAD and RMSE values of large BGMAD.

When anchor test mean difficulty differences or anchor test difficulty variability was below average—see Chapter III for details on the meaning of below average (-1),

average (0), and above average (1) ATMDD—the overall equating error was smallest as compared to average and above average ATMDD when all study conditions were considered. Specifically, this was true for systematic error (bias). Although the degree of accuracy varied across the nine test-study designs, similar patterns of RMSE were observed regardless of the study design. Stated differently, below average ATMDD condition produced the most accurate results overall vis-à-vis average and above average ATMDD whenever the rest of study condition were manipulated.

5.2.2 Research Question Number 2

How much difference between anchor test difficulty and the other four study conditions can be endured under each equating method?

The second research question focused on comparison between anchor test mean difficulty difference with other four study conditions in connection with each equating method. Under NEAT design methodology, strong underlying assumptions are made which are untenable (Kolen & Brennan, 2014; von Davier et al., 2004; Holland, Dorans, & Petersen, 2007). For this reason, forms that differ substantially in difficulty, which often is the case in a vertical scaling scenario, might not achieve a high degree of equating accuracy. In this study, difference in levels of ATMDD produced different equating errors for each equating method. This important finding is consistent with the results by Liu, Sinharay, Holland, Feigenbaum, and Curley (2011). That large between-grade (group) mean ability differences with the interaction effects of other study conditions resulted in large overall equating errors across the three different levels of ATMDD for all nine test study designs.

When DAD was below average, item discrimination (a-parameter) was moderate, and ATMDD was below average—see Chapter III for more details about levels of various factors and what they mean in regard to this dissertation—were constant and varied across small, medium, and large BGMAD the resulting overall equating error was the smallest compared with other study conditions. Furthermore, it was observed that when *b*-item difficulty for regular test in a specific grade was one unit below the mean of underlying ability for a particular grade while *a*-item discrimination was moderate and the average *b*-item difficulty for the anchor test was below average *b*-item difficulty for the regular test there was a high equating accuracy for all equating methods across test study designs.

In summary it can be noted that:

(i) Under NEAT design, KE produced more accurate equating relationships under a majority of testing conditions when paired with PSE linear and equipercentile than when paired with CE linear and equipercentile. Stated differently, keNEATPSE linear and equipercentile equating methods produced better equating results in terms of overall equating accuracy than keNEATCE linear and equipercentile in this study.

(ii) Under NEAT design, the best-performing equating methods varied significantly depending on the test study design. Of all the equating methods considered under NEAT design, regardless of linear or non-linear equating relationship, Tucker linear, Braun/Holland, Frequency estimation equipercentile equating, and both keNEATPSE linear and equipercentile

outperformed all versions of chained equating methods—that is, Chained linear, Chained equipercentile, keNEATCE linear, and keNEATCE equipercentile. The assumptions for chained linear and Levine true linear are similar; therefore, this explains why the two equating methods almost performed similarly. Levine true linear method registered the worst performance under all study conditions.

(iii) The two equating methods under EG/RG—linear and equipercentile—outperformed all NEAT equating methods.

(iv) Linear and equipercentile equating methods performed similarly depending on the equating design—NEAT or EG/RG.

(v) As BGMAD or separation of grade ability distribution increases (i.e., from .5, 1, and 1.5) systematic error (bias), random error (SEE), and overall equating error (RMSE) increased under all conditions; however, random error was more impacted than its counterpart systematic error. Fundamentally, the results of overall equating error somehow lined up with those of random error.

(vi) Although Kolen and Brennan (2014) showed that mean group ability differences of .3 or more standard deviation units oftentimes produce quite different equating results based on the equating method being applied, and problematic results are produced particularly when the magnitude of differences becomes too large (e.g., .5 or more standard deviation units), these rules of thumb could be seen as stringent guidelines when applied to

vertical scaling context, where naturally the underlying abilities of the two adjacent grades are less likely to be that close. Wang et al. (2008) labeled mean group ability differences between .05 and .1 as “relatively large” while values at .25 or above as “very large”. Even though this is the standard procedure in equating, between-grade (group) mean ability differences investigated in this dissertation are fundamentally different because they are considered within the general framework of vertical scaling.

5.2.3 Research Question Number 3

Does the use of equating introduce more errors than it can be rationalized?

This research question primarily focused on the extent the random errors could be tolerated after KENEATCE or KENEATCE equating. These two equating procedures were considered under the Kernel equating, where pre-smoothing was conducted. Overall, there are nuances of the two equating methods where KENEATCE performed slightly better than KENEATPSE. Even though there are slight differences in terms of SEE or random error introduced by the two equating procedures, this study did not gather enough evidence to support the claim that one of them is better than the other. Furthermore, the small differences between the two equating procedures suggest that their choice to use one over the other is a matter of policy or choice by the testing program, because their results produced very similar equated scores and SEE. Overall, SEE for both equating procedures is very close to zero, which implies that random error introduced across the score points can be tolerated.

5.3 Practical Implications of the Results

The concept of vertical scaling and learning-progressions are not only at the foundation of the policy and practice around systems of education where accountability is vital but also are fundamental concern for teachers as they interact and navigate through multiple pedagogical approaches with learners throughout the entire academic year. Within the educational accountability terrain—a stringent policy requirement currently adopted by most public schools across the U.S.—a student is required to demonstrate what he or she knows and can do before higher order skills are introduced to the learner or even proceeding to the next grade. In the same vein, teachers are held accountable for the performance of their learners. With this in mind, it can be argued that a well-constructed vertical scale is a necessary but not a sufficient requirement for the success of a learning-progression approach (Briggs & Peck, 2015).

By use of a vertical scale, teachers are able to know where a learner is along the learning/developmental trajectory or scale. This would substantially help in guiding the teachers, administrators, and other educational stakeholders in making decisions for early intervention for learners who are struggling or performing below passing threshold. Similarly, it can also guide in other placement decisions like for students with special academic talents and ability or achievement. Students who have demonstrated extraordinary performance can be recommended for promotion to the next grade, where he or she can get challenged appropriately. Therefore, designing vertical scales and establishing learning-progressions go hand in hand to support construct validity

(Messick, 1989) and validation (Kane, 2006, 2013) and cogent underlying presumptions about student progression along a vertical scale spanning over grades.

This research can be conceived as an attempt to renew interest in this field of research; however, caution must be exercised in interpreting vertical scales. In the vertical scale literature cases of scale shrinkage, where scale score decreased as grade level increased, and deceleration of growth in the subsequent grades have been reported. Also, this study will help test developers, psychometricians, and practitioners to select high quality items to be included when constructing anchor test items. Likewise, the notion of where equating works best or worst is also an important consideration. For instance, in this study it has been demonstrated that moderate item *a*-discrimination parameters when used in conjunction with short tests, small BGMAD, and below average for DAD and ATMDD the total equating error is smallest compared to the other conditions.

They are also used to create large-scale assessments. When this is successfully done by construction of a vertical scale that spans across grades, therefore, the question of how much a student learnt over the year becomes pertinent. It is important to note that inferences about students learning-progression is somewhat related to the quality of linking items, both content-wise and psychometric properties and the criteria used to select them within NEAT design. In this dissertation it has been demonstrated that below average ATMDD condition produced the most accurate results overall. It can be recommended that anchor items that are closely overlapping in the adjacent grades are the best choices for inclusion when constructing common item test.

5.4 Limitations

The meaning, interpretation, and conclusions concerning the results of this dissertation are to be understood within its study conditions, context, and limitations. Therefore, it would be an exaggeration to contend that this study covered every aspect and issue related to the study of vertical scales in the context of NEAT design.

One major limitation of this study was lack of consideration of content. Content was not investigated when vertical scaling was constructed— that is, the content specifications that might require a wider spread of item difficulties was never examined and that it was not an important component in this study. The second drawback of this study was that 10 replications were run for every panel. Although these ten replications could have affected overall equating results, 100 or more replications could have increased equating accuracy. Third limitation was selection of degree of between-grade mean ability difference (BGMAD). Technically, the BGMAD or the separation of grade ability distribution or group effect of 0.5, 1.0, and 1.5 could be large, but again in the context of vertical scaling the adjacent grades abilities must be different even though it is assumed that items with the same difficult parameters in the overlapping areas are ideal for sampling common items. Furthermore, in a practical situation, the students in the higher grade might have a high propensity to forget the previously covered materials in the lower grade due to time lapse and, perhaps, level of ability to retain and recall past materials in comparison with lower level grade students with whom they share the overlapping contents and where the material had been covered in recent time.

It is a common expectation in the field of vertical scaling that the construction and ultimate use of the vertical scale within the context of large-scale assessment design and analysis comes with considerable multiple psychometric challenges. To begin with, the initial design and development of a vertical scale can be tedious or time-consuming process, not always accommodating the tight deadlines encountered during large-scale assessment programs. Indeed, not all vertical scaling factors were addressed in this study. For example, use of internal common items, even though internal item sets pose the challenge of context effect and structural zeros, a fact well documented in the vertical scaling literature. Therefore, it was for this reason the external anchor was selected for this study. This is a limitation in itself because there are testing programs that use internal anchor, which means application and implication of this research in their program is less consequential. It is worthwhile to mention that spread of *b*-difficulty parameter was not factored in the study. The results of this study can only be applied in the context of the study conditions investigated in this dissertation.

5.5 Suggestions or Recommendations for Future Research Study

By the time of conducting this study, there was no known testing program that used my study or similar study to construct vertical scales. The results from this research study indicate that the design and study conditions investigated can be extended to the real world of testing, because the study examined some of empirical issues and complications that are witnessed in vertical scaling on a daily basis. It is recommended that practitioners, psychometricians, and scholars in vertical scaling widen their horizon

in searching for appropriate vertical scaling design like the one adopted in this study in order to construct a defensible vertical scale.

It is self-evident that the study was limited to a sample size of 3,000 and there were no other levels of sample size considered. On the same vein, small sample sizes like 300 or less and 2,000 sample sizes which are frequently used in testing reality should be considered in future studies. Specifically, a study involving small sample size like 300 or less examinees would be worth investigating in future study. It is a common phenomenon in testing programs to have situations where small sample size is involved. A case in point is testing students with various types of disability.

Smoothing was not considered in this study for three main reasons. First, a sample size of 3,000 was deemed to represent a relatively large sample size where performing smoothing would be unnecessary—or not be effective as it should be—when a small sample size was used. Second, in equating and vertical scaling literature, the purpose of using smoothing is to reduce the SEE or random error (Kolen & Brennan, 2014). As previously mentioned, levels or factors of sample sizes were not included in the study; therefore, it was expected the SEE would not have a significant variation either under smoothing or no smoothing analysis with a large sample size of 3,000. Third, smoothing methods are also known to introduce bias. It is recommended that in future study different levels of sample sizes would be used ranging from small to large with possible smoothing—or warranting conducting smoothing procedures. This might contribute to a more accurate estimate of the equating relationship which in turn leads to a minimal overall equating error.

Other areas that future study needs to focus on include, first, using different proportions of common items; only one proportion of 20% was considered in this study. Second, internal common items were not considered. It is suggested that the effects of internal common items be investigated in the context of vertical scaling. Third, this study did not treat issues concerning reliability and equity. Brennan (2010) has demonstrated that for curvilinear equating first-order equity—i.e., conditional expected scale scores for both old and new versions of the test are the same—and second-order equity—i.e., which holds that after equating, the conditional standard errors of measurement are the same for both the old and new forms—are more likely to be satisfied whenever reliability increases. Hence, examining the role of reliability would also give us valuable information about equating relationships and the resulting vertical scale. Also, $p(\theta)$, the generating distribution of the proficiency θ in the simulation study, which was assumed to be N(0,1) (average) could be studied together with say N (1,1) (high) or multiple levels and investigate how the two or multiple levels impact overall equating results on vertical scale. Last, although classification consistency and bias are somewhat related in that the two focus on the question of fit between generating and estimating models, the former was not explored in relation to accuracy classification of examinees. Rather, this study focused on bias statistic. Future study could investigate how the conditions of this study could affect different examinees abilities in multiple ways.

REFERENCES

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrews, K. M. (1995). *The effects of scaling design and scaling method on the primary score scale associated with a multi-level achievement test* (Unpublished doctoral dissertation, The University of Iowa).

Angoff, W. H. (1968). How we calibrate College Board scores. *College Board Review, 68,* 11–14.

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

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

Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement, 29,* 341–354.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee`s ability. In F. M. Lord & M. Novick, *Statistical theories of mental test scores* (pp. 395–549) Reading, MA: Addison-Wesley.

Bock, R. D. (1983). The mental growth curve reexamined. In D. J. Weiss (Ed.), New horizons in testing (pp. 205–209). New York: Academic Press.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13,* 1–18. doi:10.2307/1164948

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, NY: Academic Press.

Brennan, R. L. (2010). Assumptions about true scores and populations in equating. *Measurement, 8,* 1–3.

Briggs, D. C. (2010, May 3). *The problem with vertical scales.* Paper presented at the 2010 Annual Meeting of the American Educational Research Association, Denver, CO.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement, 50*(2), 204–226.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38*(6), 551–576.

Briggs, D. C., & Peck, F. A. (2015). Using Learning Progressions to Design Vertical Scales that Support Coherent Inferences about Student Growth, Measurement: *Interdisciplinary Research and Perspectives, 13*(2), 75–99.

Briggs, D. C., & Weeks, J. P. (2009a). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3–14.

Briggs, D. C., & Weeks, J. P. (2009b). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy, 4*(4), 384–414.

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

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practice, 3*(4), 15–16.

Camilli, G. (2005). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement, 36*(1), 73–78.

Camilli, G., Yamamoto, K., & Wang, M.-M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379–388.

Carlson, J. E. (2011). Statistical Models for Vertical Linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 60–62). New York, NY: Springer.

Chen, H., Yan, D., Hemat, L., Han, N., & von Davier, A. A. (2011). *LOGLIN/KE User Guide.* Princeton, NJ: Educational Testing Service. Version 3.1.

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment, 1,* 329–347.

Clemans, W. V. (1996). Reply to Yen, Burket, and Fitzpatrick. *Educational Assessment, 3,* 192–206.

Confrey, J. (2012). Better measurement of higher cognitive processes through learning trajectories and diagnostic assessments in mathematics: The challenge in adolescence. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The adolescent brain: Learning, reasoning, and decision making* (pp. 155–182). Washington, DC: American Psychological Association.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10,* 37–45.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*(3), 225–244. doi:10.1177/014662168701100302

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction.* Englewood Cliffs, NJ: Prentice-Hall.

CTB/McGraw-Hill. (1997). TerraNova, Monterey, CA: Author.

CTB/McGraw-Hill. (2001). TerraNova, Monterey, CA: Author

Custer, M., Omar, M. H., & Pomplun, M. (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education, 19,* 133–149.

Deng, N., Sukin, T., & Hambleton, R. K. (2009). *Judging the content and statistical equivalence of MCAS Operational and Linking Items.* Center for Educational Assessment MCAS Validity Report No. 20 (CEA-709). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

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

Dorans, N. J., Kubiak, A., & Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating* (ETS SR-98-02). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). New York, NY: Springer.

Dorans, N. J., Moses, T. P., & Sinharay, S. (2010). *First language of examinees and its relationship to equating* (ETS Research Rep. No. RR-10-29). Princeton, NJ: Educational Testing Service.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Science Education, 7*(3), 235–241. Retrieved from https://link.springer.com/article/10.1023/A:1021112514626

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv Health Sci Educ., 10,* 133–143.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327–333. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/14996342

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: Macmillan.

Fitzpatrick, A. R. (2008). NCME 2008 presidential address: The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice, 27*(4), 34–40.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5,* 175–186.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: John Wiley.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement, 5,* 187–201.

Gustafsson, J.-E. (1979). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement, 16,* 153–158.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150.

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

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(3), 221–235.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233–251). New York, NY: Springer.

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

Harris, D. J., & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*(2), 151–159.

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.

Hendrickson, A. B., Kolen, M. J., & Tong, Y. (2004, April). *Comparison of IRT vertical scaling from scaling-test and common item designs.* Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Hendrickson, A. B., Wei, H., & Kolen, M. J. (2005, April). *Dichotomous and polytomous scoring for IRT vertical scaling from scaling-test and common-item designs.* Paper presented at the annual meeting of the National Council for Measurement in Education, Montreal, Canada.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger Publishers.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, psychometrics* (Vol. 26, pp. 169–203). Amsterdam, The Netherlands: Elsevier.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Technical Report 89–84). Princeton, NJ: Educational Testing Service.

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

Holland, P. W., von Davier, A. A, Sinharay, S., & Han, N. (2006). *Testing the untestable assumptions of the chain and poststratification equating methods for the NEAT design* (ETS Research Rep. No. RR-06-17). Princeton, NJ: Educational Testing Service.

Holmes, S. E. (1982). Unidimesionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19*(2), 139–147.

Hoover, H. D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues & Practice, 3*(4), 8–14.

Hoover, H. D. (1984b). Rejoinder to Burket. *Educational Measurement: Issues and Practice, 3*(4), 16–18.

Hoover, H. D. (1988). Growth expectations for low-achieving students: A reply to Yen. *Educational Measurement: Issues and Practice, 7*(4), 21–23.

Hoover, H. D., Dunbar, S. D., & Frisbie, D. A. (2003). *The Iowa Tests of Basic Skills. Interpretive guide for teachers and counselors.* Forms A and B. Levels 9–14. Itasca, IL: Riverside.

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education, 21,* 187–206.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation procedures in capturing academic growth. *The Journal of Experimental Education, 71,* 229–250.

Johnson, M., & Yi, Q. (2011). *Investigating common-item screening procedures in developing a vertical scale.* Paper presented at annual meeting of the National Council of Educational Measurement, New Orleans, LA.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64), Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Kelly, T. L. (1923). *Statistical methods.* New York, NY: Macmillan.

Kish, L. (1965). *Survey sampling.* New York, NY: John Wiley and Sons, Inc.

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

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

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29–37.

Kolen, M. J. (2006). Scaling and norming. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: Praeger Publishers.

Kolen, M. J. (2011). *Issues associated with vertical scales for PARCC assessments.* Retrieved from Partnership for Assessment of Readiness for College and Careers (PARCC). Retrieved from http://www. parcconline.org/technical-advisory-committee

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

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Lei, P., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement, 36,* 21–39.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of differentability* (Research Bulletin 55–23). Princeton, NJ: Educational Testing Service.

Li, Y., & Lissitz, R. W. (2012). *Exploring the full-information bi-factor model in vertical scaling with construct shift.* NCME.

Lindquist, E. F. (1953). Selecting appropriate score scales for tests. In *Proceedings of the 1952 Invitational Conference on Testing Problems* (pp. 34–40). Princeton, NJ: Educational Testing Service.

Liou, M., Cheng, P. E., & Li, M.-Y. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement, 25,* 197–207.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken NJ: Wiley.

Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2011). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement, 71*, 346–361.

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

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, *3*(1), 73–95.

Lord, F. M. (1975). Formula scoring and numbering-right scoring. *Journal of Educational Measurement, 12*(1), 7–11.

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

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

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229–249.

Marco, G. L, Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing* (pp. 147–176). New York, NY: Academic.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Mittman, A. (1958). *An empirical study of methods of scaling achievement tests at the elementary grade level* (Unpublished doctoral dissertation, The University of Iowa, Iowa City).

Omar, M. H. (1996). *An investigation into the reasons item response theory scales show smaller variability for higher achieving groups* (Iowa Testing Programs Occasional Papers Number 39). Iowa City, IA: University of Iowa.

Omar, M. H. (1997, March). *An investigation into the reasons why IRT theta scale shrinks for higher achieving groups.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Omar, M. H. (1998, April). *Item parameter invariance assumption and its implications on vertical scaling of multilevel achievement test data.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Otis, A. S. (1916). The reliability of spelling scales, including a ‘deviation formula’ for correlation. *School of Society, 4,* 96–99.

Paek, I., & Young, M. J. (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education, 18,* 199–215.

Patz, R. J., & Yao, L. (2007). Practical issues in vertical equating. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253–272). New York, NY: Springer-Verlag

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). New York, NY: Springer.

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). Washington, DC: American Council on Education.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating method. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York, NY: Academic Press.

Phillips, S. E. (1983). Comparison of equipercentile and item response theory equating when the scaling test method is applied to a multilevel achievement battery. *Applied Psychological Measurement, 7,* 267–281.

Phillips, S. E. (1986). The effects of the deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement, 23,* 107–118.

Phillips, S. E., & Clarizio, H. F. (1988a). Conflicting growth expectations cannot both be real: A rejoinder to Yen. *Educational Measurement: Issues and Practice, 7*(4), 18–19.

Phillips, S. E., & Clarizio, H. F. (1988b). Limitations of standard scores in individual achievement testing. *Educational Measurement: Issues and Practice, 7*(1), 8–15.

Pomplun, M., Omar, M. H., & Custer, M. (2004). A comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement, 64*, 600–616.

Ricker, K., & von Davier, A. A. (2007). *The impact of anchor test length on equating results in a non-equivalent group design* (ETS Research Rept. RR-07-44). Princeton, NJ: ETS.

Schulz, E. M., & Nicewander, W. A. (1997). Grade equivalent and IRT representations of growth. *Journal of Educational Measurement, 34*, 315–331.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation & Policy Analysis, 16*, 41–49.

Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and ananchor test* (ETS RR-06-04). Princeton, NJ: Educational Testing Service.

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.

Sinharay, S., & Holland, P. W. (2008). Choice of anchor test in equating. *ETS Research Spotlight, 1,* 3–6.

Sinharay, S., & Holland, P. W. (2009). *The missing data assumptions of the NEAT design and their implications for test equating* (ETS Research Rep. No. RR-09-16). Princeton, NJ: Educational Testing Service.

Skaggs, G., & Lissitz, R. W. (1986a). An exploration of the robustness of four tests equating models. *Applied Psychological Measurement, 10,* 301–317.

Skaggs, G., & Lissitz, R. W. (1986b). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495–529.

Skaggs, G., & Lissitz, R.W. (1986c). *The effects of examinee ability on test equating invariance.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Skaggs, G., & Lissitz, R. W. (1988). The effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12,* 69–82.

Slinde, J. A., & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? *Journal of Educational Measurement, 14,* 23–32.

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement, 15,* 23–35.

Slinde, J. A., & Linn, R. L. (1979a). A note on vertical equating via the Rasch Model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement, 16,* 159–165.

Slinde, J. A., & Linn, R. L. (1979b). The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement, 3,* 437–452.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Sudman, S. (1976). *Applied sampling*. New York, NY: Academic Press.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. I, pp. 1–76). New York, NY: Wiley.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16*(7), 433–451.

Thurstone, L. L. (1927). The unit of measurement in educational scales. *The Journal of Educational Psychology, 18,* 505–524.

Thurstone, L. L. (1928). The absolute zero in intelligence measurement. *Psychological Review, 35,* 175–197.

Thurstone, L. L. (1938). Primary mental abilities. *Psychological Monographs*. No. 1.

Thurstone, L. L., & Ackerman, L. (1929). The mental growth curve for the Binet tests. *Journal of Educational Psychology, 20,* 569–583.

Tong, Y., & Kolen, M. J. (2007). Comparison of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20,* 227–253.

van der Linden, W. J. (2006). Equating error in observed-score equating. *Applied Psychological Measurement, 30*(5), 355–378.

van der Linden, W. J., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement, 34*(8), 620–640. doi:10.1177/0146621609349803

von Davier, A. A. (2011). A statistical perspective on equating test scores. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 1–17). New York, NY: Springer.

von Davier, A. A. (2011a). An observed-score equating framework. In N.J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland. (Lecture Notes in Statistics 202)* (pp. 221–238). New York, NY: Springer.

von Davier, A. A. (Ed). (2011b). *Statistics for social and behavioral sciences statistical models for test equating, scaling, and linking.* New York, NY: Springer.

von Davier, A. A., & Chen, H. (2013). *The Kernel Levine Equipercentile Observed-Score Equating Function* (Research Report No. RR-13-38). Princeton, NJ: Educational Testing Service.

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

Wang, T., Lee, W.-C., Brennan, R. L., & Kolen, M. J. (2008). A comparison of frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632–651. doi:10.1177/0146621608314943

Williams, V. S., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*(2), 93–107.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

Yan, D., von Davier, A. A., & Lewis, C. (Eds). (2014). *Computerized multistage testing: Theory and applications.* Boca Raton, FL: Chapman and Hall/CRC.

Yang, W. L. (2000). *The effects of content homogeneity and equating method on the accuracy of common-item test equating.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, *50*(4), 399–410.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299–325.

Yen, W. M. (1988). Normative growth expectations must be realistic: A response to Phillips and Clarizio. *Educational Measurement: Issues and Practice, 7*(4), 16–17.

Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*(4), 293–313.

Yen, W. M., Burket, G. R., & Fitzpatrick, A. R. (1996). Response to Clemans. *Educational Assessment, 3*, 181–190.

Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Halady (Eds.). *Handbook of test development* (pp. 472–477). Mahwah, NJ: Lawrence Erlbaum.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*, 205–218.

