Assessing the Impact of Characteristics of the Test, Common-items, and Examinees on the Preservation of Equity Properties in Mixed-format Test Equating

by

Raffaela Wolf, MS, MA

Bachelor of Science, University of Maine, 2003

Master of Science, Robert Morris University, 2008

Master of Arts, University of Pittsburgh, 2010

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Research Methodology

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Raffaela Wolf

It was defended on

November 13, 2013

and approved by

Suzanne Lane, Professor, Psychology in Education

Clement A. Stone, Professor, Psychology in Education

Kevin H. Kim, Associate Professor, Psychology in Education

Levent Kirisci, Professor, Pharmaceutical Sciences

Dissertation Advisor: Suzanne Lane, Professor, Psychology in Education

Assessing the Impact of Characteristics of the Test, Common-items, and Examinees on the Preservation of Equity Properties in Mixed-format Test Equating

Raffaela Wolf, PhD.

University of Pittsburgh, 2013

Preservation of equity properties was examined using four equating methods - IRT True Score, IRT Observed Score, Frequency Estimation, and Chained Equipercentile - in a mixed-format test under a common-item nonequivalent groups (CINEG) design. Equating of mixed-format tests under a CINEG design can be influenced by factors such as attributes of the test, the common-item set, and examinees. Additionally, unidimensionality may not hold due to the inclusion of multiple item formats. Different item formats could measure different latent constructs and thus cause a multidimensional test structure. The purpose of this study was to examine the impact of test structures (unidimensional versus within-item multidimensional as modeled through a bifactor model), differences in group ability distributions (equivalent versus nonequivalent), and characteristics of the common-item set (format representative versus non-representative) on each equating method's ability to preserve equity properties.

The major findings can be summarized as follows: IRT equating methods outperformed traditional equating methods in terms of equity preservation across all conditions. Traditional equating methods performed similarly when groups were equivalent. However, large discrepancies between the methods were found as a direct function of an increase in mean group ability differences. The IRT true score method was most successful in terms of First-Order Equity preservation regardless of test structure. All methods preserved Second-Order Equity similarly under unidimensional test structures. The IRT true score method was superior to all

other equating methods in terms of Second-Order Equity when the test structures were multidimensional. Similar results in terms of the Same Distribution property were obtained for each method when the groups were equivalent. The IRT observed score method was the best preserving when mean group ability differences increased. This was observed regardless of underlying test structure. Lower equity indices were observed when the common-item set was representative of the total test in particular when group differences were large. Similar patterns in terms of the performance of equating methods were observed regardless of the underlying test structure.

These results are discussed within the literature framework as it pertains to mixed-format test equating. Limitations of the current study are discussed and suggestions for future research are provided.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my academic advisor and chair of the dissertation committee, Dr. Suzanne Lane. I am truly thankful for her guidance and feedback throughout my studies at the University of Pittsburgh. I also would like to thank Dr. Kevin Kim for his constant assistance throughout the dissertation process and beyond. I wish to express my gratitude to the other members of my dissertation committee, Dr. Clement Stone and Dr. Levent Kirisci. The feedback provided by my committee members enhanced the quality of this study.

I also would like to thank my family and friends who have supported me throughout the chapters in my life. The wisdom and passion as expressed by these individuals inspired me during this journey and have enabled me to reach this goal.

This milestone is dedicated to my beloved grandparents, Walter and Hilde Wolf, who instilled in me the responsibility to strive for excellence in academic and non-academic endeavors.

1.0 INTRODUCTION

Mixed format test designs, composed of multiple-choice (MC) and constructed-response (CR) items, are becoming more common in large scale assessments to measure the broad range of content and skills delineated in federal and state education standards. Equating of mixed-format tests under a common-item nonequivalent groups design (CINEG) can be complicated by multiple factors such as attributes of the test, the common-item set, and examinees. Further, unidimensionality may be jeopardized due to the inclusion of multiple item formats. Different item formats could measure different latent proficiencies and thus cause a multidimensional test structure (Lane & Stone, 2006). Numerous procedures have been implemented to judge the adequacy of equating results, typically using criteria such as systematic and random errors rather than criteria that pertain to equating properties. Equating properties such as the first-order (FO), second-order (SO), and same distributions properties are used as evaluation criteria to assess whether the equating process has achieved its purpose in terms of the interchangeability of scores on alternate forms (Kolen & Brennan, 2004). The application of equity-based criteria to mixed-format tests under different equating designs as well as comparisons to equipercentile and item response theory (IRT) equating methods is scarce. This dissertation compared current equating methods in terms of their ability to preserve FO, SO, and the same distribution equity properties under various test structures, differences in group ability distributions, and characteristics of the common-item set.

Many large scale assessment programs and state assessment organizations have embraced the use of multiple item formats because a mixed-format test design affords the opportunity to assess a broader set of cognitive skills. Examples of programs that include mixed-format tests in addition to state assessments include: The College Board's Advanced Placement (AP) examinations, the National Assessment of Educational Progress (NAEP), the Graduate Record Examination (GRE), and the Test of English as a Foreign Language (TOEFL).

1.1 RATIONALE FOR AND DEFINITION OF EQUATING

Due in part to state and federal legislations, high stakes testing scenarios associated with the use and interpretation of test scores have increased over the past decade; therefore, most large scale testing programs must construct and administer multiple forms of the same test to control for item exposure and to ensure test security. The use of multiple forms also affords the opportunity for examinees to be assessed on multiple occasions. One assumption is that test developers have assembled test forms that are similar in content, and psychometric specifications including difficulty (Kolen & Brennan, 2004). Despite these efforts, it is almost inevitable that differences among forms exist in terms of difficulty, which raises a concern in terms of comparability of test scores and test fairness. In order to use the scores from different forms interchangeably, a statistical adjustment needs to be applied to place test forms onto a common scale. This adjustment process is commonly referred to as test equating or simply equating. Equating can be conceptualized as a statistical procedure that adjusts for incidental difficulty differences between forms and that relates scores on the various forms to one another and to the resulting score scale.

1.2 EQUATING DESIGNS AND METHODS

The equating process begins with the selection of an equating design, which refers to a plan in terms of how the data are collected and then analyzed by one or more equating methods. The equating designs include a single group design, a random groups design, and the CINEG design (Kolen & Brennan, 2004). In the single group and random groups designs, the samples of examinees taking different forms of an assessment are the same and thus are assumed to be equivalent in ability; therefore, the group ability differences do not need to be disentangled from form differences. In the CINEG design, group ability is not assumed to be equivalent. The CINEG design has been widely implemented in large scale assessment programs mainly due to its flexibility in terms of disentangling form difficulty and examinee's ability. A certain item pool, that should be representative of the test specifications, is chosen to be administered on two distinct test forms. These items are often referred to as common-items or anchor item sets. Two test taker populations that are likely nonequivalent in overall ability each take one of the two test forms.

Several equating methods have been developed for use with each of these equating designs. Equating methods are the building tools to derive an equating function which places scores from different test forms onto a common scale. In general, equating methods can be based on Classical Test Theory (CTT) or Item Response Theory (IRT) methods. Classical equating methods that have been used by testing programs include: Tucker method (Gulliksen, 1950), the Levine true and observed score methods (Levine, 1955), Braun-Holland linear method (Braun & Holland, 1982), frequency estimation (FE) method (Angoff, 1971), and the chained equipercentile method (Angoff, 1971). Theoretical and practical applications of these procedures are described by Kolen and Brennan (2004). When equating is conducted using the CINEG

design, strong statistical assumptions are made regarding the anchor item set since the common-items are used to adjust for differences in groups' proficiencies. The methods mentioned above differ from one another in terms of their statistical assumptions.

Technological advances have sparked the development of more sophisticated computer software programs and, as a result, the application IRT equating methods, such as IRT true- and observed score methods, are more commonly used in testing programs. Item response theory methods are popular because examinee responses are modeled at the item level rather than the test score level. The implementation of IRT models requires strong statistical assumptions such as unidimensionality and local independence. Unidimensionality implies that the likelihood of an examinee's successful performance on an item is contingent upon the item parameters and one latent construct. For the local independence assumption to hold, the performance on any two items is orthogonal for a fixed value of the latent trait estimate. Both IRT procedures (true score and observed score) entail numerous steps such as item calibration, scale transformation, and the equating to place forms onto a common scale.

1.3 EVALUATION CRITERIA

High stakes decisions are contingent upon assessment results, thus it is of great significance to ensure that the assessment results are fair, valid, and reliable and that score interpretations are suitable for all stakeholders involved in the assessment process. Thus, it is critical that equating results are evaluated for accuracy. In order to enhance the appropriateness of score interpretations, measurement experts in the field have advocated that the focus of many psychometric procedures should investigate the amount of measurement error associated with

each score scale (AERA, APA, & NCME, 1999). In general, criteria need to be identified so that equating accuracy can be evaluated. The preponderance of the psychometric literature on mixed-format test issues has focused on evaluating the equating results in terms of systematic (Bias) and random errors (e.g. RMSD, SEE). While these evaluation criteria are useful and widely implemented within the context of equating one should note that these criteria are not assessing whether the operational definition of equating has been attained. Criteria such as equity properties allow for the examination of equating accuracy in terms of the interchangeability of scores on alternate forms. In other words, equity can be conceptualized as the most important aspect of equating (Lord,1980) because equity criteria are closely linked to the adopted operational definition of equating and test fairness.

When FO equity holds, the examinees of a given latent trait estimate are anticipated to attain the same score on Form X and Form Y after equating. When the SO equity is preserved sufficiently, then the examinees with a given latent trait estimate are anticipated to be measured with the same accuracy. The equipercentile property, also known as same distributions property, focuses on the equivalence of score distributions after equating Form X and Form Y. In order to examine how well the equity properties have been preserved, a particular psychometric test theory model (e.g. unidimensional IRT) has to be assumed. The assessment of the equity properties involves several steps. First, a unidimensional IRT framework is used in that the MC-items are calibrated using a dichotomous model (e.g. three parameter logistic (3PL) model) and CR items are calibrated using a polytomous model (e.g. graded response (GR) model). With this approach, it is assumed that all test items are influenced by the same latent construct. Estimated item parameters are used to solve for the distribution of raw scores conditional on true scores for the reference test form. The raw score distribution is then typically converted to scale scores.

Once the item parameter estimates are put on the reference form scale, the conditional raw score distributions for the new form can be determined. These resulting raw scores are converted to their reference form equivalents and to scale scores based on the results from an equating method. Equity properties are applied to evaluate the comparability or interchangeability of scores between the test forms.

1.4 FACTORS INFLUENCING MIXED-FORMAT TEST EQUATING

Mixed-format test equating accuracy can be influenced by attributes of the test, common items, and examinees. When a combination of MC and CR items are used in an assessment, one concern may be that different items may be assessing different latent abilities and thus induce a more complex test structure also known as multidimensionality. In other words, diverse content areas, item format effects, speededness, guessing effect, and the confounding of examinees' cognitive skills with the items on the assessment may cause a more complex test structure and, as a result, nonequivalent constructs may be assessed unintentionally. Previous research examining the equating accuracy of various test structures for mixed-format tests showed equivocal results in different contexts. For example, researchers have suggested that MC and CR items may measure different latent abilities in writing assessments (Traub, 1993). However, similar latent constructs may be evaluated in the reading comprehension and quantitative domains (Bennett, Rock, & Wang, 1991).

Most testing programs employ unidimensional IRT models regardless of the underlying test structure because it is assumed that IRT equating methods are relatively robust to the underlying test structure. However, if the assumption of unidimensionality or essential

unidimensionality does not hold in an assessment, then bias is introduced into the equity properties when applying unidimensional equating methods to a multidimensional test structure. As a result, examinees with different latent abilities could earn the same achievement scores (Lord, 1980). The majority of previous research investigating the effects of multidimensionality on equating procedures has assumed a simple between-item multidimensional test structure (Béguin & Hanson, 2001; Béguin, Hanson, & Glas, 2000; Cao, 2008; S. Kim, 2004; S. Kim & Kolen, 2006; Tate, 2000; Wei & Yi, 2012). In practice, the underlying test structure could be far more complex; therefore, it is important to examine the impact of an underlying complex test structure when the unidimensional equity framework is applied to the data.

Since anchor items are typically utilized to place scores from different forms onto a common scale, it appears that the attributes of the common-items play an important role in the equating accuracy of mixed-format tests, particularly in terms of accounting for the different format types. While numerous equating guidelines have been established in terms of characteristics of the common-item set for MC-only assessments, it is still uncertain how well these guidelines can be extrapolated to mixed-format tests. Much of the current literature has been built on the assumption that equating guidelines for MC-only examinations also will hold true for mixed-format tests. One guideline is that common items should be representative of the total test; thus the common item set should include a combination of MC and CR items in a mixed-format test. Some researchers have argued that it is not always desirable to include CR items in the common-item set since the item pool for CR items is typically limited. The time and cost involved with administering and scoring these types of items has also been a concern for the psychometric community. The inclusion of CR items in the common-item set also could jeopardize test security due to the potential for memorization of the items (Haertel & Linn,

1996). Lastly, the inclusion of CR items could introduce another complexity in terms of systematic error due to raters scoring the items (Tate, 2000).

While the effect of common-item composition or format representativeness on mixed-format test equating has been examined in the past, most of these studies focused on whether a CR item should be included in the common-item set (S. Kim & Lee, 2006; Kirkpatrick, 2005; Tate, 2000). However, the results of these studies were based on different evaluation criteria. The few studies that examined the impact on equity properties utilized common-item sets that were composed of MC-only items (Andrews, 2011; He, 2011; E. Lee, Lee, & Brennan, 2012); therefore, it appears reasonable to examine how the accuracy of equating scores in terms of the preservation of equity properties is impacted when the common-item set is composed of MC-only items versus MC+CR items.

It is evident that the characteristics of the test and common item set impact the accuracy of the equating results for mixed-format assessments. However, several investigators (Bastari, 2000; Cao, 2008; S. Kim & Kolen, 2006; Kirkpatrick, 2005; W. Lee, Hagge, He, Kolen, & Wang, 2010; Powers & Kolen, 2011; Von Davier, Holland, & Thayer, 2004; Wang, Lee, Brennan, & Kolen, 2008; Wu, Huang, Huh, & Harris, 2009) have proposed that these equating results appear to be dependent upon group ability distributions across test administrations. Previous research has found that the various equating methods produce similar results when the test taking groups are equivalent in ability. However, in practice it is unlikely that different populations of test takers are equivalent in ability across test administrations; therefore, it is important to examine the severeness of nonequivalence in the ability distributions and its impact on the accuracy of equating outcomes in terms of preservation of equity properties. In addition, researchers who have examined the impact of group mean differences for mixed-format tests

under a multidimensional test structure have assumed that group ability distributions are identical for multiple-choice and constructed-response items. However, this may not hold true in practice and therefore group mean ability distributions varied across MC and CR subdomains in this study.

1.5 PURPOSE OF STUDY AND RESEARCH QUESTIONS

While there has been a steady increase in the psychometric literature pertaining to the utility of mixed-format tests, there are still many areas that should be investigated. There is a limited number of studies involving the assessment of equity properties for mixed-format tests. The studies that have been conducted employed operational test data that were assumed to be unidimensional and the common-item sets contained MC items only. The generalizability of the findings are limited due to the restriction of range in the factors of interest. Further, the format representativeness of the common-item sets under various test structures also appears to play an important role in the equating accuracy of mixed-format tests. Thus, it is important to assess how factors such as the type of common-item set and examinee's proficiency interact with the structure of the test in the context of equating. The primary purpose of this study was to use FO, SO, and same distribution equity criteria to evaluate the performance of four commonly used equating methods under a CINEG design with various test structures (unidimensional versus within-item multidimensionality) when a unidimensional equity framework was utilized. A secondary purpose was to examine the impact of the common-item characteristics and differences in ability group distributions on the preservation of FO, SO, and same distribution

equity properties. More specifically, this study attempted to answer the following research questions:

1. Overall, how do the Frequency Estimation (FR), Chained Equipercentile (CH), and Item Response Theory (IRT) true score (TR) and observed score (OB) equating methods compare to one another in terms of preservation of the First-Order (FO), Second-Order (SO), and same distribution equity properties?

2. Which of the above equating methods performs best in preserving equity properties when various test structures are applied to a unidimensional equity framework?

3. Which of the above equating methods performs best in preserving the equity properties when the common-item set format is representative and not representative of item types?

4. Which of the above equating methods is most accurate in preserving equity properties when groups associated with each form differ in ability?

Based on Kolen and Brennan (2004) it is hypothesized that methods may perform differently under each criterion. The equipercentile method and the IRT observed score method rely on observed score distributions in the equating process; thus these procedures may perform well under the same distribution property. The IRT true score method equates “true” scores; thus this method is expected to be superior under the FO property. However, it is difficult to predict which method will perform best in preserving the SO property, especially since a unidimensional equity framework was applied to various test structures. Item response theory equating methods could be favored since the equity framework assumed an underlying IRT model.

The above research questions were addressed through a simulation study. A bifactor model was be used to generate the response data under the multidimensional test structure. The 3PL/GRM model combination was used to calibrate the data. Concurrent calibration was conducted by placing parameters onto a common scale through the common-item set. The factors that were manipulated include: Common-item composition, differences in group ability distributions, and the underlying test structure. The impact of these factors were compared for four commonly used equating methods and the results were assessed in terms of the preservation of equity properties.

The remainder of this dissertation is organized as follows. Chapter two reviews related literature on mixed-format test equating is an attempt to provide a theoretical background. Chapter three presents the methodology employed in this study with a particular focus on the research design, the simulation procedure, and evaluation criteria. In Chapter four the results of this study are summarized and reported. Chapter five provides a discussion of the major findings, and the limitations and future directions of the current research are addressed.

2.0 REVIEW OF LITERATURE

This chapter consists of several sections. The item formats that comprise a mixed-format test are discussed in terms of their relative strengths and weaknesses, followed by an explanation of mixed-format tests. Equating designs and methods then are described followed by a discussion of equating properties. The chapter concludes with a summary of prior relevant research as it pertains to this study.

2.1 ITEM FORMATS

Test designs utilizing a variety of test item formats are widely implemented in assessment systems. It is assumed that combinations of different item formats allow for the measurement of a broader set of cognitive skills than the use of a single format design (Kim & Lee, 2006). The different formats are generally classified as multiple-choice (MC) items and constructed-response (CR) items. Multiple-choice items require the test taker to select an option from a set of response strings (e.g., four or five). In contrast, CR items necessitate the examinee to generate a response. Multiple-choice items are typically scored dichotomously whereas CR items can be either dichotomously or polytomously scored.

Each of these item formats has numerous associated strengths and weaknesses. The main differences between the two item formats highlighted in the literature are that MC items allow

for the assessment of a greater span of content coverage in a small window of testing time under small budget constraints while only a few CR items can be administered since it takes examinees more time to generate responses. The clear disadvantage in this case is that the use of a restricted number of items usually leads to an underrepresentative sample of the content domain (Linn, 1995). In addition, the scoring of CR items typically involves a group of raters who are trained on applying appropriate scoring rubrics to evaluate student responses, which increase the associated time and cost. Further, the scoring of CR items may still be considered subjective and may vary across judges and occasions despite the use of trained raters and scoring rubrics. However, proponents of CR items have contended that these items may increase content representativeness by affording the opportunity to measure content and skill objectives more directly as compared to MC items. This may be due to the fact that MC items are less likely to evoke certain types of cognitive skills due to their constrained nature. However, research has demonstrated that well-written MC items can indeed elicit complex thought processes, such as problem solving (Haladyna, 1992, 1997; Wainer & Thissen, 1993) but CR items may measure higher level cognitive skills that cannot be measured by MC items.

Constructed-response items can encompass a greater variety of tasks ranging from fill-in-the-blanks to writing an essay to producing a multi-step solution to a quantitative problem; therefore, the cognitive skills stimulated by different varieties of CR items may vary substantially. Thus, CR items may capture both the process of student learning and the final product. In addition, the use of CR type items may eliminate the random guessing effect seen in MC items since examinees are required to generate responses rather than choosing from a set of response options. It is evident that the rationale for incorporating CR items in a mixed-format assessment is that CR items may assess cognitive processes that cannot be adequately measured by MC items.

2.2 MIXED-FORMAT TESTS

Both MC and CR type items display strengths and weaknesses and it appears that one item format is not superior to another; therefore, many assessments are composed of mixed-format tests that include both item formats. It is typical that these mixed-format tests contain a large number of MC items with a few short CR items (Koretz & Hamilton, 2006) mainly due to practical purposes, such as cost and time constraints.

As with other assessment systems, mixed-format tests must be equated to ensure that test scores are comparable across test forms. Therefore, the process of equating commences with the choice of a data collection design.

2.3 DATA COLLECTION DESIGNS

Equating designs are implemented to aid in the collection of test data which is to be used within the context of equating. Three widely used data collection designs in test equating and scaling are the single group design, random groups design, and common-item nonequivalent groups design (CINEG) (Kolen & Brennan, 2004). Each of these designs poses their own strength and limitations within different contexts.

2.3.1 Single Group Design

In the single group design, one group of test takers are administered both test Form X and Y, usually in counterbalanced order to control for order effects. Since the examinees take both forms, it appears to be less challenging to determine the difficulties between the two forms. However, the concerns regarding the implementation of the single group design include differential order effects as well as administration time since each examinee needs to take both forms.

2.3.2 Random Group Design

In the random group design, examinees are administered either Form X or Form Y as determined through random assignment. However, both test forms must be proctored in the same test administration; therefore, the main limitations of the random groups design appear to be test security since both forms must be administered at once and sample size, which is assumed to be large since each examinee takes only one form.

2.3.3 Common Item Nonequivalent Design

In the Common Item Nonequivalent Group (CINEG) design, there are two different test taking populations. One group takes Form X whereas the other takes Form Y. Form X and Form Y have a set of items in common also referred to as anchor or common-item set, which is used to equate the two forms.

If the score on the anchor set is integrated in the total test score then the common-item set is referred to as being internal as opposed to external which implies that the score on the anchor set is not part of the total test score. The CINEG design improves upon the single group design by not requiring examinees to take both Forms X and Y. In addition, it improves the flexibility compared to the random group design because the CINEG design does not require for the test taking populations to be equivalent. In the random groups design there is no need to adjust for group ability differences because no common-item set is needed to estimate data for a synthetic population. Further, a relatively larger number of examinees is required in the random groups design as compared to the CINEG design to ensure the constant equating accuracy across different equating samples. The CINEG design was chosen for this study due to its wide implementation in research and practice. The next step in the equating process is the choice of equating methods to place scores onto a common scale.

2.4 CLASSICAL EQUATING METHODS

Classical equating methods typically are either based on linear methods or equipercentile methods. Procedures based on linear methods are not often applied in practice and thus this study focuses on the implementation of equipercentile equating methods. Two widely implemented equating methods by testing programs are the frequency estimation method and the chained equipercentile method. These methods are described in more detail in the following section.

2.4.1 Frequency Estimation Method

Braun and Holland (1982), Angoff (1971), and Kolen and Brennan (2004) have discussed the frequency estimation equipercentile equating method (FR). This procedure considers two scores (Form X and Form Y) to be comparable if the two scores have the same percentile rank. Since the CINEG design will be used in the current study, the population of interest is a synthetic population, which can be conceptualized as a weighted combination of the two populations from which the two groups were sampled. Thus, the synthetic distributions for Form X and Form Y can be written as

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x), \tag{2.1}$$

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y), \tag{2.2}$$

where the subscript *s* denotes the synthetic population,

w_1 and w_2 are synthetic weights for populations 1 and 2,

f(x) represents the score distribution on Form X,

g(x) represents the score distribution on Form Y.

Under this method, some assumptions about the data need to be made because samples from population 1 and 2 took Form X and Form Y, respectively. More specifically, it is assumed that the distributions conditional on the number of common items answered correctly is the same for each population. This can be expressed as

$$f_1(x|v) = f_2(x|v), \tag{2.3}$$

$$g_1(y|v) = g_2(y|v), \tag{2.4}$$

where *v* represents the common-item score. The joint distribution of the two unobservable quantities $f_2(x|v)$ and $g_1(y|v)$ can be written as

$$f_2(x|v) = f_2(x|v)h_2(v) = f_1(x|v)h_2(v), \tag{2.5}$$

$$g_1(y|v) = g_1(y|v)h_1(v) = g_2(y|v)h_1(v), \tag{2.6}$$

where *h(v)* represents the marginal distributions for the common-item scores. The marginal distributions of $g_1(y)$ and $f_2(x)$ can be estimated by summing over all the levels of *v*. These values can then be substituted into Equations 2.1 and 2.2 and the following synthetic group distributions for Form X and Y are obtained as follows

$$f_s(x) = w_1 f_1(x) + w_2 \sum_v f_1(x|v)h_2(v), \tag{2.7}$$

$$g_s(y) = w_1 \sum_v g_2(y|v)h_1(v) + w_2 g_2(y). \tag{2.8}$$

The following formula can then be applied to approximate the equipercentile relationship between test forms in the synthetic population

$$eq_X(y) = P_s^{-1}[Q(y)] \tag{2.9}$$

where $Q_s(y)$ is the percentile rank function for Form Y for the synthetic group, P_s^{-1} is the inverse percentile rank function for Form X for the synthetic group, and F_s is the cumulative distribution function of scores on Form X.

2.4.2 Chained Equipercentile Equating

Angoff (1971) described another method that incorporates equipercentile equating which is commonly known as chained equipercentile equating (CH). Several steps are performed to obtain an equating relationship in that test forms are equated using a chain that proceeds from the new form to the common-item scale and then to the reference form. More specifically, equipercentile equating methods are utilized to convert scores on Form Y to equivalents on the common-item scale, *V*, for the group that took Form Y. This equating function can be expressed as $e_{V1}(y)$. The next step entails finding the equipercentile equating relationship between the common-item scores and the Form X scores ($e_{X2}(v)$). In order to obtain Form X equivalents from Form Y, Form Y scores are converted to common-item scores using $e_{V1}(y)$ and then to Form X scores using $e_{X2}(v)$. This can be expressed as

$$e_X = e_{X2}[e_{V1}(y)]. \qquad 2.10$$

2.4.3 Smoothing

Within the context of the CINEG design, there are two observed bivariate distributions: one for the pair (X, A) of Form X and the other for the pair (Y, A) of Form Y, where *A* represents the anchor or common item set. These distributions are obtained from the two samples of examinees that take Form X and Y, respectively. Random error is induced by sampling examinees from the

overall test taking population; thus, irregularities at the extremes of the score scale often are observed, in particular when sample sizes are small. Since these obstacles may results in inaccurate equating functions, Kolen and Brennan (2004) suggested to apply smoothing methods to sample score distributions prior to equating in an attempt to mitigate these effects. Presmoothing often is employed to smooth out some of the sampling variability in an attempt to produce more stable score distributions. When samples are sufficiently large, presmoothing may not lead to a significant improvement. However, presmoothing could aid in removing some of the undesired roughness in the sample score distributions (Hanson, Zeng, & Colton, 1994).

Several presmoothing methods are available. Some of the popular models used in presmoothing include the log-linear models, the four-parameter binomial model, and the beta binomial models. The current study employed the log-linear models as these may fit a wider class of bivariate distributions (Holland & Thayer, 2000). The log-linear models considered in this dissertation are those that produce a smoothed version of a bivariate distribution of total test score and common-item score, such as (*X*, *A*) for Form X and (*Y*, *A*) for Form Y. The following log-linear model can be used to fit a bivariate distribution to the observed distribution of (*X*, *A*)

$$log_e(p_{ij}) = \beta_0 + \sum_{c=1}^{C} \beta_{xc} x_i^c + \sum_{d=1}^{D} \beta_{ad} a_j^d + \sum_{e=1}^{E} \sum_{f=1}^{F} \beta_{xaef} x_i^e a_j^f \ , \qquad (2.11)$$

where p_{ij} represents the expected joint score probability of the pair ($x_i \ on \ X, a_j \ on \ A$), β_0 can be thought of as a normalizing constant that forces the sum of the expected probability p_{ij} to equal 1, whereas the remaining β's are free parameters that are to be estimated in the process of model fitting.

Through the implementation of this model, *C* moments (i.e. means, standard deviations) in the univariate marginal distribution of *X*, and *D* moments (i.e. means, standard deviations) in the univariate marginal distribution of *A* are preserved. The number of cross moments

(covariance) in the bivariate (X, A) distribution is determined by E and F, respectively. A similar procedure can be applied to the observed bivariate distribution of Form Y.

2.5 ITEM RESPONSE THEORY

Although classical test theory (CTT) was the pillar of psychological test development for several decades, IRT has become the mainstream theoretical framework for modeling individual item responses at the item level in the context of educational and psychological measurement since the 1950's. The backbone of IRT rests upon a non-linear function that connects the item characteristics (e.g. slope, location, and/or guessing parameters) and the underlying latent proficiency responsible for the item response. In order to obtain valid and reliable IRT measurement results, several stringent statistical assumptions must be met. Further, numerous IRT models exist and the choice of model is contingent upon the nature of the data and research questions of interest. The next sections describe the assumptions underlying IRT as well as various IRT models commonly applied in practice.

2.5.1 Unidimensional Item Response Theory Assumptions and Models

An abundance of psychometric literature has focused on Item Response Theory (IRT) (Lord, 1980), which employs mathematical models to relate an examinee's latent construct (θ) and item characteristics to the likelihood of a correct response on a particular item. When using IRT to analyze data, it is pivotal that some assumptions such as monotonicity, local independence, and dimensionality have been met in order to ensure valid and reliable examinee score interpretations

within an educational measurement framework (Embretson & Reise, 2000). Monotonicity asserts that the likelihood of successful performance is a non-decreasing function of a test taker's proficiency. Local independence infers that item performance is provisionally independent given an examinee's trait level, whereas the dimensionality of an assessment refers to the quantity of latent aptitudes required to capture the construct of interest (Embretson & Reise, 2000). In an attempt to separate dominant dimensions from transient dimensions, the concept of essential unidimensionality was proposed by Stout (1987). Essential unidimensionality can be conceptualized as the least complex test structure necessary to allow for the assumptions of monotonicity and local independence to be met.

A variety of models exist for the estimation of item parameters under the assumption of unidimensionality in IRT. If dichotomous IRT models (multiple-choice items only) are utilized, then one, two, and three-parameter logistic (i.e., 1PL, 2PL and 3PL) or normal ogive models can be applied. The difference between the logistic and normal ogive mathematical functions in terms of probabilities and parameter estimates is diminutive. However, in practice logistic models are applied more frequently due to their simplicity in computation (Embretson & Reise, 2000).

For example, the probability of a successful performance on item *j* for the 3PL model can be expressed as

$$P\left(X_{ij}=1\middle|\theta_j,a_i,b_i,c_i\right)=\ c_i+\frac{1-c_i}{1+exp^{-Da_i\left(\theta_j-b_i\right)}}, \tag{2.12}$$

where

X_{ij} is the response of person *j* to item *i* (scored 0 or 1; 1 indicates a correct response);

θ_j reflects an examinees proficiency;

a_j corresponds to the item discrimination parameter;

b_j is the item difficulty parameter;

c_j is known as the pseudo guessing parameter, and

D is a scaling constant (1.7).

Equation 2.12 defines the probability of a successful performance to an item as a function of one latent trait or person parameter (θ_j) and three item parameters (a_i, b_i, and c_i). The 3PL model reduces to the 2 PL model if the pseudo guessing parameter is zero. In addition, if it is assumed that all items have the same slope, then only the difficulty parameters are utilized to describe the response data and thus a 1PL or Rasch model can be applied.

Several polytomous IRT models exist for situations in which an examinee may obtain one of several different scores, such as on a CR item which could be scored on a scale of 0 to 4. Popular models in the literature include the graded partial credit (GPC) model (Muraki, 1998), the graded response (GR) model (Samejima, 1997), and the nominal response model (Bock, 1972). Detailed theoretical discussions on these types of models are well documented in the literature (Baker & Kim, 2004; De Ayala, 2009; Embretson & Reise, 2000). In these models, the probability of an examinee receiving each possible score category is a direct function of an examinee's latent proficiency. The choice of model is typically driven by the type of item data. Samejima's GR model was used in this study. The GR model is suitable for items with ordered polytomous responses such as CR items. The implementation of this model follows a two-step procedure. The first step requires the calculation of the cumulative category response functions that represent the likelihood of a successful examinee response at or above particular category level k ($k = 0, 1, \dots, m_i$) on an item. The probability (P^*_{ijk}) that an examinee j with a latent ability θ_j earns a score on item i at or above category k can then be defined as

$$P^*_{ijk} = \frac{\exp[Da_i(\theta_j - b_{ik})]}{1+\exp[Da_i(\theta_j - b_{ik})]}, \tag{2.13}$$

θ_j is the latent trait for examinees *j*;

a_i is the discrimination parameter of item *i*;

b_{ik} is the threshold parameter for category *k* of item *i*; and

D is the scaling constant (1.7).

The b_{ik} parameters can be conceptualized as the boundaries between category levels. For an item with $(m_i + 1)$ response categories, there are m_i threshold parameters and one item discrimination parameter (a_i).

Once these cumulative category response functions (P^*_{ijk}) have been obtained, the second step follows by solving for the likelihood of attaining a particular category by taking the difference between cumulative probabilities of adjacent categories. It follows that the likelihood that an examinee responds to a particular category *k* $(k = 0, 1, \dots, m_i)$ on item *i* can be expressed as

$$P_{ijk} = P^*_{ijk} - P^*_{ij(k+1)} \tag{2.14}$$

For example for an item with four categories, three cumulative probabilities will be computed using Equation (2.13), that is, P^*_{ij1}, P^*_{ij2}, and P^*_{ij3}. Based on Equation (2.14), the probability of responding to a particular category $(P_{ijk)}$ can be calculated as follows (Embretson & Reise, 2000):

$$\begin{cases} P_{ij0} = 1 - P^*_{ij1} \\ P_{ij1} = P^*_{ij1} - P^*_{ij2} \\ P_{ij2} = P^*_{ij2} - P^*_{ij3} \\ P_{ij3} = P^*_{ij3} - 0 \end{cases} \tag{2.15}$$

Baker and Kim (2004) suggested that any combination of the dichotomous and polytomous IRT models can be utilized for item parameter estimation in a mixed-format assessment. In this study, the 3PL/GR model combination was chosen due to the successful application of the 3PL model to MC items in terms of model fit in some well-known assessment programs such as NAEP, GRE, and the TOEFL examinations. The GR and GPC models are commonly applied to CR items and several studies examined the differences between the two models in terms of model fit. Despite the mathematical differences between the two models, the overall findings suggested that both models produced similar results in terms of model fit (Maydeu-Olivares, Drasgow, & Mead, 1994; Tang & Eignor, 1997).

2.5.2 Bifactor Model

The bifactor model was introduced by Holzinger and Swineford (1937) and it is typically used within the factor analysis and structural equation modeling communities. The model allows each item response to be explained by both a general or dominant factor as well as secondary orthogonal factors (Gibbons & Hedeker, 1992). The dominant trait is the factor of interest, whereas the secondary traits may be considered as subdomains. In other words, the application of the bifactor model allows for retaining the goal of measuring a general latent construct while controlling for the variance that arises due to the assessment of different cognitive skills by MC and CR subdomains within the context of mixed-format tests.

The assumptions of the model include that each item loads on a dominant factor in addition to only one of the subdomain factors (MC or CR). In addition, the subdomains are orthogonal to each other and to the dominant factor. For example, for a test that is composed of six items with

two subdomains (MC and CR), the model can be conceptualized in terms of a factor pattern as follows

$$\Lambda = \begin{pmatrix} \lambda_{10} & \lambda_{11} & 0 \\ \lambda_{20} & \lambda_{21} & 0 \\ \lambda_{30} & \lambda_{31} & 0 \\ \lambda_{40} & \lambda_{41} & 0 \\ \lambda_{50} & 0 & \lambda_{52} \\ \lambda_{60} & 0 & \lambda_{62} \end{pmatrix} \tag{2.16}$$

λ_{ij} represents the loading of item *i* (*i*=1,2,…6) on latent factor *j* (*j*=0,1,2).

In this structure matrix the general domain items will have a nonzero value of the item discriminations or slopes along with clusters of items that belong either to the MC or CR subdomain. All other item discriminations are zero.

2.5.3 Relationship between CFA and IRT Parameters

The parameters from factor analytic approaches do not directly correspond to the IRT item parameters. However, it is possible to transform the factor loadings (λ) and threshold values (τ) to obtain the item parameter estimates for the within-item multidimensional structure with uncorrelated dimensions as follows

$$a_{ik} = \frac{(D)\lambda_{ik}}{\sqrt{1-\sum \lambda_{ik}^2}} \tag{2.17}$$

In the case of the bifactor model, each complex item would have two a_{ik} or λ_{ik} slope parameters. Similarly the item-category threshold parameters can be obtained as follows

$$d_{ik} = \frac{\tau_{ik}}{\sqrt{1-\sum \lambda_{ik}^2}} \tag{2.18}$$

2.5.4 Comparison of Bifactor and Multidimensional Item Response Theory Models

The bifactor model is a complex MIRT in that some of the items load on more than one dimension as opposed to a simple structure in which different items load on different dimensions. An attractive feature of the bifactor model is its ease of interpretation as well as its simplification of the computational complexity. Every item loads on a dominant dimension in addition to one specific dimension regardless of how many specific dimensions exist; thus, the number of integrals for any bifactor model is always two. Therefore the computational complexity of a bifactor model is in close alignment with the two-dimensional MIRT models.

A three-parameter bifactor model can be modeled within the MIRT framework as follows

$$P(X_{ij} = 1|\theta_j, a_i, c_i, d_i) = c_i + (1 - c_i)\frac{\exp[D(a_{iG}\theta_G + a_{ik}\theta_k + d_i)]}{1+\exp[D(a_{iG}\theta_G + a_{ik}\theta_k + d_i)]}\ , \qquad (2.19)$$

where θ_G represents the general or dominant proficiency, while θ_s (s=1,2,…,*k*) represents one of the *k* subdomains that are orthogonal to each other as well as to the general or dominant proficiency. Additionally, a_{iG} and a_{is} are item discrimination parameters for the general factor and one of the *k* subdomains. Lastly, d_i can be conceptualized as a scalar parameter that is related to an overall multidimensional item difficulty as found in the typical MIRT model, where

$$d_i = -b_i\sqrt{a^2_{General} + a^2_{ij}} \qquad (2.20)$$

The bifactor model for graded response items (Cai, 2010; Gibbons et al., 2007) can be expressed as

$$P\ (\mathrm{X_{ij}} \geq 1|\theta_\mathrm{G}, \theta_\mathrm{S}) = \frac{1}{1+\exp\{-[\mathrm{d_1} + \mathrm{a_G}\theta_\mathrm{G} + \mathrm{a_S}\theta_\mathrm{S}]\}}, \qquad (2.21)$$

.
.
.

$$P\,(\mathrm{X_{ij}} \geq \mathrm{K\text{-}1} | \theta_{\mathrm{G}}, \theta_{\mathrm{S}}) = \frac{1}{1+\exp\{-\left[\mathrm{d_{K\text{-}1}} + \mathrm{a_G}\theta_{\mathrm{G}} + \mathrm{a_S}\theta_{\mathrm{S}}\right]\}},$$

where $d_1,\ldots,d_{K-1}$ are strictly ordered intercepts that are related to the MIRT item difficulty parameters, $a_G\ \theta_G$ are the item discrimination and proficiency estimates for the general factor, whereas $a_S\ \theta_S$ reflect the item discrimination and proficiency estimates for the specific factor.

2.5.5 Bifactor Model Applications

Bifactor models have been applied to empirical data in diverse contexts such as achievement tests and survey instruments in health and psychological measurements (Gibbons & Hedeker, 1992; Reise, Morizot, & Hays, 2007). The relative performance has been assessed in terms of model fit. For example, Gibbons and Hedeker (1992) fit a full-information bifactor model with four group factors to a dichotomously scored ACT science assessment. A sample of 1000 examinees was tested on 20 items. The bifactor model outperformed a four factor model with promax rotation in terms of model fit. In addition, it was noted that substantial factor loadings were observed on the general latent construct, whereas factor loadings on the subdomains appeared to vary in a greater range. Reise et al. (2007) compared the fit of the bifactor model to unidimensional and multidimensional IRT models. A sample of 1000 examinees completed a five domain health outcome survey instrument consisting of 16 items. The bifactor model was superior to both unidimensional and orthogonal multidimensional IRT models in terms of model fit.

Similar results in terms of model fit were obtained when a bifactor model was fit to graded response data of a survey instrument consisting of seven subdomains and 34 items (Gibbons et al., 2007). The model fit of the bifactor model was superior compared to the unidimensional IRT model. In a more recent application, Li and Rupp (2011) conducted a simulation study to examine the performance of the extension of the multidimensional S-χ^2 statistic under various conditions such as sample size, test length, and levels of the discrimination or factor slopes. Data were generated using either a simple-structure MIRT or full information bifactor model and then a unidimensional, multidimensional and full information bifactor model were fit to the data. Results indicated that the power of the S-χ^2statistic for detecting model misfit was low for all models under investigation regardless of which model was utilized as the generating and the fitting model.

The predominant application of the bifactor model within the field of educational measurement has been to testlet-based assessments, such as reading comprehension examinations. Several researchers have investigated the relationship between bifactor, testlet, and second-order MIRT models. Li, Bolt, and Fu (2006) demonstrated that the testlet model can be modeled as a constrained version of the bifactor model if the testlet item discriminations are proportional to the item discriminations of the general latent construct. The equivalence of the testlet model to a second-order MIRT model has been established by Rijmen (2010) and it was concluded that both the testlet and second-order MIRT models can be thought of as constrained bifactor models. Rijmen (2010) used data from an international English assessment test to assess model fit of the bifactor, testlet and second-order MIRT models. A sample of 13,508 examinees

took a subset consisting of 20 reading comprehension items that were comprised of four testlets with five items within each testlet. Results indicated superior model fit of the bifactor model.

DeMars (2006) also fit a bifactor model to testlet based assessments and concluded that the latent trait and item parameter estimates recovered appropriately for simulated and real data. However, latent trait recovery appeared to be less influenced by model choice compared to item parameter recovery. In particular, choice of model had the most impact on the recovery of the item discrimination parameters. Other psychometric issues that have been successfully addressed through the application of a bifactor model to testlet based assessments. These issues include vertical scaling (Li & Rijmen, 2009), extension of bifactor model to a multi-group bifactor model (Cai, 2010; Cai, Yang, & Hansen, 2011), and differential item functioning (DIF) (Fukuhara & Kamata, 2011; Jeon & Rijmen, 2010).

The bifactor model also was utilized to address construct shift within an IRT vertical scaling framework (Li, 2011). Model fit and recovery of parameter estimates were examined in terms of systematic and random error under various conditions such as sample size, length of common-item set, and variance of grade subdomains. The bifactor model showed superior model fit compared to a unidimensional 2PL IRT model. Parameter estimation accuracy was greatly affected by sample size as a larger sample size led to more accurate parameter estimates. The variance of the grade specific subdomains also affected the accuracy of item parameter estimates in that with a larger degree of construct shift the accuracy of the parameter estimates for the general dimension decreased, whereas the stability in terms of parameter estimates increased for the grade specific subdomains. The length of the common item set did not impact the results significantly.

In general, the bifactor model has proven to be a valuable tool to tackle various psychometric issues such as vertical scaling, differential item functioning (DIF), and multi-group modeling. However, most of these applications were in context of either dichotomously or polytomously scored instruments but not mixed-format assessments. In addition, the mainstream of the reviewed studies utilized a two parameter bifactor model. In an attempt to account for a guessing effect on the MC items, a three parameter bifactor model was chosen to generate the data for the dichotomously scored items. A bifactor graded response model was utilized to generate data for polytomously scored test items.

2.6 ITEM RESPONSE THEORY EQUATING AND MIXED-FORMAT TEST

Equating procedures that rest upon IRT methodologies typically incorporate three steps. Item calibration is performed first, which entails the choice of an IRT model to estimate the item and person parameters. The next step includes scale transformation to place the estimated parameters onto a common scale. The last step is the equating of the number correct scores in terms of raw-to-scale score conversions. The majority of research involving mixed-format test equating has focused on the first two steps of the IRT equating process, focusing on IRT linking methods.

In the item calibration step, a combination of dichotomous and polytomous IRT models are used to estimate MC item responses and CR item responses. As Baker and Kim (2004) indicated, any combination of dichotomous and polytomous IRT models can be chosen for the analysis of item responses in mixed-format tests. In the current study, the 3PL model was used to estimate MC response data and the GRM was used to estimate the CR item responses due to the wide implementation of these models in testing programs. One should note that the item

parameters for a mixed format test can either be estimated separately in that one format at a time is estimated, or simultaneously across formats. Some researchers (Sykes & Yen, 2000) have suggested that a simultaneous item calibration across formats may be more appropriate in practice because item formats are placed on the same scale allowing for the comparison of performance on the different item formats. Further, by utilizing the simultaneous calibration the weighting selection issue for the different item formats in a mixed format test is eliminated.

In the scale transformation step, there is a choice of either separate calibration or concurrent calibration when placing IRT parameter estimates from two different test forms onto a common scale. In a separate calibration, the item and person parameters for the two test forms are estimated distinctly in two computer runs. Scale transformation methods then are employed, in an attempt to place parameter estimates of one form on to the scale of the other form through the common-item set. The typical scale transformation methods for dichotomous IRT models are two moment methods and two characteristics curve methods. Two moment methods include: mean/mean (Loyd & Hoover, 1980) and mean/sigma (Marco, 1977). There are also two characteristic curve methods known as Haebara (1980) and Stocking and Lord (1983). The moment and characteristic curve methods also have been extended to different polytomous IRT models (Baker, 1992, 1993; A. S. Cohen & Kim, 1998). The theoretical underpinnings in terms of how these methods have been extended to mixed-format tests can be found in S. Kim and Lee (2006).

In concurrent calibration, the item and person parameters on both test forms are estimated jointly in one computer run, which guarantees that all parameter estimates are on the same scale. This is achieved by combining data from both test taking populations and treating items not taken by a particular population as not reached or missing. Extensive research has been

conducted investigating the relative performance of concurrent calibration and separate calibration for single format tests (e.g. MC-only) with different scale transformation methods. In general, the findings suggest that concurrent calibration may outperform separate calibration under various conditions because it is believed that concurrent calibration makes complete use of the available information and may remove some bias associated with potentially inaccurate scale transformation procedures as induced by separate calibration (S. H. Kim & Cohen, 1998; Kolen & Brennan, 2004).

In the context of mixed-format tests, the concurrent calibration method also has been shown to be superior to separate calibration techniques when the groups taking each form are equivalent in ability and when the test structure is unidimensional (S. Kim, 2004; S. Kim & Kolen, 2006). Both methods appear to perform similarly in terms of systematic and random error when a more complex test structure was induced by item format effects (Hanson & Béguin, 2002). However, as groups become more nonequivalent in terms of ability and the abilities are highly correlated, separate calibration with characteristic curve methods may produce slightly more accurate item parameter recoveries compared to concurrent calibration (Béguin & Hanson, 2001; Béguin et al., 2000). Since concurrent calibration typically outperforms separate calibration in regards to linking accuracy and robustness to multidimensionality only the concurrent calibration method was utilized in this study.

2.6.1 Item Response Theory Equating Methods

2.6.1.1 Unidimensional Item Response Theory True Score Equating

Unidimensional IRT true score equating creates an association between unobserved latent variables, also called true scores, on Form X and Form Y. Item response theory true score equating can be performed in three steps:

1. A true score τ_Y on Form Y is established.
2. A latent proficiency θ that corresponds to τ_Y is identified through an iterative procedure such as the Newton-Raphson method. Details of the Newton-Raphson procedure can be found in Kolen and Brennan (2004).
3. The true score on Form X (τ_X) that is associated with the latent prficiency θ from step two is found.

Lord (1980) was the first to describe this method by relating a latent construct (θ) to true or expected raw scores through the implementation of the test characteristic curve. For example, the test characteristic curve for Form X can be expressed as

$$\tau_X(\theta) = E(X|\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m_j} U_{jkX} P_{jkX}(\theta), \quad (2.22)$$

where U_{jkX} reflects the score that is associated with category k and $P_{jkX}(\theta)$ stands for the likelihood that an examinee with the latent construct (θ) obtained that score. The total number of items is represented by n, whereas m_j reflects the largest score category for item j. For MC items which are dichotomous (0,1), the 0 corresponds to an incorrect response and 1 indicates a correct response. The polytomous items in this dissertation all have a 0 as the first score category. The test characteristic curve for Form Y can be expressed in a similar manner

$$\tau_Y(\theta) = E(Y|\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m_j} U_{jkY} P_{jkY}(\theta). \quad (2.23)$$

The relationship between true scores for the new form and the old form can be established after the item parameters have been placed on the same scale. In this process, a corresponding θ is found for each integer true score on Form Y. The true score on Form X that corresponds to each θ value is considered to be equivalent. This equation can be written as

$$e_X(\tau_Y) = \tau_X(\tau_Y^{-1}), \tag{2.24}$$

where τ_Y^{-1} symbolizes the value of θ that corresponds to the true score of τ_Y.

2.6.1.2 Unidimensional Item Response Theory Observed Score Equating

Observed score equating within the IRT framework can be described in two steps. The observed number correct score distributions for Form X and Y are estimated in the first step. Then, traditional equipercentile equating methods are applied to the estimated observed number correct score distributions. A recursive algorithm (Lord and Wingersky, 1984) is used to obtain the raw score distributions for Form X and Form Y. For example, define $f_r(x|\theta)$ as the distribution of the number correct scores over the first *r* items for examinees with latent proficiency θ, and p_{ir} as the probability for those examinees to answer r^{th} successfully. Then, for *r* > 1, the recursion formula is as follows (Kolen & Brennan, 2004):

$$\begin{aligned} f_r(x|\theta_i) &= f_{r-1}(x|\theta_i)(1-p_{ir}), && x = 0 \\ &= f_{r-1}(x|\theta_i).(1-p_{ir}) + f_{r-1}(x-1|\theta_i)\,p_{ir}, && 0 < x < r, \\ &= (x-1|\theta_i)p_{ir}, && x = r \end{aligned} \tag{2.25}$$

The formula is applied by starting with *r* = 1 and progressively increasing *r* on each repetition until *r* is equal to the total number of items on the assessment. The resulting conditional distributions then are aggregated over all ability estimates to obtain the approximated population

distribution of raw scores for Form X and Form Y after the item parameters have been placed onto the same scale. This procedure can be mathematically expressed as

$$f_X(x) = \int_\theta f_X(x|\theta) f_\theta(\theta), \tag{2.26}$$

where $f_\theta(\theta)$ reflects the ability distribution. Many testing programs utilize a posterior distribution of θ, which is obtained from the IRT calibration process. In this case, the marginal distribution can be computed by using

$$f_X(x) = \sum f_X(x|\theta) f_\theta(\theta) \tag{2.27}$$

where $f_\theta(\theta)$ represents the posterior weight found at the quadrature point θ. The same procedure can be utilized for Form Y. Once cumulative distributions ($F_X(x), F_Y(y)$) have been obtained from $f_X(x)$ and $f_Y(y)$, a conventional equipercentile equating is conducted to conclude the equating process.

2.7 EVALUATION CRITERIA AND EQUITY PROPERTIES

It is evident that the process of equating scores in order to make alternate forms comparable is of great importance since high stakes decisions are contingent upon the accuracy of the equating outcomes. Evaluation criteria have been proposed and applied in research and practice. The majority of research on mixed-format test equating has used systematic (Bias) and random error (e.g. RMSD, SEE) as the equating criterion. While some of these indices are well suited as overall summary indices, there are some limitations worth mentioning. A potential problem with systematic error is that raw-to-scale score conversions could be systematically too high for half of the score distribution and too low for the remaining half of the score distribution if the bias between the two forms was in the opposite direction, which may cause the error to cancel out

(Harris & Crouse, 1993). Another frequently used equating criterion is the standard error of equating (SEE), which is the standard deviation of score equivalents over replications of an equating function on samples from a population of examinees (Kolen & Brennan, 2004). The adequacy of using SEE as a criterion has been criticized in that it only accounts for random errors due to the sampling of examinees from the population of test takers; thus, other sources of imprecision are not taken into consideration (Harris & Crouse, 1993). Based on these limitations, it appears that none of the criteria are unequivocally preferable to the others; therefore, the use of different criteria may lead to different conclusions in terms of equating accuracy depending upon the specific situation.

The use of equity properties as evaluation criteria for equating results has been limited in the literature mainly due to difficulties associated with computations and explanations of the procedures. Equity properties as evaluation criteria may be better than the overall summary indices typically employed in that they are in alignment with Lord's equity definition of equating. Lord (1980) argued that equity is achieved if an examinee would be indifferent about which form of the test is taken. In other words, conditional on a latent construct, the distribution of equated scores on the new form should be equivalent to the conditional distribution of scores on the reference form; thus, these conditional distributions should be the same at all levels of the latent construct and between test forms. Mathematically this can be expressed as the following

$$F^*[eq_X(Y)|\tau] = F(x|\tau), \text{ for all } \tau, \tag{2.28}$$

where τ represents the latent construct, F^* is the cumulative distribution of the equated scores on Form Y, $eq_X(Y)$ is the equating function that puts a score on Form Y onto the scale of Form X, and F is the cumulative distribution of scores on Form X.

Lord (1980) showed that this condition can only be achieved if the two forms are perfectly reliable (i.e. have a value of 1) or if both forms are identical, in which case equating is redundant (Kolen & Brennan, 2004). Divgi (1981) and Morris (1982) discussed a more flexible definition of equity. According to this definition, the means are conditional on the latent construct and should be equivalent after equating each test form, which also is referred to as first-order (FO) equity. First-order equity can be expressed as

$$E[eq_X(y)|\tau] = E(X|\tau) \text{ for all } \tau. \tag{2.29}$$

Similarly, standard deviations of the conditional distributions should be equivalent, which is known as second-order (SO) equity and can be expressed as

$$SEM_{eq_X(y)|\tau} = SEM_{X|\tau}\,. \tag{2.30}$$

While these two equity properties are concerned with individual score points along the scale of scores, another property known as observed score or same distribution equity property takes into account the score distribution. This property assumes that the distribution of equated scores of the new form (Form Y) is equivalent to the score distribution of the reference form (Form X). This property can be expressed as

$$F^*[eq_X(y)] = F(X). \tag{2.31}$$

There are some practical implications regarding equity properties that are worth discussing (Andrews, 2011). For instance, if the FO and SO equity properties are not preserved, then it is no longer a matter of indifference of examinees which test form is taken. If the FO equity does not hold, then the expected scale score at a certain ability is higher for one form of a test. Examinees with that trait estimate taking that particular form would have an advantage over examinees with the same trait estimate who take a different test form. Another possible scenario arises when the FO equity is preserved but the SO equity is not. In this situation, the conditional

means for the different test forms could be equivalent for a wide range of latent abilities. However, since the SO equity is not met, lower ability students would likely prefer to take the test form with more variability as these students could benefit from the measurement imprecision. In contrast, higher ability students would feel more strongly about the form that has a smaller variability because the likelihood of the appropriate assessment of their high ability is greater.

A psychometric model can be applied in an attempt to evaluate how well these properties are being preserved. This model must fulfill several features in that true scores must be related to latent constructs; thus, the error variances conditional on true scores also must be quantified. The versatility of several psychometric models has been examined. Kolen, Hanson, and Brennan (1992) presented methodologies for assessing how well the equating properties are preserved for scale scores when a strong true score model is used to model the data. They examined equating properties preservation within a unidimensional IRT framework for dichotomous tests. These methods were extended to tests with polytomous items (Wang, Kolen, & Harris, 2000). The focus of this dissertation was the application of the unidimensional IRT equity framework.

2.8 UNIDIMENSIONAL ITEM RESPONSE THEORY EQUITY FRAMEWORK

In order to assess how well the equating properties are maintained, the expected raw/scale scores and conditional standard errors of measurement (CSEM's) must be calculated. Since mixed-format tests consist of dichotomous and polytomous item formats, raw scores are used as number correct scores. The raw score can be written as

$$X = \sum_{j=1}^{n} U_j, \tag{2.32}$$

where U_j is the examinee's score on item j. Hanson (1994) and Thissen, Pommerich, Billeaud, and Williams (1995) described a procedure for calculating the conditional raw score distributions. Let $P_r(x|\theta)$ represent the conditional raw score distribution of Form X over the first r items at a given latent trait estimate. For the first item, each score category has a probability of earning that score given the latent trait estimate. This can be expressed as

$$P_1(x = U_{11}|\theta),\ P_1(x = U_{12}|\theta),\ldots\ P_1(x = U_{1k}|\theta) \qquad (2.33)$$

for the first, second, and last categories. The recursive formula then can be applied for each additional item in an attempt to find the probability of attaining a score x on the test that now contains r items. This formula can be written as

$$P_r(x|\theta) = \sum_{k=1}^{m_j} P_{r-1}\left(x - U_{jk}\middle|\theta\right)P_{jk}(\theta) \qquad min_r < x < max_r\ , \quad (2.34)$$

where min_r and max_r are the minimum and maximum scores on the test after the r^{th} item is added. This formula also is used to calculate the distribution of raw scores for Form Y. The true or expected raw score at a given latent ability is the mean of these conditional distributions and the standard deviation of this distribution is referred to as the CSEM. Raw scores are typically transformed into scale scores. However, only raw scores are used in the current study. The expected or true scale score given IRT ability is the mean of the scale score distribution

$$\tau_{SC|\theta} = E[SC(x)|\theta] = \sum_{j=\min x}^{\max x} SC(j) \times P(x = j|\theta)\ , \qquad (2.35)$$

where max and min reflect the maximum and minimum raw scores on the test, $SC(j)$ represents a scale score that corresponds to a raw score of j, and $P(x = j|\theta)$ stands for the probability of an examinee with latent trait θ attaining a raw score of j.

The variance of the conditional scale score distribution is the conditional error variance for scale scores which can be expressed as

$$\sigma^2_{SC|\theta} = \sum_{\min X}^{\max X}[SC(j) - \tau_{SC|\theta}]^2 \times P(x = j|\theta). \qquad (2.36)$$

The CSEM can be obtained by taking the square root of equation (2.41).

2.9 REVIEW OF STUDIES ON MIXED-FORMAT TEST EQUATING

The results of mixed-format test equating can be influenced by attributes of the test, common-items, and test taking populations. Therefore relevant studies as they pertain to the current study will be reviewed first, followed by a summary of studies involving equity properties.

2.9.1 Comparison of Equating Methods for Mixed-format Tests

Although previous studies have compared equating methods for tests composed of only MC items, an increased interest in the comparison of equating methods for mixed-format tests has more recently emerged. Various researchers have compared equipercentile equating methods such as the frequency estimation equipercentile (FR) and chained equipercentile (CH) to Item Response Theory (IRT) equating methods under a common item nonequivalent groups (CINEG) design for mixed-format tests (Hagge, 2010; Hagge et al., 2011; Liu & Kolen, 2011a; Powers et al., 2011). Each of the equating methods has been demonstrated to perform differently under specific conditions based on systematic and random error comparisons. For example, when the IRT assumptions have been met, IRT methods are preferred as these methods produce the smallest amount of systematic and random error. When larger group differences exist, the CH method has been shown to outperform the FR method in terms of the smallest amount of systematic error. However, the FR method can produce the smallest random error under small group differences. Overall, the amount of random error was found to be smallest for the IRT

observed score equating method compared to IRT true score equating and the FR and CH methods. Similar results were observed in terms of systematic error. However, when the discrepancy between the group ability distributions is small across administrations, all four methods produce similar results.

The comparison of performance of traditional and IRT equating methods for mixed-format tests is scarce in the literature. Most of the studies utilized operational test data and differences in equating performance were evaluated in terms of systematic and random error. Different results could be obtained when a simulation study is conducted and when the equating accuracy is assessed in terms of preservation of equity properties.

2.9.2 Test Characteristics

When a combination of MC and CR items are used, a common question arises in terms of equivalence of latent aptitudes measured by the different item types. Numerous uncontrollable factors could cause a more complex test structure and, as a result, nonequivalent constructs may be assessed unintentionally. For example, when considering a math assessment that captures problem solving and mathematical reasoning by means of mathematical communication, it would appear logical to assume an underlying multidimensional test structure rather than a unidimensional test structure. Several researchers have investigated whether the mixture of MC and CR items lead to a multidimensional test structure. Differences in study outcomes have been attributed to discrepancies among cognitive learning domains. For example, some evidence has suggested that MC and CR items may measure different latent abilities in writing assessments (Traub, 1993). However, similar latent constructs may be evaluated in the reading comprehension and quantitative domains (Bennett et al., 1991). Some researchers have found

that there still may be an item format factor in the underlying test structure of a quantitative test, which may result in multidimensional data (Sykes, Hou, Hanson, & Wang, 2002; Thissen, Wainer, & Wang, 1994).

Sykes et al. (2002) investigated the effects of various test structures embedded in the common items on IRT equating results. Factor analytic methods were used to assess the dimensionality of a large scale mathematics achievement test. The findings demonstrated that the use of multiple item formats led to a violation of the IRT assumption of unidimensionality. In addition, it was found that the first factor was a common dimension but MC items showed higher loadings on the second factor.

Similarly, Perkhounkova and Dunbar (1999) assessed the test structure of several language and mathematics assessments for various grade levels. Test formats were MC-only, CR-only, and mixed-format. The Poly-Dimtest procedure was used to assess the assumption of unidimensionality. For the language tests, the results indicated that both the MC and CR tests measure the same latent proficiency. However, for the mathematics assessments, the MC test, the mixed-format test, and the CR test appeared to measure different latent proficiencies.

Other studies have examined whether CR items are as effective in measuring the same construct as MC items. These studies have suggested that the inclusion of CR items in an assessment does not necessarily boost the reliability of a test that already consists of MC items. For example, Wainer and Thissen (1993) utilized Advanced Placement (AP) data from various content areas to investigate whether the inclusion of CR items into a test that already consists of MC items would lead to the measurement of a construct not captured by the MC items. The CR items were not superior to MC items in measuring constructs; thus the efficiency resulting from the inclusion of CR items was questionable. Similar results were reported in a follow-up study by

Lukhele, Thissen, and Wainer (1994). Data from various AP examinations were used to integrate CR items into a test consisting of MC items. It was concluded that CR items did not measure different constructs than the MC items.

These findings suggest that equivocal results have been found in terms of the equating accuracy for mixed-format tests under various test structures. The dimensionality of a test appears to vary greatly within different contexts such as diverse cognitive learning domains.

2.9.2.1 Application of Unidimensional Equating Methods when IRT Assumptions are Violated

Researchers have contended that unidimensional equating procedures are applied although the underlying test structure may be multidimensional. In an attempt to ensure valid and reliable interpretation of test scores it is of paramount importance to develop an understanding of what effect multidimensional data may have on unidimensional equating results. Several authors have suggested that unidimensional IRT equating methods may produce accurate equating results despite the presence of a multidimensional test structure (Camilli, Wang, & Fesq, 1995; Cook, Eignor, & Taft, 1988; Dorans & Kingston, 1985).