APPENDIX A

AVERAGE DESCRIPTIVE STATISTICS FOR ALL VERTICAL SCALING PANELS BY TEST DESIGN

Table A.1

Test Study Design 30_0.5_6: (i) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	22.01	1.21	0.64	23.13	1.32	0.75
			AT	3000	6	4.92	0.32		5.43	0.29	
			RT	3000	24	17.08	1.03		17.70	1.12	
	4	F2	TT	3000	30	22.06	1.24	0.65	23.14	1.32	0.75
			AT	3000	6	4.93	0.32		5.43	0.30	
			RT	3000	24	17.13	1.06		17.71	1.11	
	5	F4	TT	3000	30	23.56	1.12	0.63	24.82	1.13	0.71
			AT	3000	6	5.19	0.29		5.67	0.22	
		F4	RT	3000	24	18.38	0.97		19.14	0.98	
	5	F6	TT	3000	30	23.59	1.14	0.63	24.82	1.12	0.70
			AT	3000	6	5.19	0.28		5.67	0.22	
		F6	RT	3000	24	18.41	0.99		19.15	0.98	
	6	F7	TT	3000	30	24.92	1.04	0.63	26.18	0.92	0.62
			AT	3000	6	5.38	0.25		5.82	0.16	
			RT	3000	24	19.54	0.90		20.36	0.83	
	6	F8	TT	3000	30	24.91	1.01	0.60	26.16	0.93	0.62
			AT	3000	6	5.38	0.25		5.82	0.16	
			RT	3000	24	19.54	0.88		20.34	0.84	
Alternate	5	F3	TT	3000	30	23.56	1.16	0.64	24.82	1.12	0.69
			AT	3000	6	5.18	0.28		5.67	0.22	
			RT	3000	24	18.38	1.01		19.15	0.98	
	5	F5	TT	3000	30	23.59	1.13	0.65	24.82	1.11	0.69
			AT	3000	6	5.18	0.29		5.67	0.22	
			RT	3000	24	18.41	0.97		19.15	0.97	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.2

Test Study Design 30_0.5_6: (ii) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	20.98	1.33	0.62	22.47	1.52	0.82
			AT	3000	6	4.27	0.34		4.46	0.42	
			RT	3000	24	16.71	1.15		18.01	1.20	
	4	F2	TT	3000	30	21.06	1.33	0.62	22.49	1.56	0.83
			AT	3000	6	4.28	0.34		4.46	0.42	
			RT	3000	24	16.78	1.15		18.03	1.23	
	5	F4	TT	3000	30	22.76	1.24	0.64	24.50	1.34	0.81
			AT	3000	6	4.54	0.31		4.93	0.37	
		F4	RT	3000	24	18.22	1.07		19.57	1.06	
	5	F6	TT	3000	30	22.79	1.25	0.62	24.56	1.31	0.81
			AT	3000	6	4.55	0.33		4.94	0.36	
		F6	RT	3000	24	18.25	1.08		19.62	1.04	
	6	F7	TT	3000	30	24.33	1.10	0.62	26.21	1.12	0.80
			AT	3000	6	4.77	0.30		5.32	0.31	
			RT	3000	24	19.56	0.95		20.89	0.89	
	6	F8	TT	3000	30	24.31	1.11	0.62	26.24	1.12	0.80
			AT	3000	6	4.76	0.30		5.32	0.30	
			RT	3000	24	19.55	0.96		20.92	0.89	
Alternate	5	F3	TT	3000	30	22.73	1.25	0.62	24.53	1.35	0.82
			AT	3000	6	4.52	0.32		4.93	0.37	
			RT	3000	24	18.21	1.08		19.60	1.07	
	5	F5	TT	3000	30	22.77	1.27	0.63	24.53	1.37	0.82
			AT	3000	6	4.53	0.32		4.94	0.37	
			RT	3000	24	18.23	1.10		19.60	1.08	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.3

Test Study Design 30_0.5_6: (iii) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	20.64	1.45	0.72	20.58	1.49	0.70
			AT	3000	6	3.61	0.43		3.41	0.34	
			RT	3000	24	17.03	1.18		17.17	1.28	
	4	F2	TT	3000	30	20.66	1.43	0.73	20.54	1.49	0.67
			AT	3000	6	3.61	0.42		3.40	0.34	
			RT	3000	24	17.05	1.16		17.14	1.29	
	5	F4	TT	3000	30	22.49	1.36	0.72	22.55	1.39	0.71
			AT	3000	6	4.00	0.40		3.71	0.34	
		F4	RT	3000	24	18.49	1.10		18.84	1.17	
	5	F6	TT	3000	30	22.51	1.32	0.72	22.58	1.35	0.70
			AT	3000	6	4.00	0.39		3.71	0.34	
		F6	RT	3000	24	18.50	1.08		18.87	1.13	
	6	F7	TT	3000	30	24.18	1.19	0.73	24.35	1.21	0.72
			AT	3000	6	4.37	0.37		4.05	0.34	
			RT	3000	24	19.81	0.95		20.30	1.00	
	6	F8	TT	3000	30	24.23	1.16	0.73	24.37	1.22	0.73
			AT	3000	6	4.37	0.37		4.04	0.34	
			RT	3000	24	19.85	0.92		20.33	1.00	
Alternate	5	F3	TT	3000	30	22.53	1.31	0.74	22.58	1.35	0.69
			AT	3000	6	4.01	0.40		3.72	0.34	
			RT	3000	24	18.52	1.04		18.87	1.15	
	5	F5	TT	3000	30	22.54	1.31	0.71	22.58	1.38	0.69
			AT	3000	6	4.02	0.39		3.72	0.34	
			RT	3000	24	18.52	1.07		18.87	1.17	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.4

Test Study Design 30_0.5_6: (iv) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	18.58	1.52	0.74	18.10	1.59	0.75
			AT	3000	6	4.43	0.41		4.07	0.37	
			RT	3000	24	14.15	1.25		14.03	1.34	
	4	F2	TT	3000	30	18.63	1.53	0.73	18.07	1.61	0.75
			AT	3000	6	4.43	0.41		4.07	0.38	
			RT	3000	24	14.20	1.26		14.00	1.35	
	5	F4	TT	3000	30	20.71	1.43	0.72	20.30	1.51	0.75
			AT	3000	6	4.85	0.37		4.46	0.35	
		F4	RT	3000	24	15.86	1.19		15.83	1.27	
	5	F6	TT	3000	30	20.66	1.44	0.70	20.33	1.53	0.75
			AT	3000	6	4.83	0.36		4.47	0.35	
		F6	RT	3000	24	15.82	1.21		15.86	1.29	
	6	F7	TT	3000	30	22.52	1.33	0.71	22.41	1.44	0.75
			AT	3000	6	5.16	0.33		4.83	0.32	
			RT	3000	24	17.37	1.12		17.58	1.22	
	6	F8	TT	3000	30	22.53	1.32	0.69	22.41	1.39	0.74
			AT	3000	6	5.16	0.32		4.82	0.32	
			RT	3000	24	17.37	1.12		17.58	1.18	
Alternate	5	F3	TT	3000	30	20.68	1.42	0.71	20.33	1.55	0.77
			AT	3000	6	4.84	0.36		4.48	0.35	
			RT	3000	24	15.84	1.19		15.85	1.29	
	5	F5	TT	3000	30	20.66	1.42	0.72	20.33	1.58	0.76
			AT	3000	6	4.83	0.37		4.46	0.35	
			RT	3000	24	15.83	1.18		15.86	1.33	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.5

Test Study Design 30_0.5_6: (v) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	17.18	1.55	0.71	17.95	1.73	0.81
			AT	3000	6	3.42	0.43		3.46	0.47	
			RT	3000	24	13.76	1.28		14.49	1.37	
	4	F2	TT	3000	30	17.27	1.53	0.72	17.96	1.70	0.81
			AT	3000	6	3.43	0.43		3.46	0.47	
			RT	3000	24	13.84	1.26		14.50	1.34	
	5	F4	TT	3000	30	19.40	1.50	0.74	20.38	1.65	0.82
			AT	3000	6	3.87	0.42		4.02	0.46	
		F4	RT	3000	24	15.53	1.22		16.36	1.29	
	5	F6	TT	3000	30	19.38	1.51	0.74	20.40	1.60	0.82
			AT	3000	6	3.86	0.42		4.03	0.46	
		F6	RT	3000	24	15.53	1.23		16.38	1.26	
	6	F7	TT	3000	30	21.37	1.44	0.76	22.55	1.47	0.84
			AT	3000	6	4.27	0.40		4.54	0.43	
			RT	3000	24	17.10	1.16		18.01	1.13	
	6	F8	TT	3000	30	21.43	1.41	0.74	22.54	1.46	0.83
			AT	3000	6	4.29	0.40		4.55	0.42	
			RT	3000	24	17.14	1.14		17.99	1.14	
Alternate	5	F3	TT	3000	30	19.37	1.50	0.73	20.40	1.62	0.82
			AT	3000	6	3.86	0.41		4.03	0.46	
			RT	3000	24	15.51	1.23		16.37	1.27	
	5	F5	TT	3000	30	19.36	1.53	0.75	20.38	1.61	0.82
			AT	3000	6	3.86	0.43		4.02	0.46	
			RT	3000	24	15.50	1.24		16.35	1.26	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.6

Test Study Design 30_0.5_6: (vi) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	16.60	1.48	0.69	17.18	1.68	0.73
			AT	3000	6	2.72	0.44		2.64	0.42	
			RT	3000	24	13.88	1.21		14.54	1.40	
	4	F2	TT	3000	30	16.61	1.52	0.71	17.11	1.63	0.71
			AT	3000	6	2.72	0.46		2.61	0.40	
			RT	3000	24	13.90	1.24		14.50	1.38	
	5	F4	TT	3000	30	18.69	1.46	0.73	19.49	1.60	0.75
			AT	3000	6	3.15	0.45		3.06	0.42	
		F4	RT	3000	24	15.53	1.18		16.44	1.31	
	5	F6	TT	3000	30	18.65	1.51	0.72	19.50	1.59	0.75
			AT	3000	6	3.15	0.46		3.06	0.42	
		F6	RT	3000	24	15.50	1.23		16.44	1.30	
	6	F7	TT	3000	30	20.66	1.42	0.74	21.68	1.49	0.79
			AT	3000	6	3.59	0.44		3.52	0.43	
			RT	3000	24	17.07	1.13		18.16	1.18	
	6	F8	TT	3000	30	20.68	1.45	0.74	21.66	1.45	0.78
			AT	3000	6	3.60	0.45		3.52	0.42	
			RT	3000	24	17.08	1.15		18.14	1.15	
Alternate	5	F3	TT	3000	30	18.65	1.46	0.72	19.50	1.57	0.74
			AT	3000	6	3.14	0.45		3.06	0.42	
			RT	3000	24	15.51	1.17		16.44	1.29	
	5	F5	TT	3000	30	18.66	1.51	0.73	19.46	1.57	0.75
			AT	3000	6	3.13	0.47		3.05	0.42	
			RT	3000	24	15.52	1.21		16.40	1.29	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.7

Test Study Design 30_0.5_6: (vii) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	14.56	1.47	0.70	13.69	1.76	0.81
			AT	3000	6	3.53	0.43		3.44	0.51	
			RT	3000	24	11.03	1.20		10.25	1.39	
	4	F2	TT	3000	30	14.54	1.49	0.73	13.65	1.78	0.82
			AT	3000	6	3.54	0.45		3.44	0.51	
			RT	3000	24	11.00	1.21		10.21	1.40	
	5	F4	TT	3000	30	16.62	1.51	0.73	16.33	1.85	0.82
			AT	3000	6	3.99	0.43		4.07	0.49	
		F4	RT	3000	24	12.64	1.23		12.27	1.47	
	5	F6	TT	3000	30	16.57	1.52	0.72	16.33	1.84	0.83
			AT	3000	6	3.98	0.43		4.06	0.50	
		F6	RT	3000	24	12.59	1.25		12.27	1.46	
	6	F7	TT	3000	30	18.70	1.50	0.72	18.90	1.78	0.82
			AT	3000	6	4.40	0.40		4.62	0.45	
			RT	3000	24	14.30	1.24		14.28	1.44	
	6	F8	TT	3000	30	18.69	1.51	0.72	18.95	1.81	0.81
			AT	3000	6	4.40	0.41		4.63	0.44	
			RT	3000	24	14.28	1.25		14.32	1.47	
Alternate	5	F3	TT	3000	30	16.66	1.48	0.71	16.34	1.88	0.83
			AT	3000	6	4.00	0.42		4.06	0.51	
			RT	3000	24	12.66	1.22		12.28	1.48	
	5	F5	TT	3000	30	16.58	1.51	0.71	16.30	1.85	0.82
			AT	3000	6	3.97	0.42		4.07	0.50	
			RT	3000	24	12.60	1.25		12.23	1.47	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.8

Test Study Design 30_0.5_6: (viii) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	13.64	1.48	0.69	13.08	1.50	0.76
			AT	3000	6	2.68	0.44		2.57	0.43	
			RT	3000	24	10.95	1.23		10.51	1.21	
	4	F2	TT	3000	30	13.58	1.45	0.71	13.00	1.49	0.73
			AT	3000	6	2.68	0.43		2.54	0.41	
			RT	3000	24	10.90	1.19		10.46	1.22	
	5	F4	TT	3000	30	15.68	1.55	0.72	15.21	1.56	0.77
			AT	3000	6	3.10	0.45		3.01	0.44	
		F4	RT	3000	24	12.58	1.27		12.21	1.26	
	5	F6	TT	3000	30	15.67	1.5	0.71	15.23	1.59	0.76
			AT	3000	6	3.10	0.44		3.01	0.43	
		F6	RT	3000	24	12.57	1.23		12.22	1.29	
	6	F7	TT	3000	30	17.75	1.57	0.72	17.45	1.55	0.79
			AT	3000	6	3.52	0.44		3.48	0.42	
			RT	3000	24	14.22	1.29		13.98	1.25	
	6	F8	TT	3000	30	17.73	1.57	0.73	17.48	1.56	0.77
			AT	3000	6	3.51	0.44		3.49	0.42	
			RT	3000	24	14.23	1.28		13.99	1.26	
Alternate	5	F3	TT	3000	30	15.68	1.52	0.70	15.19	1.59	0.77
			AT	3000	6	3.10	0.43		3.00	0.44	
			RT	3000	24	12.58	1.25		12.19	1.28	
	5	F5	TT	3000	30	15.61	1.53	0.71	15.14	1.57	0.76
			AT	3000	6	3.09	0.44		2.99	0.43	
			RT	3000	24	12.52	1.25		12.16	1.27	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.9

Test Study Design 30_0.5_6: (ix) Total Test Length =30(6); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	12.80	1.46	0.63	12.80	1.33	0.55
			AT	3000	6	1.96	0.40		2.13	0.31	
			RT	3000	24	10.84	1.25		10.67	1.19	
	4	F2	TT	3000	30	12.76	1.42	0.63	12.75	1.32	0.52
			AT	3000	6	1.94	0.39		2.13	0.32	
			RT	3000	24	10.82	1.21		10.62	1.18	
	5	F4	TT	3000	30	14.79	1.48	0.67	14.63	1.42	0.59
			AT	3000	6	2.30	0.42		2.34	0.34	
		F4	RT	3000	24	12.48	1.24		12.29	1.25	
	5	F6	TT	3000	30	14.72	1.50	0.68	14.69	1.41	0.58
			AT	3000	6	2.29	0.43		2.34	0.34	
		F6	RT	3000	24	12.43	1.25		12.35	1.24	
	6	F7	TT	3000	30	16.89	1.53	0.70	16.70	1.49	0.65
			AT	3000	6	2.70	0.44		2.61	0.36	
			RT	3000	24	14.19	1.26		14.09	1.29	
	6	F8	TT	3000	30	16.88	1.57	0.72	16.78	1.49	0.67
			AT	3000	6	2.71	0.46		2.63	0.35	
			RT	3000	24	14.18	1.28		14.15	1.28	
Alternate	5	F3	TT	3000	30	14.75	1.51	0.68	14.69	1.44	0.60
			AT	3000	6	2.29	0.44		2.35	0.34	
			RT	3000	24	12.47	1.26		12.33	1.27	
	5	F5	TT	3000	30	14.74	1.52	0.69	14.70	1.45	0.59
			AT	3000	6	2.28	0.42		2.36	0.33	
			RT	3000	24	12.45	1.26		12.34	1.28	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.10

Test Study Design 30_1.0_6: (i) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	22.41	1.31	0.66	22.43	1.36	0.73
			AT	3000	6	4.95	0.31		5.39	0.29	
			RT	3000	24	17.46	1.13		17.04	1.16	
	4	F2	TT	3000	30	22.44	1.29	0.65	22.45	1.36	0.71
			AT	3000	6	4.96	0.31		5.39	0.29	
			RT	3000	24	17.48	1.11		17.06	1.17	
	5	F4	TT	3000	30	25.50	1.02	0.62	25.87	1.09	0.62
			AT	3000	6	5.42	0.24		5.80	0.16	
		F4	RT	3000	24	20.08	0.89		20.07	0.99	
	5	F6	TT	3000	30	25.55	1.03	0.61	25.88	1.08	0.63
			AT	3000	6	5.43	0.24		5.80	0.16	
		F6	RT	3000	24	20.12	0.90		20.08	0.99	
	6	F7	TT	3000	30	27.6	0.72	0.53	28.17	0.71	0.44
			AT	3000	6	5.70	0.17		5.94	0.08	
			RT	3000	24	21.9	0.64		22.23	0.68	
	6	F8	TT	3000	30	27.61	0.70	0.55	28.16	0.71	0.49
			AT	3000	6	5.70	0.18		5.94	0.08	
			RT	3000	24	21.9	0.62		22.22	0.67	
Alternate	5	F3	TT	3000	30	25.52	1.02	0.62	25.83	1.09	0.60
			AT	3000	6	5.42	0.24		5.79	0.16	
			RT	3000	24	20.10	0.89		20.04	1.00	
	5	F5	TT	3000	30	25.50	1.02	0.61	25.81	1.07	0.60
			AT	3000	6	5.42	0.24		5.80	0.17	
			RT	3000	24	20.09	0.90		20.02	0.98	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.11

Test Study Design 30_1.0_6: (ii) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	21.54	1.40	0.70	22.06	1.62	0.79
			AT	3000	6	4.22	0.39		4.50	0.41	
			RT	3000	24	17.32	1.16		17.55	1.32	
	4	F2	TT	3000	30	21.58	1.40	0.71	21.98	1.64	0.79
			AT	3000	6	4.23	0.40		4.48	0.40	
			RT	3000	24	17.35	1.15		17.50	1.34	
	5	F4	TT	3000	30	24.98	1.09	0.70	26.07	1.20	0.78
			AT	3000	6	4.91	0.32		5.29	0.30	
		F4	RT	3000	24	20.07	0.89		20.78	0.98	
	5	F6	TT	3000	30	24.96	1.09	0.67	26.05	1.21	0.77
			AT	3000	6	4.91	0.33		5.28	0.30	
		F6	RT	3000	24	20.05	0.90		20.77	0.99	
	6	F7	TT	3000	30	27.23	0.75	0.66	28.42	0.70	0.73
			AT	3000	6	5.40	0.25		5.74	0.19	
			RT	3000	24	21.83	0.61		22.68	0.57	
	6	F8	TT	3000	30	27.20	0.75	0.67	28.40	0.73	0.72
			AT	3000	6	5.39	0.25		5.740	0.19	
			RT	3000	24	21.80	0.61		22.66	0.61	
Alternate	5	F3	TT	3000	30	25.00	1.12	0.69	26.08	1.21	0.79
			AT	3000	6	4.92	0.33		5.29	0.31	
			RT	3000	24	20.08	0.92		20.79	0.98	
	5	F5	TT	3000	30	24.98	1.11	0.70	26.03	1.25	0.79
			AT	3000	6	4.91	0.32		5.28	0.31	
			RT	3000	24	20.07	0.92		20.75	1.02	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.12

Test Study Design 30_1.0_6: (iii) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	20.55	1.34	0.73	21.38	1.55	0.82
			AT	3000	6	3.61	0.43		3.47	0.50	
			RT	3000	24	16.93	1.07		17.91	1.17	
	4	F2	TT	3000	30	20.52	1.32	0.73	21.35	1.53	0.82
			AT	3000	6	3.61	0.43		3.47	0.49	
			RT	3000	24	16.91	1.05		17.89	1.16	
	5	F4	TT	3000	30	23.78	1.10	0.74	25.18	1.19	0.85
			AT	3000	6	4.38	0.38		4.59	0.44	
		F4	RT	3000	24	19.40	0.85		20.59	0.85	
	5	F6	TT	3000	30	23.81	1.11	0.73	25.14	1.20	0.85
			AT	3000	6	4.40	0.38		4.58	0.43	
		F6	RT	3000	24	19.41	0.87		20.55	0.86	
	6	F7	TT	3000	30	26.34	0.87	0.73	27.68	0.76	0.83
			AT	3000	6	5.05	0.31		5.41	0.29	
			RT	3000	24	21.29	0.67		22.27	0.54	
	6	F8	TT	3000	30	26.38	0.87	0.73	27.71	0.76	0.84
			AT	3000	6	5.06	0.31		5.41	0.30	
			RT	3000	24	21.32	0.67		22.29	0.53	
Alternate	5	F3	TT	3000	30	23.79	1.11	0.74	25.21	1.19	0.85
			AT	3000	6	4.39	0.38		4.60	0.44	
			RT	3000	24	19.39	0.87		20.61	0.85	
	5	F5	TT	3000	30	23.82	1.11	0.74	25.21	1.18	0.84
			AT	3000	6	4.41	0.38		4.60	0.43	
			RT	3000	24	19.41	0.87		20.62	0.85	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.13

Test Study Design 30_1.0_6: (iv) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	18.29	1.40	0.66	18.15	1.64	0.70
			AT	3000	6	4.28	0.35		4.18	0.33	
			RT	3000	24	14.01	1.20		13.97	1.43	
	4	F2	TT	3000	30	18.34	1.38	0.65	18.22	1.63	0.69
			AT	3000	6	4.29	0.35		4.18	0.33	
			RT	3000	24	14.06	1.19		14.04	1.42	
	5	F4	TT	3000	30	21.94	1.26	0.63	22.74	1.42	0.71
			AT	3000	6	4.83	0.29		4.85	0.31	
		F4	RT	3000	24	17.11	1.10		17.89	1.22	
	5	F6	TT	3000	30	21.92	1.22	0.62	22.65	1.43	0.72
			AT	3000	6	4.82	0.28		4.84	0.30	
		F6	RT	3000	24	17.10	1.07		17.81	1.23	
	6	F7	TT	3000	30	24.92	1.01	0.61	26.21	1.08	0.73
			AT	3000	6	5.23	0.24		5.42	0.24	
			RT	3000	24	19.68	0.88		20.79	0.92	
	6	F8	TT	3000	30	24.89	1.00	0.60	26.22	1.09	0.72
			AT	3000	6	5.23	0.24		5.42	0.25	
			RT	3000	24	19.66	0.88		20.80	0.93	
Alternate	5	F3	TT	3000	30	21.92	1.27	0.61	22.70	1.43	0.73
			AT	3000	6	4.83	0.28		4.85	0.30	
			RT	3000	24	17.09	1.12		17.85	1.23	
	5	F5	TT	3000	30	21.91	1.25	0.62	22.67	1.44	0.73
			AT	3000	6	4.81	0.28		4.84	0.30	
			RT	3000	24	17.09	1.10		17.83	1.24	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.14

Test Study Design 30_1.0_6: (v) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	17.55	1.47	0.73	17.63	1.76	0.79
			AT	3000	6	3.27	0.44		3.58	0.47	
			RT	3000	24	14.28	1.18		14.05	1.42	
	4	F2	TT	3000	30	17.51	1.48	0.74	17.60	1.76	0.81
			AT	3000	6	3.26	0.44		3.58	0.48	
			RT	3000	24	14.26	1.20		14.02	1.41	
	5	F4	TT	3000	30	21.5	1.35	0.75	22.43	1.53	0.79
			AT	3000	6	4.18	0.42		4.57	0.39	
		F4	RT	3000	24	17.32	1.07		17.86	1.25	
	5	F6	TT	3000	30	21.47	1.36	0.76	22.49	1.50	0.79
			AT	3000	6	4.18	0.42		4.58	0.38	
		F6	RT	3000	24	17.29	1.07		17.91	1.22	
	6	F7	TT	3000	30	24.79	1.08	0.73	26.23	1.11	0.75
			AT	3000	6	4.96	0.34		5.30	0.27	
			RT	3000	24	19.83	0.87		20.93	0.92	
	6	F8	TT	3000	30	24.78	1.11	0.75	26.22	1.12	0.76
			AT	3000	6	4.95	0.33		5.29	0.27	
			RT	3000	24	19.82	0.89		20.93	0.92	
Alternate	5	F3	TT	3000	30	21.42	1.40	0.77	22.47	1.55	0.80
			AT	3000	6	4.16	0.42		4.58	0.39	
			RT	3000	24	17.25	1.11		17.89	1.26	
	5	F5	TT	3000	30	21.45	1.33	0.73	22.48	1.57	0.80
			AT	3000	6	4.17	0.40		4.57	0.39	
			RT	3000	24	17.28	1.07		17.90	1.28	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.15

Test Study Design 30_1.0_6: (vi) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	17.08	1.46	0.68	15.95	1.78	0.79
			AT	3000	6	2.79	0.40		2.24	0.49	
			RT	3000	24	14.29	1.23		13.71	1.43	
	4	F2	TT	3000	30	17.04	1.50	0.70	16.02	1.78	0.78
			AT	3000	6	2.79	0.41		2.24	0.49	
			RT	3000	24	14.25	1.24		13.78	1.43	
	5	F4	TT	3000	30	20.98	1.36	0.73	21.26	1.72	0.84
			AT	3000	6	3.55	0.41		3.47	0.53	
		F4	RT	3000	24	17.43	1.10		17.78	1.31	
	5	F6	TT	3000	30	20.98	1.34	0.70	21.31	1.83	0.85
			AT	3000	6	3.55	0.40		3.49	0.55	
		F6	RT	3000	24	17.44	1.10		17.82	1.39	
	6	F7	TT	3000	30	24.36	1.15	0.74	25.76	1.32	0.86
			AT	3000	6	4.31	0.39		4.73	0.44	
			RT	3000	24	20.05	0.90		21.03	0.96	
	6	F8	TT	3000	30	24.37	1.12	0.73	25.74	1.38	0.86
			AT	3000	6	4.30	0.38		4.73	0.46	
			RT	3000	24	20.06	0.89		21.01	1.01	
Alternate	5	F3	TT	3000	30	20.95	1.38	0.72	21.24	1.77	0.85
			AT	3000	6	3.54	0.41		3.48	0.53	
			RT	3000	24	17.42	1.13		17.76	1.35	
	5	F5	TT	3000	30	20.95	1.38	0.72	21.24	1.78	0.84
			AT	3000	6	3.54	0.41		3.49	0.54	
			RT	3000	24	17.41	1.12		17.76	1.36	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.16