Dorans and Kingston (1985) examined the impacts of test dimensionality on the equating results of four forms of the verbal Graduate Record Examination (GRE). Factors of interest were: equivalent versus nonequivalent groups design and the comparison of various unidimensional calibration procedures. LOGIST was used to estimate person and item parameters using the 3PL logistic model. Dimensionality was assessed via factor analysis methods, which revealed that two highly related factors existed. The item calibration followed a multi-step procedure. The test was calibrated as a whole and then test items were split into two homogeneous subsections which were assumed to measure different latent constructs. These

subgroups then were calibrated separately to ensure that these items would be placed on the same scale as the original test. In the next step, all items were recombined to form one whole test. Similar adequate equating results were obtained across the factors of investigation; therefore, it was concluded that unidimensional IRT equating methods may produce accurate results irrespective of the test dimensionality structure of the data.

Similar conclusions were drawn by Camilli, Wang, and Fesq (1995) who examined the effects of various test structures on unidimensional IRT true score equating results of the Law School Admissions Test (LSAT). Factor analytic methods were utilized to assess the dimensionality structure of the test. It was found that the test contained two factors that spanned across content areas and a few factors that loaded on a specific item or cluster of items. The authors utilized the calibration procedure as found in Dorans and Kingston (1985). BILOG was used for item calibration. The equating results displayed only small differences at the tails of the score distribution. Because the differences were so small, the authors concluded that unidimensional IRT equating methods seem to be able to handle various test structures accurately in terms of IRT true score equating results.

However, it should be noted that all of these studies focused on the examination of equating accuracy in terms of systematic and random errors. Lord (1980) suggested that bias might be added into the equity properties when unidimensional equating methods are applied to multidimensional test structures. Consequently examinees with different latent proficiencies could attain the same test scores. Therefore, it is of importance to examine the robustness of the unidimensional equity framework when test structures vary in complexity.

2.9.3 Common Item Characteristics

When mixed-format test equating is conducted using a CINEG design, an anchor item set is utilized to place scores from different forms onto a common scale. Therefore, the impact of common-item set attributes such as common item length, item position, and representativeness and non-representativeness of the common items on equating results has been studied extensively. In general, it was concluded that a longer common item set may lead to more accurate equating results (Bastari, 2000, Wang, Lee, Brennan, & Kolen, 2008). In addition, common items should be administered in approximately the same position across different tests or test forms to avoid having the common items function differently across groups (e.g. differential item functioning [DIF]) (Cook & Paterson, 1987).

Another factor of interest has been the composure of the common-item set (MC-only, CR-only, MC+CR) and its effects on equating accuracy for mixed-format tests. Mixed results have been found across various studies. Hagge (2010) investigated how characteristics of mixed-format tests and configuration of the common-item set impact the accuracy of equating results under the CINEG design. Operational test forms and pseudo-test forms were utilized for analysis on three mixed-format tests from an AP Examination program. For the operational test form, factors such as difference in proficiency between groups of examinees across testing occasions and the relative difficulty of MC and CR items were examined. In addition, the ratio of MC and CR items relative to the total test and the item difficulty in the common item set were manipulated in the pseudo-test form analyses. The comparison of traditional (i.e. CH, FR) and IRT equating methods were other factors of interest in this study. It was concluded that using CR items along with MC items in the common item set may improve equating relationships in

certain situations, such as when MC and CR correlations are low or examinees perform differently on MC and CR items across the forms to be equated.

Similar conclusions were drawn by Hagge and Kolen (2011) who examined the impact of the ratio of MC and CR items of the common-item set relative to the total test on equating outcomes. The data generation process was adopted from the Hagge (2010) study. The factors of interest included equivalent versus non-equivalent groups, relative difficulty of MC and CR items, equating methods (FR, CH, IRT true score and observed score methods), and the composition of the common item items (MC-only versus MC+CR). Bias, conditional standard error of equating (CSE), and root mean squared error (RMSE) were calculated to examine the results at each score point, while weighted average root mean squared bias (WARMSB), weighted average RMSE (WARMSE), and the weighted standard error of equating (WASE) were computed to examine the amount of error over the entire score scale. Results indicated that the common item set containing both the MC and CR items led to the least amount of error in terms of WARMSB as compared to MC-only common-item set. However, these results were not consistent across study conditions. For example, the inclusion of the CR item into the common-item set showed minimal differences among the equating relationships when examinees performed similarly on MC and CR items. These findings suggest that systematic and random errors tend to be smaller when CR items are included in the common-item set. However, these results are contingent upon other factors such as the discrepancy in group proficiencies and the ratio of MC to CR items in the common-item set.

Comparable results were observed by S. Kim and Lee (2006) who investigated the consequences of the structure of the common-item sets on equating results amongst other factors in a simulation study. Four factors were manipulated: equivalent versus nonequivalent groups,

sample size, proportion of MC items to CR items relative to the total test, and the formation of the common-item sets (MC+CR, MC only, and CR only). The three-parameter logistic/graded partial credit (3PL/GPC) models were used for generating item and person parameters and item calibration was performed in MULTILOG. The difference between the predicted and observed characteristic curves was used as the evaluation criterion. Results were examined in terms of bias and mean square error (MSE). The smallest bias and MSE was obtained when linking was performed with a common-item set composed of both MC and CR items compared to using MC-only or CR-only common-item sets. The accuracy of the linking coefficients was also compared between MC-only common items and CR-only common items. Lower bias and MSE were found as a function of linking with the prepotent item format.

Distorted equating results can be obtained when an MC-only design is used for the analysis. Walker and Kim (2009) examined the use of MC-only common items for a mixed-format test. The researchers generated two pseudo-test forms, each with 16 MC and 8 CR items. Each test form had internal common items (8 MC, 4 CR) and only the MC items were considered as anchor items in the analysis. The dependent variables were root mean square difference (RMSD), bias, standard errors of equating (SEE), and RMSE. Using MC-only common items led to substantial bias in the accuracy of equating results. However, MC-only common items resulted in an adequate equating function as long as there was a high degree of association between the MC and composite scores across the test taking populations. These results further strengthen the belief that the inclusion of CR items into the common-item set may lead to more accurate equating results in terms of systematic and random errors.

The effects of common-item set composition on equating results also have been studied under various test structures. Tate (2000) conducted a simulation study in which various factors

were manipulated: simple test structure in which multidimensionality was due to an item format effect, proportion of MC items and CR items relative to the total test, length of common-item sets, types of anchors (MC+CR, and MC only), sample size, and group ability differences. The two- parameter logistic model/graded response model (2PL/GRM) combination was used to generate unidimensional person and item parameter estimates. A simple structure multidimensional model with two distinct yet correlated (0.6) latent factors was used to simulate a non-unidimensional test structure. When the assumption of unidimensionality was met, linking accuracy with the use of MC-only items led to adequate results. Erroneous results in terms of linking accuracy were observed when the test structure was multidimensional. However, this bias was mitigated when the proportions of MC and CR items relative to the total test were appropriate.

Kirkpatrick (2005) also investigated the effects of the structure of common item sets on mixed-format tests in both empirical (assumed unidimensionality) and simulation studies (multidimensionality). In the empirical study, data from a large scale testing program that spanned across various content areas and grade levels was utilized. Principal factor analysis was used to assess the underlying test structure. The unidimensionality assumption was met for the data. The 3PL/2PL/GPC model combination was used to generate the IRT item and ability parameters. Separate calibration with the Stocking-Lord method was utilized to place the two forms on a common scale. Results in terms of equivalent scores varied across content areas and grade levels irrespective of whether a CR item was included or excluded from the common-item set. The findings from the empirical study were used to conduct a simulation study to investigate the consequences of a simple multidimensional test structure on equating results. The test structure induced by item format was quantified by the correlation (0.5, 0.8, and 1.0) between

two latent factors. Discrepancies between group ability distributions also were simulated. A major finding was that different correlations between item formats had different impacts on equating results regardless of whether a CR item was included or excluded from the common item set. A strong correlation between item formats led to adequate equating accuracy. Equating accuracy deteriorated as a function of weaker associations between item formats. The inclusion of CR items into the common-item set only results in minimal differences among the equating relationships when factors are held constant across the various conditions.

Cao (2008) conducted a simulation study to investigate several attributes of the common-item set under various test structures in mixed-format tests and their resulting impact on equating accuracy. The effectiveness of concurrent calibration was examined under unidimensional and multidimensional test structures. Factors of interest were: test dimensionality structure (complex test structure, where multidimensionality was simulated by incorporating two distinct content areas and two item format effects); equivalent and nonequivalent groups; ratio of MC and CR items relative to the total test; representativeness of content area; and item difficulty in the common-item set. The criterion was the difference between the populations' observed and estimated total test scores. The dependent variables were bias, RMSE, and classification consistency. Accurate equating results were obtained under a multidimensional test structure as long as the item difficulty parameters in the common-item set were a representative sample of the total test. Equating error due to moderate or severe multidimensionality in terms of item format effects were mitigated by including both MC and CR items in the common-item sets and by ensuring that the proportion of MC and CR items were appropriate relative to the total test.

In summary, the use of MC-only common items can lead to substantial bias while common-item sets containing both MC and CR items tend to produce more adequate equating

results, in particular when the test structure is multidimensional or the group ability distributions differ across item formats. In addition, it was found that higher associations between MC and CR items and smaller group differences could yield less biased equating relationships. One should keep in mind that all of these studies evaluated equating accuracy in terms of systematic and random errors rather than the preservation of equity properties. Thus different results may be obtained when utilizing different evaluation criteria.

The inclusion of CR items into the assessment poses another challenge to the measurement community in that subjectivity in scoring these response items may lead to erroneous linking and equating outcomes (Tate, 1999). Tate (1999, 2000) proposed a procedure (trend scoring) to overcome the bias introduced by scoring CR items in the context of the CINEG framework. Trend scoring involves rescoring of the same examinee responses to CR items by the same pool of raters across administrations to eliminate the group ability difference. This allows for the identification of discrepancies in scored responses across administrations; thus rater effects can be extricated from the group ability differences. Tate (2003) and Kamata and Tate (2005) used simulation studies to demonstrate the adequate performance of the trend scoring method. Results of these studies indicated that the application of the proposed method produced adequate equating relationships and it led to appropriate recovery of the latent trait estimates. The application of traditional equating methods on the same anchor item set produced distorted equating relationships and inconsistent latent proficiency estimates.

Several other studies (S. Kim, Walker, & McHale, 2008, 2010; Tan, Kim, Paek, & Xiang, 2009; Wei & Yi, 2012) have examined the accuracy of the trend scoring method in equating. In general it was found that systematic and random errors were smallest when the designs included the rescoring of CR items. Further, the MC-only common item design and the trend scoring

method may perform similarly under a multidimensional test structure. Alternative methods for detecting scoring shifts in CR items without trend scoring also have been investigated. (Paek & Kim, 2007) used differential bundle functioning (DBF) methods to examine the CR score distributions across two administrations after matching on the MC items. It was assumed that when there is notable scoring shift, the DBF analysis would show shift from zero favoring one administration against the other. However, the results demonstrated that the DBF methods perform similarly in detecting scoring shift compared to the trend scoring method. Although it is apparent that rater effects should be accounted for when a CINEG design is used, the treatment of rater severity/leniency is not of interest in this study.

2.9.4 Examinee Characteristics

It is evident that the characteristics of the test and common item set impact the accuracy of the equating results for mixed-format assessments. However, these equating results appear to be dependent upon group ability distributions across test administrations. Several investigators (Bastari, 2000; Cao, 2008; S. Kim & Kolen, 2006; Kirkpatrick, 2005; W. Lee et al., 2010; Powers & Kolen, 2011; Von Davier et al., 2004; Wang et al., 2008; Wu et al., 2009) have examined the effects of group differences on equating accuracy amongst other factors. The results of these studies indicated that the equating results tended to be more accurate when the differences between the group ability distributions for examinees taking the reference and new test form were small.

Bastari (2000) conducted a simulation study to investigate the effects of various factors (test length, proportion of MC and CR items, length of common-item set, group ability distributions, item calibration methods, and sample size) on the linking accuracy for mixed-

format tests under the CINEG design. The 3PL/GRM model combination was used to generate and estimate item and person parameter estimates. Linking accuracy showed the least amount of bias as a function of various factors such as the use of the concurrent calibration method in combination with a longer test, larger proportion of MC items in the test, more common items, a larger sample size, and equivalent groups. Other researchers have suggested that as group differences increase, equating methods are affected differently in that the systematic equating error tends to increase for the frequency estimation method whereas CH methods and IRT methods appear to be unaffected by group differences (Powers et al., 2011; Sinharay & Holland, 2007).

Under the CINEG design it is not assumed that test taking populations are equivalent in proficiency distributions. Numerous studies have examined the effects of nonequivalence in the test taking populations on mixed-format test equating outcomes. However, most research has used random and systematic errors as evaluation criteria. Therefore, how the differences in ability distributions for the test taking groups affect the preservation of equity properties for mixed-format tests has not been assessed under various conditions. Further, the majority of studies examining the effects of nonequivalence in ability distributions within a between-item multidimensional test structure kept the mean differences constant across the various subdomains. This study is different in that it was assumed that mean ability differences can vary across subdomains such as multiple-choice and constructed-response domains. A summary of previous studies on mixed-format test equating is provided in Table 2.1.

Table 2-1 Summary of Previous Studies on Mixed-format Equating

Focus of Study	Authors
Extension of Linking Methods to Mixed-format Tests	(S. Kim & Lee, 2006)
Comparison of Linking Methods for Mixed-format Tests	(Bastari, 2000; S. Kim & Kolen, 2006; S. Kim & Lee, 2006)
Characteristics of Common-item set (Length, Format Representativeness)	(Bastari, 2000; Cao, 2008; S. Kim & Kolen, 2006; S. Kim & Lee, 2006; Kirkpatrick, 2005; Tan et al., 2009; Tate, 2000; Walker & Kim, 2009)
Effects of Group or Form Difference on Linking Accuracy	(Bastari, 2000; S. Kim & Kolen, 2006; S. Kim & Lee, 2006)
Methods to Account for Rater Effects	(Kamata & Tate, 2005; Paek & Kim, 2007; Tan et al., 2009; Tate, 1999, 2000, 2003; Wei & Yi, 2012)
Impact of Between-Item Multidimensionality on Linking Accuracy	(Béguin & Hanson, 2001; Béguin et al., 2000; Cao, 2008; S. Kim, 2004; S. Kim & Kolen, 2006; Tate, 2000; Wei & Yi, 2012)
Effects of Sample Size and Test Length on Linking Accuracy	(Bastari, 2000; Proctor, Reshetar, & Patel, 2012; Wu et al., 2009)

Most of the literature on mixed format test equating has focused on the accuracy of IRT scale linking which is the second step within the IRT equating framework. The last step in the equating process includes the scale-score transformation and the accuracy of equating results in terms of the complete IRT equating framework has not received much attention in the literature. The mainstream of psychometric literature has focused on the evaluation of equating accuracy in terms of systematic and random errors rather than equity properties. Therefore, this study employed equity properties to assess the equating accuracy of traditional and IRT equating methods for a mixed-format test when various conditions are manipulated under a CINEG design.

2.10 SUMMARY OF STUDIES ON EQUITY PROPERTIES

The use of equity properties for mixed-format tests using a CINEG design in the literature is scarce. A few studies were conducted on single format tests under a random groups design. Some recent studies have used equating properties to compare equating methods under the CINEG design for mixed-format tests. Two of these studies utilized empirical test data while one of the studies used both empirical test data and a simulation study. Findings of these studies will be presented next.

Tong and Kolen (2005) used empirical and simulated data to investigate the preservation of equity properties for different equating methods under a unidimensional test framework. The assessment consisted of MC-only items and the data collection method was a random groups design. The factors of interest included the performance of equipercentile equating methods versus IRT true score and observed score equating methods, and effects of form difficulty

differences on preservation of equity properties. Similar results were found for the IRT equating methods (true and observed score) and equipercentile methods when the magnitude between the differences in form difficulties were small. However, different results were observed when the forms differed in difficulty. Item response theory true score equating method preserved the FO equity property better compared to the other methods, while IRT observed score and equipercentile equating performed better in preserving the SO equity property.

Kim, Brennan, and Kolen (2005) also applied the equity properties to assess the accuracy of the equating results underlying an unidimensional IRT framework and the beta 4 true and observed score framework. The test contained MC-only items and the data collection design was a random groups design. Consistent results were observed when the equating method had the same underlying model assumptions as the framework used to assess the equity properties. For example, the preservation of the FO equity was more accurate for the true score equating method, whereas the SO equity was preserved more accurately for observed score equating. Further, unidimensional IRT observed score equating outperformed all equating methods in terms of the preservation of the SO equity property regardless of the underlying statistical framework that was used to assess the equity properties.

Bolt (1999) examined the performance of linear and equipercentile equating to IRT true score equating for a section preequating design under various test structures. The accuracy of the equating results was examined in terms of preservation of FO and SO properties. In this study, the SO equity property was defined as the mean squared error (MSE) associated with the equating function rather than a variance conditional on true scores. Data were generated by using a multidimensional IRT model (Reckase, 1985). Simulation conditions included: the interaction of item difficulty with the test structure and four correlation levels (0.3, 0.5, 0.7, and 1.0)

between the latent constructs. Overall, the IRT true score method was superior to the traditional equating methods when the association between abilities was high (>0.7). Even though the correlation was moderate to low (0.5), the IRT true score method was nearly as accurate as the equipercentile method in terms of equity performance.

E. Lee et al. (2012) used operational test data and pseudo-test forms to assess the performance of various equating methods in terms of FO and SO equity under a CINEG design for mixed-format assessments. The findings were consistent with previous research that was based on a random groups design that utilized MC-only assessments. The IRT true-score method was found to be superior in terms of preserving the FO equity than any other equating method, whereas IRT observed score equating was better in preserving the SO equity compared to the IRT true score equating method. Overall, the two IRT equating methods outperformed traditional equating methods in terms of FO equity preservation. However, all methods (traditional and IRT methods) performed similarly in preserving the SO equity.

He (2011) used real data to compare the performance of four equating methods on the preservation of equity properties for mixed-format tests. Equipercentile equating methods were compared to IRT equating methods. The factors of interest in this study included: proportion of common items, correlation between MC and CR scores, proportion of MC item score points, and the resemblance between alternate test forms. Results indicated that the IRT true score method was more successful in preserving FO equity compared to the IRT observed score method, whereas IRT observed score equating was superior in preserving the SO property. The chained equipercentile method was better in preserving the FO property compared to the frequency estimation method. However, both methods performed similiarly in preserving SO and same

distribution properties. A higher MC-CR correlation was found to be associated with more accurate preservation of the FO and SO properties for the IRT equating methods.

Andrews (2011) used both real and simulated data to examine different equating methods in terms of the preservation of equity properties for mixed-format tests. Equipercentile methods, IRT true score and observed score methods, and multidimensional IRT observed score equating methods were compared in this study. Equity properties were assessed using both a unidimensional and multidimensional IRT framework. The real data analysis indicated that the performance of the equating methods was dependent upon the framework that was used to assess the equity properties. In terms of equating methods, the chained equating method outperformed the frequency estimation method in terms of FO equity preservation, whereas the frequency estimation method was superior in preserving the SO equity property compared to the chained equating method. The multidimensional observed score method outperformed the other methods when the test structure became more complex ($\rho \leq .5$) and when the mean differences in ability distributions increased in magnitude (>.10). However, unidimensional equating methods still performed accurately under a multidimensional test structure when the correlation between the latent traits was fairly high ($\rho \geq .8$) supporting the supposition of robustness of the methods to violations of IRT assumptions.

2.11 SUMMARY OF FINDINGS

Research on equating has taken into account many factors including equating methods and attributes of the test, common items, and examinees. Although the majority of the research has been conducted on assessments containing only MC items, an emerging body of literature has

focused on mixed-format tests. A number of consistent results have been found across equating studies. In terms of equating methods, FR, CH and IRT methods can provide similar results when groups are similar in ability across administrations. IRT methods tend to outperform traditional methods when the assumption of (essential) unidimensionality has been met. The CH method is superior to the FR method when the discrepancy between group ability distributions is large.

The characteristics of the test also have been shown to impact the equating accuracy of mixed-format tests. The findings tend to be mixed in terms of the underlying dimensionality structure of mixed-format tests. Several researchers have suggested that dimensionality is contingent upon the domain of interest. Some investigators reported that MC and CR items measure nearly the same latent ability in the quantitative and reading comprehension domains. However, different latent abilities may be measured in writing assessments.

The characteristics of the common item set such as common item length, item position, and representativeness and non-representativeness also have been shown to influence the equating results. However, this was often contingent upon differences in the group ability distributions across test forms. It was found that a longer common item set leads to more accurate equating relationships. Items also should be administered in the same position across testing sessions as this is thought to minimize differential item functioning. The content and format representativeness of common items has a minimal impact on linking accuracy when the test structure is unidimensional. In terms of examinee characteristics, more accurate equating results are obtained when the differences between the group ability distributions are small across testing occasions.

There has been published research that has examined the impact of the common item composition (MC-only, CR-only, and MC+CR) on equating accuracy. The use of MC-only common items can lead to substantial bias while common items containing both MC and CR items tend to produce more accurate equating outcomes, especially when a test is multidimensional or group proficiency differs across item formats. Higher correlations between MC and CR items, higher MC to CR point ratios, and smaller group differences may result in less biased equating results.

Studies that utilized equity properties as evaluation criteria found similar results regardless of the format of the test (single format versus mixed format) and equating design (random groups versus CINEG design). For example, IRT observed score equating tends to preserve the SO property better than the IRT true score equating. Overall, IRT methods and traditional methods produced similar results in the context of SO preservation. In terms of the FO, both IRT methods produced similar results. However, overall the IRT equating methods were superior in preserving FO compared to traditional equating methods.

The application of equity properties for the evaluation of equating accuracy for mixed-format tests under a CINEG design is sparse. Therefore, this study contributes to the literature that pertains to the evaluation of equating accuracy for mixed-format tests when equity properties are applied under various conditions.

This study differs from previous research in the following aspects:

- Test Structure: The evaluation of equity properties has been predominantly assessed under a unidimensional test structure. Two studies (Andrews, 2011; Bolt, 1999) examined the preservation of equity properties under a multidimensional test structure. However, both studies assumed a correlated traits model. None of the studies under

review have utilized a three-parameter bifactor model with uncorrelated traits to generate response data for a mixed-format test within a horizontal equating framework. In addition, this study applied a unidimensional equity framework to complex multidimensional data.

- Common-item Composition: The few studies (Andrews, 2011; He, 2011) that investigated equity properties for mixed-format tests utilized format non-representative common-item sets by incorporating only MC items in the common-item set. Due to recommendations by Kolen and Brennan (2004), the common-item set should be representative of the total test; therefore, this study compared format representative and non-representative common-items sets in terms of equating accuracy.
- Group Ability Distributions: One study (Andrews, 2011) examined the effect of group ability differences (0.05, 0.10, and 0.30) on the preservation of equity properties for mixed-format tests. Various other studies have examined the effect of group ability differences on equating accuracy in terms of systematic bias and random error for mixed-format examinations within horizontal equating. These studies simulated nonequivalent groups by using mean differences of 0.5 and/or 1 respectively (Bastari, 2000; Cao, 2008; S. Kim & Kolen, 2006; S. Kim & Lee, 2006). Within the horizontal equating framework, mean group ability differences of 0.25 or higher are considered very large effect sizes (Wang et al., 2008). Therefore, this study examined group mean ability differences of 0.15 and 0.30 to reflect medium/large and very large effect sizes. In addition, group mean ability differences were allowed to vary on the subdomains within the multidimensional framework. Most simulation studies have kept the group mean ability differences constant across subdomains. Therefore it was of interest to examine the effect on

equating accuracy when group mean ability differences vary on the MC and CR subdomains. Based on cultural backgrounds and differences in educational systems it can be hypothesized that some examinee populations perform higher or lower on the MC subdomains in comparison to the CR subdomains (Hambleton, Merenda, & Spielberger, 2005).

3.0 METHODS

The primary purpose of this study was to examine the performance of four commonly used equating methods under a common-item nonequivalent groups design (CINEG) design with various test structures, compositions of common-item set, and differences in ability distributions in regards to First-Order (FO), Second-Order (SO), and same distribution equity criteria. In order to fulfill this purpose, a simulation study was conducted. An advantage of conducting a simulation study is that it affords the opportunity to assess the effects of the factors under investigation and it provides true population values that can aid in the evaluation of the results. This chapter outlines the methodological framework for the simulation study. First, an introduction to the test design is presented, which is followed by a description of the factors of interest. A detailed outline of the data generation and validation is illustrated in the next section. The chapter ends with a description of the evaluation criteria used to judge the results.

3.1 TEST DESIGN AND FIXED FACTORS

The mixed-format test simulated in this study may reflect one potential test configuration for a statewide assessment. Two test forms (X and Y) for equating were considered. Each test form was composed of a unique item set that is specific to that particular test form and a common-item set that is the same for both test forms. Form X and Form Y contain a mixture of MC and CR

items. In this study, two test structures were examined: unidimensionality and within-item multidimensionality. In the unidimensional model it is assumed that one general construct accounts for a preponderance of the common variance. The within-item multidimensionality can be modeled through a bifactor model (Gibbons & Hedeker, 1992). Here it was presumed that two subdomains may be present which is defined by two latent abilities, one loading on MC items and one loading on CR items. In addition, it was assumed that one dominant latent construct (e.g. math knowledge) loaded on all items in the assessment. Further, it was assumed that the dominant dimension and the subdomains are orthogonal and that the subdomains are orthogonal to each to each other. The subdomains are assumed to capture the item covariation that is independent of the covariation due to the dominant latent construct.

3.1.1 Fixed Factors

Equating accuracy is influenced by numerous factors as outlined in Chapter Two. In this study some factors were held constant to reduce the number of conditions and to keep the sources of random variability at a minimum. Test length, number of common-items, examinee sample size, number of dimensions, and item parameters for the bifactor model were fixed based on the review of literature and common test practice.

3.1.1.1 Test Length

One important consideration when developing a test is the test length which has been shown to impact the accuracy of IRT parameter estimation and equating outcomes. Test length is typically defined in terms of the number of items and the number of score points. In general, it has been found that a longer test in terms of the number of items and number of score points may result in

more accurate equating outcomes (Bastari, 2000; Cao, 2008; Kolen & Brennan, 2004). Each test form in this study was comprised of 50 items, 45 MC items (0/1) and 5 (0/1/2/3/4) CR items. The ratio of the MC and CR items in the total test was 10:1 on both forms and 2.25:1 in terms of score points, which reflects ratio settings that can be found in practice in state assessment programs (Cao, 2008). It was assumed that this test length is large enough to ensure accurate item parameter estimation and adequate equating accuracy.

3.1.1.2 Number of Common-items

The length of the anchor item also impacts the equating results when a CINEG design is employed. Kolen and Brennan (2004) recommend that the common-item set should be approximately 20% of the length of the total test based on an assessment that consists of at least 40 items. Several researchers have investigated the length of the common-item set and reasonable results were obtained when the common-item set was about 20% of the length of the total test (Bastari, 2000; Cao, 2008; S. Kim & Kolen, 2006). Based on this previous research, the common-item set in this study also accounted for about 20% of the total test length.

3.1.1.3 Sample Size

There are two test forms to be equated, Form X and Form Y, one associated with Group 1 and the other associated with group 2. The groups are randomly drawn from two populations. A total of 3000 examinees were drawn from each of the two populations. This sample size is thought to be large enough to ensure stable estimation of the equating methods under investigation (Han, Kolen, & Pohlmann, 1997; Hanson & Béguin, 2002). In addition this sample size is also within the range that was utilized by other studies (Bastari, 2000; Cao, 2008; Kirkpatrick, 2005).

3.1.1.4 Number of Dimensions

Under the unidimensional condition it is assumed that there is one underlying construct accounting for the item responses. For the bifactor model, there is one general construct and two subdomains as introduced by the multiple-choice and constructed-response item formats. In the bifactor model every item loads on one subdomain only in addition to the general construct; therefore, no matter how many subdomains exist, the number of integrals for any bifactor models is always two (Li, 2011).

3.1.1.5 Bifactor Subdomain Item Parameters

In the bifactor data generation model (Gibbons & Hedeker, 1992), the general dimension and the MC and CR subdomains are orthogonal and the subdomains are orthogonal to each other. The relative importance of the general construct and the MC and CR subdimensions on examinees' item responses is manipulated by varying the magnitude of the item discrimination parameters. In this study it is assumed that the general construct has the main influence on item responses since the main purpose of most educational assessments is to measure the general construct of interest. This is in alignment with studies that applied the bifactor model within the context of educational measurement (DeMars, 2006; Li, 2011). Further, it should be noted that a condition in which the general construct is less informative than the subdimensions is not typical within educational settings (Li & Rupp, 2011). Therefore, MC and CR group factors are considered to provide less discriminating power compared to the general latent construct. One condition was considered:

- MC and CR group factors provide similar discriminating power

This fixed condition was chosen because empirical test data from large scale math assessments indicate that item discrimination parameters for CR items are similar compared to MC item discrimination parameters (e.g. Pennsylvania System of School Assessment, New York State Testing Program). In addition, this follows the procedure of other simulation studies that examined mixed-format tests within a horizontal equating context.

3.2 FACTORS UNDER INVESTIGATION

For the undimensional case, 4 equating methods, 3 common-item compositions, and 3 mean differences of group ability were examined (4x3x3 simulation conditions). Under the multidimensional test structure, 4 equating methods, 3 common-item compositions, and 5 differences of group ability were considered (4x3x5 simulation conditions).

Table 3-1 Overview of Varied Factors

Factor	Unidimensional Level	Multidimensional Level
Equating Methods	1. Pre-smoothed Equipercentile (FR) 2. Chained Equipercentile (CH) 3. IRT True Score (TR) 4. IRT Observed Score (OB)	1. Pre-smoothed Equipercentile (FR) 2. Chained Equipercentile (CH) 3. IRT True Score (TR) 4. IRT Observed Score (OB)
Common-item Composition	1. 10-MC only 2. 9MC+1CR 3. 8MC+2CR	1. 10-MC only 2. 9MC+1CR 3. 8MC+2CR
Mean Difference of Group Ability	1. μ=[0] 2. μ=[0.15] 3. μ=[.30]	1. μ=[0,0,0] 2. μ=[0.15,0.15,0.15] 3. μ=[0.30,0.30,0.30] 4. μ=[0.30,0.15,0.30] 5. μ=[0.30,0.30,0.15]

3.2.1 Equating Methods (4 levels)

Several linear and non-linear equating procedures have been well documented in the literature. As mentioned by several researchers (Brennan, 2010; Duong, 2011; Hagge et al., 2011; He, 2011), linear methods such as the Tucker and Levine methods are not often utilized by testing programs and therefore this study focused on methods that are more widely implemented in practice.

These methods include:

- Presmoothed Frequency Estimation (FR)
- Chained Equipercentile (CH)
- IRT True Score Equating (TR)
- IRT Observed Score Equating (OB)

The presmoothed frequency estimation and the chained equipercentile methods establish an equating function in that the equated scores of the new form (Form Y) have the same distribution

as the raw scores of the reference form (Form X) assuming the same group of examinees. More specifically, when the FR method is applied it is assumed that the distribution of the total score which is conditional on the score of the common-item set is equivalent for Group 1 and Group 2 for each of the two forms. Thus, the score distributions for Form X and Form Y for the same group of examinees can be obtained, and as a result the equipercentile relationship between the two distributions of the same group of examinees can be determined. The CH method begins with the equating of the total score to the common-item score for Form Y by estimating equipercentile equivalents. Then the common-item score is equated to the total score for Form X by estimation of equipercentile equivalents.