Test Study Design 30_1.0_6: (vii) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	13.99	1.40	0.69	13.78	1.65	0.77
			AT	3000	6	3.41	0.42		3.14	0.43	
			RT	3000	24	10.57	1.15		10.64	1.36	
	4	F2	TT	3000	30	13.97	1.40	0.67	13.80	1.65	0.75
			AT	3000	6	3.40	0.41		3.14	0.42	
			RT	3000	24	10.57	1.16		10.66	1.36	
	5	F4	TT	3000	30	17.90	1.46	0.68	18.59	1.69	0.80
			AT	3000	6	4.15	0.38		4.11	0.42	
		F4	RT	3000	24	13.75	1.23		14.48	1.38	
	5	F6	TT	3000	30	17.86	1.47	0.68	18.65	1.68	0.80
			AT	3000	6	4.14	0.38		4.12	0.42	
		F6	RT	3000	24	13.72	1.25		14.53	1.37	
	6	F7	TT	3000	30	21.78	1.41	0.70	23.14	1.45	0.79
			AT	3000	6	4.83	0.34		5.02	0.34	
			RT	3000	24	16.95	1.20		18.12	1.21	
	6	F8	TT	3000	30	21.78	1.34	0.67	23.19	1.42	0.78
			AT	3000	6	4.82	0.33		5.02	0.34	
			RT	3000	24	16.96	1.14		18.17	1.18	
Alternate	5	F3	TT	3000	30	17.87	1.46	0.69	18.57	1.64	0.80
			AT	3000	6	4.14	0.39		4.11	0.41	
			RT	3000	24	13.72	1.22		14.47	1.34	
	5	F5	TT	3000	30	17.83	1.45	0.67	18.62	1.68	0.80
			AT	3000	6	4.13	0.38		4.12	0.43	
			RT	3000	24	13.70	1.22		14.50	1.37	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.17

Test Study Design 30_1.0_6: (viii) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	13.64	1.42	0.70	30	18.15	0.70
			AT	3000	6	2.54	0.44		6	4.18	
			RT	3000	24	11.10	1.16		24	13.97	
	4	F2	TT	3000	30	13.63	1.42	0.68	30	18.22	0.69
			AT	3000	6	2.54	0.43		6	4.18	
			RT	3000	24	11.09	1.17		24	14.04	
	5	F4	TT	3000	30	17.55	1.46	0.73	30	22.74	0.71
			AT	3000	6	3.38	0.44		6	4.85	
		F4	RT	3000	24	14.17	1.17		24	17.89	
	5	F6	TT	3000	30	17.64	1.52	0.74	30	22.65	0.72
			AT	3000	6	3.40	0.46		6	4.84	
		F6	RT	3000	24	14.24	1.22		24	17.81	
	6	F7	TT	3000	30	21.50	1.34	0.74	30	26.21	0.73
			AT	3000	6	4.29	0.42		6	5.42	
			RT	3000	24	17.21	1.07		24	20.79	
	6	F8	TT	3000	30	21.44	1.38	0.75	30	26.22	0.72
			AT	3000	6	4.27	0.41		6	5.42	
			RT	3000	24	17.17	1.11		24	20.80	
Alternate	5	F3	TT	3000	30	17.62	1.49	0.74	30	22.70	0.73
			AT	3000	6	3.40	0.45		6	4.85	
			RT	3000	24	14.22	1.20		24	17.85	
	5	F5	TT	3000	30	17.56	1.49	0.76	30	22.67	0.73
			AT	3000	6	3.39	0.45		6	4.84	
			RT	3000	24	14.17	1.18		24	17.83	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.18

Test Study Design 30_1.0_6: (ix) Total Test Length =30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	12.82	1.34	0.59	11.91	1.50	0.70
			AT	3000	6	1.88	0.39		1.70	0.40	
			RT	3000	24	10.93	1.16		10.21	1.26	
	4	F2	TT	3000	30	12.81	1.36	0.59	11.82	1.43	0.69
			AT	3000	6	1.88	0.38		1.68	0.40	
			RT	3000	24	10.93	1.17		10.15	1.19	
	5	F4	TT	3000	30	16.55	1.45	0.68	16.35	1.69	0.78
			AT	3000	6	2.54	0.43		2.55	0.47	
		F4	RT	3000	24	14.01	1.20		13.8	1.36	
	5	F6	TT	3000	30	16.59	1.47	0.70	16.38	1.74	0.80
			AT	3000	6	2.55	0.44		2.57	0.48	
		F6	RT	3000	24	14.03	1.21		13.81	1.38	
	6	F7	TT	3000	30	20.55	1.39	0.76	21.28	1.66	0.82
			AT	3000	6	3.42	0.45		3.62	0.46	
			RT	3000	24	17.13	1.09		17.67	1.30	
	6	F8	TT	3000	30	20.54	1.39	0.74	21.29	1.64	0.82
			AT	3000	6	3.42	0.45		3.62	0.46	
			RT	3000	24	17.12	1.10		17.66	1.29	
Alternate	5	F3	TT	3000	30	16.60	1.47	0.69	16.37	1.70	0.78
			AT	3000	6	2.56	0.43		2.56	0.46	
			RT	3000	24	14.04	1.21		13.82	1.37	
	5	F5	TT	3000	30	16.60	1.48	0.70	16.37	1.75	0.79
			AT	3000	6	2.56	0.43		2.55	0.48	
			RT	3000	24	14.04	1.22		13.82	1.40	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.19

Test Study Design 30_1.5_6: (i) Total Test Length =30(6); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	22.28	1.32	0.69	22.96	1.55	0.75
			AT	3000	6	5.09	0.32		5.31	0.30	
			RT	3000	24	17.18	1.12		17.65	1.34	
	4	F2	TT	3000	30	22.30	1.35	0.69	22.98	1.47	0.75
			AT	3000	6	5.10	0.32		5.31	0.30	
			RT	3000	24	17.20	1.15		17.67	1.27	
	5	F4	TT	3000	30	26.58	0.87	0.59	27.80	0.83	0.59
			AT	3000	6	5.73	0.18		5.86	0.13	
		F4	RT	3000	24	20.86	0.78		21.94	0.76	
	5	F6	TT	3000	30	26.58	0.87	0.58	27.80	0.83	0.59
			AT	3000	6	5.72	0.18		5.86	0.13	
		F6	RT	3000	24	20.86	0.78		21.94	0.77	
	6	F7	TT	3000	30	28.77	0.45	0.38	29.57	0.31	0.39
			AT	3000	6	5.94	0.08		5.98	0.04	
			RT	3000	24	22.83	0.42		23.59	0.29	
	6	F8	TT	3000	30	28.77	0.45	0.39	29.57	0.30	0.36
			AT	3000	6	5.94	0.08		5.98	0.04	
			RT	3000	24	22.83	0.42		23.59	0.29	
Alternate	5	F3	TT	3000	30	26.63	0.88	0.57	27.83	0.85	0.61
			AT	3000	6	5.72	0.18		5.87	0.13	
			RT	3000	24	20.91	0.79		21.96	0.78	
	5	F5	TT	3000	30	26.6	0.85	0.58	27.77	0.86	0.61
			AT	3000	6	5.72	0.18		5.86	0.13	
			RT	3000	24	20.88	0.76		21.91	0.79	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.20

Test Study Design 30_1.5_6: (ii) Total Test Length=30(6); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	22.28	1.32	0.69	22.96	1.55	0.75
			AT	3000	6	5.09	0.32		5.31	0.30	
			RT	3000	24	17.18	1.12		17.65	1.34	
	4	F2	TT	3000	30	22.3	1.35	0.69	22.98	1.47	0.75
			AT	3000	6	5.10	0.32		5.31	0.30	
			RT	3000	24	17.20	1.15		17.67	1.27	
	5	F4	TT	3000	30	26.58	0.87	0.59	27.80	0.83	0.59
			AT	3000	6	5.73	0.18		5.86	0.13	
		F4	RT	3000	24	20.86	0.78		21.94	0.76	
	5	F6	TT	3000	30	26.58	0.87	0.58	27.80	0.83	0.59
			AT	3000	6	5.72	0.18		5.86	0.13	
		F6	RT	3000	24	20.86	0.78		21.94	0.77	
	6	F7	TT	3000	30	28.77	0.45	0.38	29.57	0.31	0.39
			AT	3000	6	5.94	0.08		5.98	0.04	
			RT	3000	24	22.83	0.42		23.59	0.29	
	6	F8	TT	3000	30	28.77	0.45	0.39	29.57	0.30	0.36
			AT	3000	6	5.94	0.08		5.98	0.04	
			RT	3000	24	22.83	0.42		23.59	0.29	
Alternate	5	F3	TT	3000	30	26.63	0.88	0.57	27.83	0.85	0.61
			AT	3000	6	5.72	0.18		5.87	0.13	
			RT	3000	24	20.91	0.79		21.96	0.78	
	5	F5	TT	3000	30	26.60	0.85	0.58	27.77	0.86	0.61
			AT	3000	6	5.72	0.18		5.86	0.13	
			RT	3000	24	20.88	0.76		21.91	0.79	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.21

Test Study Design 30_1.5_6: (iii) Total Test Length =30(6); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	20.38	1.32	0.68	21.48	1.36	0.77
			AT	3000	6	3.51	0.40		3.64	0.38	
			RT	3000	24	16.86	1.09		17.84	1.09	
	4	F2	TT	3000	30	20.35	1.33	0.68	21.46	1.36	0.77
			AT	3000	6	3.51	0.39		3.63	0.38	
			RT	3000	24	16.85	1.10		17.82	1.10	
	5	F4	TT	3000	30	25.18	1.00	0.70	25.88	0.87	0.77
			AT	3000	6	4.54	0.34		4.69	0.30	
		F4	RT	3000	24	20.64	0.80		21.20	0.66	
	5	F6	TT	3000	30	25.19	1.01	0.72	25.92	0.86	0.77
			AT	3000	6	4.54	0.34		4.70	0.30	
		F6	RT	3000	24	20.65	0.80		21.22	0.66	
	6	F7	TT	3000	30	28.16	0.60	0.70	28.60	0.53	0.79
			AT	3000	6	5.32	0.25		5.53	0.22	
			RT	3000	24	22.84	0.46		23.07	0.38	
	6	F8	TT	3000	30	28.16	0.60	0.71	28.61	0.53	0.78
			AT	3000	6	5.32	0.25		5.53	0.22	
			RT	3000	24	22.83	0.46		23.07	0.38	
Alternate	5	F3	TT	3000	30	25.21	1.00	0.70	25.94	0.86	0.78
			AT	3000	6	4.55	0.34		4.69	0.31	
			RT	3000	24	20.67	0.80		21.24	0.65	
	5	F5	TT	3000	30	25.18	1.06	0.73	25.91	0.88	0.78
			AT	3000	6	4.55	0.36		4.70	0.31	
			RT	3000	24	20.63	0.84		21.21	0.67	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.22

Test Study Design 30_1.5_6: (iv) Total Test Length =30(6); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	18.04	1.44	0.65	18.55	1.66	0.76
			AT	3000	6	4.19	0.34		4.31	0.36	
			RT	3000	24	13.85	1.24		14.24	1.41	
	4	F2	TT	3000	30	18.06	1.44	0.65	18.57	1.65	0.73
			AT	3000	6	4.20	0.36		4.33	0.36	
			RT	3000	24	13.86	1.24		14.24	1.41	
	5	F4	TT	3000	30	23.49	1.21	0.64	24.66	1.19	0.73
			AT	3000	6	4.99	0.29		5.28	0.25	
		F4	RT	3000	24	18.50	1.05		19.38	1.02	
	5	F6	TT	3000	30	23.52	1.22	0.63	24.65	1.16	0.73
			AT	3000	6	4.99	0.28		5.28	0.25	
		F6	RT	3000	24	18.53	1.06		19.37	0.99	
	6	F7	TT	3000	30	27.33	0.77	0.58	28.19	0.63	0.63
			AT	3000	6	5.55	0.20		5.82	0.13	
			RT	3000	24	21.78	0.67		22.37	0.55	
	6	F8	TT	3000	30	27.34	0.77	0.57	28.19	0.65	0.65
			AT	3000	6	5.55	0.20		5.82	0.14	
			RT	3000	24	21.79	0.68		22.37	0.57	
Alternate	5	F3	TT	3000	30	23.54	1.23	0.62	24.67	1.15	0.71
			AT	3000	6	5.00	0.28		5.28	0.25	
			RT	3000	24	18.54	1.08		19.39	0.99	
	5	F5	TT	3000	30	23.54	1.20	0.62	24.63	1.16	0.72
			AT	3000	6	4.99	0.28		5.28	0.25	
			RT	3000	24	18.55	1.05		19.36	1.00	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.23

Test Study Design 30_1.5_6: (v) Total Test Length =30(6); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	17.68	1.58	0.75	18.08	1.72	0.81
			AT	3000	6	3.50	0.46		3.41	0.47	
			RT	3000	24	14.18	1.27		14.67	1.37	
	4	F2	TT	3000	30	17.62	1.55	0.74	18.06	1.68	0.80
			AT	3000	6	3.48	0.46		3.42	0.46	
			RT	3000	24	14.14	1.24		14.64	1.34	
	5	F4	TT	3000	30	23.43	1.22	0.71	24.56	1.22	0.81
			AT	3000	6	4.78	0.37		4.98	0.37	
		F4	RT	3000	24	18.65	0.99		19.58	0.95	
	5	F6	TT	3000	30	23.44	1.22	0.73	24.55	1.23	0.82
			AT	3000	6	4.79	0.38		4.97	0.37	
		F6	RT	3000	24	18.65	0.98		19.58	0.96	
	6	F7	TT	3000	30	27.18	0.74	0.63	28.09	0.60	0.70
			AT	3000	6	5.55	0.23		5.78	0.17	
			RT	3000	24	21.63	0.63		22.31	0.50	
	6	F8	TT	3000	30	27.21	0.74	0.64	28.09	0.61	0.72
			AT	3000	6	5.56	0.23		5.77	0.17	
			RT	3000	24	21.65	0.62		22.32	0.50	
Alternate	5	F3	TT	3000	30	23.49	1.21	0.72	24.56	1.25	0.81
			AT	3000	6	4.79	0.37		4.97	0.37	
			RT	3000	24	18.70	0.98		19.58	0.98	
	5	F5	TT	3000	30	23.45	1.24	0.74	24.53	1.27	0.83
			AT	3000	6	4.78	0.38		4.97	0.38	
			RT	3000	24	18.67	0.99		19.57	0.98	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.24

Test Study Design 30_1.5_6: (vi) Total Test Length =30(6); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	16.82	1.47	0.71	16.54	1.66	0.77
			AT	3000	6	2.83	0.44		2.38	0.47	
			RT	3000	24	13.99	1.2		14.16	1.33	
	4	F2	TT	3000	30	16.79	1.44	0.69	16.53	1.64	0.77
			AT	3000	6	2.82	0.44		2.39	0.48	
			RT	3000	24	13.97	1.18		14.14	1.31	
	5	F4	TT	3000	30	22.48	1.28	0.72	23.29	1.39	0.84
			AT	3000	6	4.08	0.41		4.14	0.45	
		F4	RT	3000	24	18.41	1.02		19.15	1.04	
	5	F6	TT	3000	30	22.47	1.29	0.75	23.29	1.41	0.84
			AT	3000	6	4.07	0.42		4.13	0.46	
		F6	RT	3000	24	18.41	1.01		19.15	1.06	
	6	F7	TT	3000	30	26.72	0.86	0.70	27.72	0.79	0.81
			AT	3000	6	5.08	0.31		5.4	0.27	
			RT	3000	24	21.65	0.68		22.33	0.60	
	6	F8	TT	3000	30	26.7	0.86	0.71	27.72	0.79	0.80
			AT	3000	6	5.07	0.31		5.38	0.28	
			RT	3000	24	21.63	0.68		22.33	0.59	
Alternate	5	F3	TT	3000	30	22.51	1.35	0.75	23.3	1.37	0.83
			AT	3000	6	4.08	0.43		4.13	0.46	
			RT	3000	24	18.44	1.07		19.17	1.02	
	5	F5	TT	3000	30	22.48	1.28	0.72	23.32	1.38	0.85
			AT	3000	6	4.07	0.41		4.14	0.46	
			RT	3000	24	18.41	1.03		19.18	1.02	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.25

Test Study Design 30_1.5_6: (vii) Total Test Length =30(6); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	14.05	1.30	0.61	14.13	1.60	0.80
			AT	3000	6	3.54	0.35		3.58	0.48	
			RT	3000	24	10.51	1.12		10.55	1.25	
	4	F2	TT	3000	30	14.00	1.29	0.62	14.11	1.63	0.80
			AT	3000	6	3.53	0.34		3.56	0.49	
			RT	3000	24	10.46	1.11		10.55	1.27	
	5	F4	TT	3000	30	19.53	1.40	0.63	21.19	1.56	0.77
			AT	3000	6	4.40	0.32		5.08	0.35	
		F4	RT	3000	24	15.14	1.23		16.11	1.31	
	5	F6	TT	3000	30	19.56	1.38	0.63	21.19	1.51	0.75
			AT	3000	6	4.40	0.31		5.08	0.35	
		F6	RT	3000	24	15.16	1.20		16.11	1.27	
	6	F7	TT	3000	30	24.63	1.10	0.61	26.30	0.95	0.63
			AT	3000	6	5.09	0.24		5.79	0.16	
			RT	3000	24	19.55	0.97		20.51	0.85	
	6	F8	TT	3000	30	24.61	1.09	0.60	26.31	0.95	0.61
			AT	3000	6	5.10	0.25		5.80	0.16	
			RT	3000	24	19.51	0.96		20.51	0.86	
Alternate	5	F3	TT	3000	30	19.50	1.37	0.62	21.22	1.54	0.77
			AT	3000	6	4.39	0.31		5.08	0.36	
			RT	3000	24	15.10	1.20		16.14	1.29	
	5	F5	TT	3000	30	19.52	1.39	0.62	21.22	1.52	0.76
			AT	3000	6	4.40	0.31		5.08	0.35	
			RT	3000	24	15.12	1.22		16.14	1.28	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.26

Test Study Design 30_1.5_6: (viii) Total Test Length =30(6); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	13.89	1.43	0.63	12.72	1.53	0.76
			AT	3000	6	2.77	0.37		2.47	0.45	
			RT	3000	24	11.12	1.23		10.25	1.23	
	4	F2	TT	3000	30	13.88	1.43	0.65	12.72	1.48	0.76
			AT	3000	6	2.78	0.37		2.47	0.45	
			RT	3000	24	11.10	1.23		10.24	1.18	
	5	F4	TT	3000	30	19.80	1.37	0.68	19.81	1.64	0.82
			AT	3000	6	3.79	0.37		4.08	0.45	
		F4	RT	3000	24	16.01	1.15		15.73	1.30	
	5	F6	TT	3000	30	19.80	1.38	0.69	19.75	1.67	0.81
			AT	3000	6	3.80	0.36		4.07	0.46	
		F6	RT	3000	24	16.00	1.16		15.68	1.33	
	6	F7	TT	3000	30	24.60	1.02	0.69	25.76	1.10	0.77
			AT	3000	6	4.73	0.30		5.37	0.29	
			RT	3000	24	19.88	0.84		20.39	0.90	
	6	F8	TT	3000	30	24.62	1.02	0.68	25.75	1.09	0.78
			AT	3000	6	4.72	0.30		5.37	0.30	
			RT	3000	24	19.90	0.84		20.38	0.88	
Alternate	5	F3	TT	3000	30	19.83	1.39	0.69	19.80	1.65	0.81
			AT	3000	6	3.80	0.37		4.08	0.45	
			RT	3000	24	16.04	1.16		15.72	1.31	
	5	F5	TT	3000	30	19.78	1.38	0.68	19.83	1.65	0.80
			AT	3000	6	3.78	0.36		4.09	0.45	
			RT	3000	24	16.00	1.17		15.74	1.31	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.27

Test Study Design 30_1.5_6: (ix) Total Test Length =30(6); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	30	13.11	1.36	0.64	11.75	1.59	0.69
			AT	3000	6	2.27	0.41		1.76	0.41	
			RT	3000	24	10.84	1.14		9.99	1.34	
	4	F2	TT	3000	30	13.04	1.35	0.64	11.75	1.53	0.66
			AT	3000	6	2.26	0.41		1.75	0.40	
			RT	3000	24	10.78	1.14		10.01	1.30	
	5	F4	TT	3000	30	18.93	1.44	0.69	18.80	1.63	0.80
			AT	3000	6	3.40	0.41		3.17	0.48	
		F4	RT	3000	24	15.53	1.19		15.62	1.28	
	5	F6	TT	3000	30	18.92	1.42	0.70	18.79	1.66	0.81
			AT	3000	6	3.39	0.41		3.17	0.49	
		F6	RT	3000	24	15.53	1.17		15.62	1.30	
	6	F7	TT	3000	30	24.06	1.10	0.66	24.98	1.19	0.80
			AT	3000	6	4.37	0.33		4.70	0.36	
			RT	3000	24	19.69	0.92		20.29	0.93	
	6	F8	TT	3000	30	24.03	1.10	0.67	24.98	1.17	0.80
			AT	3000	6	4.37	0.33		4.69	0.36	
			RT	3000	24	19.66	0.91		20.29	0.91	
Alternate	5	F3	TT	3000	30	18.92	1.46	0.70	18.78	1.63	0.80
			AT	3000	6	3.39	0.41		3.17	0.47	
			RT	3000	24	15.52	1.21		15.61	1.28	
	5	F5	TT	3000	30	18.92	1.44	0.69	18.83	1.63	0.81
			AT	3000	6	3.39	0.41		3.18	0.48	
			RT	3000	24	15.53	1.20		15.65	1.27	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.28

Test Study Design 60_0.5_12: (i) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	44.14	2.38	0.75	46.18	2.69	0.85
			AT	3000	12	9.94	0.52		10.53	0.52	
			RT	3000	48	34.20	2.02		35.65	2.27	
	4	F2	TT	3000	60	44.14	2.52	0.78	46.24	2.68	0.85
			AT	3000	12	9.93	0.54		10.54	0.52	
			RT	3000	48	34.21	2.13		35.70	2.26	
	5	F4	TT	3000	60	47.45	2.18	0.75	49.80	2.24	0.82
			AT	3000	12	10.47	0.45		11.07	0.40	
		F4	RT	3000	48	36.98	1.86		38.72	1.92	
	5	F6	TT	3000	60	47.53	2.11	0.73	49.74	2.29	0.82
			AT	3000	12	10.50	0.44		11.07	0.41	
		F6	RT	3000	48	37.03	1.81		38.67	1.97	
	6	F7	TT	3000	60	50.35	1.88	0.73	52.55	1.83	0.78
			AT	3000	12	10.90	0.38		11.44	0.31	
			RT	3000	48	39.45	1.63		41.10	1.60	
	6	F8	TT	3000	60	50.31	1.87	0.71	52.59	1.84	0.78
			AT	3000	12	10.89	0.38		11.45	0.31	
			RT	3000	48	39.41	1.62		41.15	1.61	
Alternate	5	F3	TT	3000	60	47.49	2.18	0.76	49.72	2.29	0.83
			AT	3000	12	10.48	0.46		11.06	0.42	
			RT	3000	48	37.01	1.86		38.65	1.96	
	5	F5	TT	3000	60	47.47	2.20	0.75	49.79	2.23	0.82
			AT	3000	12	10.48	0.45		11.08	0.40	
			RT	3000	48	36.99	1.88		38.70	1.91	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.29

Test Study Design 60_0.5_12: (ii) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	43.19	2.58	0.81	44.16	3.10	0.87
			AT	3000	12	8.50	0.64		8.66	0.69	
			RT	3000	48	34.69	2.09		35.5	2.52	
	4	F2	TT	3000	60	43.17	2.58	0.80	44.15	3.12	0.87
			AT	3000	12	8.48	0.65		8.66	0.71	
			RT	3000	48	34.68	2.09		35.49	2.52	
	5	F4	TT	3000	60	46.70	2.32	0.81	48.43	2.73	0.87
			AT	3000	12	9.23	0.60		9.51	0.64	
		F4	RT	3000	48	37.47	1.86		38.92	2.20	
	5	F6	TT	3000	60	46.71	2.36	0.81	48.45	2.71	0.87
			AT	3000	12	9.24	0.60		9.51	0.61	
		F6	RT	3000	48	37.47	1.91		38.94	2.20	
	6	F7	TT	3000	60	49.86	2.03	0.78	51.94	2.24	0.86
			AT	3000	12	9.89	0.53		10.22	0.53	
			RT	3000	48	39.97	1.65		41.72	1.80	
	6	F8	TT	3000	60	49.82	2.05	0.81	51.87	2.23	0.86
			AT	3000	12	9.90	0.53		10.22	0.53	
			RT	3000	48	39.93	1.65		41.65	1.79	
Alternate	5	F3	TT	3000	60	46.75	2.32	0.81	48.39	2.69	0.86
			AT	3000	12	9.24	0.60		9.50	0.61	
			RT	3000	48	37.51	1.86		38.89	2.19	
	5	F5	TT	3000	60	46.68	2.40	0.82	48.44	2.63	0.87
			AT	3000	12	9.23	0.61		9.51	0.62	
			RT	3000	48	37.46	1.93		38.93	2.12	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.30