Both TR and OB methods entail numerous steps including item parameter estimation, scale transformation, and equating. Detailed steps of these procedures were presented in Chapter 2. After items have been calibrated and the item parameters from Form X and Form Y have been placed onto a common scale the IRT true score equating can be conducted by relating the true score on Form Y to its corresponding latent trait estimate, and then the latent trait estimate can be related to the true score on Form X. IRT observed score equating is conducted by establishing the observed score distribution for Form X and Form Y by utilizing the estimated item parameters and latent trait estimates. The observed score distributions are then used to establish the equipercentile relationship between Form Y and Form X.

All equating methods were applied as outlined in Chapter 2. Presmoothing was conducted before the actual equating for the FR and CH methods.

3.2.2 Test Dimensionality Structure (2 levels)

3.2.2.1 Unidimensionality

The test structure is assumed to be unidimensional in that the test is truly measuring an examinee's general latent ability as one dominant dimension. Figure 3.1 displays this unidimensional test structure.

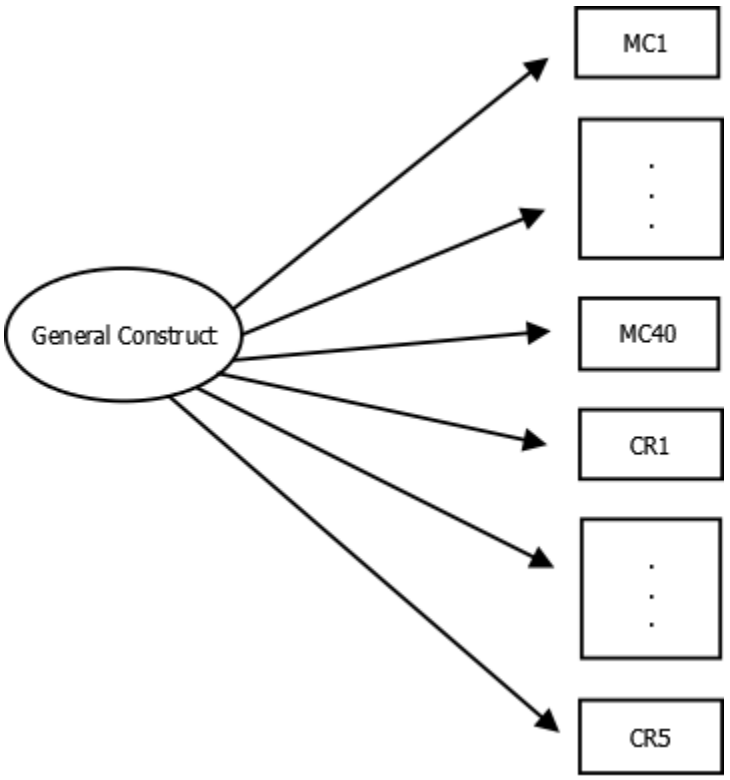


Figure 3.1 Example of Unidimensional Model

3.2.2.2 Multidimensionality

Mixed-format examinations appear to be more susceptible to underlying complex test structures due to the inclusion of different item formats. Different item formats may measure different constructs which could cause a multidimensional test structure. For example, a common

conception is that CR items may assess knowledge or skills that cannot be adequately captured by MC items. Therefore, the different formats may measure different levels of cognitive processing (Lane & Stone, 2006). In this study, a bifactor model was used to generate data with a complex within-item multidimensional test structure. It was assumed that one general construct loads on all items, while the subdomains, which capture the potential different skills assessed by MC and CR items, load on a subset of items. Figure 3.2 illustrates this within item multidimensional test structure.

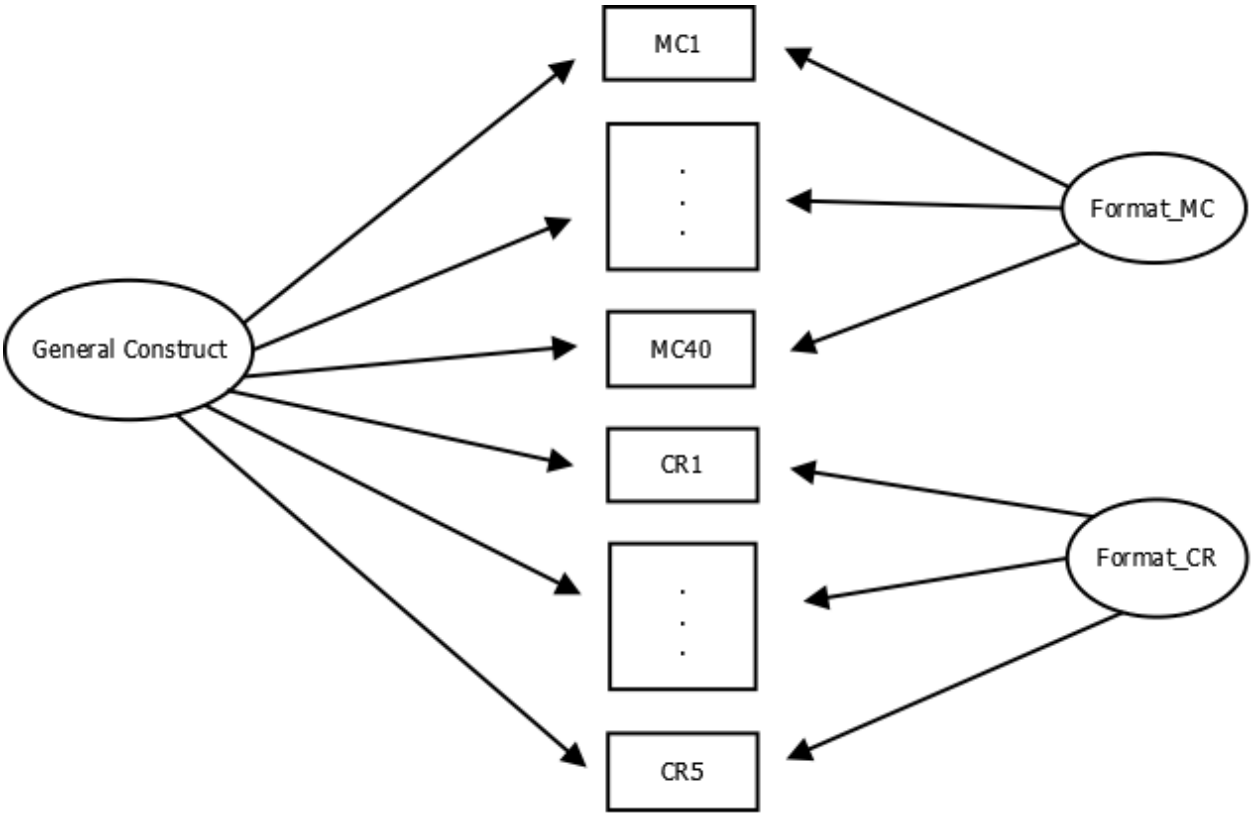


Figure 3.2 Example of a Bifactor Model

3.2.3 Common-item Composition (3 levels)

The common-item composition can be referred to as format representativeness and format non-representativeness. Under format representativeness, the proportion of MC items to CR items in the common-item set should correspond to the ratio of MC to CR items in the total test in terms of the item number and score points.

3.2.3.1 Format Representativeness

The ratio of MC items to CR items in the total test was 10:1 in terms of the number of items and 2.25:1 in terms of the number of score points. Similar ratios were reflected in the common-item set, which results in 9 MC items and 1 CR item and 8 MC and 2 CR items. This could be of interest since stakeholders are concerned about maintaining test security; thus, the inclusion of only one CR item may lead to adequate equating outcomes. Further, it may minimize the systematic variability as introduced by rater effects.

3.2.3.2 Format Non-representativeness

One situation was considered in the format non-representativeness condition. In this condition, the common-item set consists of only 10 MC items. This scenario has been applied to advanced placement examinations (e.g. College Board). However, the adequacy of the results has been questionable in particular when the test structure is complex (Cao, 2008).

3.2.4 Group Ability Distributions (8 levels)

Two groups were randomly drawn from the population of examinees, which were referred to as Group 1 and Group 2. Examinees in Group 1 are associated with Form X, whereas examinees in Group 2 are associated with Form Y. Ability parameters of examinees in Group 2 were placed on the scale of Group 1 in the equating process; therefore, the ability distribution of Group 1 is fixed with a mean of 0 and a standard deviation of 1. These values were drawn from a standard normal distribution. The ability of Group 2 could be manipulated by either varying the mean and/or the standard deviation. However, only mean differences were simulated while the standard deviation was fixed at 1. This is in alignment with other simulation studies that have examined ability group differences in a horizontal equating framework (Bastari, 2000; Cao, 2008; S. Kim & Kolen, 2006; Kirkpatrick, 2005).

3.2.4.1 Equivalent Groups

It is assumed that Group 1 and Group 2 are equal in ability; thus, the two groups are both drawn from a standard normal distribution with mean of 0 and a standard deviation of 1. However, it is important to note that the CINEG design does not assume that the two groups are equal in ability.

3.2.4.2 Non-equivalent Groups

Examinees in Group 2 were assumed to be more competent than the test takers in Group 1. Two conditions were examined. The ability distribution for Group 2 was modeled with a higher population mean of 0.15, and 0.30, with a standard deviation of 1. Additional levels were considered within the multidimensional framework in that population mean differences varied among the two subdomains.

These values were chosen because they were considered to reflect large (0.15) and very large (0.30) effect sizes within the context of horizontal test equating (Wang et al., 2008). Similar mean differences have been utilized in other test equating simulation studies (Andrews, 2011; Kirkpatrick, 2005). Wang et al. (2008) noted that even group mean differences of .05 to 0.1 may be considered large for test equating purposes.

3.3 DATA GENERATION

Examinees in Group 1 were associated with Form X and examinees in Group 2 are associated with Form Y. The resulting item responses were generated separately. More specifically, the unique item set for Form X, the unique item set for Form Y, and the anchor item set were generated separately.

The 3PL/GR model combination was used to simulate 3000 examinees' responses for each group under each design condition assuming unidimensionality. The multidimensional alternative to the unidimensional 3PL model was utilized to generate 3000 examinees' responses for each group under each design condition when a complex test structure was assumed. The data generation process is summarized in three steps as follows:

3.3.1 Step 1: Ability Parameter Generation

Considering the factors of interest in this study, the test structure and the group ability differences were taken into account when generating the examinee proficiency parameters for Groups 1 and 2.

3.3.1.1 Unidimensional Test Structure

Under the unidimensional condition, one latent construct (θ) affects the test takers' responses to all items without considering the influence of different item formats; therefore, the examinees proficiency parameters (θ) in Group 1 were randomly drawn from a standard normal distribution ($N(0,1)$). The examinee proficiency parameters in Group 2 were randomly drawn from a normal distribution ($N(0,1)$) for an equivalent group, and $N(0.15,1)$, and $N(0.30,1)$ for nonequivalent groups.

3.3.1.2 Multidimensional Test Structure

Under the multidimensional condition, the examinees proficiency parameters ($\theta_{General}, \theta_{\mathrm{MC,}}$ and θ_{CR}) were randomly drawn from a multivariate normal distribution with a pre-specified mean and variance-covariance matrix (see Table 3.2). The difference between the group ability distributions was once again taken into consideration in that the mean difference between the equivalent Groups 1 and 2 was 0, whereas the mean difference between the nonequivalent groups was 0.15, and 0.30, respectively. These mean differences were reflected for the general construct as well as the MC and CR subdomains. Two additional scenarios were considered in that the ability distributions on the MC and CR subdomains vary from the general construct. More specifically, it was assumed that group mean ability distributions for the MC subdomain were either lower or higher than the general construct and the CR subdomain.

Table 3-2 Generating Ability under Complex Test Structure

	Group 1	Group 2-1	Group 2-2	Group 2-3	Group 2-4	Group 2-5
Bifactor θ_{General} θ_{MC} θ_{CR}	$\mu = \begin{bmatrix}0\\0\\0\end{bmatrix}$ $\sum = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$	$\mu = \begin{bmatrix}0\\0\\0\end{bmatrix}$ $\sum = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$	$\mu = \begin{bmatrix}.30\\.30\\.30\end{bmatrix}$ $\sum = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$	$\mu = \begin{bmatrix}.10\\.10\\.10\end{bmatrix}$ $\sum = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$	$\mu = \begin{bmatrix}.30\\.10\\.30\end{bmatrix}$ $\sum = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$	$\mu = \begin{bmatrix}.30\\.30\\.10\end{bmatrix}$ $\sum = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$

3.3.2 Step 2: Item Parameter Estimation

Test forms X and Y were generated for equating. Form X consisted of a unique item set and a set of anchor items. Form Y also consisted of a set of unique items that were specific to Form Y and the set of anchor items. The anchor items in forms X and Y were identical. The unique item set pertaining to Form X, the unique item set that pertains to Form Y, and the anchor item set were generated in separate runs. The item parameter generation process was influenced by all factors under investigation.

3.3.2.1 Unidimensional Test Structure

The 3PL/GR model combination was used for item parameter calibration. The 3PL model for MC items can be written as follows

$$P\left(x_{ij}=1\middle|\theta_j,a_i,b_i,c_i\right)= c_i+\frac{1-c_i}{1+exp^{-Da_i\left(\theta_j-b_i\right)}}, \tag{3.1}$$

where θ corresponds to the person parameter, D is a scaling constant (1.7), a_i is the item discrimination parameter, whereas b_i represents the item difficulty parameter, and c_i is the lower asymptote or guessing parameter.

The MC items in the unique item set for Form X were generated using the following steps: The item discrimination parameters (*a*) are typically generated from either a uniform distribution (*U*(0.2,2.0)) or a lognormal distribution with a mean of 0 and different levels of variances (Desa, 2012). The default setting in BILOG is *LN*(0,0.5). However, in this study the discrimination parameters were generated from a uniform distribution, which has been used by several researchers (DeMars, 2006; Desa, 2012). The discrimination parameter was specified to be *U*(0.7,1.3) which reflects average item discrimination parameters that are commonly observed in many educational and psychological measures. The mean of this distribution is 1. The item difficulty parameters were sampled from a normal distribution with a mean of 0 and standard deviation of 1 (*N*(0,1)). The range for the item difficulty parameters were set from -2 to 2 to ensure that low and high ability levels were modeled. These values were also used by various researchers (Cao, 2008; Finch, 2006). The guessing parameters for the multiple-choice items were sampled from a beta distribution with α = 8 and β = 32 (*M* = .2, *SD* = .062). These values were chosen such that the mean of the beta distribution was approximately equal to the desired probability of a correct response assuming dichotomous items with five response choices. These values were also utilized by other researchers (S. Kim & Lee, 2006).

A similar procedure was used for the CR items. The unidimensional GR can be expressed as

$$P_{ijk}^{*} = \frac{\exp[Da_i(\theta_j - b_{ik})]}{1+\exp[Da_i(\theta_j - b_{ik})]}, \tag{3.2}$$

where θ_j is the latent trait for examinees *j*, a_i is the discrimination parameter of item *i* ,b_{ik} is the threshold parameter for category *k* of item *i*, and *D* is the scaling constant (1.7). The value of the

between category threshold parameter represents the point on the θ_j continuum where examinees have a 50% chance of responding in or above category *k*.

The CR items in the unique item set for Form X were generated using the following steps. The item slope parameters were generated from a *U*(0.7,1.3), which is the same as was used for the 3PL model. The CR items were fixed in terms of the number of categories. In this study, all CR items were assumed to have five categories; therefore, four category threshold parameters were sampled from *N*(-1.5,0.2), *N*(-0.5,0.2), *N*(0.5,0.2), *N*(1.5,0.2). It should be noted that the between category threshold parameters in GRM are always in the ascending order in that $b_{i1} < b_{i2} < b_{i3} < b_{i4}$. Such configurations of the threshold parameters were intended to reflect not only a wide range of proficiency levels (-2 to 2) covered by the items but also variability of the thresholds across items. These values were also used in previous simulation studies (Cao, 2008; S. Kim & Lee, 2006). The same item parameter distributions were used to generate the item parameters for the common-item set.

A similar process was followed to generate the item parameters for the unique items for Form Y. The item discrimination parameters were sampled from the same item parameter distributions as the unique items for Form X.

3.3.2.2 Multidimensional Test Structure

Under the multidimensional test structure an examinee's response to each item was determined by a general construct and one of the specific format factors. Examinee's responses to MC items under a complex structure can be modeled by a three-parameter bifactor model

$$P\left(X_i = 1 \middle| \theta_j, a_i, c_i, d_i\right) = c_i + (1 - c_i)\frac{\exp[D(a_{iG}\theta_G + a_{is}\theta_s + d_i)]}{1+\exp[D(a_{iG}\theta_G + a_{is}\theta_s + d_i)]}, \quad (3.3)$$

where θ_G represents the general or dominant proficiency, while θ_s (s=1,2,…,*k*) represents one of the *k* subdomains that are orthogonal to each other as well as to the general or dominant proficiency. Additionally, a_{iG} and a_{is} are item discrimination parameters for the general factor and one of the *k* subdomains. Lastly, d_i can be conceptualized as a scalar parameter that is related to an overall multidimensional item difficulty as found in the typical MIRT model and *D* (1.7) is the scaling constant.

The item discrimination and the unidimensional-like difficulty parameters for the unique item set in Form X were generated first. Similar to procedures for unidimensional models, the item discrimination parameters for the general construct were sampled from a uniform distribution (*U*(0.7-1.3)). Similar values were also used by (Li, 2011). Within the factor analytic or structural equation modeling framework these parameters would correspond to fairly high factor loadings that range from about 0.6 to 0.8. The discrimination parameters for the MC and CR group factors were sampled from *U*(0.6,0.9) and corresponding factor loadings for these parameters range from 0.35-0.50. The item difficulty was sampled from a standard normal distribution with the range of -2.0 to 2.0 to reflect a range of difficulty levels as found in practice. The scalar d_i from Equation (3.3) can be calculated as follows

$$d_i = -b_i\sqrt{a_{General}^2 + a_{ij}^2} \tag{3.4}$$

using the b_i parameter and the discrimination parameters from the general dimension and one of the format factors such as MC or CR format factors. Similar to the unidimensional condition the guessing parameters were sampled from a beta distribution with α = 8 and β = 32.

Meanwhile, data for the CR items were generated based on the bifactor graded response model

$$P\left(X_{ij} \geq 1|\theta_G, \theta_S\right) = \frac{1}{1+\exp\{-[d_1+a_G\theta_G+a_S\theta_S]\}}, \quad (3.5)$$

$$\vdots$$

$$P\left(X_{ij} \geq K-1|\theta_G, \theta_S\right) = \frac{1}{1+\exp\{-[d_{K-1}+a_G\theta_G+a_S\theta_S]\}},$$

where $d_1,\ldots,d_{K-1}$ are strictly ordered intercepts that are related to the MIRT item difficulty parameters, a_G θ_G are the item discrimination and proficiency estimates for the general factor, whereas a_S θ_S reflect the item discrimination and proficiency estimates for the specific factor.

The same item parameter generation process was applied to CR items in the unique item set for Form X. More specifically, the item discrimination and difficulty parameters were generated first. The item discrimination parameter for the general construct follows a uniform distribution (*U*(0.7-1.3)). All the CR items are fixed to have five categories. The thresholds were sequentially drawn from *N*(-1.5,0.2), *N*(-0.5,0.2), *N*(0.5,0.2) and *N*(1.5,0.2). A similar process was applied to generate the item parameters for the unique items for Form Y. The item discrimination parameter and the unidimensional-like item difficulty parameters were sampled from the same item parameter distributions as the unique items for Form X.

3.3.3 Step 3: Response Data Generation

The item parameters for Form X and Form Y and the ability parameters for each group was used to generate appropriate correct response probabilities using applicable IRT models. The values of the correct response probabilities were compared to the values of the uniform random number which were in the range (0,1) to assign the item responses. Different IRT models were used to generate the item responses under the various test structures.

3.3.3.1 Unidimensional Test Structure

The unidimensional 3PL model expressed in Equation (3.1) was utilized to compute the likelihood that an examinee with latent construct (θ) successfully responds to a MC item, which can be denoted as $P_i(\theta)$. The value of the $P_i(\theta)$ then is compared to a value of the uniform random number (U_i) to generate a dichotomous item response of an examinee with latent ability (θ)to MC item i (R_i) given the following rule:

$$R_i = \begin{cases} 0, & P_i(\theta) \leq U_i \\ 1, & P_i(\theta) > U_i \end{cases} \tag{3.6}$$

The GRM requires an additional step. In the first step, operating characteristic curves $P_{ij}^*(\theta)$ are estimated, which stand for the conditional probability of an examinee's response falling in or above a given item category. Following the estimation of $P_{ij}^*(\theta)$, the actual category response curves must be computed. This is done by using the following formula:

$$P_{ij}(\theta) = P_{ij}^*(\theta) - P_{i,j+1}^*(\theta). \tag{3.7}$$

Then the value of the $P_{ij}^*(\theta)$ can be compared to a uniform random number (U_i) to generate a polytomous item response of an examinee with latent construct (θ) to CR item i (R_i) based on the following rule for a CR item with 5 response categories:

$$R_i = \begin{cases} 0, & P_{i2}^*(\theta) \leq U_i < 1 \\ 1, & P_{i3}^*(\theta) \leq U_i < P_{i2}^*(\theta) \\ 2, & P_{i4}^*(\theta) \leq U_i < P_{i3}^*(\theta) \\ 3, & P_{i5}^*(\theta) \leq U_i < P_{i4}^*(\theta) \\ 4, & 0 \leq U_i < P_{i5}^*(\theta). \end{cases} \tag{3.8}$$

3.3.3.2 Multidimensional Test Structure

A three parameter bifactor compensatory MIRT model was used to generate examinee response data. The compensatory nature of the model allows an examinee's high proficiency on one trait to potentially compensate for a lower proficiency on another factor. The three parameter bifactor

model as found in Equation (3.3) was applied to compute the likelihood that an examinee with multiple latent constructs successfully performs on MC item i $(P_i(\theta))$. The value of $P_i(\theta)$ then was compared to a value of the uniform random number (U_i) to generate a dichotomous item response to MC item i (R_i) based on the same rule as described under the unidimensional condition.

Similarly, the bifactor graded response model as expressed in Equation (3.5) was used to compute the operating characteristic curves $(P_{ij}^*(\theta))$, and Equation (3.7) was used to compute the actual category response curves $(P_{ij}(\theta))$. The value of the $P_{ij}^*(\theta)$ was compared to a value of the uniform random number (U_i) to generate a polytomous item response to CR item i (R_i) based on the same rule as demonstrated under the unidimensional condition.

3.4 DATA VALIDATION

The process of data validation affords the opportunity to examine the adequacy of the simulated data based on the intended criteria of the factors of investigation. This process is of importance since the purpose is to prevent the analysis of distorted results. A few random datasets from various conditions were drawn to perform the data validation. Means and variances were examined in order to verify the adequacy of the generated latent proficiencies. The simulated item responses were assessed in terms of the difference between the observed and model-based proportions of examinees' responses. A small difference indicated that the item response generation was suitable. In addition, the structure of the simulated data was assessed to ensure that the intended model structures have been met. Exploratory factor analysis (EFA) was conducted using the mean and variance adjusted weighted least squares estimation (WLSMV)

method in MPlus (Muthén & Muthén) to establish whether the intended model structures were simulated appropriately under the unidimensional condition. Model fit indices (SRMR) and factor loadings were examined to verify the adequate number of dimensions. Mplus was used to validate the appropriateness of the multidimensional test structure by estimation of a bifactor model.

3.5 EQUATING

Once the item response data for Group 1 taking Form X and for Group 2 taking Form Y were generated as outlined in the steps above, the two sets of parameter estimates were placed on a common scale through the common item set. In this study, concurrent calibration was utilized to fulfill this purpose. Concurrent calibration involved numerous steps. First, response data from both test taking populations was combined by treating items not taken by a particular group as missing or not reached. In addition, the group membership was specified before proceeding to the next step. Secondly, it was determined which form served as the reference form. In this study, Form X was the old or reference Form; thus, Form Y was placed onto Form X's scale by using the common-item set. Finally, the unidimensional 3PL model was used to calibrate MC items, whereas the unidimensional GR model was used to calibrate the CR items. PARSCALE (Muraki & Bock, 1993) was utilized to estimate the parameters on Form X and Form Y simultaneously. Several estimation methods are available in Parscale to estimate the item and ability parameters. In this study, Marginal Maximum Likelihood (MML) was used to estimate the item parameters, whereas Maximum Likelihood Estimation (MLE) was utilized to estimate the ability parameters. After parameter estimates were placed onto a common scale, the equating

was performed by utilizing the equipercentile and IRT methods as outlined in Chapter 2. The function "equate" (Albano, 2010) within the R software environment was used to conduct the non-IRT equating, while the POLYEQUATE (Kolen, 2004b) package was used to perform IRT true score and observed score equating.

3.6 REPLICATIONS

There were 4x3x3 (equating method x common-item composition x group mean ability differences) simulation conditions in the unidimensional case. In addition there were 4x3x5 (equating method x common-item composition x group mean ability differences) simulation cells under the multidimensional test structure. The data generation and equating process for each cell in the design was replicated 100 times.

3.7 EVALUATION CRITERIA

First-order equity holds if an examinee obtains the same equated score on Form Y when compared to Form X. Second-order equity holds if the conditional SEM of the equated score on Form Y is the same as compared to the conditional SEM on Form X. In order to examine how well the equating methods preserve the equity properties two different indices were calculated. These indices were used by several researchers (Andrews, 2011; He, 2011; Tong & Kolen, 2005). Tong and Kolen (2005) defined index D_1 as

$$D_1 = \sqrt{\sum_i q_i \left\{\frac{E[X|\theta_i] - E[e\hat{q}_x(y)|\theta_i]}{SD_x}\right\}^2}, \quad (3.9)$$

where $e\hat{q}_x(y)|\theta$ represents the expected raw score on Form Y after equating, $E[X|\theta]$ stands for the expected raw score on Form X, q is the density at a latent construct of θ, and SD_Y represents the standard deviation of scale scores for Form X. The summation is taken over all quadrature points. Large differences between expected scale scores over all score points will increase the magnitude in D_1. An increased magnitude in D_1 indicates that the FO equity is less preserved.

A similar index called D_2 was used to examine the preservation of SO property:

$$D_2 = \sqrt{\sum_i \frac{q_i(SEM_x|\theta_i - SEM_{eq_x(y)}|\theta_i)^2}{SD_x}}, \tag{3.10}$$

where $SEM_x|\theta$ is the raw score CSEM for the reference form and $SEM_{eq_x(y)}|\theta$ is the raw score CSEM for the new form after equating. SD_X represents the standard deviation of scores on Form X. Similar to the D_1 index, larger values in D_2 imply that SO is not preserved as accurate.

The same distributions property examines whether the distribution of scores on the reference form is the same compared to the distribution of scores on Form Y for the same group of examinees. In order to examine how well the same distributions property has been preserved, the Kolmogorov-Smirnov *T* (Conover, 1999) was obtained, which is a nonparametric statistic that allows for the comparison of two score distributions. The Kolmogorov-Smirnov *T* was calculated as the discrepancy between the empirical distribution function (*edf*) of the equated scores on Form Y ($G_1[eq_X(y)]$) and the *edf* of the scores on Form X ($F_1[x]$) for the synthetic group.

$$T = \sup_{eq_X(y)=x} |G_1[eq_X(y)] - F_1[x]| \tag{3.11}$$

It should be noted that both G_1(y) and F_1(x) are obtained from the marginal distributions of g_1(y) and f_1(x) respectively. The computer program POLYCSEM (Kolen, 2004a) was used to

obtain the conditional expected scores and standard errors of measurement. For each of the 100 replications there were D_1, D_2, and T statistics. The performance of the equating methods were evaluated based on the mean of the D_1, D_2, and T values over 100 replications. Meaningful differences were examined in terms of mixed ANOVA's.

4.0 RESULTS

The purpose of the current study was to examine which equating method performed best in preserving the same distribution and equity properties under various attributes of the test, examinees, and common-item set. It should be noted that under the multidimensional condition, a mixed bifactor IRT model was used to generate the response data, which was then applied to the undimensional equity framework as described in Chapter 2. In order to examine whether there were any meaningful differences among the methods, a mixed ANOVA was performed for each evaluation index. In this design, equating methods were treated as the within-subject effect because each method was applied to the same set of data in each replication whereas the group and anchor conditions were treated as between subject-effects. Results for each evaluation criterion are presented in the following order: mixed ANOVA results; follow-up analyses; and comparisons among IRT methods, traditional equating methods and across all equating methods. It should be noted that it was not of interest to compare across all methods due to the potential advantage given to the IRT equating methods given that the data generation, item calibration and equity framework were based on the same underlying psychometric model; however, the comparisons across all methods were included for the sake of completeness. General guidelines of effect size in ANOVA (partial η^2: small: 0.01; medium: 0.06; large: 0.14) as suggested by J. Cohen (1988) were used as cut-off values and only effects (particularly for interaction effects) with moderate or large effect size were further analyzed in the current study.

When the common-item set was composed of multiple-choice items only, the PARSCALE calibrations converged successfully for all conditions in this study. However, when the anchor set included constructed-response items, the calibration runs did not converge successfully for some conditions. Lack of convergence occurred most often for the condition with the anchor set consisting of 8 multiple-choice + 2 constructed-response items and large mean group differences under the unidimensional and multidimensional conditions. Under the unidimensional test structure, 12 out of 100 data sets were not successfully calibrated (common-item composition: 8MC+2CR, group mean difference: $\mu = .30$) whereas 16 out of 100 data sets did not converge successfully under the multidimensional test structure (common-item composition: 8MC+2CR, group mean difference: $\mu = .30, \mu = .30, \mu = .30$).

4.1 FIRST-ORDER EQUITY

First-order equity assumes the equivalence of expected values of conditional distributions, which implies that conditioning on ability, the distribution of the reference form and the equated form are comparable. Similar to previous research studies examining the preservation of equity properties (Andrews, 2011; He, 2011; Tong & Kolen, 2005), a D_1 index was used to evaluate the first-order equity property. Overall, lower D_1 values indicate better preservation of first-order equity. Table 4-1 shows the mean D_1 values over 100 replications for each condition under the unidimensional test structure.

Table 4-1 D_1 Values over 100 Replications

Group	Anchor	FR Mean	FR SD	CH Mean	CH SD	TR Mean	TR SD	OB Mean	OB SD
$\mu = 0$	10MC	.806	.920	.718	.892	.067	.053	.332	.227
	9MC+1CR	.726	.734	.656	.778	.074	.124	.307	.216
	8MC+2CR	.814	1.014	.702	1.010	.054	.043	.286	.221
$\mu = 0.15$	10MC	2.157	.967	.772	.861	.083	.113	.311	.231
	9MC+1CR	1.985	.771	.541	.536	.082	.151	.333	.267
	8MC+2CR	1.913	.800	.602	.465	.053	.050	.260	.234
$\mu = 0.30$	10MC	4.449	1.061	.733	.723	.073	.102	.352	.290
	9MC+1CR	3.918	1.011	.677	.651	.067	.124	.293	.195
	8MC+2CR	3.478	.974	.890	.795	.056	.078	.306	.262

FR: Frequency Estimation
CH: Chained Equipercentile
TR: IRT True Score
OB: IRT Observed Score

Overall the IRT methods produced lower D_1 values compared to the traditional equating methods. Among the IRT methods the IRT true score method was superior to the IRT observed score method across all group difference and common-item composition conditions. The IRT true score method produced the lowest values when the common-item set consisted of 8MC + 2 CR items. However, no such pattern was observed for the other equating methods. Comparisons among the traditional methods indicate that the chained equipercentile method produced lower D_1 values compared to the frequency estimation method across all conditions. A noticeable pattern was observed for the frequency estimation method in that the conditional expected raw scores differed in a great extent for the alternate forms in particular when differences between group mean ability increased in magnitude. Similiar results have been found by other researchers (Andrews, 2011; Duong, 2011). These severe discrepancies between the classical equating methods can be attributed to the differences in assumptions that each method presumes about the

data. More specifically, the frequency estimation method makes a stringent assumption about the synthetic population whereas the chained equipercentile method does not make such an overt assumption.