Test Study Design 60_0.5_12: (iii) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	40.91	2.68	0.82	42.33	3.13	0.88
			AT	3000	12	7.00	0.75		7.02	0.79	
			RT	3000	48	33.9	2.11		35.31	2.46	
	4	F2	TT	3000	60	41.09	2.71	0.82	42.38	3.04	0.87
			AT	3000	12	7.05	0.75		7.02	0.79	
			RT	3000	48	34.04	2.14		35.36	2.38	
	5	F4	TT	3000	60	44.70	2.51	0.83	46.67	2.71	0.88
			AT	3000	12	7.89	0.74		8.04	0.74	
		F4	RT	3000	48	36.81	1.95		38.63	2.09	
	5	F6	TT	3000	60	44.73	2.47	0.82	46.69	2.75	0.89
			AT	3000	12	7.91	0.72		8.05	0.75	
		F6	RT	3000	48	36.82	1.92		38.64	2.11	
	6	F7	TT	3000	60	48.13	2.21	0.83	50.3	2.27	0.88
			AT	3000	12	8.77	0.68		8.99	0.67	
			RT	3000	48	39.36	1.69		41.31	1.71	
	6	F8	TT	3000	60	48.07	2.25	0.84	50.31	2.31	0.88
			AT	3000	12	8.75	0.69		8.99	0.67	
			RT	3000	48	39.32	1.71		41.32	1.75	
Alternate	5	F3	TT	3000	60	44.78	2.48	0.83	46.59	2.72	0.87
			AT	3000	12	7.91	0.73		8.04	0.74	
			RT	3000	48	36.87	1.92		38.55	2.11	
	5	F5	TT	3000	60	44.74	2.46	0.82	46.62	2.72	0.87
			AT	3000	12	7.92	0.72		8.03	0.74	
			RT	3000	48	36.82	1.91		38.59	2.11	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.31

Test Study Design 60_0.5_12: (iv) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	28.38	2.62	0.74	37.18	3.16	0.86
			AT	3000	12	7.20	0.60		9.20	0.71	
			RT	3000	48	21.18	2.21		27.98	2.57	
	4	F2	TT	3000	60	28.33	2.64	0.74	37.11	3.20	0.86
			AT	3000	12	7.19	0.59		9.17	0.72	
			RT	3000	48	21.14	2.24		27.94	2.60	
	5	F4	TT	3000	60	32.26	2.70	0.74	41.89	3.05	0.85
			AT	3000	12	7.82	0.58		10.03	0.61	
		F4	RT	3000	48	24.44	2.30		31.86	2.55	
	5	F6	TT	3000	60	32.25	2.68	0.72	41.85	3.01	0.85
			AT	3000	12	7.82	0.56		10.03	0.62	
		F6	RT	3000	48	24.43	2.32		31.83	2.51	
	6	F7	TT	3000	60	36.29	2.79	0.72	46.29	2.77	0.82
			AT	3000	12	8.41	0.53		10.73	0.49	
			RT	3000	48	27.88	2.43		35.57	2.39	
	6	F8	TT	3000	60	36.21	2.78	0.73	46.27	2.72	0.83
			AT	3000	12	8.39	0.54		10.71	0.50	
			RT	3000	48	27.82	2.42		35.56	2.32	
Alternate	5	F3	TT	3000	60	32.26	2.78	0.73	41.82	3.04	0.84
			AT	3000	12	7.82	0.58		10.02	0.62	
			RT	3000	48	24.44	2.39		31.79	2.54	
	5	F5	TT	3000	60	32.24	2.71	0.72	41.91	3.01	0.85
			AT	3000	12	7.83	0.56		10.05	0.61	
			RT	3000	48	24.41	2.34		31.87	2.51	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.32

Test Study Design 60_0.5_12: (v) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	35.34	2.86	0.82	35.50	3.39	0.85
			AT	3000	12	7.05	0.73		7.08	0.73	
			RT	3000	48	28.29	2.30		28.42	2.80	
	4	F2	TT	3000	60	35.27	2.85	0.81	35.61	3.32	0.85
			AT	3000	12	7.04	0.72		7.13	0.72	
			RT	3000	48	28.22	2.30		28.48	2.74	
	5	F4	TT	3000	60	39.37	2.71	0.82	40.47	3.24	0.86
			AT	3000	12	7.88	0.70		8.04	0.72	
		F4	RT	3000	48	31.49	2.18		32.44	2.65	
	5	F6	TT	3000	60	39.45	2.73	0.81	40.38	3.16	0.86
			AT	3000	12	7.90	0.70		8.02	0.69	
		F6	RT	3000	48	31.55	2.20		32.36	2.58	
	6	F7	TT	3000	60	43.28	2.58	0.82	44.85	2.80	0.86
			AT	3000	12	8.71	0.66		8.91	0.65	
			RT	3000	48	34.57	2.08		35.94	2.26	
	6	F8	TT	3000	60	43.24	2.63	0.82	44.82	2.85	0.87
			AT	3000	12	8.71	0.66		8.88	0.66	
			RT	3000	48	34.53	2.12		35.94	2.30	
Alternate	5	F3	TT	3000	60	39.38	2.74	0.82	40.34	3.17	0.85
			AT	3000	12	7.90	0.70		8.01	0.70	
			RT	3000	48	31.47	2.20		32.32	2.59	
	5	F5	TT	3000	60	39.27	2.77	0.82	40.47	3.11	0.86
			AT	3000	12	7.87	0.71		8.03	0.69	
			RT	3000	48	31.40	2.23		32.43	2.55	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.33

Test Study Design 60_0.5_12: (vi) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	33.32	2.73	0.77	33.26	2.97	0.80
			AT	3000	12	5.42	0.70		5.30	0.66	
			RT	3000	48	27.89	2.23		27.96	2.48	
	4	F2	TT	3000	60	33.27	2.71	0.77	33.17	3.00	0.80
			AT	3000	12	5.41	0.70		5.28	0.66	
			RT	3000	48	27.86	2.21		27.89	2.50	
	5	F4	TT	3000	60	37.21	2.71	0.80	37.82	2.99	0.84
			AT	3000	12	6.21	0.71		6.12	0.69	
		F4	RT	3000	48	31.00	2.19		31.70	2.45	
	5	F6	TT	3000	60	37.17	2.65	0.79	37.78	3.05	0.84
			AT	3000	12	6.19	0.71		6.11	0.72	
		F6	RT	3000	48	30.98	2.14		31.67	2.47	
	6	F7	TT	3000	60	40.99	2.51	0.81	42.16	2.95	0.86
			AT	3000	12	7.04	0.70		7.00	0.73	
			RT	3000	48	33.95	1.99		35.17	2.36	
	6	F8	TT	3000	60	40.97	2.57	0.81	42.15	2.87	0.86
			AT	3000	12	7.03	0.71		6.99	0.70	
			RT	3000	48	33.94	2.03		35.17	2.30	
Alternate	5	F3	TT	3000	60	37.14	2.67	0.80	37.71	3.08	0.84
			AT	3000	12	6.20	0.72		6.11	0.71	
			RT	3000	48	30.94	2.14		31.61	2.51	
	5	F5	TT	3000	60	37.22	2.64	0.79	37.68	3.04	0.84
			AT	3000	12	6.22	0.70		6.08	0.69	
			RT	3000	48	31.00	2.12		31.60	2.49	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.34

Test Study Design 60_0.5_12: (vii) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	28.38	2.62	0.74	27.67	3.02	0.83
			AT	3000	12	7.20	0.60		7.10	0.7	
			RT	3000	48	21.18	2.21		20.57	2.47	
	4	F2	TT	3000	60	28.33	2.64	0.74	27.69	3.02	0.84
			AT	3000	12	7.19	0.59		7.10	0.71	
			RT	3000	48	21.14	2.24		20.59	2.46	
	5	F4	TT	3000	60	32.26	2.70	0.74	32.32	3.06	0.83
			AT	3000	12	7.82	0.58		7.97	0.66	
		F4	RT	3000	48	24.44	2.30		24.35	2.54	
	5	F6	TT	3000	60	32.25	2.68	0.72	32.35	3.09	0.82
			AT	3000	12	7.82	0.56		7.98	0.64	
		F6	RT	3000	48	24.43	2.32		24.37	2.59	
	6	F7	TT	3000	60	36.29	2.79	0.72	36.99	3.06	0.84
			AT	3000	12	8.41	0.53		8.77	0.59	
			RT	3000	48	27.88	2.43		28.22	2.59	
	6	F8	TT	3000	60	36.21	2.78	0.73	37.00	3.07	0.83
			AT	3000	12	8.39	0.54		8.76	0.61	
			RT	3000	48	27.82	2.42		28.24	2.58	
Alternate	5	F3	TT	3000	60	32.26	2.78	0.73	32.27	3.17	0.84
			AT	3000	12	7.82	0.58		7.96	0.67	
			RT	3000	48	24.44	2.39		24.31	2.63	
	5	F5	TT	3000	60	32.24	2.71	0.72	32.37	3.11	0.84
			AT	3000	12	7.83	0.56		8.00	0.65	
			RT	3000	48	24.41	2.34		24.38	2.58	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.35

Test Study Design 60_0.5_12: (viii) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	26.94	2.76	0.79	26.61	2.67	0.79
			AT	3000	12	5.29	0.72		5.52	0.62	
			RT	3000	48	21.65	2.23		21.09	2.22	
	4	F2	TT	3000	60	26.96	2.73	0.78	26.55	2.70	0.80
			AT	3000	12	5.31	0.71		5.51	0.63	
			RT	3000	48	21.65	2.23		21.04	2.23	
	5	F4	TT	3000	60	31.08	2.87	0.80	30.57	2.88	0.80
			AT	3000	12	6.15	0.73		6.23	0.63	
		F4	RT	3000	48	24.92	2.32		24.34	2.40	
	5	F6	TT	3000	60	31.04	2.82	0.81	30.58	2.86	0.80
			AT	3000	12	6.14	0.73		6.23	0.63	
		F6	RT	3000	48	24.90	2.26		24.35	2.38	
	6	F7	TT	3000	60	35.16	2.85	0.81	34.9	2.88	0.82
			AT	3000	12	7.01	0.73		6.98	0.63	
			RT	3000	48	28.15	2.29		27.92	2.39	
	6	F8	TT	3000	60	35.19	2.84	0.81	34.96	2.85	0.81
			AT	3000	12	7.01	0.74		6.98	0.61	
			RT	3000	48	28.18	2.28		27.98	2.38	
Alternate	5	F3	TT	3000	60	31.08	2.79	0.80	30.64	2.85	0.80
			AT	3000	12	6.15	0.72		6.23	0.63	
			RT	3000	48	24.93	2.26		24.41	2.37	
	5	F5	TT	3000	60	30.98	2.83	0.82	30.61	2.79	0.80
			AT	3000	12	6.13	0.74		6.23	0.61	
			RT	3000	48	24.85	2.27		24.39	2.33	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.36

Test Study Design 60_0.5_12: (ix) Total Test Length=60(12); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	25.54	2.55	0.71	23.39	2.78	0.72
			AT	3000	12	4.05	0.62		3.86	0.56	
			RT	3000	48	21.48	2.16		19.53	2.41	
	4	F2	TT	3000	60	25.58	2.56	0.71	23.39	2.82	0.72
			AT	3000	12	4.06	0.61		3.85	0.58	
			RT	3000	48	21.52	2.17		19.54	2.43	
	5	F4	TT	3000	60	29.28	2.73	0.75	27.67	3.17	0.76
			AT	3000	12	4.68	0.67		4.47	0.60	
		F4	RT	3000	48	24.60	2.27		23.21	2.74	
	5	F6	TT	3000	60	29.36	2.74	0.74	27.63	3.12	0.76
			AT	3000	12	4.70	0.68		4.45	0.61	
		F6	RT	3000	48	24.66	2.28		23.17	2.68	
	6	F7	TT	3000	60	33.54	2.85	0.78	32.50	3.38	0.80
			AT	3000	12	5.45	0.69		5.18	0.65	
			RT	3000	48	28.09	2.36		27.32	2.88	
	6	F8	TT	3000	60	33.41	2.87	0.78	32.53	3.41	0.80
			AT	3000	12	5.43	0.70		5.18	0.66	
			RT	3000	48	27.98	2.37		27.34	2.91	
Alternate	5	F3	TT	3000	60	29.43	2.69	0.75	27.66	3.15	0.77
			AT	3000	12	4.71	0.65		4.46	0.63	
			RT	3000	48	24.72	2.25		23.19	2.70	
	5	F5	TT	3000	60	29.35	2.78	0.76	27.65	3.22	0.77
			AT	3000	12	4.69	0.66		4.47	0.62	
			RT	3000	48	24.66	2.32		23.18	2.77	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.37

Test Study Design 60_1.0_12: (i) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	44.24	2.45	0.76	44.91	2.76	0.82
			AT	3000	12	9.83	0.53		10.19	0.49	
			RT	3000	48	34.42	2.08		34.71	2.38	
	4	F2	TT	3000	60	44.23	2.53	0.77	44.94	2.74	0.81
			AT	3000	12	9.83	0.52		10.20	0.49	
			RT	3000	48	34.40	2.16		34.73	2.36	
	5	F4	TT	3000	60	50.44	1.92	0.70	52.00	2.06	0.77
			AT	3000	12	10.80	0.40		11.12	0.34	
		F4	RT	3000	48	39.64	1.67		40.88	1.82	
	5	F6	TT	3000	60	50.40	1.90	0.70	51.98	2.05	0.74
			AT	3000	12	10.80	0.40		11.12	0.32	
		F6	RT	3000	48	39.60	1.64		40.86	1.82	
	6	F7	TT	3000	60	54.70	1.30	0.64	56.49	1.32	0.69
			AT	3000	12	11.40	0.27		11.62	0.20	
			RT	3000	48	43.31	1.15		44.88	1.19	
	6	F8	TT	3000	60	54.74	1.29	0.63	56.55	1.29	0.68
			AT	3000	12	11.39	0.27		11.63	0.21	
			RT	3000	48	43.35	1.14		44.92	1.16	
Alternate	5	F3	TT	3000	60	50.48	1.88	0.71	52.11	2.04	0.76
			AT	3000	12	10.82	0.40		11.13	0.33	
			RT	3000	48	39.66	1.62		40.98	1.81	
	5	F5	TT	3000	60	50.39	1.93	0.73	51.92	2.08	0.76
			AT	3000	12	10.79	0.40		11.11	0.34	
			RT	3000	48	39.60	1.66		40.81	1.84	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.38

Test Study Design 60_1.0_12: (ii) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	43.25	2.56	0.78	44.79	2.89	0.85
			AT	3000	12	8.74	0.60		8.98	0.61	
			RT	3000	48	34.51	2.12		35.81	2.39	
	4	F2	TT	3000	60	43.18	2.56	0.78	44.63	2.94	0.85
			AT	3000	12	8.71	0.59		8.95	0.63	
			RT	3000	48	34.47	2.13		35.68	2.43	
	5	F4	TT	3000	60	49.59	1.99	0.74	52.13	2.14	0.84
			AT	3000	12	9.84	0.46		10.31	0.48	
		F4	RT	3000	48	39.75	1.68		41.82	1.76	
	5	F6	TT	3000	60	49.65	2.00	0.75	52.14	2.19	0.84
			AT	3000	12	9.85	0.47		10.31	0.48	
		F6	RT	3000	48	39.81	1.68		41.83	1.81	
	6	F7	TT	3000	60	54.21	1.40	0.72	56.61	1.29	0.79
			AT	3000	12	10.70	0.37		11.19	0.32	
			RT	3000	48	43.51	1.16		45.42	1.06	
	6	F8	TT	3000	60	54.23	1.40	0.72	56.64	1.31	0.80
			AT	3000	12	10.70	0.37		11.20	0.33	
			RT	3000	48	43.53	1.17		45.44	1.06	
Alternate	5	F3	TT	3000	60	49.65	1.96	0.75	52.16	2.11	0.83
			AT	3000	12	9.86	0.46		10.30	0.47	
			RT	3000	48	39.8	1.65		41.86	1.74	
	5	F5	TT	3000	60	49.64	1.99	0.75	52.11	2.14	0.83
			AT	3000	12	9.86	0.47		10.30	0.47	
			RT	3000	48	39.78	1.67		41.80	1.77	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.39

Test Study Design 60_1.0_12: (iii) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	41.59	2.61	0.83	42.12	3.05	0.90
			AT	3000	12	7.23	0.74		7.04	0.88	
			RT	3000	48	34.37	2.05		35.08	2.29	
	4	F2	TT	3000	60	41.52	2.65	0.83	42.13	3.01	0.90
			AT	3000	12	7.21	0.75		7.05	0.89	
			RT	3000	48	34.31	2.07		35.09	2.25	
	5	F4	TT	3000	60	48.36	2.08	0.82	50.25	2.48	0.91
			AT	3000	12	8.85	0.62		9.33	0.78	
		F4	RT	3000	48	39.50	1.61		40.92	1.80	
	5	F6	TT	3000	60	48.27	2.13	0.83	50.25	2.44	0.91
			AT	3000	12	8.83	0.64		9.30	0.77	
		F6	RT	3000	48	39.44	1.63		40.95	1.77	
	6	F7	TT	3000	60	53.36	1.52	0.79	55.67	1.55	0.90
			AT	3000	12	10.09	0.48		10.92	0.51	
			RT	3000	48	43.27	1.18		44.75	1.11	
	6	F8	TT	3000	60	53.29	1.53	0.79	55.69	1.49	0.89
			AT	3000	12	10.08	0.47		10.92	0.50	
			RT	3000	48	43.21	1.19		44.77	1.06	
Alternate	5	F3	TT	3000	60	48.33	2.13	0.83	50.35	2.38	0.90
			AT	3000	12	8.85	0.64		9.34	0.77	
			RT	3000	48	39.48	1.64		41.01	1.72	
	5	F5	TT	3000	60	48.30	2.11	0.81	50.32	2.39	0.91
			AT	3000	12	8.84	0.63		9.34	0.76	
			RT	3000	48	39.46	1.64		40.98	1.73	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.40

Test Study Design 60_1.0_12: (iv) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	36.42	2.85	0.78	36.84	2.91	0.80
			AT	3000	12	8.77	0.62		8.58	0.56	
			RT	3000	48	27.64	2.40		28.26	2.48	
	4	F2	TT	3000	60	36.44	2.82	0.78	36.92	2.93	0.80
			AT	3000	12	8.78	0.62		8.60	0.56	
			RT	3000	48	27.66	2.37		28.31	2.51	
	5	F4	TT	3000	60	44.42	2.56	0.73	45.09	2.53	0.79
			AT	3000	12	10.02	0.49		9.78	0.46	
		F4	RT	3000	48	34.41	2.23		35.31	2.19	
	5	F6	TT	3000	60	44.34	2.55	0.74	45.03	2.53	0.78
			AT	3000	12	10.00	0.49		9.78	0.46	
		F6	RT	3000	48	34.34	2.21		35.26	2.19	
	6	F7	TT	3000	60	50.89	1.90	0.67	51.57	1.93	0.78
			AT	3000	12	10.86	0.36		10.72	0.37	
			RT	3000	48	40.04	1.68		40.85	1.66	
	6	F8	TT	3000	60	50.87	1.97	0.69	51.62	1.98	0.78
			AT	3000	12	10.86	0.36		10.72	0.37	
			RT	3000	48	40.01	1.74		40.90	1.71	
Alternate	5	F3	TT	3000	60	44.30	2.57	0.74	44.97	2.55	0.79
			AT	3000	12	9.99	0.50		9.77	0.45	
			RT	3000	48	34.31	2.22		35.19	2.21	
	5	F5	TT	3000	60	44.45	2.53	0.75	45.01	2.55	0.79
			AT	3000	12	10.01	0.50		9.78	0.46	
			RT	3000	48	34.44	2.18		35.23	2.21	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.41

Test Study Design 60_1.0_12: (v) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	34.66	2.86	0.82	35.36	3.34	0.89
			AT	3000	12	6.73	0.72		7.03	0.84	
			RT	3000	48	27.94	2.31		28.33	2.62	
	4	F2	TT	3000	60	34.78	2.83	0.81	35.27	3.25	0.88
			AT	3000	12	6.75	0.71		7.03	0.80	
			RT	3000	48	28.03	2.29		28.25	2.58	
	5	F4	TT	3000	60	42.84	2.64	0.82	44.90	2.93	0.88
			AT	3000	12	8.41	0.67		9.11	0.71	
		F4	RT	3000	48	34.43	2.13		35.79	2.33	
	5	F6	TT	3000	60	43.00	2.65	0.82	44.88	2.88	0.88
			AT	3000	12	8.43	0.67		9.10	0.70	
		F6	RT	3000	48	34.57	2.14		35.78	2.29	
	6	F7	TT	3000	60	49.86	2.12	0.80	52.01	2.08	0.86
			AT	3000	12	9.89	0.55		10.63	0.50	
			RT	3000	48	39.98	1.71		41.38	1.67	
	6	F8	TT	3000	60	49.86	2.12	0.82	52.06	2.08	0.86
			AT	3000	12	9.88	0.55		10.63	0.51	
			RT	3000	48	39.98	1.70		41.43	1.67	
Alternate	5	F3	TT	3000	60	42.96	2.68	0.82	44.88	2.95	0.89
			AT	3000	12	8.44	0.67		9.10	0.71	
			RT	3000	48	34.52	2.16		35.78	2.34	
	5	F5	TT	3000	60	42.81	2.67	0.82	44.84	2.94	0.89
			AT	3000	12	8.41	0.66		9.09	0.71	
			RT	3000	48	34.40	2.15		35.75	2.33	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.42

Test Study Design 60_1.0_12: (vi) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	33.00	2.73	0.77	32.75	2.88	0.78
			AT	3000	12	5.40	0.70		5.01	0.66	
			RT	3000	48	27.60	2.24		27.74	2.40	
	4	F2	TT	3000	60	33.01	2.70	0.77	32.69	2.91	0.79
			AT	3000	12	5.41	0.71		4.98	0.66	
			RT	3000	48	27.60	2.20		27.70	2.42	
	5	F4	TT	3000	60	40.88	2.58	0.81	41.54	2.85	0.86
			AT	3000	12	7.07	0.71		6.71	0.72	
		F4	RT	3000	48	33.81	2.05		34.83	2.26	
	5	F6	TT	3000	60	40.84	2.60	0.81	41.54	2.87	0.85
			AT	3000	12	7.05	0.72		6.71	0.75	
		F6	RT	3000	48	33.80	2.06		34.83	2.27	
	6	F7	TT	3000	60	47.92	2.17	0.82	49.40	2.32	0.87
			AT	3000	12	8.72	0.65		8.59	0.68	
			RT	3000	48	39.20	1.69		40.81	1.75	
	6	F8	TT	3000	60	47.80	2.14	0.82	49.29	2.34	0.88
			AT	3000	12	8.70	0.65		8.57	0.70	
			RT	3000	48	39.10	1.65		40.72	1.76	
Alternate	5	F3	TT	3000	60	40.91	2.56	0.80	41.54	2.82	0.85
			AT	3000	12	7.06	0.71		6.72	0.73	
			RT	3000	48	33.85	2.04		34.82	2.24	
	5	F5	TT	3000	60	40.83	2.63	0.82	41.52	2.77	0.85
			AT	3000	12	7.07	0.73		6.71	0.72	
			RT	3000	48	33.75	2.07		34.81	2.19	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.43

Test Study Design 60_1.0_12: (vii) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	28.52	2.77	0.79	27.45	2.84	0.81
			AT	3000	12	7.10	0.70		6.77	0.63	
			RT	3000	48	21.42	2.26		20.68	2.35	
	4	F2	TT	3000	60	28.56	2.87	0.80	27.45	2.81	0.81
			AT	3000	12	7.11	0.73		6.76	0.62	
			RT	3000	48	21.45	2.32		20.69	2.33	
	5	F4	TT	3000	60	36.94	2.85	0.79	36.53	3.11	0.83
			AT	3000	12	8.72	0.63		8.35	0.61	
		F4	RT	3000	48	28.22	2.38		28.18	2.63	
	5	F6	TT	3000	60	36.91	2.85	0.80	36.4	3.14	0.84
			AT	3000	12	8.71	0.64		8.33	0.61	
		F6	RT	3000	48	28.20	2.37		28.07	2.65	
	6	F7	TT	3000	60	44.83	2.53	0.77	45.34	2.75	0.82
			AT	3000	12	10.03	0.52		9.77	0.52	
			RT	3000	48	34.81	2.15		35.57	2.34	
	6	F8	TT	3000	60	44.77	2.51	0.78	45.37	2.73	0.83
			AT	3000	12	10.02	0.51		9.78	0.53	
			RT	3000	48	34.75	2.14		35.60	2.31	
Alternate	5	F3	TT	3000	60	36.96	2.93	0.80	36.48	3.16	0.84
			AT	3000	12	8.72	0.66		8.34	0.63	
			RT	3000	48	28.24	2.43		28.14	2.66	
	5	F5	TT	3000	60	36.89	2.85	0.79	36.48	3.11	0.82
			AT	3000	12	8.71	0.63		8.35	0.60	
			RT	3000	48	28.18	2.38		28.13	2.64	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.44

Test Study Design 60_1.0_12: (viii) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	27.42	2.48	0.73	25.10	2.97	0.85
			AT	3000	12	5.53	0.62		4.77	0.79	
			RT	3000	48	21.89	2.07		20.34	2.33	
	4	F2	TT	3000	60	27.54	2.53	0.74	25.15	3.06	0.86
			AT	3000	12	5.55	0.62		4.75	0.81	
			RT	3000	48	21.99	2.12		20.40	2.40	
	5	F4	TT	3000	60	34.96	2.64	0.76	34.86	3.33	0.88
			AT	3000	12	6.87	0.62		7.02	0.86	
		F4	RT	3000	48	28.09	2.20		27.83	2.60	
	5	F6	TT	3000	60	34.99	2.65	0.76	34.92	3.44	0.89
			AT	3000	12	6.88	0.63		7.04	0.86	
		F6	RT	3000	48	28.11	2.22		27.88	2.70	
	6	F7	TT	3000	60	42.48	2.49	0.78	44.88	3.14	0.89
			AT	3000	12	8.24	0.61		9.25	0.77	
			RT	3000	48	34.24	2.06		35.63	2.48	
	6	F8	TT	3000	60	42.55	2.53	0.78	44.93	3.15	0.89
			AT	3000	12	8.25	0.62		9.25	0.76	
			RT	3000	48	34.30	2.08		35.67	2.50	
Alternate	5	F3	TT	3000	60	35.00	2.68	0.78	34.82	3.45	0.89
			AT	3000	12	6.88	0.65		7.01	0.89	
			RT	3000	48	28.12	2.22		27.82	2.69	
	5	F5	TT	3000	60	35.01	2.66	0.77	34.89	3.36	0.88
			AT	3000	12	6.89	0.63		7.02	0.86	
			RT	3000	48	28.12	2.21		27.87	2.63	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.45