4.1.1 Unidimensional

Results from the mixed ANOVA analyses for D_1 values under a unidimensional test structure are presented in Table 4-2.

Table 4-2 Mixed ANOVA Results for First-Order Equity

Effect	Source	F	p	Partial η^2
Within-Subjects	Method	2395.02	<.001	.745
	Method*Group	517.953	<.001	.558
	Method*Anchor	8.453	<.001	.020
	Method*Group*Anchor	5.794	<.001	.028
Between-Subjects	Group	360.833	<.001	.468
	Anchor	9.124	<.001	.022
	Group*Anchor	1.892	.110	.009

Overall, the pattern of difference on the D_1 index among group mean differences was significantly different among equating methods (p<.001, partial η^2= .558). Although there is a significant three way interaction, the effect sizes that involve the anchor composition are negligible compared to the method by group interaction and therefore a simple main effect analysis of methods was performed for each group as a follow-up analysis (Table 4-3).

Table 4-3 Simple Main Effects of Methods for each Group Difference

Group	F	p	Partial η^2
1: $\mu = 0$	129.806	<.001	.323
2: $\mu = .15$	779.767	<.001	.738
3: $\mu = .30$	1747.717	<.001	.864

Equating methods performed significantly different in each group difference in terms of first-order equity preservation (p<.001, partial $\eta^2 = .323 - .864$). In order to examine which equating methods differed from each other, pairwise comparisons among the equating methods were performed (Table 4-4) for each group. Marginal means are shown in Table 4-5.

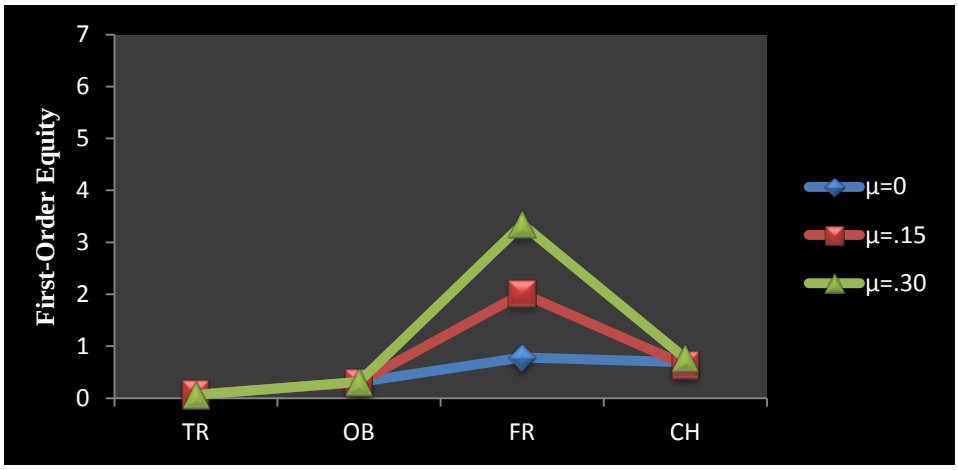
Table 4-4 Comparisons among Equating Methods for each Group Difference

Group	Methods	F	p	Partial η^2
1: $\mu = 0$	TR vs OB	316.199	<.001	.538
	FR vs CH	11.833	<.001	.042
	TR vs FR	187.039	<.001	.407
	OB vs FR	70.896	<.001	.207
	TR vs CH	144.598	<.001	.347
	OB vs CH	40.081	<.001	.147
2: $\mu = .15$	TR vs OB	233.566	<.001	.457
	FR vs CH	499.348	<.001	.643
	TR vs FR	1424.197	<.001	.837
	OB vs FR	1089.555	<.001	.797
	TR vs CH	249.726	<.001	.474
	OB vs CH	72.483	<.001	.207
3: $\mu = .30$	TR vs OB	302.388	<.001	.523
	FR vs CH	1506.899	<.001	.845
	TR vs FR	3376.127	<.001	.924
	OB vs FR	2861.805	<.001	.912
	TR vs CH	279.043	<.001	.503
	OB vs CH	97.084	<.001	.260

TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Table 4-5 Marginal Means for each Method by Group Difference

Group	Method	*M*	*SE*
1: $\mu = 0$	IRT True	.065	.011
	IRT Observed	.308	.015
	Frequency Estimation	.781	.043
	Chained Equipercentile	.692	.043
2: $\mu = .15$	IRT True	.073	.011
	IRT Observed	.302	.013
	Frequency Estimation	2.018	.041
	Chained Equipercentile	.638	.042
3: $\mu = .30$	IRT True	.078	.013
	IRT Observed	.317	.014
	Frequency Estimation	3.349	.042
	Chained Equipercentile	.767	.043



TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Figure 3 Differences among Equating Methods for each Group Difference

4.1.2 Comparison of IRT Methods

Overall, IRT true score equating (Equivalent group: M = .065, SE = .011; small group difference: M = .073, SE = .015; large group difference: M = .075, SE = .013), had lower D_1 values compared to IRT observed score equating method (Equivalent group: M = .308, SE = .015; small group difference: M = .302, SE = .013; large group difference: M = .317, SE = .014) across all group difference conditions. Differences between the two methods increased in magnitude as the groups became more nonequivalent.

4.1.3 Comparison of Equipercentile Equating Methods

The chained equipercentile method (Equivalent group: *M* = .692, *SE* = .043; small group difference: *M* = .638, *SE* = .042; large group difference: *M* = .767, *SE* = .043) outperformed the frequency estimation method (Equivalent group: *M* = .781, *SE* = .043; small group difference: *M* = 2.018, *SE* = .041; large group difference: *M* = 3.349, *SE* = .042) in terms of first-order equity preservation, in particular when group differences increased. The frequency estimation procedure became less accurate as differences between the group mean ability distributions increased. The methods performed similar when groups were equivalent (chained equipercentile: *M* = .692, *SE* = .043; frequency estimation: *M* = .781, *SE* = .043).

4.1.4 Comparison across all Methods

Comparison across all methods showed that IRT equating methods typically outperformed the traditional equating procedures in terms of first-order equity preservation. These results were to be expected in particular since the equity properties rely heavily on IRT methodologies in the evaluation of results and thus the IRT equating methods were likely favored over the traditional equating methods (Andrews, 2011).

4.1.5 Multidimensional

Table 4-6 shows the mean D_1 values over 100 replications for each condition under the multidimensional test structure.

Table 4-6 D_1 Values over 100 Replications

Group	Anchor	FR Mean	FR SD	CH Mean	CH SD	TR Mean	TR SD	OB Mean	OB SD
$\mu = 0,0,0$	10MC	.978	.716	.900	.646	.091	.071	.267	.233
	9MC+1CR	.747	.566	.662	.528	.058	.042	.213	.173
	8MC+2CR	.799	.610	.770	.605	.062	.049	.226	.171
$\mu = .3,.3,.3$	10MC	6.458	1.33	.988	.853	.102	.085	.340	.317
	9MC+1CR	6.355	1.193	.991	.657	.071	.059	.260	.205
	8MC+2CR	6.378	1.266	1.022	.837	.065	.048	.218	.167
$\mu = .15,.15,.15$	10MC	3.084	1.036	.825	.775	.093	.076	.247	.212
	9MC+1CR	3.235	1.114	.835	.635	.071	.056	.218	.168
	8MC+2CR	3.041	1.070	.854	.607	.062	.050	.169	.134
$\mu = .3,.15,.3$	10MC	4.963	1.274	1.078	1.027	.117	.097	.335	.317
	9MC+1CR	5.506	1.566	1.093	1.086	.090	.071	.221	.185
	8MC+2CR	5.767	1.231	1.074	.769	.063	.043	.176	.148
$\mu = .3,.3,.15$	10MC	6.502	1.304	1.011	.839	.088	.072	.323	.252
	9MC+1CR	6.023	1.046	.887	.698	.085	.060	.251	.194
	8MC+2CR	5.683	.908	.796	.590	.057	.052	.229	.180

FR: Frequency Estimation
CH: Chained Equipercentile
TR: IRT True Score
OB: IRT Observed Score

When a multidimensional test structure was applied to a unidimensional equity framework it was found that the IRT true score method produced lower D_1 values compared to the IRT observed score across all conditions of mean ability group differences and anchor item composition. Both IRT methods produced the lowest D_1 and D_2 values when the anchor set was representative of the total test across most mean group ability differences. However, when groups were equivalent then D_1 values for both procedures were lowest when the anchor set was composed of 9 MC + 1 CR rather than 8 MC + 2 CR. When group differences were constant across the dimensions but small and when group differences varied across the MC and CR domains, the D_1 statistic was lowest for both methods when the anchor set consisted of 8 MC + 2 CR instead of 9 MC + 1 CR.

Comparisons among the traditional methods showed that the chained equipercentile method was more successful in preserving first-order equity compared to the frequency estimation method across all conditions. Similar to what was observed under the unidimensional condition; the frequency estimation method produced higher D_1 values as a direct function of an increase in the mean ability group differences suggesting that the conditional expected scores after equating differ in a great extent on the alternate forms. When groups were equivalent both methods produced the lowest D_1 values when the anchor set was representative of the total test with 9 MC + 1 CR items rather than 8 MC + 2 CR items. Table 4-7 presents the results from the mixed ANOVA analyses for D_1 values under a multidimensional test structure.

Table 4-7 Mixed ANOVA Results for First-Order Equity

Effect	Source	F	p	Partial η^2
Within-Subjects	Method	10948.577	<.001	.899
	Method*Group	753.811	<.001	.709
	Method*Anchor	.277	.948	.000
	Method*Group*Anchor	4.225	<.001	.027
Between-Subjects	Group	750.312	<.001	.708
	Anchor	3.705	.025	.006
	Group*Anchor	4.732	<.001	.030

The pattern of difference on the D_1 index among group mean differences was significantly different among equating methods (p<.001, partial $\eta^2 = .709$). A simple main effect analyses was performed to detect the pattern of difference on the D_1 index for each group difference (Table 4-8).

Table 4-8 Simple Main Effects of Methods for each Group Difference

Group	F	p	Partial η^2
1: μ=0,0,0	301.791	<.001	.531
2: μ=.3,.3,.3	3649.373	<.001	.941
3: μ=.15,.15,.15	1309.450	<.001	.828
4: μ=.3,.15,.3	1878.591	<.001	.892
5: μ=.3,.3,.15	4233.240	<.001	.944

Equating methods performed differently in terms of first-order equity preservation in each group difference (p<.001, partial $\eta^2 = .531 - .944$). Pairwise comparisons among the equating methods were performed to examine which equating methods differed from each other (Table 4-9). Marginal means are shown in Table 4-10.

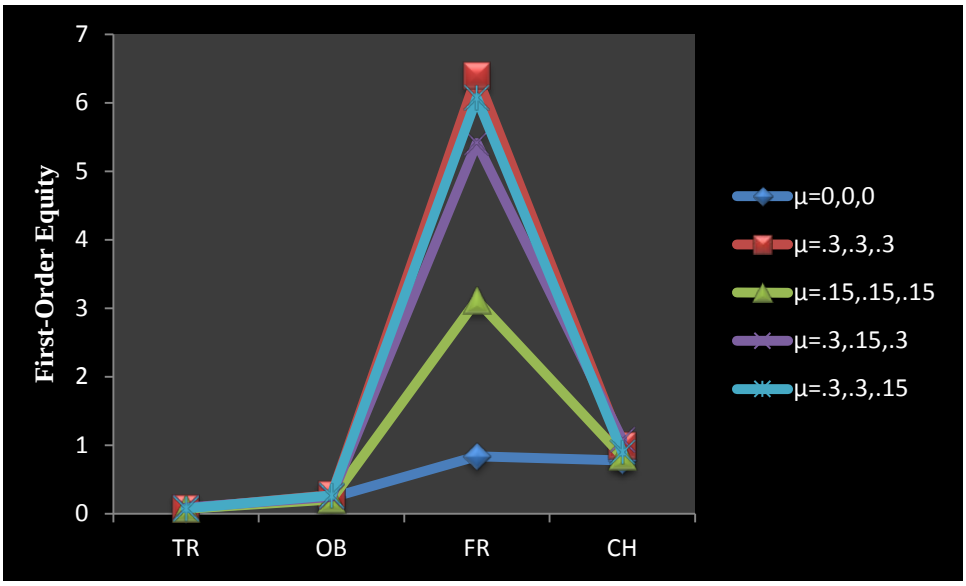
Table 4-9 Comparisons among Methods for each Group Difference

Group	Methods	F	p	partial η^2
1: $\mu = 0,0,0$	TR vs OB	215.572	<.001	.447
	TR vs FR	13.167	<.001	.607
	OB vs FR	275.761	<.001	.508
	TR vs CH	394.419	<.001	.596
	OB vs CH	266.462	<.001	.499
2: $\mu = .3,.3,.3$	TR vs OB	140.573	<.001	.380
	FR vs CH	2846.824	<.001	.926
	TR vs FR	5569.270	<.001	.961
	OB vs FR	5692.122	<.001	.961
	TR vs CH	315.269	<.001	.579
	OB vs CH	212.493	<.001	.481
3: $\mu = .15,.15,.15$	TR vs OB	173.493	<.001	.389
	FR vs CH	791.454	<.001	.744
	TR vs FR	2188.145	<.001	.889
	OB vs FR	1999.192	<.001	.880
	TR vs CH	358.774	<.001	.569
	OB vs CH	260.354	<.001	.489
4: $\mu = .3,.15,.3$	TR vs OB	94.186	<.001	.293
	FR vs CH	1340.333	<.001	.855
	TR vs FR	3111.473	<.001	.932
	OB vs FR	3004.690	<.001	.930
	TR vs CH	243.377	<.001	.517
	OB vs CH	183.944	<.001	.448
5: $\mu = .3,.3,.15$	TR vs OB	231.252	<.001	.481
	FR vs CH	3207.664	<.001	.928
	TR vs FR	6641.737	<.001	.964
	OB vs FR	6252.128	<.001	.962
	TR vs CH	340.671	<.001	.577
	OB vs CH	230.497	<.001	.480

TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Table 4-10 Marginal Means (and Standard Errors) for each Group Difference

Method	$\mu = 0,0,0$	$\mu = .3,.3,.3$	$\mu = .15,.15,.15$	$\mu = .3,.15,.3$	$\mu = .3,.3,.15$
IRT True	.070 (.004)	.079 (.004)	.075 (.004)	.090 (.004)	.077 (.004)
IRT Observed	.235 (.013)	.273 (.015)	.211 (.013)	.244 (.015)	.268 (.014)
Frequency Estimation	.841 (.068)	6.397 (.075)	3.120 (.067)	5.412 (.075)	6.069 (.070)
Chained Equipercentile	.777 (.047)	1.000 (.051)	.838 (.046)	1.082 (.051)	.898 (.048)



TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Figure 4 Differences among Equating Methods for each Group Difference

4.1.6 Comparison of IRT Methods

The IRT true score method (*M* = .070-.090, *SE* = .004) was superior to the IRT observed score method (*M* = .211-.273, *SE* = .013-.015) in terms of first-order equity preservation across all conditions of mean ability group differences. Overall, higher D_1 values were observed when

differences between groups were large ($\mu = .3, .3, .3$) and when the ability between the groups varied across the multiple-choice and constructed-response dimensions ($\mu = .3, .15, .3$; $\mu = .3, .3, .15$). Overall, the IRT true score method produced lower D_1 values when the ability was higher on the multiple-choice dimension (*M* = .077, *SE* = .004) compared to the constructed-response dimension (*M* = .090, *SE* = .004).

4.1.7 Comparison of Equipercentile Equating Methods

The chained equipercentile method (*M* = .777-1.082, *SE* = .046-.051) was more successful in preserving first-order equity compared to the frequency estimation method (*M* = .841-6.397, *SE* = .067-.075)) when group differences increased in magnitude. There was no meaningful difference between the chained equipercentile method (*M* = .777, *SE* = .047) and the frequency estimation method (*M* = .841, *SE* = .068) when groups were equivalent.

The frequency estimation method became less accurate in terms of first-order equity preservation as a direct function of an increase in the mean ability group differences, in particular when differences were large and group mean differences varied across multiple-choice and constructed-response dimensions. The frequency estimation method produced lower D_1 values when the mean group ability was lower on the multiple-choice dimension (*M* = 5.412, *SE* = .075) compared to the constructed-response dimension (*M* = 6.069, *SE* = .070) whereas the chained equipercentile method produced lower values when the mean group ability difference was larger on the multiple-choice dimension (*M* = .898, *SE* = .048) compared to the constructed-response dimension (*M* = 1.082, *SE* = .051).

4.1.8 Comparison across all Methods

Comparison across all methods showed that IRT equating methods outperformed the traditional equating procedures in terms of first-order equity preservation in all conditions regardless of the underlying test structure.

4.2 SECOND-ORDER EQUITY

The second-order equity requires the equivalence of standard deviations of conditional distributions; thus, conditioning on ability the distribution of the reference form and the equated form are comparable. In order to examine second-order equity, a D_2 index was used as an indicator of how well the second-order property was preserved. Overall, lower D_2 values denote better preservation of second- order equity. The mean D_2 values over 100 replications for each condition under the undimensional test structure are shown in Table 4-11.

Table 4-11 D_2 Values over 100 Replications

Group	Anchor	FR Mean	FR SD	CH Mean	CH SD	TR Mean	TR SD	OB Mean	OB SD
$\mu = 0$	10MC	.300	.208	.297	.211	.231	.176	.257	.190
	9MC+1CR	.283	.193	.291	.203	.234	.320	.246	.155
	8MC+2CR	.271	.227	.280	.224	.182	.158	.212	.181
$\mu = 0.15$	10MC	.305	.220	.315	.239	.224	.226	.261	.208
	9MC+1CR	.303	.197	.295	.200	.251	.226	.260	.184
	8MC+2CR	.287	.223	.296	.231	.195	.151	.233	.171
$\mu = 0.30$	10MC	.317	.236	.342	.249	.282	.230	.291	.213
	9MC+1CR	.276	.209	.301	.237	.247	.283	.242	.171
	8MC+2CR	.286	.264	.326	.302	.266	.430	.257	.306

FR: Frequency Estimation
CH: Chained Equipercentile
TR: IRT True Score
OB: IRT Observed Score

Overall, all methods produced similar D_2 values over 100 replications across all conditions indicating that the conditional raw score standard errors of measurement are similar on alternate forms. Among the IRT methods the IRT true score method produced slightly lower values than the IRT observed score method. In general slightly lower values were obtained for all methods when the anchor composition was representative of the total test (9MC + 1CR or 8MC + 2CR). Comparisons among the traditional equating methods show that the FR method produced slightly lower values compared to the chained equipercentile method.

4.2.1 Unidimensional

Table 4-12 reflects the mixed ANOVA results of the D_2 index under a unidimensional test structure.

Table 4-12 Mixed ANOVA Results of Second-Order Equity

Effect	Source	F	p	Partial η^2
Within-Subjects	Method	44.074	<.001	.051
	Method*Group	1.544	.160	.004
	Method*Anchor	0.808	0.564	.002
	Method*Group*Anchor	0.250	0.996	.001
Between-Subjects	Group	1.588	<.001	.004
	Anchor	1.444	0.205	.004
	Group*Anchor	0.469	0.237	.002

There was a significant difference among the methods averaged across group differences and anchor representativeness (p<.001, partial $\eta^2 = .051$). However, based on the effect size these differences may not be practically significant.

4.2.2 Multidimensional

Table 4-13 shows the D_2 values averaged over 100 replications for all conditions under a multidimensional test structure.

Table 4-13 D_2 Values over 100 Replications

Group	Anchor	FR Mean	FR SD	CH Mean	CH SD	TR Mean	TR SD	OB Mean	OB SD
$\mu = 0,0,0$	10MC	.333	.249	.343	.255	.194	.170	.310	.250
	9MC+1CR	.276	.201	.289	.199	.167	.120	.223	.152
	8MC+2CR	.250	.160	.248	.164	.168	.138	.212	.177
$\mu = .3,.3,.3$	10MC	.389	.290	.477	.327	.233	.182	.365	.283
	9MC+1CR	.304	.212	.416	.290	.167	.143	.220	.202
	8MC+2CR	.329	.215	.477	.289	.182	.130	.262	.183
$\mu = .15,.15,.15$	10MC	.318	.219	.331	.257	.184	.151	.308	.231
	9MC+1CR	.243	.182	.298	.208	.208	.149	.267	.180
	8MC+2CR	.264	.212	.300	.233	.183	.158	.241	.178
$\mu = .3,.15,.3$	10MC	.331	.268	.379	.279	.231	.215	.343	.332
	9MC+1CR	.281	.205	.388	.268	.196	.152	.279	.205
	8MC+2CR	.244	.208	.327	.252	.178	.142	.202	.161
$\mu = .3,.3,.15$	10MC	.354	.246	.426	.292	.213	.173	.324	.249
	9MC+1CR	.292	.224	.373	.285	.223	.169	.281	.210
	8MC+2CR	.243	.195	.325	.246	.198	.136	.220	.160

FR: Frequency Estimation
CH: Chained Equipercentile
TR: IRT True Score
OB: IRT Observed Score

Among the IRT equating methods, IRT true score method produced lower D_2 values compared to the IRT observed score method across all conditions. The IRT true score method produced higher D_2 values when the group differences were large across the dimensions and when the mean group ability differences varied across the multiple choice and constructed response dimensions. The true score method produced lower D_2 values when the mean ability difference was larger on the constructed response items. Comparisons among the classical equating methods showed that the frequency estimation method produced lower D_2 values than the chained equipercentile method. When group differences varied across the multiple choice and constructed response dimensions, the frequency estimation method produced lower D_2 values when the ability mean differences were higher on the constructed response items. In general

lower D_2 values were observed when the common-item set was representative of the total test (9MC + 1 CR and 8MC + 2 CR) compared to format non-representativeness (10MC). Table 4-14 reflects the mixed ANOVA results of the D_2 index under a multidimensional test structure.

Table 4-14 Mixed ANOVA Results of Second-Order Equity

Effect	Source	F	p	Partial η^2
Within-Subjects	Method	233.929	<.001	.159
	Method*Group	7.919	<.001	.025
	Method*Anchor	5.193	<.001	.008
	Method*Group* Anchor	.984	.484	.006
Between-Subjects	Group	5.445	<.001	.017
	Anchor	17.432	<.001	.027
	Group*Anchor	.866	.545	.006

There was a significant difference on second-order equity preservation among methods averaged across group differences and anchor representativeness (p<.001, partial $\eta^2 = .159$). Marginal comparisons were conducted as follow-up analyses to examine which methods differed from each other (Table 4-15). Marginal means are presented in Table 4-16.

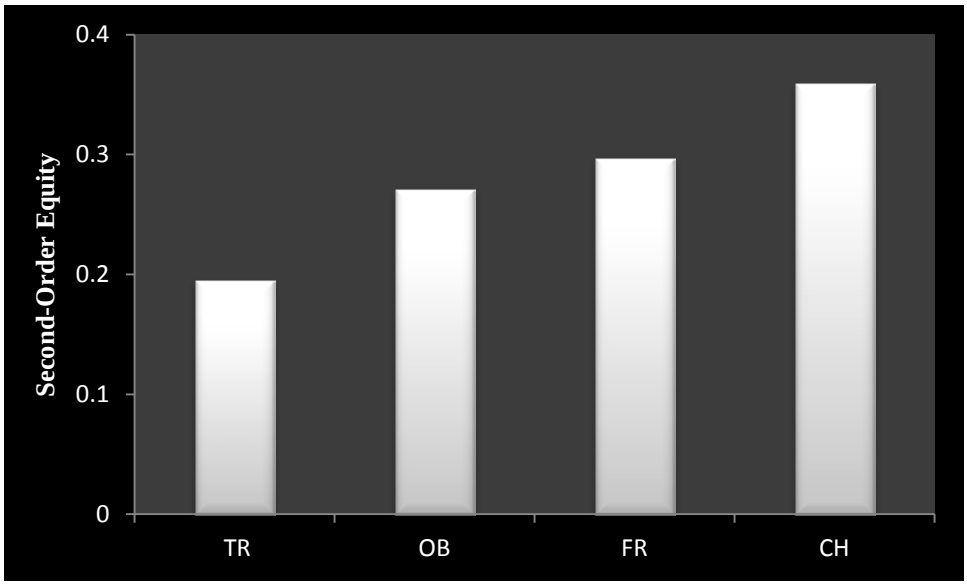
Table 4-15 Marginal Comparisons of Methods averaged across Group and Anchor

Comparison	F	p	Partial η^2
TR vs OB	313.201	<.001	.201
FR vs CH	184.629	<.001	.129
TR vs FR	229.560	<.001	.156
OB vs FR	18.433	<.001	.015
TR vs CH	419.663	<.001	.252
OB vs CH	142.396	<.001	.103

TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Table 4-16 Marginal Means and Standard Errors for Methods

Method	*M*	*SE*
IRT True	.195	.005
IRT Observed	.271	.006
Frequency Estimation	.297	.006
Chained Equipercentile	.360	.007



TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Figure 5 Marginal Means of Equating Methods of D_2 Index

4.2.3 Comparison of IRT Methods

The IRT true score method (*M* = .195, *SE* = .005) was superior to the IRT observed score method (*M* = .271, *SE* = .006) in terms of second-order equity. Given the large effect size (partial η^2 = .201) it is fair to conclude that these differences are practically meaningful.

4.2.4 Comparison of Equipercentile Equating Methods

The frequency estimation method (*M* = .297, *SE* = .006) outperformed the chained equipercentile method (*M* = .360, *SE* = .007) in regards to second-order equity preservation. These differences are found to be statistically and practically significant (p<.001, partial η^2 = .129).

4.2.5 Comparison across all Methods

Comparisons across all methods showed that IRT equating methods typically outperformed the traditional equating procedures in terms of second-order equity preservation averaged across group and common-item composition.

4.3 SAME DISTRIBUTIONS PROPERTY

The Kolmogorov-Smirnov *T* statistic was used to examine the same distributions property, which calculates the largest discrepancy in the empirical distribution functions between the reference form and the new equated form for the synthetic group. In the current study, the synthetic group was the group taking the new form. Lower *T* values indicate that the distributions are comparable after equating. Table 4-17 presents the *T* values over 100 replications for each condition under the unidimensional test structure.

Table 4-17 *T* Values over 100 Replications

Group	Anchor	FR Mean	FR SD	CH Mean	CH SD	TR Mean	TR SD	OB Mean	OB SD
$\mu = 0$	10MC	.038	.015	.039	.015	.041	.019	.040	.019
	9MC+1CR	.037	.015	.037	.014	.038	.015	.036	.015
	8MC+2CR	.038	.014	.037	.015	.038	.018	.037	.017
$\mu = 0.15$	10MC	.046	.021	.039	.015	.040	.017	.038	.017
	9MC+1CR	.045	.019	.038	.017	.041	.023	.038	.022
	8MC+2CR	.044	.020	.037	.014	.039	.016	.038	.015
$\mu = 0.30$	10MC	.066	.023	.040	.015	.039	.015	.037	.016
	9MC+1CR	.063	.025	.041	.015	.039	.019	.038	.018
	8MC+2CR	.060	.024	.042	.016	.037	.016	.037	.016

FR: Frequency Estimation
CH: Chained Equipercentile
TR: IRT True Score
OB: IRT Observed Score

Among the IRT methods, the IRT observed score method produced lower *T* values compared to the IRT true score method across all conditions. The traditional equating methods produced similar results when groups were equivalent. However, the chained equipercentile method produced lower *T* values compared to the frequency estimation method when mean group ability differences increased. Both IRT methods produced lower *T* values when the anchor set was representative of the total test (8MC + 2 CR) compared to 9MC + 1CR and 10MC. However, there was no such pattern observed for the traditional equating methods.

4.3.1 Unidimensional

A mixed ANOVA was performed to examine overall differences among the equating methods. Table 4-18 presents the numerical results of the mixed ANOVA analysis for the unidimensional test structure.

Table 4-18 Mixed ANOVA Results for the Same Distribution Property

Effect	Source	F	p	Partial η^2
Within-Subjects	Method	161.212	<.001	.165
	Method*Group	90.905	<.001	.182
	Method*Anchor	0.667	0.676	.002
	Method*Group*Anchor	1.448	0.137	.007
Between-Subjects	Group	15.229	<.001	.036
	Anchor	1.090	0.337	.003
	Group*Anchor	0.373	0.828	.002

The pattern of difference on the preservation of the same distribution property among group differences was significantly different among the equating methods under investigation (p<.001, partial $\eta^2 = .182$). In order to find the pattern of difference on *T* values among group differences and equating methods, a simple main effect of methods was performed for each group (Table 4-19).

Table 4-19 Simple Main Effects of Methods for each Group Difference

Group	F	P	Partial η^2
1: $\mu = 0$	2.943	.032	.011
2: $\mu = .15$	19.769	<.001	.068
3: $\mu = .30$	213.157	<.001	.434

Tables 4-20 and 4-21 provide the pairwise comparisons of the methods and the marginal means for each group difference.

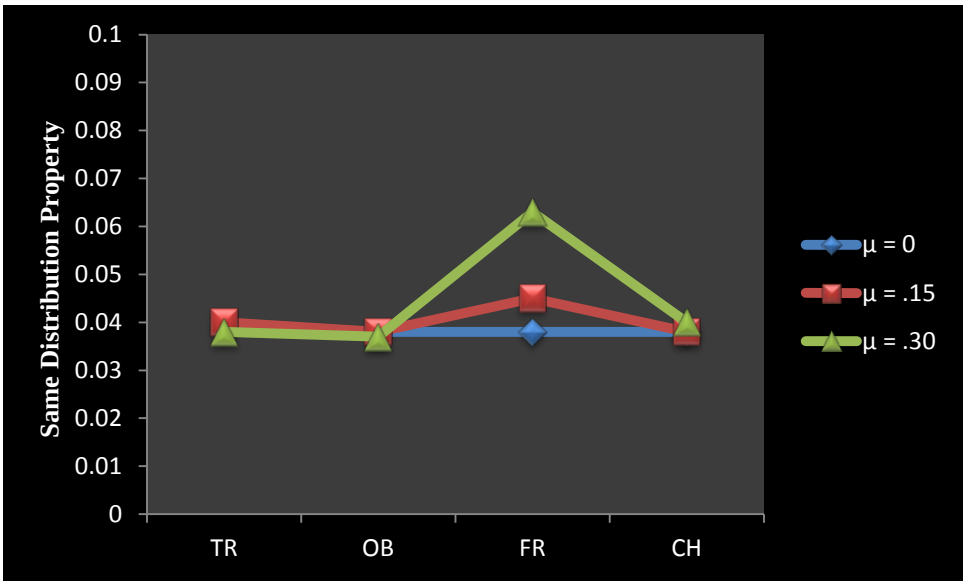
Table 4-20 Comparisons among Groups for each Group Difference

Group	Methods	F	p	Partial η^2
1: $\mu = 0$	TR vs OB	27.367	<.001	.091
	FR vs CH	.003	.955	.000
	TR vs FR	4.307	.039	.015
	OB vs FR	.056	.814	.000
	TR vs CH	4.965	.027	.018
	OB vs CH	.082	.775	.000
2: $\mu = .15$	TR vs OB	42.619	<.001	.136
	FR vs CH	46.735	<.001	.147
	TR vs FR	13.346	<.001	.047
	OB vs FR	27.422	<.001	.092
	TR vs CH	3.523	.062	.013
	OB vs CH	.111	.739	.000
3: $\mu = .30$	TR vs OB	20.932	<.001	.070
	FR vs CH	272.008	<.001	.495
	TR vs FR	237.114	<.001	.460
	OB vs FR	259.482	<.001	.483
	TR vs CH	60.007	.015	.021
	OB vs CH	15.938	<.001	.054

TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Table 4-21 Marginal Means (and Standard Errors) for each Method by Group Difference

Method	$\mu = 0$	$\mu = .15$	$\mu = .30$
IRT True	.039 (.001)	.040 (.001)	.038 (.001)
IRT Observed	.038 (.001)	.038 (.001)	.037 (.001)
Frequency Estimation	.038 (.001)	.045 (.001)	.063 (.001)
Chained Equipercentile	.038 (.001)	.038 (.001)	.040 (.001)



TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Figure 6 Differences among Equating Methods for each Group Difference

4.3.2 Comparison of IRT Methods

The IRT observed score method (Equivalent Groups: *M* = .038, *SE* = .001; Small Group Difference: *M* = .038, *SE* = .001; Large Group Difference: *M* = .037, *SE* = .001) produced slightly lower values compared to the IRT true score method (Equivalent Groups: *M* = .039, *SE* = .001; Small Group Difference: *M* = .040, *SE* = .001; Large Group Difference: *M* = .038, *SE* = .001) across all group difference conditions. The medium effect sizes (partial η^2 = .070-.136) suggest that these differences are practically meaningful.

4.3.3 Comparison of Equipercentile Methods

The chained equipercentile method (*M* = .038, *SE* = .001) and the frequency estimation (*M* = .038, *SE* = .001) methods produced similar results in terms of same distribution preservation when the groups were equivalent. However, the chained equipercentile method outperformed the frequency estimation method as the groups became nonequivalent. For example, when there was a small group difference ($\mu = .15$) the chained equipercentile method (*M* = .038, *SE* = .001) produced lower values than the frequency estimation method (*M* = .045, *SE* = .001). The discrepancy between the two methods became larger when the group difference was large ($\mu = .30$) (chained equipercentile method (*M* = .040, *SE* = .001), frequency estimation method (*M* = .063, *SE* = .001)).

4.3.4 Comparison across all Methods

Comparisons across all methods indicate that when groups are equivalent all methods perform similarly. However, when group differences are large ($\mu = .30$) IRT methods produce slightly lower *T* values than traditional equipercentile equating methods. Large differences were found between the frequency estimation and chained equipercentile method, IRT true score and frequency estimation method, and IRT observed score and frequency estimation method.

4.3.5 Multidimensional

Table 4-22 presents the T values over 100 replications for each condition under a multidimensional test structure.

Table 4-22 T Values over 100 Replications

Group	Anchor	FR Mean	FR SD	CH Mean	CH SD	TR Mean	TR SD	OB Mean	OB SD
$\mu = 0,0,0$	10MC	.042	.015	.042	.013	.050	.016	.043	.016
	9MC+1CR	.040	.017	.041	.017	.041	.019	.041	.019
	8MC+2CR	.039	.017	.038	.017	.040	.019	.040	.018
$\mu = .3,.3,.3$	10MC	.091	.028	.047	.018	.044	.018	.042	.018
	9MC+1CR	.086	.028	.049	.019	.044	.017	.042	.018
	8MC+2CR	.085	.026	.048	.016	.044	.018	.042	.018
$\mu = .15,.15,.15$	10MC	.055	.021	.040	.016	.040	.014	.038	.014
	9MC+1CR	.055	.021	.041	.017	.040	.016	.038	.015
	8MC+2CR	.052	.020	.044	.016	.035	.014	.038	.018
$\mu = .3,.15,.3$	10MC	.080	.027	.046	.018	.047	.018	.044	.017
	9MC+1CR	.078	.026	.043	.016	.042	.015	.040	.014
	8MC+2CR	.076	.025	.050	.015	.039	.016	.038	.016
$\mu = .3,.3,.15$	10MC	.084	.028	.044	.016	.044	.019	.042	.019
	9MC+1CR	.082	.028	.045	.016	.042	.017	.040	.016
	8MC+2CR	.078	.027	.047	.018	.043	.017	.041	.018

FR: Frequency Estimation
CH: Chained Equipercentile
TR: IRT True Score
OB: IRT Observed Score

Similar to the undimensional test structure, the IRT observed score method produced lower *T* values than the IRT true score method in particular when group differences were large across all dimensions and when group differences varied across dimensions. The two methods produced similar results when the groups were equivalent. Comparisons among the classical equating methods showed that the chained equipercentile method and the frequency estimation methods produced similar *T* values when the groups were equivalent However, overall the chained equipercentile method outperformed the frequency estimation method when the groups became nonequivalent. The frequency estimation method performed worst when group ability differences were large and when group ability differences varied across dimensions. There was no clear pattern for neither of the

equating methods under investigation in terms of the effect of the common-item composition on the estimation of *T* values. Table 4-23 presents the numerical results of the mixed ANOVA analysis for the multidimensional test structure.

Table 4-23 Mixed ANOVA Results for Same Distribution Property

Effect	Source	F	p	Partial η^2
Within-Subjects	Method	1192.068	<.001	.491
	Method*Group	128.806	<.001	.294
	Method*Anchor	1.444	.194	.002
	Method*Group*Anchor	1.128	.302	.007
Between-Subjects	Group	43.764	<.001	.124
	Anchor	2.176	.114	.004
	Group*Anchor	.268	.976	.002

The pattern of difference on the same distribution equity preservation among group differences was significantly different among the equating methods (p<.001 partial $\eta^2 = .294$). In an attempt to find the pattern of difference on the equipercentile index among group differences and equating methods, a simple main effect analysis of methods was performed for each group difference (Table 4-24).

Table 4-24 Simple Main Effects of Methods for each Group Difference

Group	F	p	Partial η^2
1: $\mu = 0,0,0$	4.434	.004	.016
2: $\mu = .3,.3,.3$	430.557	<.001	.653
3: $\mu = .15,.15,.15$	111.315	<.001	.290
4: $\mu = .3,.15,.3$	389.866	<.001	.632
5: $\mu = .3,.3,.15$	325.462	<.001	.566

There were significant differences on the equipercentile criterion among the equating methods ($p<.001$) for each group difference. The significant differences on the preservation of the same distribution property among methods for each group difference were followed by simple comparisons among methods (Table 4-25). Marginal means can be found in Table 4-26.

Table 4-25 Comparisons among Methods for each Group Difference

Group	Methods	F	p	Partial η^2
1: $\mu = 0,0,0$	TR vs OB	19.469	<.001	.068
	FR vs CH	1.150	.285	.004
	TR vs FR	8.448	.004	.031
	OB vs FR	1.834	.177	.007
	TR vs CH	5.721	.017	.021
	OB vs CH	.475	.491	.002
2: $\mu = .3,.3,.3$	TR vs OB	29.140	<.001	.113
	FR vs CH	506.192	<.001	.689
	TR vs FR	471.618	<.001	.673
	OB vs FR	502.833	<.001	.687
	TR vs CH	18.732	<.001	.076
	OB vs CH	38.831	<.001	.145
3: $\mu = .15,.15,.15$	TR vs OB	37.345	<.001	.121
	FR vs CH	124.562	<.001	.314
	TR vs FR	122.877	<.001	.311
	OB vs FR	156.159	<.001	.365
	TR vs CH	2.934	.088	.011
	OB vs CH	17.578	<.001	.061
4: $\mu = .3,.15,.3$	TR vs OB	37.021	<.001	.140
	FR vs CH	501.170	<.001	.688
	TR vs FR	429.545	<.001	.654
	OB vs FR	507.544	<.001	.691
	TR vs CH	2.621	.107	.011
	OB vs CH	15.850	<.001	.065
5: $\mu = .3,.3,.15$	TR vs OB	32.771	<.001	.116
	FR vs CH	448.849	<.001	.642
	TR vs FR	333.288	<.001	.571
	OB vs FR	361.582	<.001	.591
	TR vs CH	4.477	.035	.018
	OB vs CH	17.783	<.001	.066

TR: IRT True Score

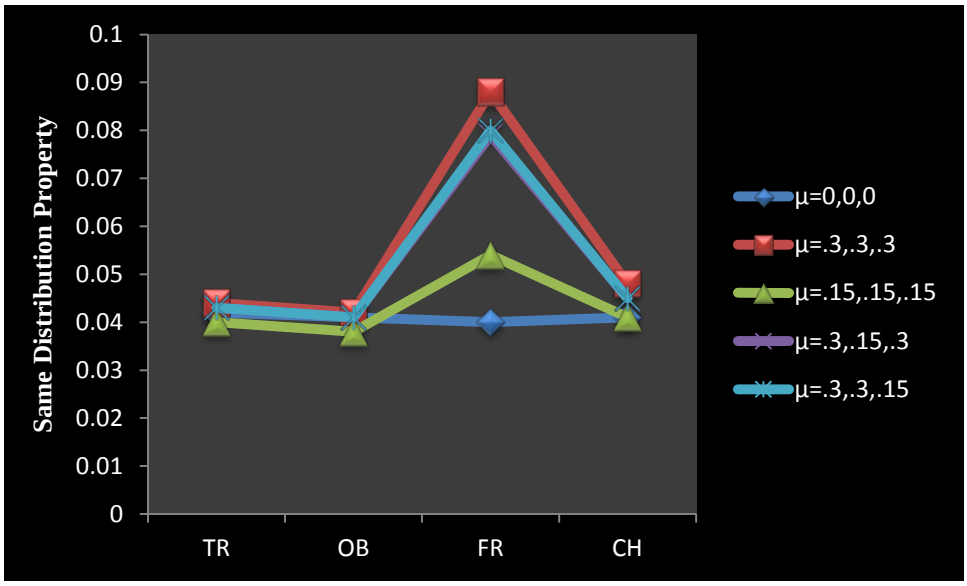
OB: IRT Observed Score

FR: Frequency Estimation

CH: Chained Equipercentile

Table 4-26 Marginal Means (and Standard Errors) for each Method by Group Difference

Method	$\mu = 0,0,0$	$\mu = .3,.3,.3$	$\mu = .15,.15,.15$	$\mu = .3,.15,.3$	$\mu = .3,.3,.15$
IRT True	.042 (.001)	.044 (.001)	.040 (.001)	.043 (.001)	.043 (.001)
IRT Observed	.041 (.001)	.042 (.001)	.038 (.001)	.041 (.001)	.041 (.001)
Frequency Estimation	.040 (.001)	.088 (.002)	.054 (.001)	.079 (.002)	.080 (.002)
Chained Equipercentile	.041 (.001)	.048 (.001)	.041 (.001)	.045 (.001)	.045 (.001)



TR: IRT True Score
OB: IRT Observed Score
FR: Frequency Estimation
CH: Chained Equipercentile

Figure 7 Differences among Equating Methods for each Group Difference

4.3.6 Comparison of IRT methods

The IRT observed score method (*M* = .038-.042, *SE* = .001) produced slightly lower values compared to the IRT true score method (*M* = .040-.044, *SE* = .001) across all group differences. Both methods produced the highest *T* values when the group differences were large ($\mu = .3, .3, .3$). IRT true score and observed score methods produced similar results when the group differences were varied across the multiple-choice and constructed-response dimensions. In other words, it did not make a difference whether the mean group ability was higher on the multiple-choice items or constructed-response items.

4.3.7 Comparison of Equipercentile Methods

The chained equipercentile method (*M* = .041, *SE* = .001) and the frequency estimation (*M* = .040, *SE* = .001) methods produced similar results in terms of same distribution preservation when the groups were equivalent. However, the chained equipercentile method outperformed the frequency estimation method when the groups became nonequivalent. For example, when there was a small group difference ($\mu = .15, .15, .15$) the chained equipercentile method (*M* = .041, *SE* = .001) produced lower values than the frequency estimation method (*M* = .054, *SE* = .001). The discrepancy between the two methods increased when the group difference was large ($\mu = .3, .3, .3$) (chained equipercentile method (*M* = .048, *SE* = .001), frequency estimation method (*M* = .088, *SE* = .002)). Large discrepancies were also observed when the group differences varied across multiple-choice and constructed-response dimensions. Although it did

not matter whether group differences were larger on the multiple-choice items compared to the constructed-response items as both methods produced similar results across these two conditions.

4.3.8 Comparison across all Methods

Overall, comparisons across all methods indicated that the mean *T* values were similar for all methods except for the frequency estimation method, in particular when groups became increasingly nonequivalent, regardless of the underlying test structure. When groups became nonequivalent, meaningful differences were found between the frequency estimation and chained equipercentile method, IRT true score and frequency estimation method, and IRT observed score and frequency estimation method.

4.4 SUMMARY OF RESULTS

In general, IRT equating methods outperformed traditional equating methods under each evaluation criterion and across all conditions. In summary, large differences were found among the methods in terms of first-order equity preservation. In general, IRT methods were superior to traditional equating methods across all conditions regardless of test structure. The order of equating methods from best preserving to least preserving in terms of first-order equity was as follows: IRT true score, IRT observed score, Chained Equipercentile, and Frequency Estimation. The same order also was observed under a multidimensional test structure. There was no meaningful difference between the traditional equating methods when groups were equivalent, regardless of the underlying test structure.

In terms of second-order equity, all methods produced similar D_2 values under the unidimensional test structure meaning that none of the equating methods was superior in terms of measurement precision. In contrast meaningful differences were found among the IRT equating and traditional equating methods under the multidimensional test structure. The IRT true score method produced lower D_2 values compared to the IRT observed score method whereas the frequency estimation method produced lower D_2 values compared to the chained equipercentile method.

Comparisons across all methods indicated that the mean *T* values were similar for all methods, except for the frequency estimation method, in particular when groups became increasingly nonequivalent regardless of the underlying test structure. *T* values were observed in the following order from best preserving to least preserving: IRT observed score, Chained Equipercentile, IRT True Score, and Frequency Estimation. This trend was apparent for a unidimensional and multidimensional test structure. Differences between the traditional methods had no practical significance when groups were equivalent. Differences between the IRT methods were practically significant.

For all evaluation indices, the anchor item composition did not have a significant effect on the observed criterion under investigation. In general, discrepancies between the methods increased when groups became increasingly nonequivalent and when the mean group ability varied across the multiple-choice and constructed-response dimensions. Overall, there was more variability among the methods when the mean group ability was lower on the multiple-choice dimension compared to the constructed-response dimension. Comparisons across test structures indicated that methods performed similarly regardless of the test dimensionality. However,

slightly lower values for each evaluation index were observed under the unidimensional test structure.

While a direct comparison across test structures was not feasible due to the constraints of the design of the current study, an indirect comparison across test structures indicated that all equating methods performed similarly regardless of the underlying test structure. In other words, the mean D_1, D_2, and T indices across 100 replications were similar for the undimensional and multidimensional test structures. However, the undimensional test structure yielded slightly lower evaluation indices for most conditions. Since the design in the current study is not fully crossed, it is uncertain whether any of the observed differences represent practically significant differences.

5.0 DISCUSSION

The purpose of this study was to compare traditional and IRT equating methods in terms of their ability to preserve first-order, second-order, and same distribution properties under various conditions such as attributes of the test, examinees, and the common-item set. The research questions of interest were addressed through a simulation study. A 3PL/GRM model combination was used to generate the unidimensional test structure whereas a bifactor model was applied to generate the multidimensional test structure. The two 3PL/GRM model combination also was used to calibrate the data. Concurrent calibration was employed to place parameters onto a common scale. The factors that were manipulated include: common-item composition, differences in group ability distributions, and the underlying test structure. The impact of these factors was compared for four commonly used equating methods and the results were addressed in terms of the preservation of equity properties.

This chapter summarizes the results based on each of the research questions. Lastly, limitations and directions for future research are discussed.

5.1 RESEARCH QUESTION 1

Overall, how do the Frequency Estimation (FR), Chained Equipercentile (CH), and Item Response Theory (IRT) true score (TR) and observed score (OB) equating methods compare to one another in terms of preservation of the First-Order (FO), Second-Order (SO), and same distribution equity properties?

In general, the IRT methods performed better than the traditional methods for first-order equity, second-order equity, and same distribution property. However, discrepancies between the methods were small for the same distribution property and second-order equity compared to the first-order equity. The IRT observed score method produced lower values compared to the IRT true score method in regards to the preservation of the same distribution property and second-order equity. The IRT true score method was more successful in preserving first-order equity compared to the IRT observed score method. In general, all equating methods produced similar results in terms of second-order equity as none of the differences were found to be practically significant under the undimensional test structure. This suggests that the equating methods under investigation performed similar in terms of measurement precision. One possible explanation could be due to the sufficiently large sample size that was applied in the current study which could have reduced the amount of random error. There were no significant differences between the traditional equating methods when groups were equivalent regardless of evaluation criterion under investigation. However, even when groups were equivalent the IRT methods still superseded the traditional equating methods. When group differences became more apparent, the chained equipercentile method was superior to the frequency estimation method in terms of first-order equity and same distribution property. Both methods performed similarly in regards to preserving second-order equity.

The findings that IRT true score equating outperformed IRT observed score equating in terms of preserving first-order equity, and IRT observed score equating is more successful in

preserving the same distributions property and second-order equity compared to the IRT true score method are in alignment with previous research under a random groups design (W. Lee et al., 2010; Tong & Kolen, 2005) and common-item nonequivalent groups design (Andrews, 2011; He, 2011; E. Lee et al., 2012). The conclusion that the chained equipercentile method outperformed the frequency estimation method in terms of first-order equity and same distribution property also has been confirmed by previous research (Duong, 2011; He, 2011).

The superior performance of the IRT equating methods was to be expected due to the application of IRT methodologies for data generation, item calibration and framework used to assess the equity properties; therefore, these methods likely had an advantage over the traditional equating methods (Andrews, 2011). Further, it is not surprising that the IRT true score method was superior in first-order equity preservation because the IRT true score method equates the expected scores conditioning on ability and the first-order equity property is defined in terms of the relationship between true scores on alternate test forms. This also explains why the IRT true score method did not perform well in terms of the same distribution property because the IRT true score method is not based on the observed score distributions whereas the IRT observed score method applies equipercentile equating to the observed score distributions of the test forms. Further, the equality of the traditional equating methods when groups taking the reference and new form are equivalent is not a new finding. Von Davier et al. (2004) showed mathematically that the frequency estimation and chained equipercentile methods can produce very similar equating functions when the group ability distributions are similar or when the scores on the anchor set are perfectly correlated with the scores on the total test.

5.2 RESEARCH QUESTION 2

Which of the above equating methods performs best in preserving equity properties when a complex multidimensional test structure is applied to a unidimensional equity framework?

In general, IRT methods outperformed traditional equating methods in terms of first-order equity, second-order equity, and same distributions property. The IRT observed score method produced slightly lower *T* values than the IRT true score method. However, the IRT true score method was more successful in preserving first-order and second-order equity compared to the IRT observed score method. In regards to the traditional equating methods, the chained equipercentile method outperformed the frequency estimation method in terms of the same distribution property and the preservation of the first-order equity whereas the frequency estimation method was superior in second-order equity preservation. In terms of first-order and second-order equity, the order from best to worst preserving was: IRT true score, IRT observed score, Chained Equipercentile, and Frequency Estimation, respectively. The order of values from lowest to highest for the same distribution property was as follows: IRT observed score, Chained Equipercentile, IRT True Score, and Frequency Estimation. Essentially identical patterns were observed between the unidimensional and multidimensional test structures with one exception: the IRT true score method produced slightly lower D_2 values compared to the IRT observed score method. Overall, similar results were obtained under each test structure.

Although a complex multidimensional model was used to generate the data, the equating methods produced similar results compared to the undimensional test structure. These results were expected because the bifactor model as applied in the current study is more or less

equivalent to a between-item multidimensional model with highly correlated factors. Yung, Mcleod, and Thissen (1999) showed mathematically that the bifactor model can be thought of as an unconstrained second-order factor model. All test items in the current study loaded on a general factor and, in addition, the MC items loaded on the MC subdimension whereas CR items loaded on the CR subdimension. The general factor and the MC and CR subdimensions were uncorrelated; therefore, the variance of the general factor exhibits the amount of the variability shared by all the items. The general factor in this study was more discriminating than the MC and CR dimensions, so the correlations among the different traits as found in a between-item multidimensional model were accounted for in the bifactor model due to the dominant presence of the general factor. Research has shown that equating methods still produce acceptable equating results as long as the correlation between the dimensions is high enough to ensure that the underlying IRT assumptions of essential undimensionality and local independence are satisfied (Camilli et al., 1995; Cook et al., 1988; Dorans & Kingston, 1985)

5.3 RESEARCH QUESTION 3

> *Which of the above equating methods performs best in preserving the equity properties when the common-item set format is representative and not representative of item types?*

In general, the common-item set composition did not have a statistically significant effect on the results. Based on the outcome of this study, it cannot be asserted that one method outperformed the other under the conditions of investigation because the practical significance is limited. Only general observations can be outlined.

A non-representative anchor set (10 MC items) resulted in the highest D_1 and D_2 values for all methods across all group difference conditions. When group differences were large, a representative anchor set still led to the lowest D_1 and D_2 values for most procedures. When the ability mean group differences varied across the dimensions, the inclusion of an additional CR item into the anchor set led to lower D_1 and D_2 for all methods except for the frequency estimation method. The lowest T values were obtained by all methods when the anchor set was representative and groups were equivalent (9 MC + 1 CR). Differences among the methods emerged as a function of an increase in the magnitude of the group differences. When group differences were large, the inclusion of a second CR item into the common-item set led to lower values for all methods regardless of test structure.

When group differences varied across the MC and CR items, the accuracy of the equating relationships in terms of *T* values was a direct function of the anchor item composition. For example, when the group ability was higher on the CR items compared to the MC items, then both methods produced lower *T* values when the anchor set consisted of 8 MC + 2 CR items rather than 9 MC + 1 CR item for the IRT methods. However, no such pattern was observed for the traditional equating methods. The IRT true score method produced the lowest values regardless of the common-item set composition, the underlying test structure, and group differences.

Previous studies investigating the effects of anchor set composition on the accuracy of equating results in terms of systematic and random error found mixed results. For example, Tate (2000) found that adequate equating results can be obtained when the anchor set consists of only MC items as long as the assumption of undimensionality is met. Other researchers found that the use of MC-only common items can lead to substantial bias (Walker & Kim, 2009) while

common-item sets containing both MC and CR items typically produce more adequate equating results, particularly if the assumption of unidimensionality is violated and/or the group ability distributions differ across MC and CR items (Cao, 2008; Kirkpatrick, 2005). This trend was observed in the current study, although the observed differences in this study were small and thus the findings may not represent meaningful practical differences. These observations suggest that the accuracy of the equating methods is dependent upon factors that go beyond the construction of the test form and the implementation of the guidelines in terms of the composition of the common-item set.

5.4 RESEARCH QUESTION 4

> *Which of the above equating methods is most accurate in preserving equity properties when groups associated with each form differ in ability?*

Overall, equating methods produced similar results when groups were equivalent under a unidimensional and multidimensional test structure. However, the equating methods differed more in terms of how well they preserved the equity properties as a result of increasing group differences. For all group differences under a unidimensional test structure, the IRT true score outperformed the IRT observed score method in terms of first-order equity whereas the second-order equity and same distribution property was better preserved by the IRT observed score method. These differences were also observed under a multidimensional test structure, except for the IRT true score method which produced lower D_2 values compared to the IRT observed score method.

In terms of the performance of the traditional equating methods, the chained equipercentile method outperformed the frequency estimation method in the preservation of first-order equity and same distribution property when group differences were small and large. These findings were consistent for a unidimensional and multidimensional test structure. The fact that the frequency estimation method performed poorly when groups become increasingly nonequivalent also has been confirmed by several researchers (Andrews, 2011; Duong, 2011; Wang et al., 2008). Possible explanations for these differences in the equating methods can be attributable to violations of assumptions of the equating design itself as well as the equating methods. For example, it is plausible that as reference and new form group ability differences increase, groups may differ in more ways than just on total test scores. Group differences might cause the anchor items to perform differently in relationship to the total test score. However, traditional and IRT methods assume that the common item to total test relationship remains constant across the groups.

The difference in performance between the two traditional methods likely can be attributed to the differences in the underlying assumptions of each method. The frequency estimation method makes a stringent assumption resting upon a synthetic population and the equivalence of the conditional cumulative score distributions involved in the equating process whereas the chained equipercentile method makes no clear definition of the test taking population. However, the chained equipercentile method assumes that the linking relationship between the alternate test forms and the common-item set is group invariant. When groups differ substantially it is unlikely that the assumption based on the synthetic population will hold for the frequency estimation method. For example, it is assumed that the conditional distribution of X on

the common-item set (V) in population Q is equivalent to the conditional distribution of X on V in Population S as shown in equation 5.1.

$$f(\mathrm{X}|\mathrm{V},\mathrm{Q}) = f(\mathrm{X}|\mathrm{V},\mathrm{S}) \tag{5.1}$$

Then let $f_Q(\theta)$ = cumulative distribution function of θ for population Q; $f(X|\theta)$ = score distribution of X given θ; and $f(V|\theta)$ = score distribution of the common-item set given θ. The conditional distribution of X on V and Q can be expressed as follows

$$f(\mathrm{X}|\mathrm{V},\mathrm{Q}) = \frac{\int f(X|\theta) f(V|\theta) df_Q(\theta)}{\int f(V|\theta) df_Q(\theta)} \tag{5.2}$$

Equation 5.2 shows that $f(X|\theta)$ is dependent upon $f_Q(\theta)$ which implies that $f(X|\theta)$ is most likely not population invariant when groups differ in ability. Therefore it is possible that

$$f(\mathrm{X}|\mathrm{V},\mathrm{Q}) \neq f(\mathrm{X}|\mathrm{V},\mathrm{S}), \tag{5.3}$$

in which case the assumption of the frequency estimation method does not hold.

However, the frequency estimation method produced lower D_2 values for all group differences under unidimensional and multidimensional test structures. Studies that examined the sensitivity of equating methods to group differences in terms of systematic and random error found that the chained method always was less biased than the frequency estimation method when group differences increased, a finding corroborated by (Holland, Sinharay, Von Davier, & Han, 2008). However, the frequency estimation method showed less random equating error compared to the chained equipercentile method (Sinharay & Holland, 2007; Wang et al., 2008). This could explain why the frequency estimation method produced lower D_2 values than the chained equipercentile method since second-order equity is concerned with measurement precision.

The fact that IRT methods appear less sensitive to group differences compared to the traditional equating methods may be explained through the IRT framework. In theory, item and

ability parameters are assumed to be population invariant; therefore, IRT equating satisfied the underlying assumptions of population invariance symmetry and equity regardless of group differences (Kolen & Brennan, 2004). However, it should be noted that the sensitivity to group differences can differ across examinations (Livingston, Dorans, & Wright, 1990) because large group differences could negatively impact the unidimensionality assumption as well as item parameter estimation, which, in turn, impacts the equating outcomes. In contrast to IRT methods, the classical equating methods are sample dependent; thus, the criterion of equity is less likely to be met.

An additional scenario was considered in that groups differed across MC and CR dimensions under the multidimensional test structure. Overall, IRT and traditional equating methods produced higher values in terms of equity preservation when the group differences were larger for the MC dimension. IRT true score and FE methods produced lower values when the group ability differences were higher on the MC dimension compared to the CR dimension. The opposite was true for the IRT observed score and chained equipercentile methods.

5.5 CONCLUSIONS AND PRACTICAL IMPLICATIONS

5.5.1 Choice of Equating Method

Comparisons between the IRT methods suggest that, if the goal is to ensure test fairness at the individual test taker level (First-Order Equity), then the IRT true score method would be preferred over the IRT observed score method. However, when the focus is on measurement

precision (Second-Order Equity) and test fairness in regards to the population of test takers (Same Distribution Property), the IRT observed score method produced lower values compared IRT true score method. Between the equipercentile equating methods, the chained method was more successful in preserving the first-order and the same distribution property compared to the frequency estimation method when group differences increased in magnitude. However, the frequency estimation method produced lower values in terms of second-order equity. Only small differences were found among the equating methods in terms of second-order equity for the unidimensional test structure; thus, these differences may have no practical significance. In other words, all equating methods under investigation performed similarly.

The results of the current study suggest that the selection of an equating method should take into consideration the magnitude of the group ability differences and the probability that the assumptions of particular equating methods have been violated because the quality of any equating method can vary depending upon the equating condition.

5.5.2 Test Structure

The within-item multidimensional test structure as modeled through the bifactor model produced similar results in terms of first-order, second-order, and same distributions property compared to the unidimensional test structure; therefore, it seems feasible to apply unidimensional equating methods presuming that the assumptions of essential undimensionality and local independence have been satisfied. This might be reassuring given that more innovative assessments could be anticipated in the near future in an attempt to measure 21st century skills. Innovative assessments could introduce different combinations of item formats and, as a result, the underlying test structure could be far more complex. This more complex structure could induce a

multidimensional test structure, thereby jeopardizing the accuracy of the equating methods that are based on unidimensional response data.

5.5.3 Composition of the Common-Item Set

The importance of the composition of the common-item set in the common-item nonequivalent groups design has long been recognized. This study showed that a representative common-item set led to lower evaluation indices, in particular when group differences increased in magnitude and when group differences varied across MC and CR dimensions. This finding became more apparent under the multidimensional test structure.

However, results of the ANOVA analysis indicated that these differences might not be meaningful; thus, interpretation of these results should be made with caution. In the current study, it can be assumed that the underlying psychometric models fit the data since the data were simulated; therefore, it is plausible that shifts in ability in any subset of items relative to the remaining items could influence the equating outcomes. This argument is supported by the general advice that the anchor set should be representative of the overall test because ability shifts on subgroups of items are then accounted for in the equating results, in particular when the test structure is multidimensional in nature. If the test structure is truly unidimensional, it can be assumed that different item formats measure the same construct; thus, the representativeness of the common-item set relative to the total test becomes less influential.

5.5.4 Group Differences

Mixed-format test equating under the common-item nonequivalent groups design has sparked an interest in the measurement community due in part to initiatives such as the Common Core State Standards and the Race to the Top, which called for assessments that are more authentic in nature in order to measure 21st century skills. Under the common-item nonequivalent group design, the population of examinees taking a test on different administration dates are typically not considered equivalent from the same population; therefore, before choosing an equating method it is imperative that group difference should be assessed as this will help determine which equating method to choose. If groups are similar, then either IRT true score, IRT observed score, frequency estimation, or chained equipercentile method can be utilized. However, when group differences exist, then the frequency estimation method is not recommended.

The inclusion of group differences under the multidimensional test structure led to similar conclusions compared to the unidimensional test structure when the groups were equivalent/nonequivalent and when the group differences in performance were held constant across dimensions. Two additional scenarios were considered where group differences could vary across dimensions. This situation was considered due to the practical concerns associated with equating mixed-format tests. For example, educators may modify instructional material to stress the content measured by one of the item formats more rigorously than the other item format. As a consequence, shifts in performance that are related to item format could be anticipated. This study showed that equity properties are less preserved under this condition. The composition of the common-item set did not have a statistically significant effect on the equating outcomes in this study. However, it can be concluded that any shift in performance on a subgroup of items may lead to different equating results depending on the composition of the

common-item set; therefore, the guidelines that exist to compose the anchor set are likely to produce different outcomes within different contexts.

6.0 LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

Despite the feasibility of using simulation studies to investigate the impact of a variety of factors simultaneously there is a major drawback in doing so in that the generalizability of the results is limited. It has long been recognized that simulation studies can't necessarily capture all features of operational test data.