Test Study Design 60_1.0_12: (ix) Total Test Length=60(12); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	25.67	2.53	0.69	23.74	2.84	0.72
			AT	3000	12	4.25	0.59		3.44	0.61	
			RT	3000	48	21.42	2.16		20.30	2.44	
	4	F2	TT	3000	60	25.68	2.65	0.71	23.67	2.82	0.73
			AT	3000	12	4.25	0.61		3.44	0.62	
			RT	3000	48	21.44	2.25		20.24	2.41	
	5	F4	TT	3000	60	33.53	2.87	0.76	32.77	3.23	0.83
			AT	3000	12	5.51	0.66		4.96	0.77	
		F4	RT	3000	48	28.02	2.40		27.81	2.62	
	5	F6	TT	3000	60	33.53	2.84	0.75	32.76	3.17	0.81
			AT	3000	12	5.52	0.66		4.94	0.74	
		F6	RT	3000	48	28.02	2.39		27.81	2.60	
	6	F7	TT	3000	60	41.81	2.67	0.79	42.3	3.09	0.88
			AT	3000	12	7.05	0.68		7.00	0.83	
			RT	3000	48	34.76	2.17		35.30	2.39	
	6	F8	TT	3000	60	41.88	2.67	0.78	42.26	3.01	0.87
			AT	3000	12	7.07	0.68		6.98	0.80	
			RT	3000	48	34.81	2.18		35.28	2.35	
Alternate	5	F3	TT	3000	60	33.56	2.90	0.77	32.81	3.22	0.82
			AT	3000	12	5.52	0.67		4.96	0.76	
			RT	3000	48	28.04	2.43		27.84	2.63	
	5	F5	TT	3000	60	33.68	2.85	0.77	32.71	3.16	0.83
			AT	3000	12	5.54	0.67		4.93	0.76	
			RT	3000	48	28.14	2.37		27.78	2.57	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.46

Test Study Design 60_1.5_12: (i) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	44.85	2.40	0.77	45.67	2.54	0.83
			AT	3000	12	10.14	0.53		9.94	0.49	
			RT	3000	48	34.72	2.02		35.73	2.16	
	4	F2	TT	3000	60	44.90	2.45	0.79	45.79	2.56	0.83
			AT	3000	12	10.14	0.54		9.96	0.50	
			RT	3000	48	34.76	2.05		35.83	2.16	
	5	F4	TT	3000	60	53.17	1.49	0.67	54.19	1.46	0.79
			AT	3000	12	11.37	0.29		11.32	0.30	
		F4	RT	3000	48	41.79	1.32		42.87	1.24	
	5	F6	TT	3000	60	53.23	1.53	0.68	54.19	1.45	0.78
			AT	3000	12	11.38	0.30		11.32	0.30	
		F6	RT	3000	48	41.85	1.35		42.86	1.23	
	6	F7	TT	3000	60	57.48	0.77	0.51	58.06	0.66	0.60
			AT	3000	12	11.84	0.14		11.87	0.13	
			RT	3000	48	45.64	0.71		46.19	0.60	
	6	F8	TT	3000	60	57.51	0.77	0.48	58.07	0.66	0.61
			AT	3000	12	11.84	0.14		11.86	0.13	
			RT	3000	48	45.66	0.71		46.21	0.59	
Alternate	5	F3	TT	3000	60	53.19	1.54	0.68	54.16	1.44	0.77
			AT	3000	12	11.38	0.30		11.32	0.29	
			RT	3000	48	41.81	1.36		42.84	1.23	
	5	F5	TT	3000	60	53.21	1.56	0.68	54.22	1.43	0.77
			AT	3000	12	11.38	0.30		11.33	0.29	
			RT	3000	48	41.83	1.38		42.89	1.22	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.47

Test Study Design 60_1.5_12: (ii) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	42.47	2.49	0.78	44.69	3.01	0.88
			AT	3000	12	8.66	0.58		8.87	0.71	
			RT	3000	48	33.82	2.07		35.82	2.41	
	4	F2	TT	3000	60	42.51	2.49	0.78	44.62	2.97	0.88
			AT	3000	12	8.66	0.58		8.83	0.70	
			RT	3000	48	33.85	2.07		35.78	2.37	
	5	F4	TT	3000	60	51.65	1.77	0.76	54.64	1.67	0.85
			AT	3000	12	10.33	0.44		10.98	0.41	
		F4	RT	3000	48	41.32	1.46		43.66	1.34	
	5	F6	TT	3000	60	51.67	1.74	0.74	54.64	1.66	0.84
			AT	3000	12	10.34	0.43		10.98	0.42	
		F6	RT	3000	48	41.33	1.45		43.66	1.33	
	6	F7	TT	3000	60	56.80	0.94	0.69	58.56	0.62	0.72
			AT	3000	12	11.35	0.28		11.81	0.17	
			RT	3000	48	45.45	0.78		46.75	0.51	
	6	F8	TT	3000	60	56.80	0.94	0.70	58.55	0.61	0.72
			AT	3000	12	11.34	0.28		11.80	0.17	
			RT	3000	48	45.46	0.77		46.75	0.51	
Alternate	5	F3	TT	3000	60	51.62	1.77	0.76	54.62	1.64	0.84
			AT	3000	12	10.33	0.44		10.98	0.41	
			RT	3000	48	41.29	1.46		43.64	1.31	
	5	F5	TT	3000	60	51.64	1.74	0.75	54.64	1.66	0.85
			AT	3000	12	10.33	0.44		10.98	0.42	
			RT	3000	48	41.31	1.44		43.65	1.32	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.48

Test Study Design 60_1.5_12: (iii) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	41.57	2.59	0.78	41.53	2.88	0.85
			AT	3000	12	6.87	0.68		6.54	0.76	
			RT	3000	48	34.70	2.10		34.99	2.27	
	4	F2	TT	3000	60	41.59	2.50	0.78	41.65	2.90	0.86
			AT	3000	12	6.86	0.67		6.56	0.77	
			RT	3000	48	34.73	2.02		35.09	2.28	
	5	F4	TT	3000	60	50.85	1.79	0.81	52.55	1.94	0.90
			AT	3000	12	9.08	0.58		9.60	0.67	
		F4	RT	3000	48	41.76	1.36		42.95	1.37	
	5	F6	TT	3000	60	50.84	1.84	0.82	52.48	1.95	0.89
			AT	3000	12	9.09	0.60		9.58	0.66	
		F6	RT	3000	48	41.75	1.40		42.91	1.39	
	6	F7	TT	3000	60	56.14	0.98	0.78	57.98	0.87	0.85
			AT	3000	12	10.63	0.40		11.33	0.33	
			RT	3000	48	45.51	0.71		46.64	0.61	
	6	F8	TT	3000	60	56.16	0.98	0.77	57.98	0.86	0.85
			AT	3000	12	10.65	0.39		11.33	0.33	
			RT	3000	48	45.51	0.73		46.65	0.60	
Alternate	5	F3	TT	3000	60	50.91	1.76	0.79	52.50	1.96	0.89
			AT	3000	12	9.11	0.57		9.59	0.68	
			RT	3000	48	41.8	1.35		42.91	1.38	
	5	F5	TT	3000	60	50.86	1.79	0.81	52.50	1.98	0.90
			AT	3000	12	9.08	0.59		9.58	0.68	
			RT	3000	48	41.77	1.35		42.92	1.40	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.49

Test Study Design 60_1.5_12: (iv) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	36.46	2.94	0.82	36.54	3.28	0.86
			AT	3000	12	8.81	0.69		9.06	0.73	
			RT	3000	48	27.65	2.40		27.48	2.68	
	4	F2	TT	3000	60	36.46	2.90	0.82	36.72	3.36	0.87
			AT	3000	12	8.82	0.70		9.08	0.77	
			RT	3000	48	27.65	2.36		27.64	2.72	
	5	F4	TT	3000	60	48.25	2.32	0.75	50.09	2.44	0.80
			AT	3000	12	10.83	0.44		11.22	0.41	
		F4	RT	3000	48	37.42	2.01		38.86	2.13	
	5	F6	TT	3000	60	48.31	2.30	0.75	50.05	2.54	0.80
			AT	3000	12	10.83	0.43		11.21	0.41	
		F6	RT	3000	48	37.48	1.99		38.84	2.22	
	6	F7	TT	3000	60	55.66	1.29	0.61	57.32	1.11	0.60
			AT	3000	12	11.68	0.21		11.88	0.14	
			RT	3000	48	43.98	1.18		45.45	1.03	
	6	F8	TT	3000	60	55.58	1.31	0.62	57.27	1.14	0.63
			AT	3000	12	11.67	0.21		11.88	0.14	
			RT	3000	48	43.92	1.19		45.40	1.06	
Alternate	5	F3	TT	3000	60	48.3	2.36	0.77	50.10	2.46	0.79
			AT	3000	12	10.83	0.45		11.22	0.40	
			RT	3000	48	37.47	2.04		38.88	2.16	
	5	F5	TT	3000	60	48.28	2.31	0.75	50.05	2.51	0.80
			AT	3000	12	10.83	0.43		11.22	0.40	
			RT	3000	48	37.45	2.00		38.84	2.20	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.50

Test Study Design 60_1.5_12: (v) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	35.09	2.63	0.78	33.86	3.16	0.85
			AT	3000	12	6.95	0.67		6.76	0.72	
			RT	3000	48	28.14	2.16		27.10	2.57	
	4	F2	TT	3000	60	35.12	2.68	0.78	33.94	3.15	0.85
			AT	3000	12	6.95	0.67		6.76	0.73	
			RT	3000	48	28.16	2.19		27.18	2.56	
	5	F4	TT	3000	60	45.84	2.26	0.79	47.71	2.71	0.87
			AT	3000	12	9.18	0.60		9.53	0.63	
		F4	RT	3000	48	36.65	1.83		38.17	2.18	
	5	F6	TT	3000	60	45.85	2.25	0.80	47.61	2.66	0.87
			AT	3000	12	9.17	0.60		9.52	0.62	
		F6	RT	3000	48	36.68	1.81		38.10	2.14	
	6	F7	TT	3000	60	53.64	1.47	0.75	56.14	1.41	0.83
			AT	3000	12	10.82	0.41		11.28	0.35	
			RT	3000	48	42.81	1.19		44.86	1.14	
	6	F8	TT	3000	60	53.64	1.49	0.76	56.16	1.37	0.83
			AT	3000	12	10.84	0.40		11.28	0.34	
			RT	3000	48	42.81	1.21		44.88	1.11	
Alternate	5	F3	TT	3000	60	45.85	2.25	0.80	47.63	2.68	0.87
			AT	3000	12	9.19	0.60		9.52	0.64	
			RT	3000	48	36.66	1.81		38.11	2.15	
	5	F5	TT	3000	60	45.85	2.21	0.80	47.73	2.66	0.87
			AT	3000	12	9.19	0.60		9.54	0.62	
			RT	3000	48	36.67	1.77		38.19	2.15	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.51

Test Study Design 60_1.5_12: (vi) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	33.66	2.69	0.75	32.15	3.29	0.81
			AT	3000	12	5.41	0.66		4.97	0.69	
			RT	3000	48	28.24	2.24		27.18	2.76	
	4	F2	TT	3000	60	33.63	2.73	0.76	32.22	3.38	0.83
			AT	3000	12	5.41	0.67		4.98	0.70	
			RT	3000	48	28.22	2.25		27.24	2.82	
	5	F4	TT	3000	60	45.10	2.40	0.82	46.94	2.94	0.87
			AT	3000	12	7.80	0.70		7.82	0.75	
		F4	RT	3000	48	37.30	1.87		39.13	2.31	
	5	F6	TT	3000	60	45.10	2.4	0.82	47.07	2.84	0.86
			AT	3000	12	7.80	0.70		7.84	0.75	
		F6	RT	3000	48	37.31	1.86		39.23	2.22	
	6	F7	TT	3000	60	53.17	1.51	0.81	56.16	1.54	0.88
			AT	3000	12	9.88	0.52		10.45	0.56	
			RT	3000	48	43.29	1.13		45.71	1.07	
	6	F8	TT	3000	60	53.17	1.57	0.82	56.2	1.47	0.88
			AT	3000	12	9.87	0.54		10.46	0.54	
			RT	3000	48	43.30	1.17		45.73	1.03	
Alternate	5	F3	TT	3000	60	45.07	2.36	0.81	46.98	2.88	0.87
			AT	3000	12	7.79	0.68		7.82	0.75	
			RT	3000	48	37.28	1.85		39.15	2.26	
	5	F5	TT	3000	60	45.05	2.33	0.80	46.87	2.90	0.86
			AT	3000	12	7.77	0.69		7.80	0.75	
			RT	3000	48	37.28	1.82		39.07	2.29	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.52

Test Study Design 60_1.5_12: (vii) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	28.80	2.69	0.82	27.28	2.81	0.85
			AT	3000	12	6.89	0.75		6.51	0.74	
			RT	3000	48	21.91	2.12		20.77	2.22	
	4	F2	TT	3000	60	28.83	2.70	0.81	27.22	2.85	0.85
			AT	3000	12	6.90	0.75		6.49	0.74	
			RT	3000	48	21.93	2.14		20.72	2.25	
	5	F4	TT	3000	60	40.89	2.71	0.80	40.45	2.94	0.87
			AT	3000	12	9.45	0.64		9.35	0.67	
		F4	RT	3000	48	31.44	2.24		31.10	2.38	
	5	F6	TT	3000	60	40.89	2.69	0.81	40.49	2.91	0.87
			AT	3000	12	9.45	0.63		9.36	0.67	
		F6	RT	3000	48	31.44	2.21		31.12	2.36	
	6	F7	TT	3000	60	50.81	1.93	0.70	51.70	2.03	0.79
			AT	3000	12	11.07	0.37		11.23	0.36	
			RT	3000	48	39.74	1.69		40.47	1.76	
	6	F8	TT	3000	60	50.86	1.91	0.71	51.70	2.04	0.80
			AT	3000	12	11.07	0.37		11.23	0.36	
			RT	3000	48	39.78	1.66		40.47	1.76	
Alternate	5	F3	TT	3000	60	40.81	2.67	0.80	40.59	2.98	0.87
			AT	3000	12	9.45	0.63		9.38	0.67	
			RT	3000	48	31.36	2.20		31.21	2.41	
	5	F5	TT	3000	60	40.90	2.66	0.80	40.61	2.92	0.86
			AT	3000	12	9.47	0.62		9.38	0.68	
			RT	3000	48	31.44	2.19		31.22	2.36	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.53

Test Study Design 60_1.5_12: (viii) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	27.30	2.61	0.79	25.55	3.07	0.86
			AT	3000	12	5.43	0.72		5.40	0.80	
			RT	3000	48	21.87	2.08		20.14	2.42	
	4	F2	TT	3000	60	27.37	2.57	0.80	25.45	3.09	0.86
			AT	3000	12	5.45	0.72		5.39	0.80	
			RT	3000	48	21.92	2.04		20.06	2.43	
	5	F4	TT	3000	60	38.97	2.64	0.80	40.18	3.27	0.87
			AT	3000	12	7.98	0.69		8.44	0.72	
		F4	RT	3000	48	31.00	2.12		31.74	2.67	
	5	F6	TT	3000	60	38.98	2.62	0.81	40.25	3.34	0.86
			AT	3000	12	7.97	0.68		8.45	0.73	
		F6	RT	3000	48	31.01	2.11		31.80	2.73	
	6	F7	TT	3000	60	49.36	2.05	0.76	52.47	2.15	0.80
			AT	3000	12	10.02	0.49		10.48	0.42	
			RT	3000	48	39.34	1.71		41.98	1.83	
	6	F8	TT	3000	60	49.32	2.06	0.77	52.46	2.15	0.80
			AT	3000	12	10.02	0.51		10.48	0.41	
			RT	3000	48	39.30	1.71		41.98	1.83	
Alternate	5	F3	TT	3000	60	39.07	2.71	0.81	40.29	3.26	0.86
			AT	3000	12	8.00	0.70		8.45	0.71	
			RT	3000	48	31.07	2.17		31.84	2.67	
	5	F5	TT	3000	60	39.01	2.73	0.81	40.29	3.30	0.87
			AT	3000	12	7.97	0.70		8.45	0.71	
			RT	3000	48	31.04	2.20		31.84	2.71	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.54

Test Study Design 60_1.5_12: (ix) Total Test Length=60(12); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	60	26.34	2.67	0.74	24.12	2.74	0.78
			AT	3000	12	4.37	0.65		3.89	0.64	
			RT	3000	48	21.98	2.24		20.23	2.28	
	4	F2	TT	3000	60	26.25	2.63	0.74	24.09	2.76	0.77
			AT	3000	12	4.33	0.64		3.90	0.63	
			RT	3000	48	21.92	2.20		20.19	2.31	
	5	F4	TT	3000	60	38.37	2.76	0.81	37.64	3.09	0.83
			AT	3000	12	6.64	0.68		6.32	0.68	
		F4	RT	3000	48	31.73	2.24		31.32	2.56	
	5	F6	TT	3000	60	38.41	2.68	0.79	37.62	3.09	0.84
			AT	3000	12	6.66	0.66		6.30	0.68	
		F6	RT	3000	48	31.76	2.19		31.32	2.54	
	6	F7	TT	3000	60	48.91	2.09	0.81	49.81	2.21	0.83
			AT	3000	12	8.83	0.56		8.72	0.59	
			RT	3000	48	40.08	1.67		41.09	1.75	
	6	F8	TT	3000	60	48.88	2.04	0.79	49.78	2.25	0.84
			AT	3000	12	8.81	0.55		8.70	0.59	
			RT	3000	48	40.07	1.63		41.08	1.78	
Alternate	5	F3	TT	3000	60	38.41	2.67	0.80	37.75	3.08	0.84
			AT	3000	12	6.67	0.68		6.35	0.69	
			RT	3000	48	31.74	2.16		31.41	2.53	
	5	F5	TT	3000	60	38.37	2.75	0.80	37.64	3.14	0.84
			AT	3000	12	6.67	0.70		6.31	0.69	
			RT	3000	48	31.70	2.23		31.33	2.59	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.55

Test Study Design 120_0.5_24: (i) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	87.99	4.81	0.85	90.74	5.37	0.89
			AT	3000	24	19.86	0.93		20.78	0.94	
			RT	3000	96	68.13	4.04		69.96	4.56	
	4	F2	TT	3000	120	87.98	4.92	0.86	90.76	5.37	0.89
			AT	3000	24	19.87	0.96		20.78	0.95	
			RT	3000	96	68.10	4.12		69.98	4.54	
	5	F4	TT	3000	120	94.74	4.33	0.84	98.52	4.61	0.87
			AT	3000	24	20.95	0.80		21.9	0.74	
		F4	RT	3000	96	73.78	3.68		76.62	3.98	
	5	F6	TT	3000	120	94.92	4.35	0.83	98.51	4.80	0.88
			AT	3000	24	20.97	0.80		21.89	0.77	
		F6	RT	3000	96	73.95	3.71		76.62	4.14	
	6	F7	TT	3000	120	100.9	3.75	0.82	104.79	3.93	0.86
			AT	3000	24	21.87	0.66		22.68	0.57	
			RT	3000	96	79.04	3.24		82.12	3.45	
	6	F8	TT	3000	120	100.8	3.80	0.81	104.68	4.00	0.86
			AT	3000	24	21.82	0.68		22.66	0.59	
			RT	3000	96	78.93	3.27		82.01	3.50	
Alternate	5	F3	TT	3000	120	94.69	4.34	0.84	98.58	4.70	0.88
			AT	3000	24	20.94	0.81		21.9	0.75	
			RT	3000	96	73.75	3.69		76.68	4.05	
	5	F5	TT	3000	120	94.76	4.39	0.84	98.59	4.76	0.89
			AT	3000	24	20.95	0.82		21.91	0.76	
			RT	3000	96	73.81	3.72		76.68	4.10	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.56

Test Study Design 120_0.5_24: (ii) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	86.76	4.77	0.87	88.25	5.61	0.92
			AT	3000	24	17.51	1.07		17.49	1.17	
			RT	3000	96	69.25	3.87		70.76	4.56	
	4	F2	TT	3000	120	86.65	4.72	0.87	88.39	5.54	0.91
			AT	3000	24	17.51	1.06		17.52	1.14	
			RT	3000	96	69.14	3.83		70.87	4.52	
	5	F4	TT	3000	120	93.38	4.24	0.87	96.08	4.91	0.91
			AT	3000	24	18.81	0.95		19.02	1.04	
		F4	RT	3000	96	74.57	3.45		77.06	3.99	
	5	F6	TT	3000	120	93.55	4.18	0.86	96.09	4.96	0.92
			AT	3000	24	18.83	0.95		19.02	1.06	
		F6	RT	3000	96	74.72	3.40		77.07	4.01	
	6	F7	TT	3000	120	99.11	3.59	0.85	102.87	4.12	0.90
			AT	3000	24	19.91	0.81		20.34	0.90	
			RT	3000	96	79.20	2.93		82.53	3.33	
	6	F8	TT	3000	120	99.10	3.63	0.85	102.84	4.08	0.91
			AT	3000	24	19.94	0.82		20.34	0.89	
			RT	3000	96	79.17	2.96		82.5	3.29	
Alternate	5	F3	TT	3000	120	93.45	4.24	0.86	96.17	4.87	0.91
			AT	3000	24	18.82	0.96		19.01	1.03	
			RT	3000	96	74.62	3.44		77.15	3.95	
	5	F5	TT	3000	120	93.41	4.20	0.86	96.26	4.86	0.91
			AT	3000	24	18.81	0.95		19.04	1.04	
			RT	3000	96	74.60	3.42		77.23	3.94	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.57

Test Study Design 120_0.5_24: (iii) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	81.25	4.98	0.88	85.20	5.54	0.90
			AT	3000	24	13.66	1.27		13.90	1.25	
			RT	3000	96	67.60	3.91		71.30	4.45	
	4	F2	TT	3000	120	81.33	5.17	0.89	85.40	5.47	0.91
			AT	3000	24	13.68	1.31		13.96	1.24	
			RT	3000	96	67.65	4.04		71.45	4.38	
	5	F4	TT	3000	120	88.83	4.84	0.90	93.41	4.88	0.91
			AT	3000	24	15.44	1.29		15.69	1.2	
		F4	RT	3000	96	73.38	3.72		77.72	3.83	
	5	F6	TT	3000	120	88.84	4.89	0.90	93.21	4.92	0.91
			AT	3000	24	15.43	1.31		15.64	1.19	
		F6	RT	3000	96	73.41	3.75		77.57	3.87	
	6	F7	TT	3000	120	95.64	4.37	0.90	100.04	4.27	0.91
			AT	3000	24	17.12	1.21		17.30	1.14	
			RT	3000	96	78.52	3.33		82.74	3.27	
	6	F8	TT	3000	120	95.72	4.29	0.90	99.90	4.23	0.90
			AT	3000	24	17.13	1.19		17.26	1.13	
			RT	3000	96	78.59	3.25		82.64	3.24	
Alternate	5	F3	TT	3000	120	88.76	4.79	0.90	93.24	4.96	0.91
			AT	3000	24	15.41	1.27		15.65	1.20	
			RT	3000	96	73.34	3.7		77.59	3.90	
	5	F5	TT	3000	120	88.64	4.71	0.89	93.24	4.92	0.91
			AT	3000	24	15.37	1.26		15.65	1.20	
			RT	3000	96	73.27	3.64		77.59	3.87	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.58

Test Study Design 120_0.5_24: (iv) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	74.26	5.47	0.89	75.28	6.30	0.90
			AT	3000	24	17.66	1.24		18.10	1.21	
			RT	3000	96	56.60	4.41		57.18	5.23	
	4	F2	TT	3000	120	74.33	5.44	0.89	75.24	6.25	0.90
			AT	3000	24	17.70	1.21		18.11	1.20	
			RT	3000	96	56.63	4.40		57.13	5.20	
	5	F4	TT	3000	120	82.62	5.35	0.88	84.88	5.86	0.89
			AT	3000	24	19.24	1.12		19.67	1.02	
		F4	RT	3000	96	63.38	4.39		65.21	4.97	
	5	F6	TT	3000	120	82.64	5.24	0.87	84.87	5.71	0.89
			AT	3000	24	19.26	1.10		19.66	1.00	
		F6	RT	3000	96	63.38	4.32		65.21	4.85	
	6	F7	TT	3000	120	90.08	4.90	0.87	93.27	5.20	0.87
			AT	3000	24	20.53	0.96		20.85	0.82	
			RT	3000	96	69.55	4.10		72.42	4.50	
	6	F8	TT	3000	120	89.98	4.74	0.85	93.12	5.27	0.88
			AT	3000	24	20.51	0.91		20.82	0.83	
			RT	3000	96	69.47	4.00		72.31	4.56	
Alternate	5	F3	TT	3000	120	82.33	5.33	0.88	84.81	5.96	0.89
			AT	3000	24	19.21	1.10		19.63	1.04	
			RT	3000	96	63.12	4.41		65.18	5.05	
	5	F5	TT	3000	120	82.53	5.17	0.87	84.81	5.87	0.89
			AT	3000	24	19.24	1.07		19.65	1.02	
			RT	3000	96	63.29	4.28		65.16	4.99	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.59

Test Study Design 120_0.5_24: (v) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	69.27	5.23	0.87	70.27	6.39	0.93
			AT	3000	24	14.28	1.17		14.40	1.44	
			RT	3000	96	54.99	4.26		55.87	5.09	
	4	F2	TT	3000	120	69.26	5.20	0.87	70.34	6.38	0.93
			AT	3000	24	14.29	1.15		14.42	1.43	
			RT	3000	96	54.98	4.24		55.92	5.08	
	5	F4	TT	3000	120	77.06	5.09	0.87	79.80	6.00	0.92
			AT	3000	24	15.76	1.10		16.38	1.34	
		F4	RT	3000	96	61.30	4.17		63.42	4.78	
	5	F6	TT	3000	120	77.09	5.23	0.87	79.79	5.97	0.92
			AT	3000	24	15.77	1.12		16.36	1.32	
		F6	RT	3000	96	61.32	4.29		63.43	4.77	
	6	F7	TT	3000	120	84.72	4.91	0.86	88.77	5.43	0.92
			AT	3000	24	17.18	1.04		18.18	1.19	
			RT	3000	96	67.55	4.05		70.59	4.36	
	6	F8	TT	3000	120	84.57	4.89	0.87	88.68	5.62	0.92
			AT	3000	24	17.14	1.05		18.15	1.23	
			RT	3000	96	67.42	4.01		70.53	4.51	
Alternate	5	F3	TT	3000	120	77.05	5.10	0.87	79.69	5.98	0.92
			AT	3000	24	15.75	1.09		16.35	1.32	
			RT	3000	96	61.31	4.19		63.35	4.79	
	5	F5	TT	3000	120	77.22	5.17	0.87	79.79	6.16	0.93
			AT	3000	24	15.80	1.11		16.37	1.36	
			RT	3000	96	61.42	4.24		63.42	4.92	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.60

Test Study Design 120_0.5_24: (vi) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	67.49	5.39	0.85	65.21	6.16	0.89
			AT	3000	24	10.90	1.17		10.10	1.26	
			RT	3000	96	56.59	4.44		55.10	5.07	
	4	F2	TT	3000	120	67.61	5.47	0.85	65.46	5.98	0.88
			AT	3000	24	10.93	1.20		10.11	1.25	
			RT	3000	96	56.68	4.50		55.34	4.91	
	5	F4	TT	3000	120	75.88	5.46	0.87	74.91	6.06	0.90
			AT	3000	24	12.49	1.23		11.89	1.33	
		F4	RT	3000	96	63.39	4.43		63.02	4.90	
	5	F6	TT	3000	120	75.91	5.37	0.87	74.74	6.10	0.90
			AT	3000	24	12.50	1.22		11.85	1.33	
		F6	RT	3000	96	63.41	4.34		62.89	4.95	
	6	F7	TT	3000	120	83.89	5.12	0.87	83.92	5.78	0.91
			AT	3000	24	14.11	1.25		13.74	1.36	
			RT	3000	96	69.79	4.08		70.18	4.58	
	6	F8	TT	3000	120	83.82	5.15	0.88	83.92	5.81	0.91
			AT	3000	24	14.10	1.26		13.73	1.36	
			RT	3000	96	69.72	4.09		70.20	4.60	
Alternate	5	F3	TT	3000	120	75.67	5.49	0.87	74.70	6.30	0.91
			AT	3000	24	12.43	1.24		11.84	1.36	
			RT	3000	96	63.23	4.46		62.86	5.10	
	5	F5	TT	3000	120	75.73	5.54	0.87	74.58	6.12	0.91
			AT	3000	24	12.47	1.24		11.83	1.35	
			RT	3000	96	63.26	4.51		62.75	4.93	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.61

Test Study Design 120_0.5_24: (vii) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	57.60	4.91	0.86	55.80	5.98	0.92
			AT	3000	24	14.34	1.17		14.40	1.55	
			RT	3000	96	43.27	3.95		41.40	4.58	
	4	F2	TT	3000	120	57.52	4.95	0.86	55.61	5.96	0.92
			AT	3000	24	14.33	1.17		14.33	1.54	
			RT	3000	96	43.19	3.99		41.27	4.58	
	5	F4	TT	3000	120	65.36	5.15	0.86	64.82	6.17	0.92
			AT	3000	24	15.92	1.11		16.51	1.48	
		F4	RT	3000	96	49.43	4.23		48.31	4.86	
	5	F6	TT	3000	120	65.07	5.17	0.87	64.98	6.06	0.92
			AT	3000	24	15.85	1.15		16.55	1.45	
		F6	RT	3000	96	49.22	4.21		48.43	4.76	
	6	F7	TT	3000	120	73.24	5.23	0.86	74.42	6.07	0.90
			AT	3000	24	17.34	1.05		18.49	1.28	
			RT	3000	96	55.90	4.36		55.92	4.94	
	6	F8	TT	3000	120	73.07	5.24	0.87	74.09	6.06	0.90
			AT	3000	24	17.30	1.07		18.44	1.29	
			RT	3000	96	55.77	4.35		55.65	4.93	
Alternate	5	F3	TT	3000	120	65.40	5.27	0.87	64.79	6.09	0.91
			AT	3000	24	15.92	1.14		16.53	1.44	
			RT	3000	96	49.49	4.32		48.27	4.80	
	5	F5	TT	3000	120	65.34	5.25	0.87	65.11	6.15	0.91
			AT	3000	24	15.89	1.15		16.60	1.44	
			RT	3000	96	49.45	4.30		48.52	4.86	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.62

Test Study Design 120_0.5_24: (viii) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	53.54	5.04	0.87	50.96	5.45	0.88
			AT	3000	24	11.12	1.26		10.45	1.16	
			RT	3000	96	42.42	3.99		40.50	4.47	
	4	F2	TT	3000	120	53.65	4.96	0.87	51.14	5.40	0.88
			AT	3000	24	11.15	1.24		10.49	1.18	
			RT	3000	96	42.50	3.92		40.65	4.40	
	5	F4	TT	3000	120	61.30	5.19	0.87	59.77	5.75	0.89
			AT	3000	24	12.79	1.24		12.12	1.21	
		F4	RT	3000	96	48.50	4.16		47.65	4.70	
	5	F6	TT	3000	120	61.39	5.37	0.88	59.62	5.93	0.90
			AT	3000	24	12.81	1.29		12.10	1.25	
		F6	RT	3000	96	48.58	4.28		47.52	4.83	
	6	F7	TT	3000	120	69.31	5.45	0.88	69.01	6.06	0.91
			AT	3000	24	14.47	1.26		13.85	1.25	
			RT	3000	96	54.84	4.39		55.16	4.96	
	6	F8	TT	3000	120	69.50	5.47	0.89	68.66	6.19	0.91
			AT	3000	24	14.52	1.25		13.77	1.26	
			RT	3000	96	54.99	4.40		54.89	5.07	
Alternate	5	F3	TT	3000	120	61.39	5.29	0.87	59.52	5.90	0.89
			AT	3000	24	12.83	1.27		12.08	1.23	
			RT	3000	96	48.56	4.23		47.44	4.84	
	5	F5	TT	3000	120	61.41	5.36	0.88	59.67	5.77	0.90
			AT	3000	24	12.81	1.28		12.11	1.21	
			RT	3000	96	48.60	4.28		47.56	4.71	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.63

Test Study Design 120_0.5_24: (ix) Total Test Length=120(24); BGMAD=0.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	51.26	4.79	0.79	47.88	5.57	0.80
			AT	3000	24	8.37	1.01		6.77	1.02	
			RT	3000	96	42.89	4.04		41.11	4.79	
	4	F2	TT	3000	120	51.11	4.78	0.80	48.01	5.51	0.80
			AT	3000	24	8.37	0.99		6.77	1.01	
			RT	3000	96	42.74	4.03		41.23	4.74	
	5	F4	TT	3000	120	58.49	5.14	0.83	56.47	6.19	0.86
			AT	3000	24	9.59	1.07		8.02	1.21	
		F4	RT	3000	96	48.89	4.29		48.45	5.20	
	5	F6	TT	3000	120	58.49	5.07	0.82	56.33	6.06	0.86
			AT	3000	24	9.58	1.08		7.99	1.19	
		F6	RT	3000	96	48.91	4.23		48.35	5.08	
	6	F7	TT	3000	120	66.43	5.38	0.86	66.11	6.30	0.89
			AT	3000	24	10.99	1.16		9.68	1.35	
			RT	3000	96	55.45	4.43		56.43	5.14	
	6	F8	TT	3000	120	66.40	5.26	0.85	65.91	6.25	0.88
			AT	3000	24	10.99	1.15		9.62	1.34	
			RT	3000	96	55.41	4.32		56.29	5.10	
Alternate	5	F3	TT	3000	120	58.71	5.18	0.83	56.54	5.93	0.85
			AT	3000	24	9.61	1.10		8.01	1.17	
			RT	3000	96	49.09	4.31		48.52	4.98	
	5	F5	TT	3000	120	58.49	5.20	0.84	56.35	5.99	0.86
			AT	3000	24	9.58	1.10		7.98	1.17	
			RT	3000	96	48.92	4.32		48.37	5.03	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.64

Test Study Design 120_1.0_24: (i) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	87.79	4.59	0.85	91.93	5.18	0.89
			AT	3000	24	20.03	0.92		20.96	0.91	
			RT	3000	96	67.76	3.83		70.97	4.38	
	4	F2	TT	3000	120	87.93	4.59	0.85	91.95	5.13	0.89
			AT	3000	24	20.05	0.93		20.97	0.92	
			RT	3000	96	67.89	3.82		70.99	4.34	
	5	F4	TT	3000	120	100.00	3.59	0.82	105.03	3.69	0.86
			AT	3000	24	21.91	0.65		22.75	0.56	
		F4	RT	3000	96	78.09	3.08		82.28	3.22	
	5	F6	TT	3000	120	100.10	3.59	0.80	104.98	3.81	0.86
			AT	3000	24	21.94	0.64		22.75	0.57	
		F6	RT	3000	96	78.18	3.09		82.23	3.32	
	6	F7	TT	3000	120	108.40	2.50	0.73	113.14	2.2	0.77
			AT	3000	24	22.98	0.41		23.61	0.28	
			RT	3000	96	85.46	2.22		89.53	1.99	
	6	F8	TT	3000	120	108.40	2.52	0.74	113.15	2.23	0.78
			AT	3000	24	22.99	0.41		23.61	0.28	
			RT	3000	96	85.45	2.23		89.54	2.02	
Alternate	5	F3	TT	3000	120	99.92	3.60	0.80	105.12	3.74	0.86
			AT	3000	24	21.90	0.65		22.77	0.56	
			RT	3000	96	78.02	3.10		82.34	3.28	
	5	F5	TT	3000	120	100.00	3.53	0.80	105.16	3.82	0.86
			AT	3000	24	21.91	0.64		22.78	0.58	
			RT	3000	96	78.10	3.05		82.38	3.34	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.65

Test Study Design 120_1.0_24: (ii) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	86.50	5.09	0.89	86.95	5.62	0.93
			AT	3000	24	16.91	1.17		17.13	1.32	
			RT	3000	96	69.59	4.09		69.82	4.42	
	4	F2	TT	3000	120	86.52	5.03	0.88	86.81	5.63	0.93
			AT	3000	24	16.91	1.13		17.10	1.32	
			RT	3000	96	69.61	4.08		69.71	4.43	
	5	F4	TT	3000	120	99.51	3.82	0.87	101.87	4.36	0.93
			AT	3000	24	19.61	0.92		20.44	1.04	
		F4	RT	3000	96	79.90	3.06		81.44	3.42	
	5	F6	TT	3000	120	99.60	3.77	0.87	101.94	4.27	0.93
			AT	3000	24	19.62	0.91		20.46	1.02	
		F6	RT	3000	96	79.99	3.01		81.48	3.34	
	6	F7	TT	3000	120	108.70	2.62	0.84	111.60	2.60	0.89
			AT	3000	24	21.57	0.67		22.53	0.62	
			RT	3000	96	87.13	2.08		89.08	2.06	
	6	F8	TT	3000	120	108.60	2.60	0.84	111.55	2.67	0.89
			AT	3000	24	21.55	0.68		22.51	0.63	
			RT	3000	96	87.04	2.07		89.04	2.12	
Alternate	5	F3	TT	3000	120	99.61	3.85	0.87	101.93	4.26	0.92
			AT	3000	24	19.61	0.92		20.46	1.02	
			RT	3000	96	80.00	3.08		81.46	3.34	
	5	F5	TT	3000	120	99.53	3.81	0.87	101.81	4.34	0.93
			AT	3000	24	19.61	0.91		20.44	1.04	
			RT	3000	96	79.92	3.06		81.36	3.40	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.66

Test Study Design 120_1.0_24: (iii) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	81.77	4.99	0.88	85.36	5.40	0.91
			AT	3000	24	14.05	1.23		14.67	1.25	
			RT	3000	96	67.72	3.96		70.69	4.29	
	4	F2	TT	3000	120	81.80	4.98	0.88	85.29	5.41	0.92
			AT	3000	24	14.05	1.22		14.64	1.26	
			RT	3000	96	67.74	3.94		70.65	4.29	
	5	F4	TT	3000	120	95.64	4.14	0.88	100.06	4.18	0.90
			AT	3000	24	17.24	1.12		17.89	1.05	
		F4	RT	3000	96	78.40	3.19		82.17	3.26	
	5	F6	TT	3000	120	95.55	4.21	0.89	100.11	4.17	0.91
			AT	3000	24	17.21	1.14		17.90	1.05	
		F6	RT	3000	96	78.34	3.24		82.20	3.25	
	6	F7	TT	3000	120	105.90	3.11	0.88	109.77	2.72	0.90
			AT	3000	24	19.88	0.93		20.39	0.82	
			RT	3000	96	86.05	2.33		89.38	2.02	
	6	F8	TT	3000	120	106.00	3.12	0.88	109.90	2.74	0.89
			AT	3000	24	19.90	0.94		20.43	0.82	
			RT	3000	96	86.08	2.33		89.47	2.03	
Alternate	5	F3	TT	3000	120	95.37	4.16	0.88	99.93	4.16	0.91
			AT	3000	24	17.16	1.13		17.85	1.06	
			RT	3000	96	78.21	3.20		82.08	3.22	
	5	F5	TT	3000	120	95.55	4.13	0.89	100.11	4.20	0.91
			AT	3000	24	17.22	1.11		17.90	1.06	
			RT	3000	96	78.33	3.19		82.21	3.26	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.67

Test Study Design 120_1.0_24: (iv) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	73.70	5.23	0.86	71.68	5.92	0.89
			AT	3000	24	17.42	1.12		16.83	1.06	
			RT	3000	96	56.28	4.31		54.85	5.00	
	4	F2	TT	3000	120	73.71	5.41	0.87	71.81	6.05	0.89
			AT	3000	24	17.40	1.14		16.87	1.08	
			RT	3000	96	56.31	4.45		54.95	5.11	
	5	F4	TT	3000	120	89.24	4.74	0.84	89.72	5.57	0.89
			AT	3000	24	20.04	0.88		19.60	0.92	
		F4	RT	3000	96	69.20	4.03		70.12	4.77	
	5	F6	TT	3000	120	89.18	4.78	0.85	89.87	5.56	0.90
			AT	3000	24	20.03	0.90		19.62	0.91	
		F6	RT	3000	96	69.15	4.04		70.25	4.76	
	6	F7	TT	3000	120	101.70	3.56	0.79	104.37	4.23	0.88
			AT	3000	24	21.85	0.61		21.72	0.67	
			RT	3000	96	79.85	3.10		82.65	3.65	
	6	F8	TT	3000	120	101.72	3.63	0.80	104.60	4.15	0.88
			AT	3000	24	21.86	0.61		21.75	0.68	
			RT	3000	96	79.86	3.16		82.85	3.57	
Alternate	5	F3	TT	3000	120	89.35	4.75	0.84	89.94	5.52	0.89
			AT	3000	24	20.04	0.89		19.64	0.92	
			RT	3000	96	69.30	4.03		70.30	4.72	
	5	F5	TT	3000	120	89.22	4.72	0.84	89.81	5.59	0.90
			AT	3000	24	20.04	0.87		19.61	0.92	
			RT	3000	96	69.18	4.01		70.21	4.79	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.68

Test Study Design 120_1.0_24: (v) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	70.10	5.10	0.86	69.03	6.33	0.92
			AT	3000	24	14.46	1.17		13.88	1.37	
			RT	3000	96	55.64	4.13		55.15	5.10	
	4	F2	TT	3000	120	70.05	4.97	0.86	69.03	6.45	0.92
			AT	3000	24	14.45	1.15		13.86	1.38	
			RT	3000	96	55.61	4.02		55.17	5.20	
	5	F4	TT	3000	120	85.00	4.66	0.86	87.96	5.83	0.92
			AT	3000	24	17.29	1.01		17.65	1.29	
		F4	RT	3000	96	67.70	3.83		70.31	4.67	
	5	F6	TT	3000	120	84.91	4.74	0.86	88.04	5.81	0.92
			AT	3000	24	17.26	1.01		17.69	1.26	
		F6	RT	3000	96	67.65	3.91		70.34	4.67	
	6	F7	TT	3000	120	97.84	3.91	0.84	103.35	4.35	0.92
			AT	3000	24	19.57	0.83		20.89	0.99	
			RT	3000	96	78.27	3.25		82.46	3.46	
	6	F8	TT	3000	120	97.91	3.89	0.84	103.31	4.33	0.92
			AT	3000	24	19.58	0.82		20.88	0.98	
			RT	3000	96	78.33	3.24		82.43	3.45	
Alternate	5	F3	TT	3000	120	84.94	4.66	0.85	88.11	5.79	0.92
			AT	3000	24	17.29	1.02		17.71	1.27	
			RT	3000	96	67.65	3.83		70.40	4.64	
	5	F5	TT	3000	120	85.01	4.75	0.86	88.00	5.88	0.93
			AT	3000	24	17.29	1.02		17.69	1.29	
			RT	3000	96	67.72	3.90		70.31	4.71	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.69

Test Study Design 120_1.0_24: (vi) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	67.03	5.43	0.86	64.91	6.29	0.88
			AT	3000	24	11.62	1.22		10.10	1.23	
			RT	3000	96	55.41	4.42		54.81	5.24	
	4	F2	TT	3000	120	67.09	5.31	0.86	65.14	6.38	0.89
			AT	3000	24	11.61	1.19		10.15	1.26	
			RT	3000	96	55.48	4.34		54.99	5.29	
	5	F4	TT	3000	120	82.97	5.07	0.86	84.30	6.06	0.91
			AT	3000	24	14.67	1.13		13.76	1.35	
		F4	RT	3000	96	68.30	4.14		70.54	4.86	
	5	F6	TT	3000	120	83.23	5.16	0.86	84.31	6.01	0.91
			AT	3000	24	14.71	1.15		13.73	1.36	
		F6	RT	3000	96	68.52	4.21		70.59	4.80	
	6	F7	TT	3000	120	96.82	4.17	0.86	100.63	4.63	0.92
			AT	3000	24	17.41	1.00		17.54	1.27	
			RT	3000	96	79.41	3.35		83.08	3.49	
	6	F8	TT	3000	120	96.91	4.15	0.86	100.58	4.63	0.92
			AT	3000	24	17.44	1.00		17.51	1.25	
			RT	3000	96	79.48	3.33		83.07	3.52	
Alternate	5	F3	TT	3000	120	82.97	5.00	0.86	84.28	5.97	0.91
			AT	3000	24	14.66	1.14		13.74	1.34	
			RT	3000	96	68.31	4.06		70.54	4.78	
	5	F5	TT	3000	120	82.97	5.19	0.87	84.21	5.99	0.91
			AT	3000	24	14.66	1.19		13.73	1.34	
			RT	3000	96	68.30	4.19		70.48	4.79	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.70

Test Study Design 120_1.0_24: (vii) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	57.32	4.99	0.86	55.38	6.01	0.92
			AT	3000	24	14.07	1.20		13.74	1.42	
			RT	3000	96	43.24	4.00		41.64	4.74	
	4	F2	TT	3000	120	57.55	5.13	0.87	55.36	5.96	0.91
			AT	3000	24	14.13	1.20		13.73	1.40	
			RT	3000	96	43.41	4.12		41.62	4.71	
	5	F4	TT	3000	120	72.92	5.17	0.86	74.48	6.17	0.92
			AT	3000	24	17.13	1.08		17.65	1.27	
		F4	RT	3000	96	55.79	4.27		56.83	5.04	
	5	F6	TT	3000	120	72.89	5.29	0.86	74.44	6.33	0.92
			AT	3000	24	17.15	1.11		17.65	1.29	
		F6	RT	3000	96	55.74	4.37		56.79	5.17	
	6	F7	TT	3000	120	88.21	4.72	0.84	92.71	5.37	0.90
			AT	3000	24	19.73	0.88		20.78	0.96	
			RT	3000	96	68.48	4.01		71.93	4.52	
	6	F8	TT	3000	120	88.11	4.83	0.85	92.74	5.34	0.90
			AT	3000	24	19.70	0.90		20.79	0.94	
			RT	3000	96	68.41	4.09		71.95	4.51	
Alternate	5	F3	TT	3000	120	72.90	5.18	0.86	74.67	6.32	0.92
			AT	3000	24	17.15	1.09		17.69	1.29	
			RT	3000	96	55.75	4.28		56.99	5.16	
	5	F5	TT	3000	120	72.85	5.27	0.86	74.41	6.28	0.92
			AT	3000	24	17.14	1.09		17.64	1.28	
			RT	3000	96	55.71	4.36		56.77	5.13	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.71

Test Study Design 120_1.0_24: (viii) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	53.98	5.02	0.86	51.24	5.68	0.89
			AT	3000	24	10.48	1.18		10.02	1.24	
			RT	3000	96	43.50	4.05		41.22	4.61	
	4	F2	TT	3000	120	53.93	5.13	0.86	51.29	5.64	0.88
			AT	3000	24	10.47	1.20		10.05	1.25	
			RT	3000	96	43.46	4.14		41.24	4.57	
	5	F4	TT	3000	120	69.97	5.57	0.89	69.81	6.14	0.92
			AT	3000	24	13.80	1.32		13.73	1.37	
		F4	RT	3000	96	56.17	4.43		56.08	4.91	
	5	F6	TT	3000	120	69.92	5.36	0.88	69.86	6.21	0.92
			AT	3000	24	13.76	1.27		13.74	1.39	
		F6	RT	3000	96	56.16	4.28		56.11	4.96	
	6	F7	TT	3000	120	85.99	5.05	0.89	88.2	5.64	0.93
			AT	3000	24	17.15	1.18		17.59	1.29	
			RT	3000	96	68.83	4.04		70.61	4.47	
	6	F8	TT	3000	120	86.08	5.06	0.89	88.25	5.73	0.93
			AT	3000	24	17.18	1.20		17.60	1.32	
			RT	3000	96	68.90	4.03		70.65	4.53	
Alternate	5	F3	TT	3000	120	69.95	5.56	0.88	69.76	6.10	0.92
			AT	3000	24	13.79	1.31		13.71	1.37	
			RT	3000	96	56.16	4.44		56.05	4.86	
	5	F5	TT	3000	120	70.04	5.44	0.89	69.69	6.17	0.92
			AT	3000	24	13.81	1.29		13.70	1.39	
			RT	3000	96	56.23	4.34		55.99	4.93	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.72

Test Study Design 120_1.0_24: (ix) Total Test Length=120(24); BGMAD=1.0; μ(a) = (0.6, 1.0); DAD=1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	50.54	4.65	0.79	47.88	5.34	0.83
			AT	3000	24	8.03	1.01		6.82	1.06	
			RT	3000	96	42.51	3.90		41.07	4.50	
	4	F2	TT	3000	120	50.55	4.70	0.79	47.72	5.38	0.83
			AT	3000	24	8.03	1.02		6.81	1.06	
			RT	3000	96	42.51	3.93		40.91	4.54	
	5	F4	TT	3000	120	65.83	5.33	0.85	65.69	6.26	0.90
			AT	3000	24	10.78	1.20		9.93	1.36	
		F4	RT	3000	96	55.05	4.35		55.77	5.08	
	5	F6	TT	3000	120	65.86	5.40	0.85	65.84	6.36	0.90
			AT	3000	24	10.78	1.21		9.96	1.38	
		F6	RT	3000	96	55.07	4.41		55.88	5.15	
	6	F7	TT	3000	120	82.17	5.13	0.88	85.00	5.90	0.93
			AT	3000	24	14.09	1.25		14.06	1.46	
			RT	3000	96	68.09	4.07		70.94	4.58	
	6	F8	TT	3000	120	82.04	5.24	0.88	85.06	5.97	0.93
			AT	3000	24	14.10	1.27		14.08	1.49	
			RT	3000	96	67.94	4.16		70.98	4.62	
Alternate	5	F3	TT	3000	120	65.78	5.45	0.86	65.93	6.37	0.90
			AT	3000	24	10.77	1.22		9.97	1.39	
			RT	3000	96	55.01	4.44		55.96	5.15	
	5	F5	TT	3000	120	65.93	5.30	0.85	65.84	6.38	0.90
			AT	3000	24	10.82	1.20		9.96	1.38	
			RT	3000	96	55.11	4.32		55.88	5.16	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.73

Test Study Design 120_1.5_24: (i) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	86.42	4.80	0.89	91.64	5.07	0.90
			AT	3000	24	17.47	1.17		21.06	0.96	
			RT	3000	96	68.95	3.81		70.58	4.23	
	4	F2	TT	3000	120	86.24	4.86	0.88	91.49	5.01	0.90
			AT	3000	24	17.43	1.15		21.04	0.97	
			RT	3000	96	68.81	3.87		70.45	4.16	
	5	F4	TT	3000	120	104.00	3.13	0.85	108.88	2.88	0.79
			AT	3000	24	21.20	0.78		23.34	0.39	
		F4	RT	3000	96	82.82	2.50		85.54	2.58	
	5	F6	TT	3000	120	104.00	3.24	0.86	108.91	2.85	0.79
			AT	3000	24	21.19	0.78		23.35	0.38	
		F6	RT	3000	96	82.83	2.60		85.56	2.56	
	6	F7	TT	3000	120	113.80	1.67	0.76	116.80	1.25	0.61
			AT	3000	24	23.10	0.39		23.91	0.12	
			RT	3000	96	90.73	1.39		92.90	1.19	
	6	F8	TT	3000	120	113.70	1.67	0.76	116.81	1.21	0.60
			AT	3000	24	23.07	0.40		23.90	0.11	
			RT	3000	96	90.66	1.39		92.91	1.15	
Alternate	5	F3	TT	3000	120	104.00	3.17	0.85	108.85	2.87	0.80
			AT	3000	24	21.19	0.77		23.34	0.39	
			RT	3000	96	82.83	2.55		85.51	2.57	
	5	F5	TT	3000	120	104.00	3.25	0.85	108.79	2.87	0.80
			AT	3000	24	21.19	0.78		23.33	0.38	
			RT	3000	96	82.80	2.62		85.46	2.57	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.74