A clear limitation of the present study is the heavy reliance on IRT methodology, which may have introduced bias into the evaluation criteria. Data were generated using IRT models and the IRT framework was used in assessing the equity properties. It is fair to conclude that when the evaluation criterion is based on a certain psychometric model it will give an advantage to the method that is based on the same psychometric model. Future research could apply a different psychometric model to the current equity framework.

Further, concurrent calibration was used in this study to place the item parameters onto a common scale through the common-item set. The comparison of different scale transformation methods in terms of equity preservation was not of interest in the current study. Concurrent calibration was chosen because it makes complete use of the available information and may remove some equating errors yielded by potentially inaccurate scale transformation procedures that are used by separate calibration. However, since group nonequivalence could negatively impact the accuracy of concurrent calibration, future research could examine the impact of separate scale linking methods on the preservation of equity properties.

Presmoothing through the application of a loglinear model was conducted to smooth out irregularities of the observed score distributions. Several models were fit in a trial run. A loglinear model that preserved the first three moments was chosen based on satisfactory results. The same model was used for all equipercentile equating methods to avoid effects due to the use of different loglinear models. While the choice of the presmoothing model appeared reasonable for these data, other smoothing options (e.g., postsmoothing) exist. Future research could examine the effect of different smoothing methods on the preservation of equity properties.

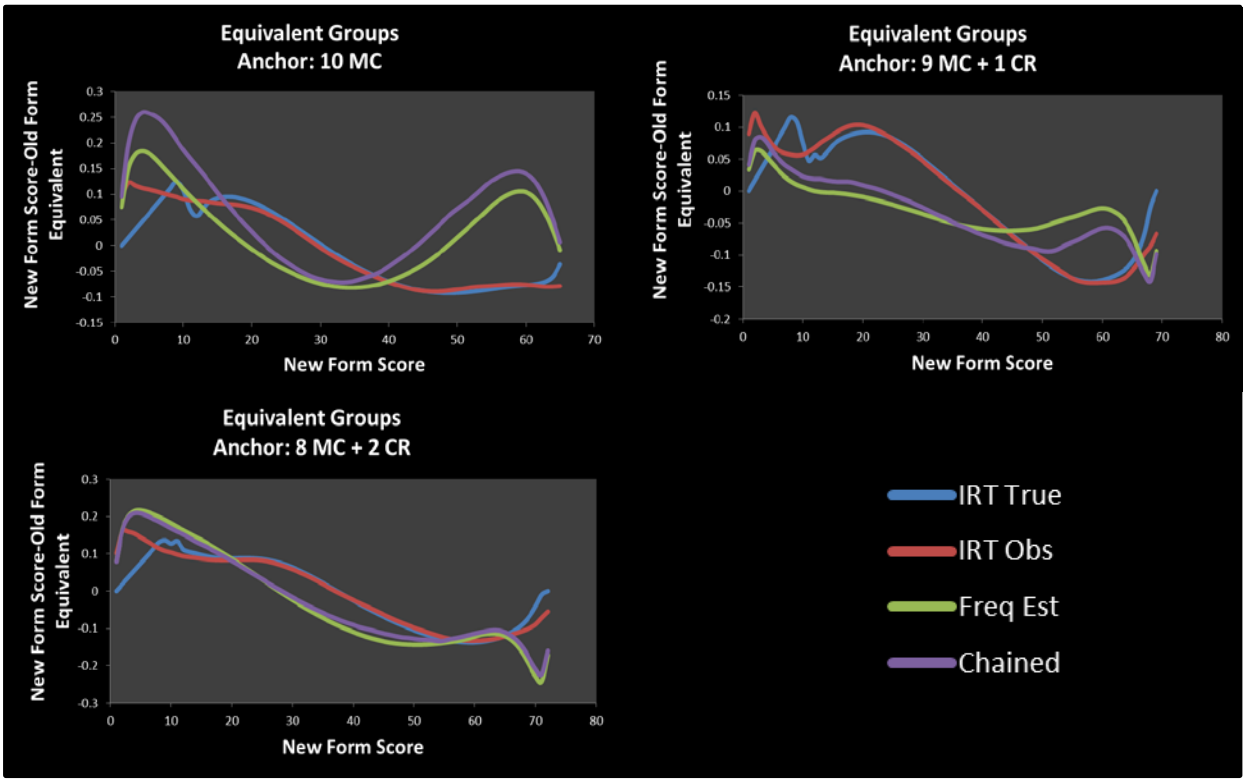
In the current study, the multidimensional test structure chosen was based on a bifactor model which was extended for a mixed-format test. It was assumed that the general construct had the main influence on item responses since the main purpose of most educational assessments is to measure the general construct of interest and because a condition in which the general construct is less informative than the subdimensions is not typical within educational settings. The MC and CR subdimensions provided similar discriminating power in this study. Future research could investigate the impact of the strength of the subdimensions on the preservation of equity properties as well as the discriminating power between the MC and CR subdimensions could be varied. Assuming that more innovative assessments are on the rise, statistical and content related pieces of evidence could be gathered to arrive at a different test structure. For example, it would be interesting to generate data based on a multi-trait multi-method model and examine the preservation of equity properties under this test structure. In this instance, the mixed-format test could include different item formats such as MC, short free response items, and long free response items.

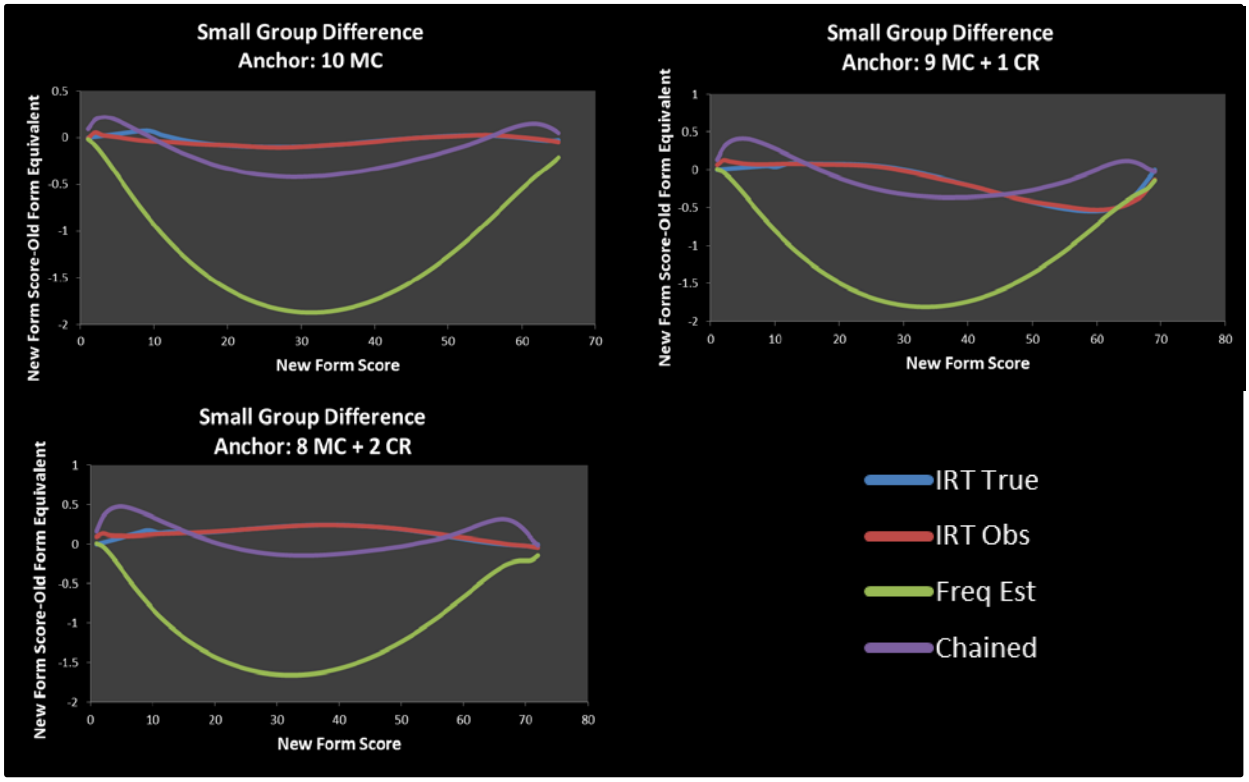
Additionaly, this study used a horizontal equating framework. However, the bifactor model as was utilized in the current study also could be applied within a vertical scaling context; therefore, it could be of interest to to assess the equity properties within this framework.

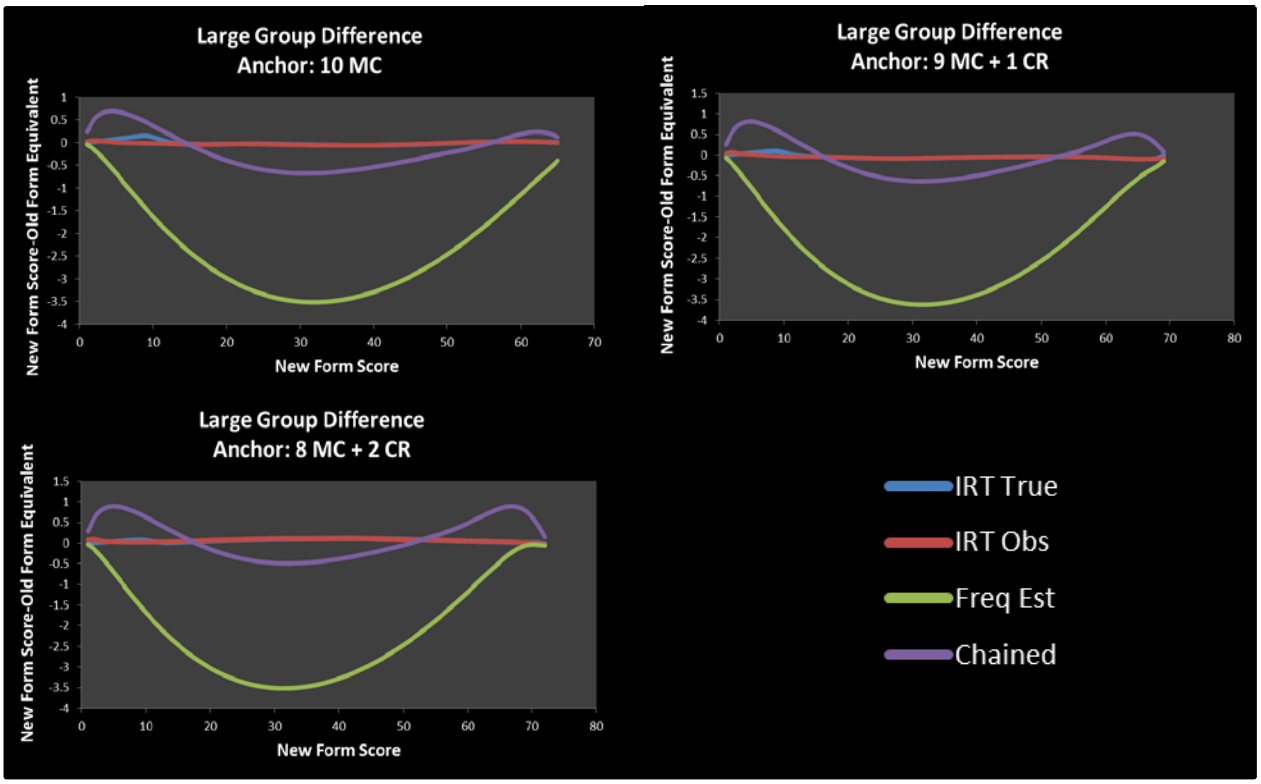
In the present study, the synthetic group was defined as the group taking the new form. Future studies could investigate whether the definition of the synthetic group has an effect on the preservation of equity property. Other factors that could be studied include different equating methods (such as linear methods and kernel equating methods), sample sizes, form differences, alternative indices to evaluate equating properties, and statistical and content representativeness of the anchor set.

Appendix A

EQUATING FUNCTIONS FOR EACH CONDITION AVERAGED ACROSS 100 REPLICATIONS UNDER UNIDIMENSIONAL TEST STRUCTURE

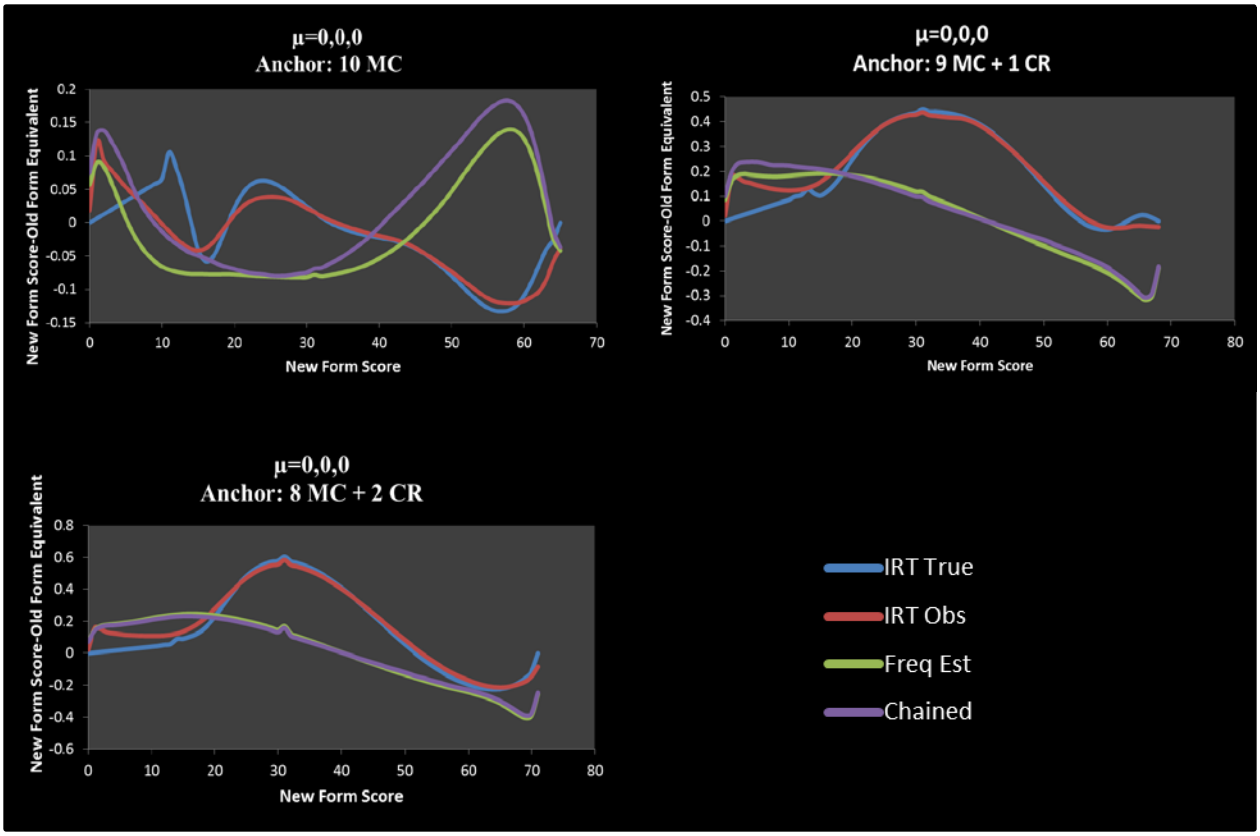


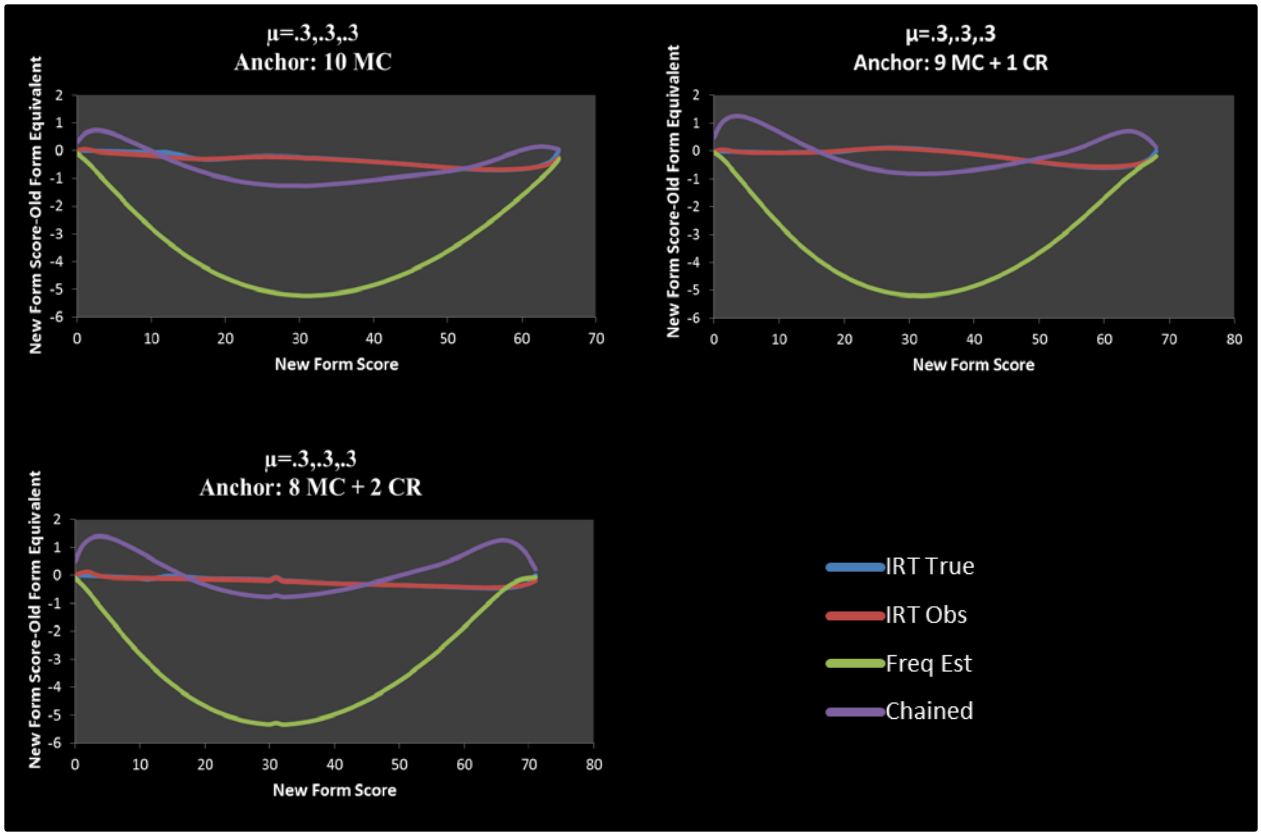

Small Group Difference
Anchor: 10 MC
New Form Score-Old Form Equivalent
New Form Score
Small Group Difference
Anchor: 9 MC + 1 CR
New Form Score-Old Form Equivalent
New Form Score
Small Group Difference
Anchor: 8 MC + 2 CR
New Form Score-Old Form Equivalent
New Form Score
IRT True
IRT Obs
Freq Est
Chained

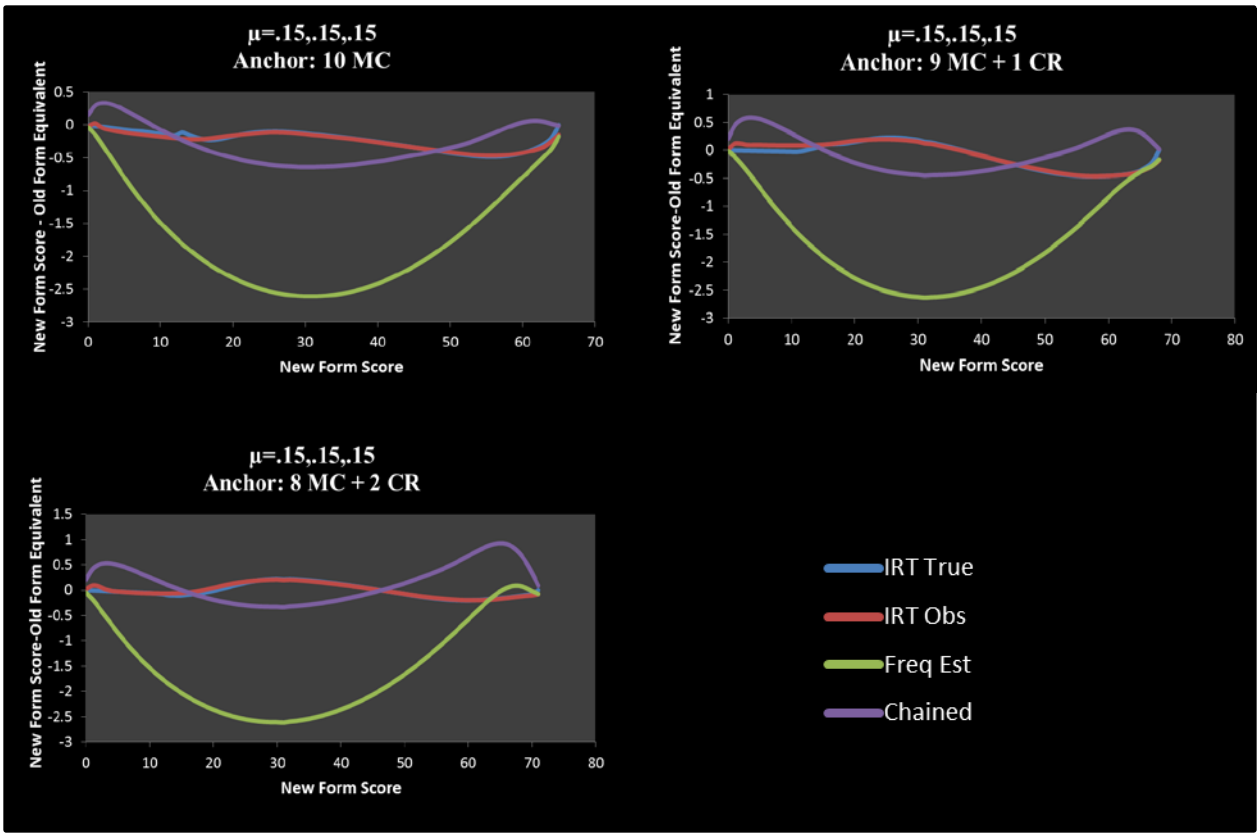

Large Group Difference
Anchor: 10 MC
New Form Score-Old Form Equivalent
New Form Score
Large Group Difference
Anchor: 9 MC + 1 CR
New Form Score-Old Form Equivalent
New Form Score
Large Group Difference
Anchor: 8 MC + 2 CR
New Form Score-Old Form Equivalent
New Form Score
IRT True
IRT Obs
Freq Est
Chained

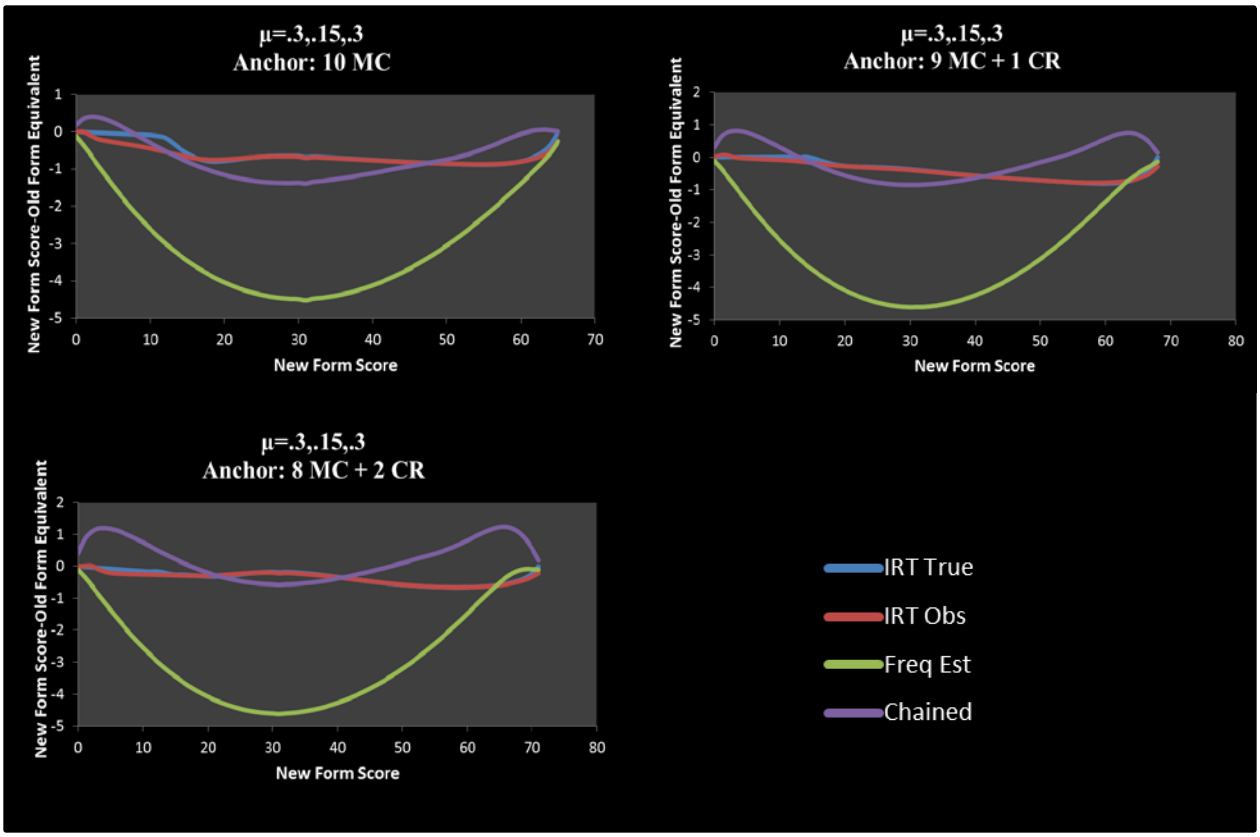
Appendix B

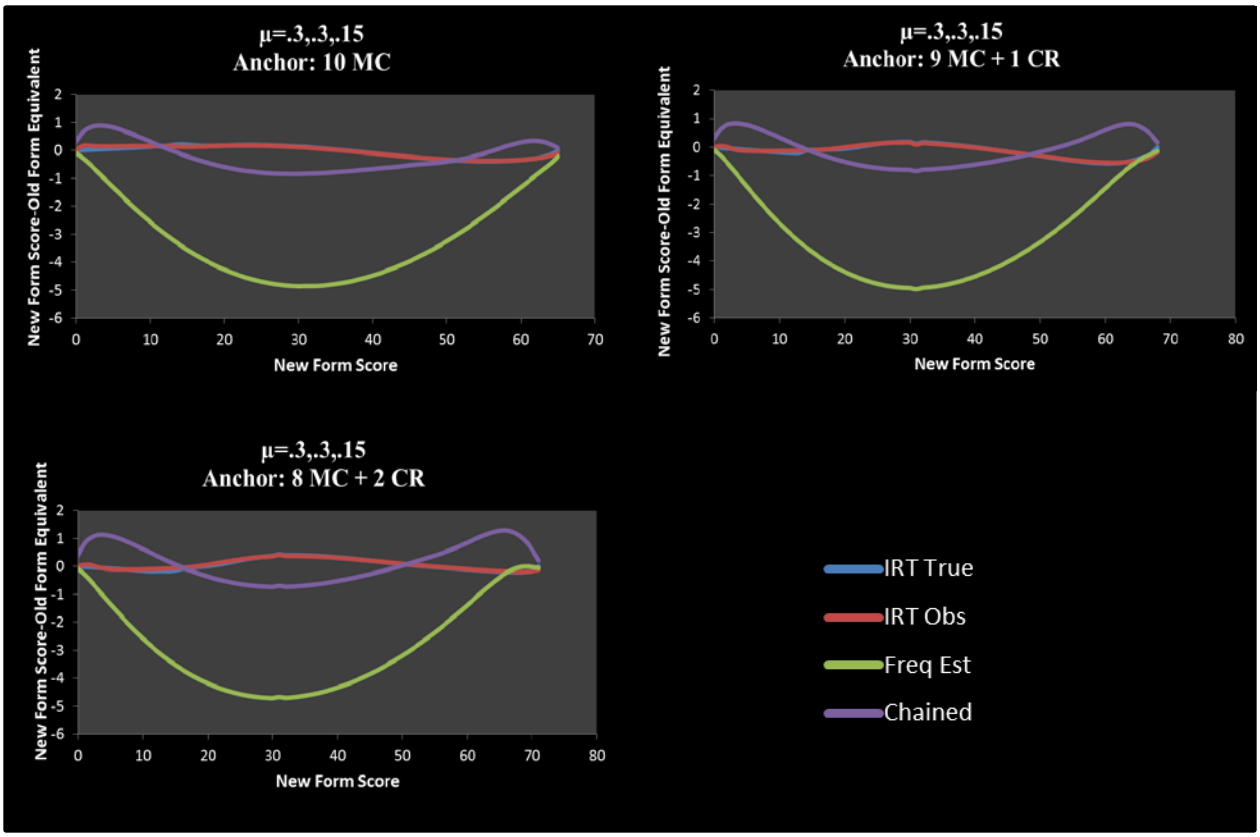
EQUATING FUNCTIONS FOR EACH CONDITION AVERAGED ACROSS 100 REPLICATIONS UNDER MULTIDIMENSIONAL TEST STRUCTURE


μ=0,0,0
Anchor: 10 MC
μ=0,0,0
Anchor: 9 MC + 1 CR
μ=0,0,0
Anchor: 8 MC + 2 CR
New Form Score-Old Form Equivalent
New Form Score
IRT True
IRT Obs
Freq Est
Chained


μ=.3,.3,.3
Anchor: 10 MC
μ=.3,.3,.3
Anchor: 9 MC + 1 CR
μ=.3,.3,.3
Anchor: 8 MC + 2 CR
New Form Score-Old Form Equivalent
New Form Score
IRT True
IRT Obs
Freq Est
Chained


μ=.15,.15,.15
Anchor: 10 MC
New Form Score - Old Form Equivalent
New Form Score
μ=.15,.15,.15
Anchor: 9 MC + 1 CR
New Form Score-Old Form Equivalent
New Form Score
μ=.15,.15,.15
Anchor: 8 MC + 2 CR
New Form Score-Old Form Equivalent
New Form Score
IRT True
IRT Obs
Freq Est
Chained


μ=.3,.15,.3
Anchor: 10 MC
New Form Score-Old Form Equivalent
New Form Score
μ=.3,.15,.3
Anchor: 9 MC + 1 CR
New Form Score-Old Form Equivalent
New Form Score
μ=.3,.15,.3
Anchor: 8 MC + 2 CR
New Form Score-Old Form Equivalent
New Form Score
IRT True
IRT Obs
Freq Est
Chained


μ=.3,.3,.15
Anchor: 10 MC
New Form Score-Old Form Equivalent
New Form Score
μ=.3,.3,.15
Anchor: 9 MC + 1 CR
New Form Score-Old Form Equivalent
New Form Score
μ=.3,.3,.15
Anchor: 8 MC + 2 CR
New Form Score-Old Form Equivalent
New Form Score
IRT True
IRT Obs
Freq Est
Chained