Test Study Design 120_1.5_24: (ii) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	86.42	4.80	0.89	89.90	5.46	0.92
			AT	3000	24	17.47	1.17		17.14	1.22	
			RT	3000	96	68.95	3.81		72.76	4.37	
	4	F2	TT	3000	120	86.24	4.86	0.88	89.99	5.39	0.91
			AT	3000	24	17.43	1.15		17.15	1.22	
			RT	3000	96	68.81	3.87		72.84	4.30	
	5	F4	TT	3000	120	104.02	3.13	0.85	108.60	3.09	0.91
			AT	3000	24	21.20	0.78		21.49	0.82	
		F4	RT	3000	96	82.82	2.50		87.11	2.36	
	5	F6	TT	3000	120	104.03	3.24	0.86	108.45	3.12	0.92
			AT	3000	24	21.19	0.78		21.44	0.84	
		F6	RT	3000	96	82.83	2.60		87.01	2.37	
	6	F7	TT	3000	120	113.83	1.67	0.76	116.79	1.26	0.84
			AT	3000	24	23.10	0.39		23.44	0.35	
			RT	3000	96	90.73	1.39		93.36	0.99	
	6	F8	TT	3000	120	113.72	1.67	0.76	116.77	1.30	0.84
			AT	3000	24	23.07	0.40		23.43	0.35	
			RT	3000	96	90.66	1.39		93.34	1.02	
Alternate	5	F3	TT	3000	120	104.02	3.17	0.85	108.46	3.12	0.91
			AT	3000	24	21.19	0.77		21.44	0.82	
			RT	3000	96	82.83	2.55		87.03	2.40	
	5	F5	TT	3000	120	103.99	3.25	0.85	108.52	3.04	0.91
			AT	3000	24	21.19	0.78		21.45	0.82	
			RT	3000	96	82.80	2.62		87.06	2.32	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.75

Test Study Design 120_1.5_24: (iii) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=-1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	82.44	5.05	0.88	85.11	5.47	0.91
			AT	3000	24	13.86	1.24		14.03	1.35	
			RT	3000	96	68.58	4.00		71.08	4.28	
	4	F2	TT	3000	120	82.45	5.04	0.87	84.95	5.53	0.91
			AT	3000	24	13.83	1.22		13.99	1.38	
			RT	3000	96	68.62	4.02		70.97	4.30	
	5	F4	TT	3000	120	101.60	3.57	0.88	105.67	3.53	0.92
			AT	3000	24	18.49	1.05		19.58	1.14	
		F4	RT	3000	96	83.14	2.69		86.09	2.52	
	5	F6	TT	3000	120	101.50	3.70	0.89	105.63	3.57	0.93
			AT	3000	24	18.45	1.10		19.57	1.14	
		F6	RT	3000	96	83.04	2.77		86.06	2.54	
	6	F7	TT	3000	120	112.90	1.90	0.86	115.87	1.56	0.90
			AT	3000	24	21.75	0.67		22.87	0.56	
			RT	3000	96	91.17	1.37		93.00	1.08	
	6	F8	TT	3000	120	112.90	1.92	0.86	115.83	1.61	0.90
			AT	3000	24	21.75	0.68		22.86	0.58	
			RT	3000	96	91.14	1.38		92.97	1.11	
Alternate	5	F3	TT	3000	120	101.50	3.59	0.88	105.67	3.59	0.93
			AT	3000	24	18.45	1.06		19.58	1.15	
			RT	3000	96	83.03	2.70		86.09	2.56	
	5	F5	TT	3000	120	101.50	3.57	0.89	105.70	3.51	0.92
			AT	3000	24	18.45	1.06		19.59	1.12	
			RT	3000	96	83.08	2.68		86.11	2.51	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.76

Test Study Design 120_1.5_24: (iv) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	72.76	5.36	0.86	72.94	5.99	0.92
			AT	3000	24	16.88	1.08		17.24	1.26	
			RT	3000	96	55.88	4.46		55.70	4.86	
	4	F2	TT	3000	120	72.76	5.42	0.85	73.02	6.13	0.92
			AT	3000	24	16.88	1.08		17.25	1.30	
			RT	3000	96	55.89	4.54		55.77	4.95	
	5	F4	TT	3000	120	95.25	4.35	0.83	97.73	4.59	0.89
			AT	3000	24	20.44	0.79		21.55	0.81	
		F4	RT	3000	96	74.81	3.72		76.18	3.88	
	5	F6	TT	3000	120	95.26	4.37	0.83	97.76	4.54	0.90
			AT	3000	24	20.45	0.79		21.55	0.81	
		F6	RT	3000	96	74.81	3.73		76.21	3.84	
	6	F7	TT	3000	120	110.04	2.53	0.76	112.29	2.37	0.80
			AT	3000	24	22.57	0.47		23.47	0.33	
			RT	3000	96	87.47	2.19		88.82	2.11	
	6	F8	TT	3000	120	110.03	2.52	0.77	112.27	2.29	0.79
			AT	3000	24	22.57	0.47		23.46	0.33	
			RT	3000	96	87.46	2.17		88.81	2.04	
Alternate	5	F3	TT	3000	120	95.29	4.46	0.84	97.81	4.55	0.89
			AT	3000	24	20.43	0.80		21.55	0.80	
			RT	3000	96	74.85	3.81		76.26	3.86	
	5	F5	TT	3000	120	95.40	4.41	0.84	97.80	4.57	0.89
			AT	3000	24	20.46	0.79		21.56	0.81	
			RT	3000	96	74.94	3.78		76.24	3.87	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.77

Test Study Design 120_1.5_24: (v) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	70.26	5.35	0.88	70.29	6.37	0.92
			AT	3000	24	14.11	1.27		14.24	1.38	
			RT	3000	96	56.15	4.27		56.04	5.13	
	4	F2	TT	3000	120	70.14	5.31	0.88	70.45	6.51	0.92
			AT	3000	24	14.07	1.27		14.28	1.41	
			RT	3000	96	56.07	4.24		56.17	5.24	
	5	F4	TT	3000	120	92.92	4.33	0.88	97.71	4.98	0.92
			AT	3000	24	18.87	1.05		19.64	1.06	
		F4	RT	3000	96	74.05	3.44		78.07	4.03	
	5	F6	TT	3000	120	92.82	4.43	0.88	97.62	5.00	0.92
			AT	3000	24	18.86	1.06		19.62	1.06	
		F6	RT	3000	96	73.96	3.53		78.00	4.04	
	6	F7	TT	3000	120	107.95	2.65	0.84	113.15	2.47	0.87
			AT	3000	24	21.99	0.63		22.62	0.54	
			RT	3000	96	85.97	2.14		90.53	2.02	
	6	F8	TT	3000	120	107.88	2.59	0.83	113.12	2.41	0.87
			AT	3000	24	21.97	0.63		22.61	0.53	
			RT	3000	96	85.91	2.10		90.51	1.97	
Alternate	5	F3	TT	3000	120	92.64	4.48	0.88	97.74	4.89	0.91
			AT	3000	24	18.80	1.06		19.66	1.05	
			RT	3000	96	73.84	3.58		78.09	3.96	
	5	F5	TT	3000	120	92.64	4.36	0.89	97.55	4.96	0.92
			AT	3000	24	18.83	1.05		19.61	1.04	
			RT	3000	96	73.81	3.47		77.94	4.02	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.78

Test Study Design 120_1.5_24: (vi) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=0; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	67.19	5.40	0.86	66.80	6.1	0.89
			AT	3000	24	10.85	1.24		10.00	1.37	
			RT	3000	96	56.34	4.39		56.80	4.91	
	4	F2	TT	3000	120	67.34	5.36	0.86	66.74	6.09	0.89
			AT	3000	24	10.88	1.22		10.00	1.37	
			RT	3000	96	56.46	4.36		56.74	4.91	
	5	F4	TT	3000	120	90.57	4.65	0.89	93.39	5.07	0.93
			AT	3000	24	15.80	1.21		16.27	1.45	
		F4	RT	3000	96	74.78	3.62		77.12	3.76	
	5	F6	TT	3000	120	90.38	4.67	0.89	93.58	5.09	0.94
			AT	3000	24	15.75	1.22		16.30	1.45	
		F6	RT	3000	96	74.63	3.63		77.27	3.76	
	6	F7	TT	3000	120	107.13	2.95	0.88	110.30	2.73	0.92
			AT	3000	24	19.95	0.90		21.12	0.87	
			RT	3000	96	87.18	2.20		89.16	1.96	
	6	F8	TT	3000	120	107.01	2.96	0.88	110.20	2.74	0.93
			AT	3000	24	19.93	0.89		21.08	0.87	
			RT	3000	96	87.09	2.22		89.08	1.96	
Alternate	5	F3	TT	3000	120	90.53	4.70	0.89	93.47	5.16	0.94
			AT	3000	24	15.78	1.22		16.27	1.46	
			RT	3000	96	74.75	3.65		77.20	3.83	
	5	F5	TT	3000	120	90.48	4.64	0.89	93.58	5.14	0.93
			AT	3000	24	15.77	1.21		16.31	1.43	
			RT	3000	96	74.71	3.60		77.27	3.84	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.79

Test Study Design 120_1.5_24: (vii) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=-1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	56.82	5.01	0.86	55.51	5.89	0.90
			AT	3000	24	13.75	1.17		14.78	1.40	
			RT	3000	96	43.07	4.05		40.73	4.67	
	4	F2	TT	3000	120	56.99	5.06	0.87	55.79	5.92	0.90
			AT	3000	24	13.78	1.15		14.82	1.40	
			RT	3000	96	43.21	4.10		40.97	4.69	
	5	F4	TT	3000	120	80.74	5.19	0.88	83.27	5.85	0.88
			AT	3000	24	18.20	1.05		19.65	0.95	
		F4	RT	3000	96	62.53	4.29		63.63	5.04	
	5	F6	TT	3000	120	80.64	5.17	0.87	83.34	5.81	0.87
			AT	3000	24	18.19	1.03		19.66	0.92	
		F6	RT	3000	96	62.44	4.30		63.68	5.03	
	6	F7	TT	3000	120	100.80	3.67	0.84	105.25	3.77	0.82
			AT	3000	24	21.53	0.69		22.23	0.49	
			RT	3000	96	79.27	3.11		83.02	3.38	
	6	F8	TT	3000	120	100.90	3.77	0.84	105.36	3.79	0.82
			AT	3000	24	21.55	0.70		22.23	0.49	
			RT	3000	96	79.33	3.21		83.12	3.40	
Alternate	5	F3	TT	3000	120	80.86	5.21	0.87	83.18	5.84	0.87
			AT	3000	24	18.24	1.03		19.63	0.93	
			RT	3000	96	62.62	4.35		63.55	5.04	
	5	F5	TT	3000	120	80.84	5.09	0.87	83.27	5.66	0.87
			AT	3000	24	18.22	1.02		19.67	0.92	
			RT	3000	96	62.62	4.24		63.60	4.87	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.80

Test Study Design 120_1.5_24: (viii) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=0

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	54.80	4.74	0.85	51.32	5.77	0.89
			AT	3000	24	10.89	1.11		10.15	1.24	
			RT	3000	96	43.91	3.84		41.17	4.70	
	4	F2	TT	3000	120	54.96	4.87	0.85	51.44	5.75	0.89
			AT	3000	24	10.93	1.14		10.15	1.21	
			RT	3000	96	44.02	3.94		41.29	4.71	
	5	F4	TT	3000	120	77.89	5.02	0.87	79.46	6.10	0.92
			AT	3000	24	15.49	1.16		15.67	1.34	
		F4	RT	3000	96	62.40	4.05		63.78	4.89	
	5	F6	TT	3000	120	77.79	5.09	0.87	79.54	6.08	0.92
			AT	3000	24	15.47	1.15		15.65	1.34	
		F6	RT	3000	96	62.32	4.13		63.89	4.87	
	6	F7	TT	3000	120	98.33	3.92	0.86	102.80	4.11	0.91
			AT	3000	24	19.58	0.92		20.49	0.92	
			RT	3000	96	78.75	3.17		82.31	3.29	
	6	F8	TT	3000	120	98.18	3.91	0.87	102.90	4.18	0.92
			AT	3000	24	19.55	0.93		20.48	0.95	
			RT	3000	96	78.63	3.15		82.38	3.33	
Alternate	5	F3	TT	3000	120	78.05	5.10	0.87	79.59	5.96	0.92
			AT	3000	24	15.52	1.17		15.69	1.33	
			RT	3000	96	62.53	4.12		63.90	4.76	
	5	F5	TT	3000	120	77.84	5.11	0.88	79.69	5.97	0.92
			AT	3000	24	15.47	1.18		15.70	1.32	
			RT	3000	96	62.37	4.12		64.00	4.78	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

Table A.81

Test Study Design 120_1.5_24: (ix) Total Test Length=120(24); BGMAD=1.5; μ(a) = (0.6, 1.0); DAD=1; ATMDD=1

Form Details				Statistics							
						μ(a) = 0.6			μ(a) = 1.0		
	Grade	Form	Test	*N*	Items	*M*	*SD*	Corr	*M*	*SD*	Corr
Base	4	F1	TT	3000	120	51.62	4.68	0.79	47.68	5.84	0.82
			AT	3000	24	7.89	1.03		7.12	1.00	
			RT	3000	96	43.73	3.92		40.56	5.06	
	4	F2	TT	3000	120	51.72	4.77	0.79	47.78	5.75	0.82
			AT	3000	24	7.90	1.03		7.11	1.01	
			RT	3000	96	43.82	4.00		40.66	4.95	
	5	F4	TT	3000	120	74.61	5.20	0.88	76.97	6.46	0.90
			AT	3000	24	12.23	1.27		11.79	1.38	
		F4	RT	3000	96	62.38	4.14		65.19	5.25	
	5	F6	TT	3000	120	74.58	5.14	0.87	77.12	6.49	0.90
			AT	3000	24	12.23	1.24		11.82	1.36	
		F6	RT	3000	96	62.35	4.12		65.30	5.30	
	6	F7	TT	3000	120	96.16	4.10	0.88	101.99	4.36	0.92
			AT	3000	24	17.18	1.15		17.53	1.25	
			RT	3000	96	78.98	3.14		84.46	3.25	
	6	F8	TT	3000	120	95.97	4.13	0.88	101.98	4.31	0.92
			AT	3000	24	17.13	1.16		17.52	1.25	
			RT	3000	96	78.84	3.15		84.46	3.20	
Alternate	5	F3	TT	3000	120	74.69	5.15	0.87	76.91	6.41	0.90
			AT	3000	24	12.25	1.27		11.80	1.33	
			RT	3000	96	62.44	4.09		65.11	5.25	
	5	F5	TT	3000	120	74.57	5.20	0.87	77.06	6.51	0.90
			AT	3000	24	12.24	1.28		11.80	1.36	
			RT	3000	96	62.33	4.14		65.26	5.32	

Note. The correlation indicates the strength of relationship between the total test (TT) and the anchor test (AT). F3 and F5 do not require equating; they are already on the F3 scale.

APPENDIX B

STANDARD ERROR OF EQUATING FOR ALL TEST STUDY DESIGNS

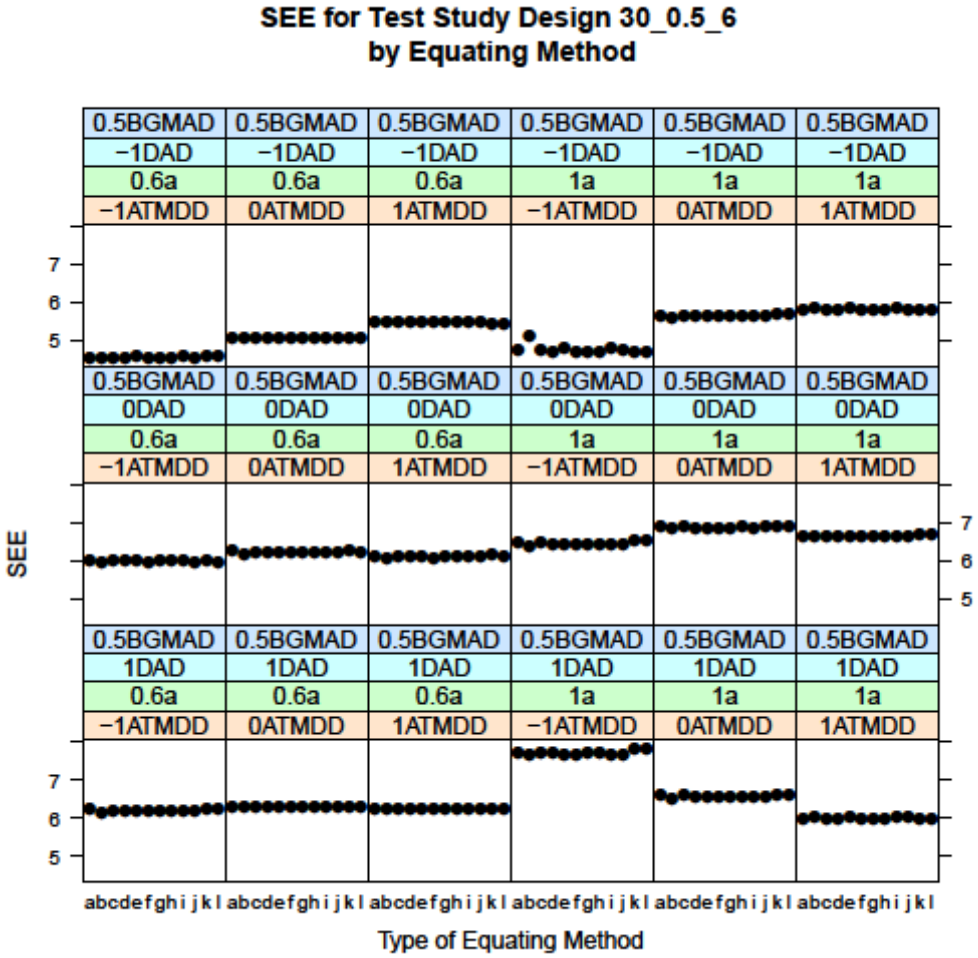

Figure B.1. Standard Error of Equating (SEE) for Test Study Design 30_0.5_6 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

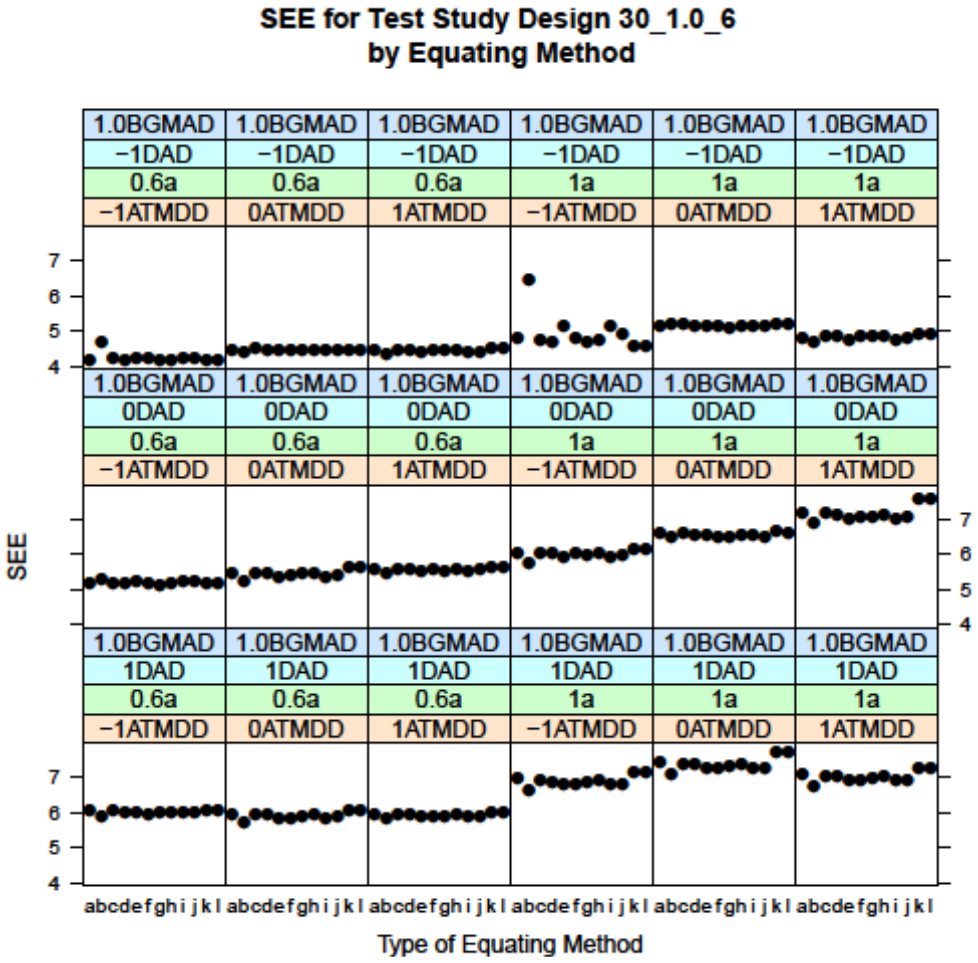

Figure B.2. Standard Error of Equating (SEE) for Test Study Design 30_1.0_6 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

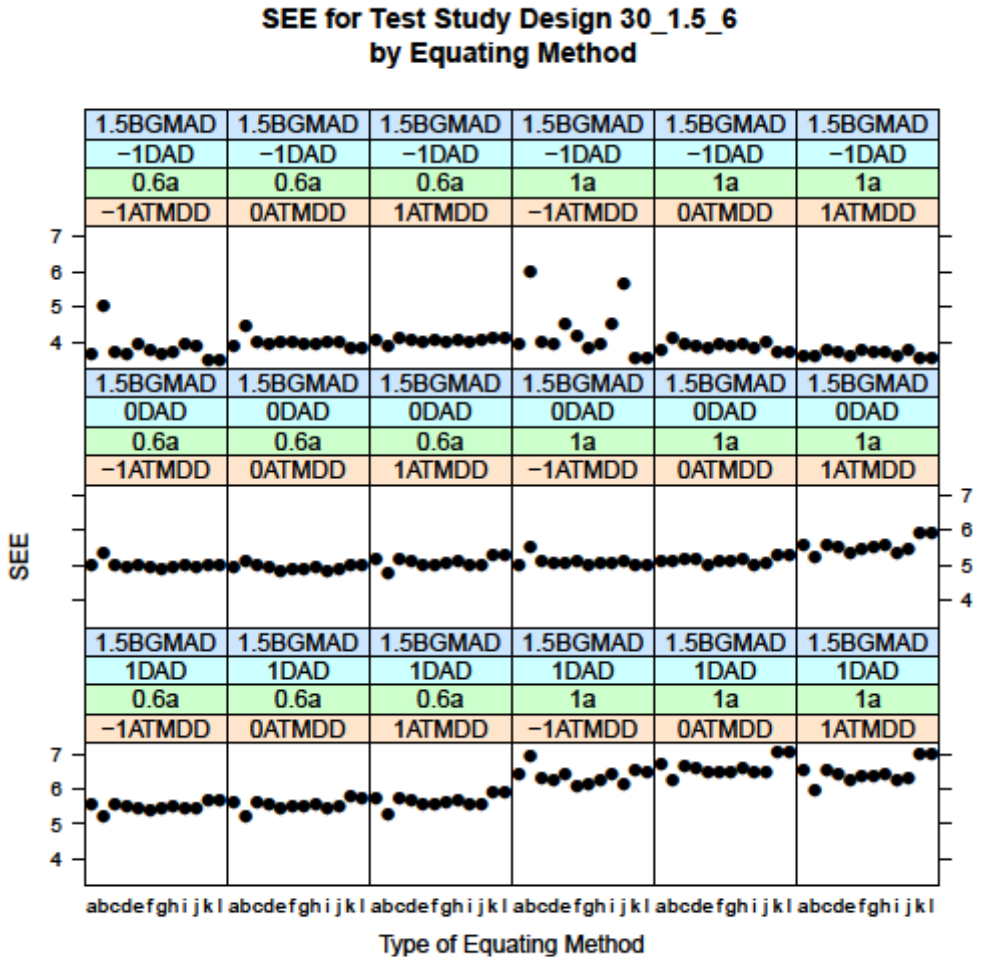

Figure B.3. Standard Error of Equating (SEE) for Test Study Design 30_1.5_6 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

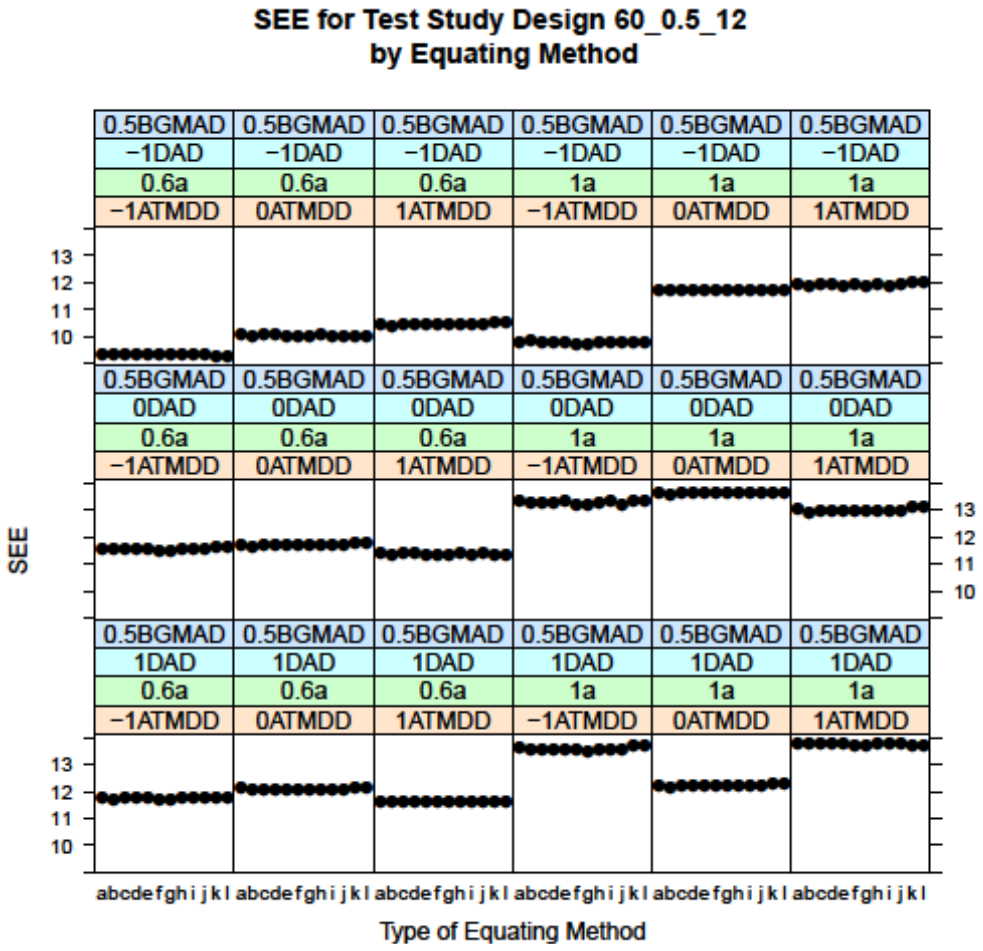

Figure B.4. Standard Error of Equating (SEE) for Test Study Design 60_0.5_12 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

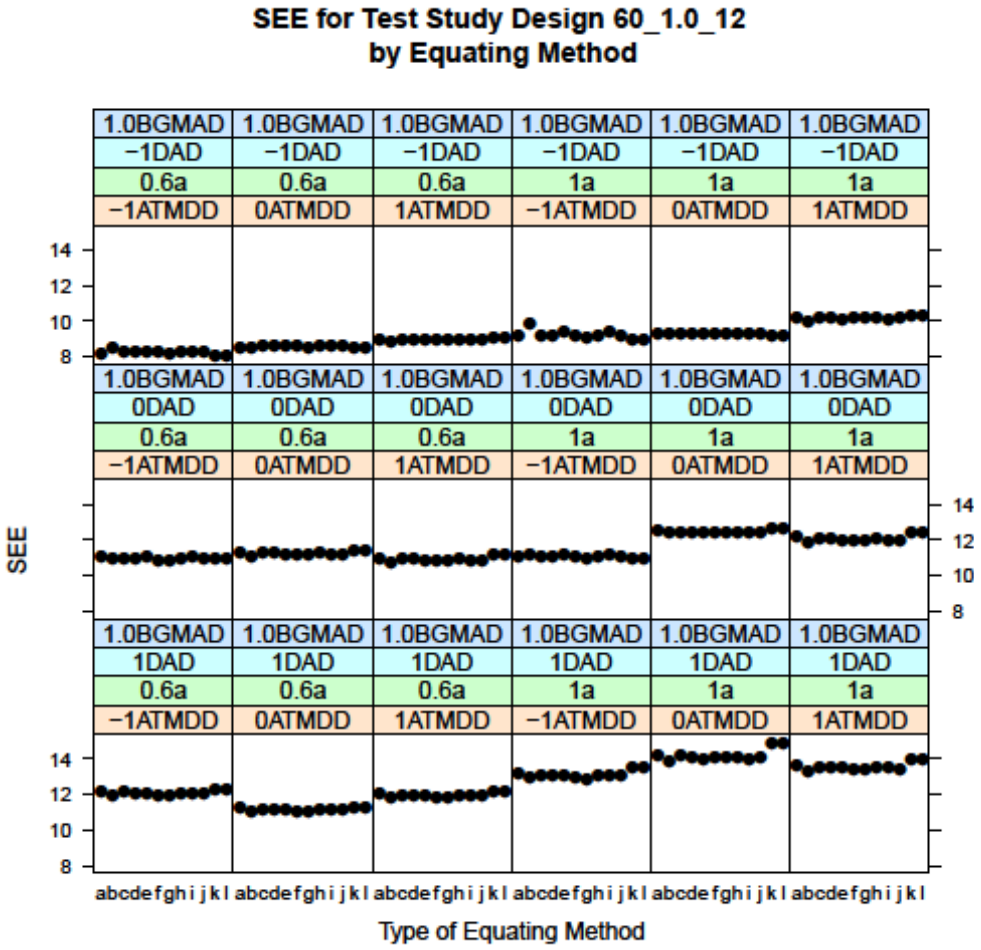

Figure B.5. Standard Error of Equating (SEE) for Test Study Design 60_1.0_12 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

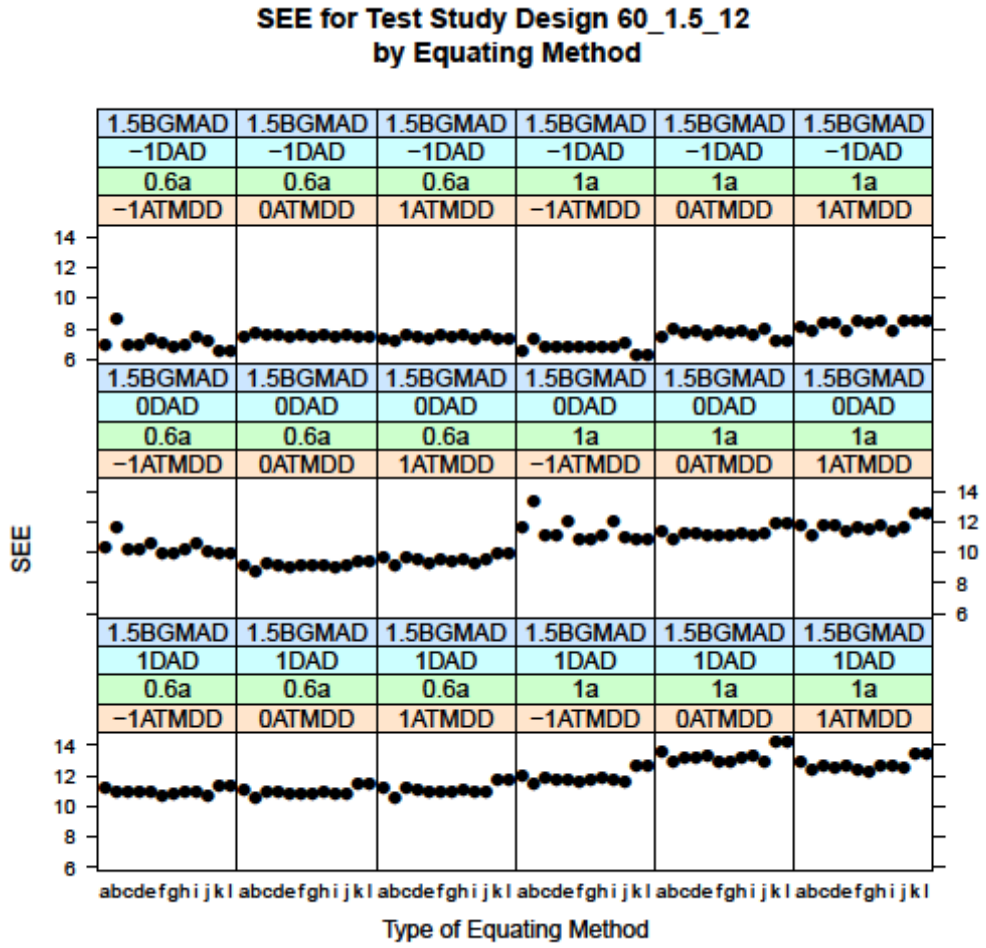

Figure B.6. Standard Error of Equating (SEE) for Test Study Design 60_1.5_12 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

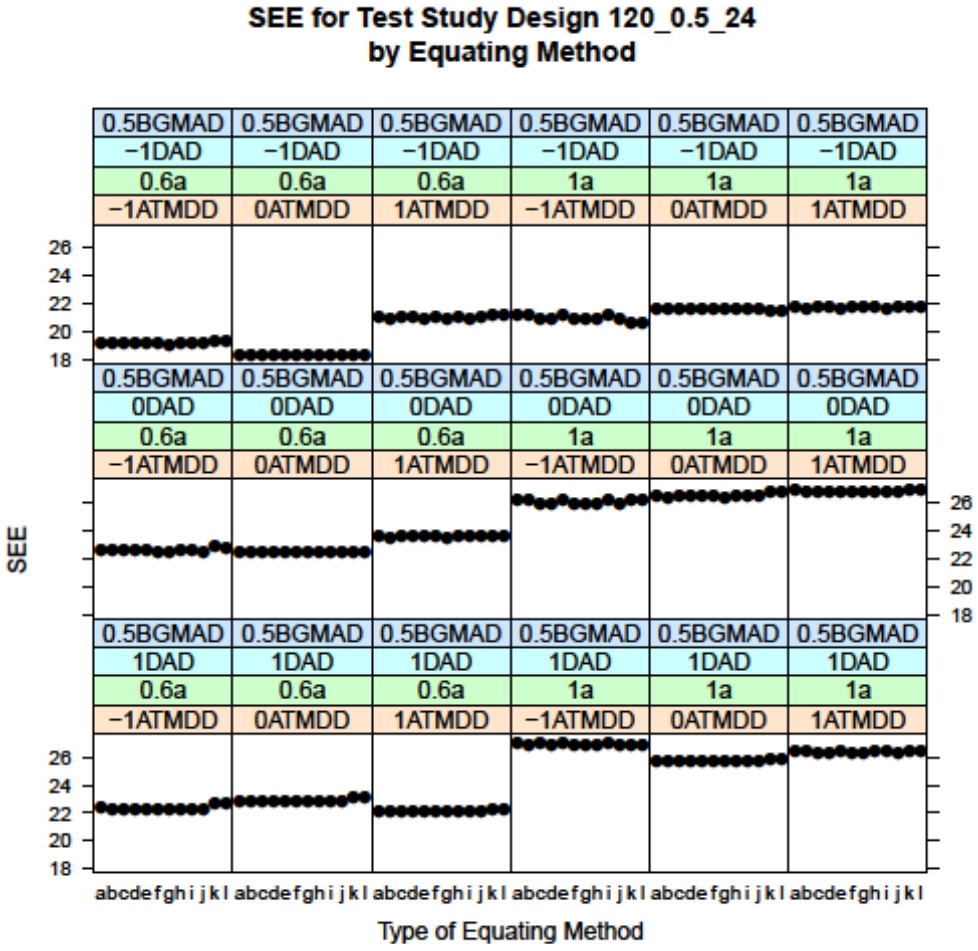

Figure B.7. Standard Error of Equating (SEE) for Test Study Design 120_0.5_24 for Small Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

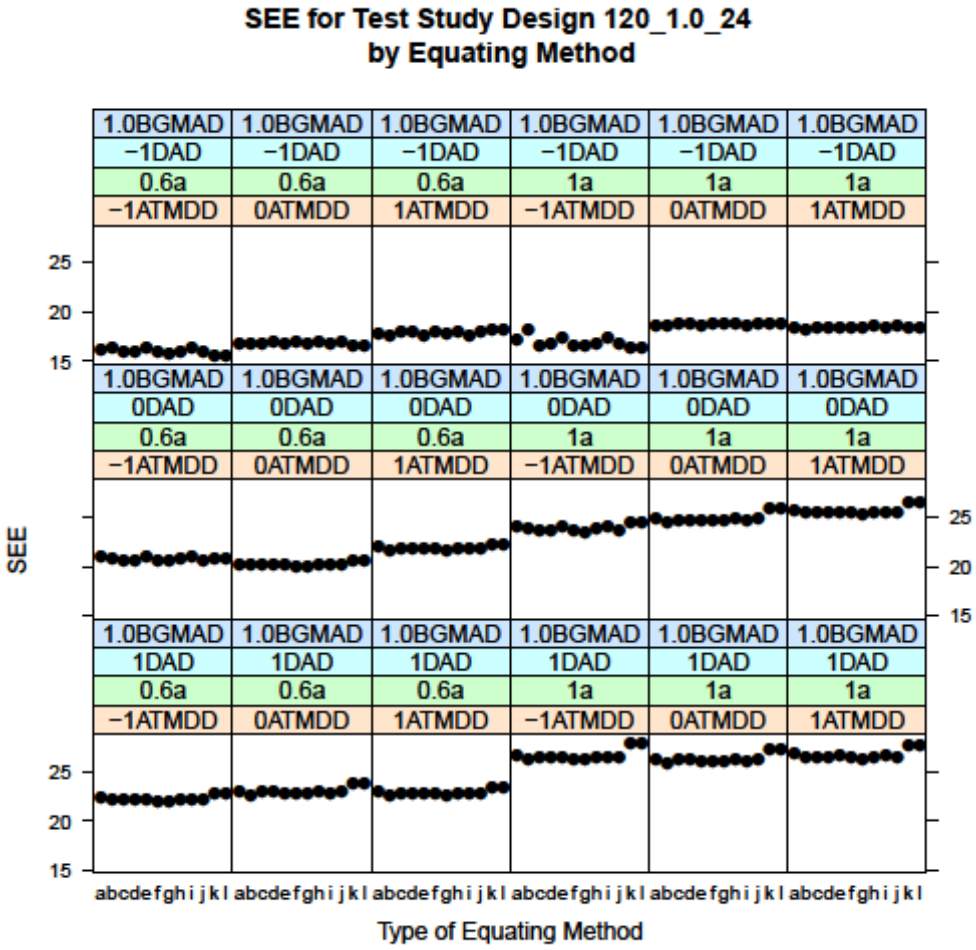

Figure B.8. Standard Error of Equating (SEE) for Test Study Design 120_1.0_24 for Medium Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating,**e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

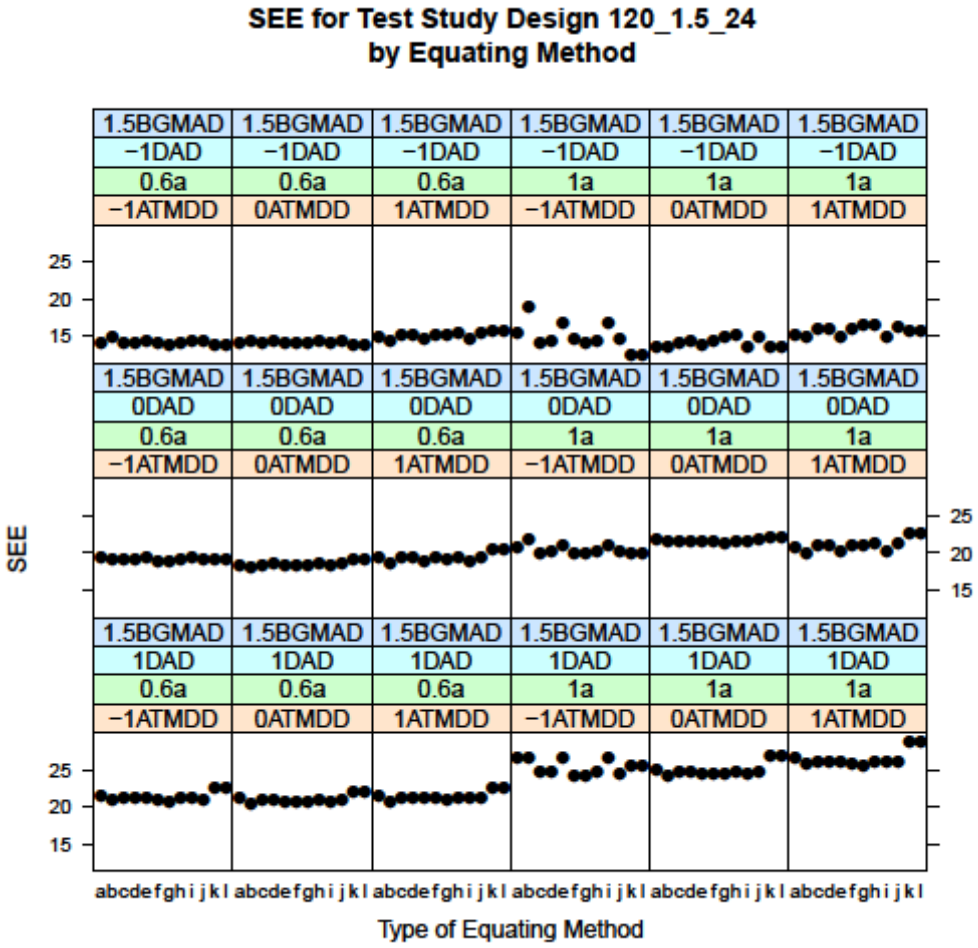

Figure B.9. Standard Error of Equating (SEE) for Test Study Design 120_1.5_24 for Large Between-grade Mean Ability Difference (BGMAD) Conditions under Different Equating Methods. **a**=Tucker Linear, **b**=Levine True Linear, **c**=Braun/Holland, **d**=Frequency Estimation Equipercentile Equating, **e**=Chained Linear, **f**=Chained Equipercentile, **g**=keNEATPSE Linear, **h**=keNEATPSE Equipercentile, **i**=keNEATCE Linear, **j**=keNEATCE Equipercentile, **k**=Linear, **l**=Equipercentile.

APPENDIX C

TEST FORMS AND EQUATING METHODS UNDER NEAT AND RG/EG DESIGNS

Table C.1

Test Forms Specification for the Equating Study (F=Base, G=Alternate, RT=Regular Test, AT=Anchor Test)

Grade	Group within Grade	Base Test Forms	Alternate Form Specification
4	1	F(4.1)=RT(4.1) + AT(4.1)	G(4.1)=RT(5.1)+AT(4.1)
4	2	F(4.2)=RT(4.2) + AT(5.1)	G(4.2)=RT(5.1)+AT(5.1)
5	1	F(5.1)=RT(5.1) + AT(4.1)	G(5.1)=RT(5.1)+AT(4.1)
5	2	F(5.2)=RT(5.2) + AT(5.1)	G(5.2)=RT(5.1)+AT(5.1)
5	3	F(5.3)=RT(5.1) + AT(5.2)	G(5.3)=RT(5.1)+AT(5.2)
5	4	F(5.4)=RT(5.2) + AT(6.1)	G(5.4)=RT(5.1)+AT(6.1)
6	1	F(6.1)=RT(6.1) + AT(5.2)	G(6.1)=RT(5.1)+AT(5.2)
6	2	F(6.2)=RT(6.2) + AT(6.1)	G(6.2)=RT(5.1)+AT(6.1)

Note. Technically speaking, the need to equate everything to the RT(5.1) scale (observed and true scores) can be eliminated; however, unless RT(5.1) and RT(5.2) are strictly parallel—a stringent requirement which is not practically possible—then there is need to adjust any bias statistics relative to the test form. By specifying RT(5.1) as the alternate form for all grades 4 and 5 tests, the problem is avoided all-together.

Table C.2

Equating Methods by Grade and Form

Grade	*Group within Grade*	*Equating Type*
4	1	NEAT via AT(4.1) to 5.1 Scale
4	2	NEAT via AT(5.1) to 5.1 Scale
5	1	No Equating, on 5.1 Scale
5	2	Random Groups to 5.1 Scale: RT(5.2)→RG→RT(5.1)
5	3	No Equating, on 5.1 Scale
5	4	Random Groups to 5.1 Scale: RT(5.2)→RG→RT(5.1)
6	1	NEAT via AT(5.2) to 5.1 Scale
6	2	NEAT via AT(6.1) to 5.1 Scale

Table C.3

Regular Test (RT) Observed Score Variables on the Base Forms to be Equated to RT(5.1) using NEAT and the AT Scores or Random Groups to Equate RT(5.2) to RT(5.1)

Grade	*Group within Grade*	Base Form Observed Score to Which to Apply Equating
4	1	$xr(4.1)=xt.F(4.1) - xa.AT(4.1)$
4	2	$xr(4.2)=xt.F(4.2) - xa.AT(5.1)$
5	1	$xr(5.1)=xt.F(5.1) - xa.AT(4.1)$
5	2	$xr(5.2)=xt.F(5.2) - xa.AT(5.1)$
5	3	$xr(5.3)=xt.F(5.1) - xa.AT(5.2)$
5	4	$xr(5.4)=xt.F(5.2) - xa.AT(6.1)$
6	1	$xr(6.1)=xt.F(6.1) - xa.AT(5.2)$
6	2	$xr(6.2)=xt.F(6.2) - xa.AT(6.1)$

Table C.4

Equated Regular Test (RT) Observed Score Variables: Scores on the Base Forms That Have Been Equated to RT(5.1) using NEAT and the AT Scores or Random Groups to Equate RT(5.2) to RT(5.1)

Grade	*Group within Grade*	Equated Observed Scores on the Regular Test (*eqxr*)
4	1	$eqxr(4.1)=NEAT.Y[xr(4.1),AT(4.1)]$
4	2	$eqxr(4.2)=NEAT.Y[xr(4.2),AT(5.1)]$
5	1	$eqxr(5.1)=xr(5.1)$
5	2	$eqxr(5.2)=RG[xr(5.1),xr(5.2)]$
5	3	$eqxr(5.3)=xr(5.3)$
5	4	$eqxr(5.4)=RG[xr(5.1),xr(5.2)]$
6	1	$eqxr(6.1)=NEAT.Y[xr(6.1),AT(5.2)]$
6	2	$eqxr(6.1)=NEAT.Y[xr(6.1),AT(6.1)]$

Table C.5

Comparative Regular Test (RT) True Score Variables: Scores on the Alternate Forms to Compare to the Equated Scores in Table 5

Grade	*Group within Grade*	Alternate Form True Score, *ur*, to Which to Compare Equated *eqxr* Scores
4	1	*ur*(4.1)=*ut*.G(5.1)-*ua*.AT(4.1)
4	2	*ur*(4.2)=*ut*.G(5.1)-*ua*.AT(5.1)
5	1	*ur*(5.1)=*ut*.G(5.1)-*ua*.AT(4.1)
5	2	*ur*(5.2)=*ut*.G(5.1)-*ua*.AT(5.1)
5	3	*ur*(5.3)=*ut*.G(5.1)-*ua*.AT(5.2)
5	4	*ur*(5.4)=*ut*.G(5.1)-*ua*.AT(6.1)
6	1	*ur*(6.1)=*ut*.G(5.1)-*ua*.AT(5.2)
6	2	*ur*(6.2)=*ut*.G(5.1)-*ua*.AT(6.1)

Table C.6

Comparative Residuals for Regular Test (RT) Variables: Equated Observed Scores vs. True Score for Alternate Forms under NEAT and RG/EG Equating Designs

Grade	Group within Grade	Residual to be Analyzed	Equating Method	Equating Design
4	1	e(4.1)=eqxr(4.1)-ur(4.1)	Tucker Linear	NEAT
			Levine True-score Linear	NEAT
			Braun/Holland	NEAT
			Frequency estimation equipercentile	NEAT
			Chained Linear	NEAT
			Chained Equipercentile	NEAT
			KeNEATCE_Linear	NEAT
			KeNEATCE_Equipercentile	NEAT
			KeNEATPSE_Linear	NEAT
			KeNEATPSE_Equipercentile	NEAT

Table C.6

Cont.

Grade	Group within Grade	Residual to be Analyzed	Equating Method	Equating Design
4	2	e(4.2)=eqxr(4.2)-ur(4.2)	Tucker Linear	NEAT
			Levine True-score Linear	NEAT
			Braun/Holland	NEAT
			Frequency estimation equipercentile	NEAT
			Chained Linear	NEAT
			Chained Equipercentile	NEAT
			KeNEATCE_Linear	NEAT
			KeNEATCE_Equipercentile	NEAT
			KeNEATPSE_Linear	NEAT
			KeNEATPSE_Equipercentile	NEAT
5	1	e(5.1)=eqxr(5.1)-tr(5.1)	No Equating, on scale	
5	2	e(5.2)=eqxr(5.2)-ur(5.2)	Linear	RG/EG
			Equipercentile	RG/EG
5	3	e(5.3)=eqxr(5.3)-tr(5.3)	No Equating, on scale	
5	4	e(5.4)=eqxr(5.4)-ur(5.4)	Linear	RG/EG
			Equipercentile	RG/EG
6	1	e(6.1)=eqxr(6.1)-ur(6.1)	Tucker Linear	NEAT
			Levine True-score Linear	NEAT
			Braun/Holland	NEAT
			Frequency estimation equipercentile	NEAT
			Chained Linear	NEAT
			Chained Equipercentile	NEAT
			KeNEATCE_Linear	NEAT
			KeNEATCE_Equipercentile	NEAT
			KeNEATPSE_Linear	NEAT
			KeNEATPSE_Equipercentile	NEAT
6	2	e(6.2)=eqxr(6.2)-ur(6.2)	Tucker Linear	NEAT
			Levine True-score Linear	NEAT
			Braun/Holland	NEAT
			Frequency estimation equipercentile	NEAT
			Chained Linear	NEAT

Table C.6

Cont.

Grade	Group within Grade	Residual to be Analyzed	Equating Method	Equating Design
			Chained Equipercentile	NEAT
			KeNEATCE_Linear	NEAT
			KeNEATCE_Equipercentile	NEAT
			KeNEATPSE_Linear	NEAT
			KeNEATPSE_Equipercentile	NEAT