Appendix C

SAS SYNTAX

```
ods results off;
ods graphics off;
ods listing;

proc printto log=log print=print new;
run;


%LET workdir=c:\raw59 ;
options noxwait ;

data junk524 ;
input newid type $ numresp c1-c5 ;
datalines ;
0 TT 0 . . . . .
1 L3 2 0 1 . . .
2 L3 2 0 1 . . .
3 L3 2 0 1 . . .
4 L3 2 0 1 . . .
5 L3 2 0 1 . . .
6 L3 2 0 1 . . .
7 L3 2 0 1 . . .
8 L3 2 0 1 . . .
9 L3 2 0 1 . . .
10 L3 2 0 1 . . .
11 L3 2 0 1 . . .
12 L3 2 0 1 . . .
13 L3 2 0 1 . . .
14 L3 2 0 1 . . .
15 L3 2 0 1 . . .
16 L3 2 0 1 . . .
17 L3 2 0 1 . . .
18 L3 2 0 1 . . .
19 L3 2 0 1 . . .
20 L3 2 0 1 . . .
21 L3 2 0 1 . . .
```

```
22 L3 2 0 1 . . .
23 L3 2 0 1 . . .
24 L3 2 0 1 . . .
25 L3 2 0 1 . . .
26 L3 2 0 1 . . .
27 L3 2 0 1 . . .
28 L3 2 0 1 . . .
29 L3 2 0 1 . . .
30 L3 2 0 1 . . .
31 L3 2 0 1 . . .
32 L3 2 0 1 . . .
33 L3 2 0 1 . . .
34 L3 2 0 1 . . .
35 L3 2 0 1 . . .
36 SL 5 0 1 2 3 4
37 SL 5 0 1 2 3 4
38 SL 5 0 1 2 3 4
39 SL 5 0 1 2 3 4
40 SL 5 0 1 2 3 4
41 L3 2 0 1 . . .
42 L3 2 0 1 . . .
43 L3 2 0 1 . . .
44 L3 2 0 1 . . .
45 L3 2 0 1 . . .
46 L3 2 0 1 . . .
47 L3 2 0 1 . . .
48 L3 2 0 1 . . .
49 L3 2 0 1 . . .
50 L3 2 0 1 . . .
 ;
run ;

data rawscore_sc ;
input r1-r3 ;
datalines ;
0 65 0
0 0 0
1 1 1
2 2 2
3 3 3
4 4 4
5 5 5
6 6 6
7 7 7
8 8 8
9 9 9
10 10 10
11 11 11
12 12 12
13 13 13
14 14 14
15 15 15
16 16 16
17 17 17
18 18 18
19 19 19
20 20 20
```

```
21 21 21
22 22 22
23 23 23
24 24 24
25 25 25
26 26 26
27 27 27
28 28 28
29 29 29
30 30 30
31 31 31
32 32 32
33 33 33
34 34 34
35 35 35
36 36 36
37 37 37
38 38 38
39 39 39
40 40 40
41 41 41
42 42 42
43 43 43
44 44 44
45 45 45
46 46 46
47 47 47
48 48 48
49 49 49
50 50 50
51 51 51
52 52 52
53 53 53
54 54 54
55 55 55
56 56 56
57 57 57
58 58 58
59 59 59
60 60 60
61 61 61
62 62 62
63 63 63
64 64 64
65 65 65
 ;
run ;

data _null_ ;
set rawscore_sc ;
file "&workdir\raw_c1" ;
put @1 r1 @5 r2 @10 r3 ;
run ;

data _null_ ;
file "c:\raw59\control_con.txt" ;
put @1 'output dissertation.out' ;
```

```
put @1 'old_para' @15 'testrun' ;
put @1 'old_quad' @15 'qfunx1' ;
put @1 'raw2scale' @15 'raw_c1' ;
put @1 'new_para' @15 'testruny' ;
put @1 'new_quad' @15 'qfuny1' ;
put @1 'end' ;
run ;

data polycsem ;
input item type $ numresp c1-c5 ;
datalines ;
0 TT 0 . . . . .
1 L3 2 0 1 . . .
2 L3 2 0 1 . . .
3 L3 2 0 1 . . .
4 L3 2 0 1 . . .
5 L3 2 0 1 . . .
6 L3 2 0 1 . . .
7 L3 2 0 1 . . .
8 L3 2 0 1 . . .
9 L3 2 0 1 . . .
10 L3 2 0 1 . . .
11 L3 2 0 1 . . .
12 L3 2 0 1 . . .
13 L3 2 0 1 . . .
14 L3 2 0 1 . . .
15 L3 2 0 1 . . .
16 L3 2 0 1 . . .
17 L3 2 0 1 . . .
18 L3 2 0 1 . . .
19 L3 2 0 1 . . .
20 L3 2 0 1 . . .
21 L3 2 0 1 . . .
22 L3 2 0 1 . . .
23 L3 2 0 1 . . .
24 L3 2 0 1 . . .
25 L3 2 0 1 . . .
26 L3 2 0 1 . . .
27 L3 2 0 1 . . .
28 L3 2 0 1 . . .
29 L3 2 0 1 . . .
30 L3 2 0 1 . . .
31 L3 2 0 1 . . .
32 L3 2 0 1 . . .
33 L3 2 0 1 . . .
34 L3 2 0 1 . . .
35 L3 2 0 1 . . .
36 SL 5 0 1 2 3 4
37 SL 5 0 1 2 3 4
38 SL 5 0 1 2 3 4
39 SL 5 0 1 2 3 4
40 SL 5 0 1 2 3 4
41 L3 2 0 1 . . .
42 L3 2 0 1 . . .
43 L3 2 0 1 . . .
44 L3 2 0 1 . . .
45 L3 2 0 1 . . .
```

```
46 L3 2 0 1 . . .
47 L3 2 0 1 . . .
48 L3 2 0 1 . . .
49 L3 2 0 1 . . .
50 L3 2 0 1 . . .
 ;
run ;

data rawscore_con1 ;
input r1-r3 ;
datalines ;
0 0 0
1 1 1
2 2 2
3 3 3
4 4 4
5 5 5
6 6 6
7 7 7
8 8 8
9 9 9
10 10 10
11 11 11
12 12 12
13 13 13
14 14 14
15 15 15
16 16 16
17 17 17
18 18 18
19 19 19
20 20 20
21 21 21
22 22 22
23 23 23
24 24 24
25 25 25
26 26 26
27 27 27
28 28 28
29 29 29
30 30 30
31 31 31
32 32 32
33 33 33
34 34 34
35 35 35
36 36 36
37 37 37
38 38 38
39 39 39
40 40 40
41 41 41
42 42 42
43 43 43
44 44 44
45 45 45
```

```
46 46 46
47 47 47
48 48 48
49 49 49
50 50 50
51 51 51
52 52 52
53 53 53
54 54 54
55 55 55
56 56 56
57 57 57
58 58 58
59 59 59
60 60 60
61 61 61
62 62 62
63 63 63
64 64 64
65 65 65
 ;
run ;

options nonumber nodate nocenter  PS = 100;
title 'Raw Score Equivalents Form X' ;

data _null_ ;
set rawscore_con1 ;
file "&workdir\rawscx1.rs" print title ;
put @1 r1 @5 r2 @10 r3 ;
run ;

data _null_ ;
file "&workdir\control_poly1x.txt" ;
put @1 'output disx.out' ;
put @1 'para' @15 'testrun' ;
put @1 'raw2scale' @15 'rawscx1.rs' ;
put @1 'quad' @15 'funx1.pst' ;
put @1 'end' ;
run ;

data _null_ ;
file "&workdir\control_polyy1.txt" ;
put @1 'output  disyfr.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy1.rs' ;
put @1 'quad' @15 'funy1.pst' ;
put @1 'end' ;
run ;

data _null_ ;
file "&workdir\control_polyy2.txt" ;
put @1 'output  disych.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy2.rs' ;
put @1 'quad' @15 'funy1.pst' ;
put @1 'end' ;
```

```
run ;
data _null_ ;
file "&workdir\control_polyy3.txt" ;
put @1 'output  disytr.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy3.rs' ;
put @1 'quad' @15 'funy1.pst' ;
put @1 'end' ;
run ;
data _null_ ;
file "&workdir\control_polyy4.txt" ;
put @1 'output  disyob.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy4.rs' ;
put @1 'quad' @15 'funy1.pst' ;
put @1 'end' ;
run ;

data fun524 ;
input newid type $ numresp c1-c5 ;
datalines ;
0 TT 0 . . . . .
1 L3 2 0 1 . . .
2 L3 2 0 1 . . .
3 L3 2 0 1 . . .
4 L3 2 0 1 . . .
5 L3 2 0 1 . . .
6 L3 2 0 1 . . .
7 L3 2 0 1 . . .
8 L3 2 0 1 . . .
9 L3 2 0 1 . . .
10 L3 2 0 1 . . .
11 L3 2 0 1 . . .
12 L3 2 0 1 . . .
13 L3 2 0 1 . . .
14 L3 2 0 1 . . .
15 L3 2 0 1 . . .
16 L3 2 0 1 . . .
17 L3 2 0 1 . . .
18 L3 2 0 1 . . .
19 L3 2 0 1 . . .
20 L3 2 0 1 . . .
21 L3 2 0 1 . . .
22 L3 2 0 1 . . .
23 L3 2 0 1 . . .
24 L3 2 0 1 . . .
25 L3 2 0 1 . . .
26 L3 2 0 1 . . .
27 L3 2 0 1 . . .
28 L3 2 0 1 . . .
29 L3 2 0 1 . . .
30 L3 2 0 1 . . .
31 L3 2 0 1 . . .
32 L3 2 0 1 . . .
33 L3 2 0 1 . . .
34 L3 2 0 1 . . .
35 L3 2 0 1 . . .
```

```
36 SL 5 0 1 2 3 4
37 SL 5 0 1 2 3 4
38 SL 5 0 1 2 3 4
39 SL 5 0 1 2 3 4
40 SL 5 0 1 2 3 4
41 L3 2 0 1 . . .
42 L3 2 0 1 . . .
43 L3 2 0 1 . . .
44 L3 2 0 1 . . .
45 L3 2 0 1 . . .
46 L3 2 0 1 . . .
47 L3 2 0 1 . . .
48 L3 2 0 1 . . .
49 L3 2 0 1 . . .
50 SL 5 0 1 2 3 4
 ;
run ;


data rawscore_cond2 ;
input r1-r3 ;
datalines ;
0 68 0
0 0 0
1 1 1
2 2 2
3 3 3
4 4 4
5 5 5
6 6 6
7 7 7
8 8 8
9 9 9
10 10 10
11 11 11
12 12 12
13 13 13
14 14 14
15 15 15
16 16 16
17 17 17
18 18 18
19 19 19
20 20 20
21 21 21
22 22 22
23 23 23
24 24 24
25 25 25
26 26 26
27 27 27
28 28 28
29 29 29
30 30 30
31 31 31
32 32 32
33 33 33
```

```
34 34 34
35 35 35
36 36 36
37 37 37
38 38 38
39 39 39
40 40 40
41 41 41
42 42 42
43 43 43
44 44 44
45 45 45
46 46 46
47 47 47
48 48 48
49 49 49
50 50 50
51 51 51
52 52 52
53 53 53
54 54 54
55 55 55
56 56 56
57 57 57
58 58 58
59 59 59
60 60 60
61 61 61
62 62 62
63 63 63
64 64 64
65 65 65
66 66 66
67 67 67
68 68 68
 ;
run ;

data _null_ ;
set rawscore_cond2 ;
file "&workdir\raw_c2" ;
put @1 r1 @5 r2 @10 r3 ;
run ;


data _null_ ;
file "&workdir\control_con2.txt" ;
put @1 'output dissertation2.out' ;
put @1 'old_para' @15 'testrun' ;
put @1 'old_quad' @15 'qfunx2' ;
put @1 'raw2scale' @15 'raw_c2' ;
put @1 'new_para' @15 'testruny' ;
put @1 'new_quad' @15 'qfuny2' ;
put @1 'end' ;
run ;

data polycsem2 ;
```

```
input item type $ numresp c1-c5 ;
datalines ;
0 TT 0 . . . . .
1 L3 2 0 1 . . .
2 L3 2 0 1 . . .
3 L3 2 0 1 . . .
4 L3 2 0 1 . . .
5 L3 2 0 1 . . .
6 L3 2 0 1 . . .
7 L3 2 0 1 . . .
8 L3 2 0 1 . . .
9 L3 2 0 1 . . .
10 L3 2 0 1 . . .
11 L3 2 0 1 . . .
12 L3 2 0 1 . . .
13 L3 2 0 1 . . .
14 L3 2 0 1 . . .
15 L3 2 0 1 . . .
16 L3 2 0 1 . . .
17 L3 2 0 1 . . .
18 L3 2 0 1 . . .
19 L3 2 0 1 . . .
20 L3 2 0 1 . . .
21 L3 2 0 1 . . .
22 L3 2 0 1 . . .
23 L3 2 0 1 . . .
24 L3 2 0 1 . . .
25 L3 2 0 1 . . .
26 L3 2 0 1 . . .
27 L3 2 0 1 . . .
28 L3 2 0 1 . . .
29 L3 2 0 1 . . .
30 L3 2 0 1 . . .
31 L3 2 0 1 . . .
32 L3 2 0 1 . . .
33 L3 2 0 1 . . .
34 L3 2 0 1 . . .
35 L3 2 0 1 . . .
36 SL 5 0 1 2 3 4
37 SL 5 0 1 2 3 4
38 SL 5 0 1 2 3 4
39 SL 5 0 1 2 3 4
40 SL 5 0 1 2 3 4
41 L3 2 0 1 . . .
42 L3 2 0 1 . . .
43 L3 2 0 1 . . .
44 L3 2 0 1 . . .
45 L3 2 0 1 . . .
46 L3 2 0 1 . . .
47 L3 2 0 1 . . .
48 L3 2 0 1 . . .
49 L3 2 0 1 . . .
50 SL 5 0 1 2 3 4
 ;
run ;

data rawscore_con2 ;
```

```
input r1-r3 ;
datalines ;
0 0 0
1 1 1
2 2 2
3 3 3
4 4 4
5 5 5
6 6 6
7 7 7
8 8 8
9 9 9
10 10 10
11 11 11
12 12 12
13 13 13
14 14 14
15 15 15
16 16 16
17 17 17
18 18 18
19 19 19
20 20 20
21 21 21
22 22 22
23 23 23
24 24 24
25 25 25
26 26 26
27 27 27
28 28 28
29 29 29
30 30 30
31 31 31
32 32 32
33 33 33
34 34 34
35 35 35
36 36 36
37 37 37
38 38 38
39 39 39
40 40 40
41 41 41
42 42 42
43 43 43
44 44 44
45 45 45
46 46 46
47 47 47
48 48 48
49 49 49
50 50 50
51 51 51
52 52 52
53 53 53
54 54 54
```

```
55 55 55
56 56 56
57 57 57
58 58 58
59 59 59
60 60 60
61 61 61
62 62 62
63 63 63
64 64 64
65 65 65
66 66 66
67 67 67
68 68 68
 ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents Form X" ;
data _null_ ;
set rawscore_con2 ;
file "&workdir\rawscx2.rs" print title ;
put @1 r1 @5 r2 @10 r3 ;
run ;

data _null_ ;
file "&workdir\control_poly2x.txt" ;
put @1 'output disx.out' ;
put @1 'para' @15 'testrun' ;
put @1 'raw2scale' @15 'rawscx2.rs' ;
put @1 'quad' @15 'funx2.pst' ;
put @1 'end' ;
run ;

data polycsemy2 ;
input item type $ numresp c1-c5 ;
datalines ;
0 TT 0 . . . . .
1 L3 2 0 1 . . .
2 L3 2 0 1 . . .
3 L3 2 0 1 . . .
4 L3 2 0 1 . . .
5 L3 2 0 1 . . .
6 L3 2 0 1 . . .
7 L3 2 0 1 . . .
8 L3 2 0 1 . . .
9 L3 2 0 1 . . .
10 L3 2 0 1 . . .
11 L3 2 0 1 . . .
12 L3 2 0 1 . . .
13 L3 2 0 1 . . .
14 L3 2 0 1 . . .
15 L3 2 0 1 . . .
16 L3 2 0 1 . . .
17 L3 2 0 1 . . .
18 L3 2 0 1 . . .
19 L3 2 0 1 . . .
```

```
20 L3 2 0 1 . . .
21 L3 2 0 1 . . .
22 L3 2 0 1 . . .
23 L3 2 0 1 . . .
24 L3 2 0 1 . . .
25 L3 2 0 1 . . .
26 L3 2 0 1 . . .
27 L3 2 0 1 . . .
28 L3 2 0 1 . . .
29 L3 2 0 1 . . .
30 L3 2 0 1 . . .
31 L3 2 0 1 . . .
32 L3 2 0 1 . . .
33 L3 2 0 1 . . .
34 L3 2 0 1 . . .
35 L3 2 0 1 . . .
36 SL 5 0 1 2 3 4
37 SL 5 0 1 2 3 4
38 SL 5 0 1 2 3 4
39 SL 5 0 1 2 3 4
40 SL 5 0 1 2 3 4
41 L3 2 0 1 . . .
42 L3 2 0 1 . . .
43 L3 2 0 1 . . .
44 L3 2 0 1 . . .
45 L3 2 0 1 . . .
46 L3 2 0 1 . . .
47 L3 2 0 1 . . .
48 L3 2 0 1 . . .
49 L3 2 0 1 . . .
50 SL 5 0 1 2 3 4
 ;
run ;

data rawscore_con2 ;
input r1-r3 ;
datalines ;
0 0 0
1 1 1
2 2 2
3 3 3
4 4 4
5 5 5
6 6 6
7 7 7
8 8 8
9 9 9
10 10 10
11 11 11
12 12 12
13 13 13
14 14 14
15 15 15
16 16 16
17 17 17
18 18 18
19 19 19
```

```
20 20 20
21 21 21
22 22 22
23 23 23
24 24 24
25 25 25
26 26 26
27 27 27
28 28 28
29 29 29
30 30 30
31 31 31
32 32 32
33 33 33
34 34 34
35 35 35
36 36 36
37 37 37
38 38 38
39 39 39
40 40 40
41 41 41
42 42 42
43 43 43
44 44 44
45 45 45
46 46 46
47 47 47
48 48 48
49 49 49
50 50 50
51 51 51
52 52 52
53 53 53
54 54 54
55 55 55
56 56 56
57 57 57
58 58 58
59 59 59
60 60 60
61 61 61
62 62 62
63 63 63
64 64 64
65 65 65
66 66 66
67 67 67
68 68 68
 ;
run ;

data _null_ ;
file "&workdir\control_poly2x.txt" ;
put @1 'output disx.out' ;
put @1 'para' @15 'testrun' ;
put @1 'raw2scale' @15 'rawscx2.rs' ;
```

```
put @1 'quad' @15 'funx2.pst' ;
put @1 'end' ;
run ;

data _null_ ;
file "&workdir\control_polyy12.txt" ;
put @1 'output  disyfr.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy12.rs' ;
put @1 'quad' @15 'funy2.pst' ;
put @1 'end' ;
run ;

data _null_ ;
file "&workdir\control_polyy22.txt" ;
put @1 'output  disych.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy22.rs' ;
put @1 'quad' @15 'funy2.pst' ;
put @1 'end' ;
run ;
data _null_ ;
file "&workdir\control_polyy32.txt" ;
put @1 'output  disytr.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy32.rs' ;
put @1 'quad' @15 'funy2.pst' ;
put @1 'end' ;
run ;
data _null_ ;
file "&workdir\control_polyy42.txt" ;
put @1 'output  disyob.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy42.rs' ;
put @1 'quad' @15 'funy2.pst' ;
put @1 'end' ;
run ;

data fun613 ;
input newid type $ numresp c1-c5 ;
datalines ;
0 TT 0 . . . . .
1 L3 2 0 1 . . .
2 L3 2 0 1 . . .
3 L3 2 0 1 . . .
4 L3 2 0 1 . . .
5 L3 2 0 1 . . .
6 L3 2 0 1 . . .
7 L3 2 0 1 . . .
8 L3 2 0 1 . . .
9 L3 2 0 1 . . .
10 L3 2 0 1 . . .
11 L3 2 0 1 . . .
12 L3 2 0 1 . . .
13 L3 2 0 1 . . .
14 L3 2 0 1 . . .
15 L3 2 0 1 . . .
```

```
16 L3 2 0 1 . . .
17 L3 2 0 1 . . .
18 L3 2 0 1 . . .
19 L3 2 0 1 . . .
20 L3 2 0 1 . . .
21 L3 2 0 1 . . .
22 L3 2 0 1 . . .
23 L3 2 0 1 . . .
24 L3 2 0 1 . . .
25 L3 2 0 1 . . .
26 L3 2 0 1 . . .
27 L3 2 0 1 . . .
28 L3 2 0 1 . . .
29 L3 2 0 1 . . .
30 L3 2 0 1 . . .
31 L3 2 0 1 . . .
32 L3 2 0 1 . . .
33 L3 2 0 1 . . .
34 L3 2 0 1 . . .
35 L3 2 0 1 . . .
36 SL 5 0 1 2 3 4
37 SL 5 0 1 2 3 4
38 SL 5 0 1 2 3 4
39 SL 5 0 1 2 3 4
40 SL 5 0 1 2 3 4
41 L3 2 0 1 . . .
42 L3 2 0 1 . . .
43 L3 2 0 1 . . .
44 L3 2 0 1 . . .
45 L3 2 0 1 . . .
46 L3 2 0 1 . . .
47 L3 2 0 1 . . .
48 L3 2 0 1 . . .
49 SL 5 0 1 2 3 4
50 SL 5 0 1 2 3 4
 ;
run ;

data rawscore_sc ;
input r1-r3 ;
datalines ;
0 71 0
0 0 0
1 1 1
2 2 2
3 3 3
4 4 4
5 5 5
6 6 6
7 7 7
8 8 8
9 9 9
10 10 10
11 11 11
12 12 12
13 13 13
14 14 14
```

```
15 15 15
16 16 16
17 17 17
18 18 18
19 19 19
20 20 20
21 21 21
22 22 22
23 23 23
24 24 24
25 25 25
26 26 26
27 27 27
28 28 28
29 29 29
30 30 30
31 31 31
32 32 32
33 33 33
34 34 34
35 35 35
36 36 36
37 37 37
38 38 38
39 39 39
40 40 40
41 41 41
42 42 42
43 43 43
44 44 44
45 45 45
46 46 46
47 47 47
48 48 48
49 49 49
50 50 50
51 51 51
52 52 52
53 53 53
54 54 54
55 55 55
56 56 56
57 57 57
58 58 58
59 59 59
60 60 60
61 61 61
62 62 62
63 63 63
64 64 64
65 65 65
66 66 66
67 67 67
68 68 68
69 69 69
70 70 70
71 71 71
```

```
 ;
run ;

data _null_ ;
set rawscore_sc ;
file "&workdir\raw_c3" ;
put @1 r1 @5 r2 @10 r3 ;
run ;

/*create control file*/

data _null_ ;
file "&workdir\control_con3.txt" ;
put @1 'output dissertation3.out' ;
put @1 'old_para' @15 'testrun' ;
put @1 'old_quad' @15 'qfunx3' ;
put @1 'raw2scale' @15 'raw_c3' ;
put @1 'new_para' @15 'testruny' ;
put @1 'new_quad' @15 'qfuny3' ;
put @1 'end' ;
run ;

data polycsem3 ;
input item type $ numresp c1-c5 ;
datalines ;
0 TT 0 . . . . .
1 L3 2 0 1 . . .
2 L3 2 0 1 . . .
3 L3 2 0 1 . . .
4 L3 2 0 1 . . .
5 L3 2 0 1 . . .
6 L3 2 0 1 . . .
7 L3 2 0 1 . . .
8 L3 2 0 1 . . .
9 L3 2 0 1 . . .
10 L3 2 0 1 . . .
11 L3 2 0 1 . . .
12 L3 2 0 1 . . .
13 L3 2 0 1 . . .
14 L3 2 0 1 . . .
15 L3 2 0 1 . . .
16 L3 2 0 1 . . .
17 L3 2 0 1 . . .
18 L3 2 0 1 . . .
19 L3 2 0 1 . . .
20 L3 2 0 1 . . .
21 L3 2 0 1 . . .
22 L3 2 0 1 . . .
23 L3 2 0 1 . . .
24 L3 2 0 1 . . .
25 L3 2 0 1 . . .
26 L3 2 0 1 . . .
27 L3 2 0 1 . . .
28 L3 2 0 1 . . .
29 L3 2 0 1 . . .
30 L3 2 0 1 . . .
31 L3 2 0 1 . . .
```

```
32 L3 2 0 1 . . .
33 L3 2 0 1 . . .
34 L3 2 0 1 . . .
35 L3 2 0 1 . . .
36 SL 5 0 1 2 3 4
37 SL 5 0 1 2 3 4
38 SL 5 0 1 2 3 4
39 SL 5 0 1 2 3 4
40 SL 5 0 1 2 3 4
41 L3 2 0 1 . . .
42 L3 2 0 1 . . .
43 L3 2 0 1 . . .
44 L3 2 0 1 . . .
45 L3 2 0 1 . . .
46 L3 2 0 1 . . .
47 L3 2 0 1 . . .
48 L3 2 0 1 . . .
49 SL 5 0 1 2 3 4
50 SL 5 0 1 2 3 4
 ;
run ;

data rawscore_con3 ;
input r1-r3 ;
datalines ;
0 0 0
1 1 1
2 2 2
3 3 3
4 4 4
5 5 5
6 6 6
7 7 7
8 8 8
9 9 9
10 10 10
11 11 11
12 12 12
13 13 13
14 14 14
15 15 15
16 16 16
17 17 17
18 18 18
19 19 19
20 20 20
21 21 21
22 22 22
23 23 23
24 24 24
25 25 25
26 26 26
27 27 27
28 28 28
29 29 29
30 30 30
31 31 31
```

```
32 32 32
33 33 33
34 34 34
35 35 35
36 36 36
37 37 37
38 38 38
39 39 39
40 40 40
41 41 41
42 42 42
43 43 43
44 44 44
45 45 45
46 46 46
47 47 47
48 48 48
49 49 49
50 50 50
51 51 51
52 52 52
53 53 53
54 54 54
55 55 55
56 56 56
57 57 57
58 58 58
59 59 59
60 60 60
61 61 61
62 62 62
63 63 63
64 64 64
65 65 65
66 66 66
67 67 67
68 68 68
69 69 69
70 70 70
71 71 71
 ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents Form X" ;
data _null_ ;
set rawscore_con3 ;
file "&workdir\rawscx3.rs" print title ;
put @1 r1 @5 r2 @10 r3 ;
run ;

data _null_ ;
file "&workdir\control_poly3x.txt" ;
put @1 'output disx.out' ;
put @1 'para' @15 'testrun' ;
put @1 'raw2scale' @15 'rawscx3.rs' ;
put @1 'quad' @15 'funx3.pst' ;
```

```
put @1 'end' ;
run ;

data _null_ ;
file "&workdir\control_polyy13.txt" ;
put @1 'output  disyfr.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy13.rs' ;
put @1 'quad' @15 'funy3.pst' ;
put @1 'end' ;
run ;
data _null_ ;
file "&workdir\control_polyy23.txt" ;
put @1 'output  disych.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy23.rs' ;
put @1 'quad' @15 'funy3.pst' ;
put @1 'end' ;
run ;
data _null_ ;
file "&workdir\control_polyy33.txt" ;
put @1 'output  disytr.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy33.rs' ;
put @1 'quad' @15 'funy3.pst' ;
put @1 'end' ;
run ;
data _null_ ;
file "&workdir\control_polyy43.txt" ;
put @1 'output  disyob.out' ;
put @1 'para' @15 'testruny' ;
put @1 'raw2scale' @15 'rawscy43.rs' ;
put @1 'quad' @15 'funy3.pst' ;
put @1 'end' ;
run ;

data resetks;
input KStr ptr KSob pob KSfr pfr KSch pch ;
datalines;
-999 -999 -999 -999 -999 -999 -999 -999
;
run;

data resetd1d2 ;
input D1_fr D1_ch D1_tr D1_ob D2_fr D2_ch D2_tr D2_ob ;
datalines;
-999 -999 -999 -999 -999 -999 -999 -999
;
run;

%let lowest_a = .7 ;
%let range_a = 0.6 ;
%let d=1.7 ;

/*Multidimensional Condition

%let lowest_a = .7 ;
```

```
%let range_a = 0.6 ;
%let d=1.7 ;
%let lowest_as=0.6;
%let range_as=0.3;


data varcov;
           input v1-v3;
           datalines;
1 0 0
0 1 0
0 0 1

        ;

data means;
           input m;
           datalines;
         0
             0
             0
        ;
data means1;
           input m1;
           datalines;
         0.30
             0.30
             0.30
        ;

data means2;
           input m2;
           datalines;
         0.15
             0.15
             0.15
        ;

data means3;
           input m3;
           datalines;
         0.30
             0.15
             0.30
        ;
data means4;
           input m4;
           datalines;
         0.30
             0.30
             0.15
        ;
run;

%include 'c:\raw59\mvn.sas';
*/
```

```
/************** HERE START ***************/


%macro toomuchfun ;


%do anchor = 1 %to 3;

%if &anchor = 1 %then %do ;
%let anchormc = 50 ;
%let anchorcr = 0 ;
%let commonmc = 10 ;
%let commoncr = 0 ;


%end ;

%if &anchor = 2 %then %do ;
%let anchormc = 49 ;
%let anchorcr = 50 ;
%let commonmc = 9 ;
%let commoncr = 1 ;
%end ;

%if &anchor = 3 %then %do ;
%let anchormc = 48 ;
%let anchorcr= 49 ;
%let commonmc = 8 ;
%let commoncr= 2 ;
%end ;

/*Unidimensional Condition*/

%do group = 1 %to 3 ;

%if &group = 1 %then %let groupdiff = 0 ;
%if &group = 2 %then %let groupdiff = .15 ;
%if &group = 3 %then %let groupdiff = .30 ;

/*Multidimensional Condition

%do group = 1 %to 5 ;

%do replication = 1 %to 100 ;


%mvn(varcov=varcov,means=means, n=3000, sample=xtheta, seed = &seedx);

%if &group = 1 %then %do;
 %mvn(varcov=varcov,means=means, n=3000, sample=ytheta, seed = &seedy);
%end;
%if &group = 2 %then %do;
 %mvn(varcov=varcov,means=means1, n=3000, sample=ytheta, seed = &seedy);
                   %end;

%if &group = 3 %then %do;
```

```
 %mvn(varcov=varcov,means=means2, n=3000, sample=ytheta, seed = &seedy);
                    %end;

%if &group = 4 %then %do;
 %mvn(varcov=varcov,means=means3, n=3000, sample=ytheta, seed = &seedy);
                    %end;
%if &group = 5 %then %do;
 %mvn(varcov=varcov,means=means4, n=3000, sample=ytheta, seed = &seedy);
                    %end;

*/

%do replication = 1 %to 100 ;

proc printto log="&workdir\log.txt" print="&workdir\out.txt" new;
run;


%let seedx = 1000000 + &anchor*100000 + &group*10000 + &replication ;
%let seedy = 2000000 + &anchor*100000 + &group*10000 + &replication ;
%let seedc = 3000000 + &anchor*100000 + &group*10000 + &replication ;
%let seed =  4000000 + &anchor*100000 + &group*10000 + &replication ;
%let seedtx = 5000000 + &anchor*100000 + &group*10000 + &replication ;
%let seedty = 6000000 + &anchor*100000 + &group*10000 + &replication ;


data xunique ;
do item = 1 to 35 ;
a = &range_a*(ranuni(&seedx))+ &lowest_a ;
b=rannor(&seedx) ;
call streaminit(&seedx) ;
c = Rand('Beta',8,32) ;
output ;
end ;
run ;


data xuniquecr ;
do item = 36 to 40 ;
a = &range_a*(ranuni(&seedx))+ &lowest_a ;
call streaminit(&seedx) ;
b1=Rand('normal', -1.5,0.20) ;
b2=Rand('normal', -0.5,0.20) ;
b3=Rand('normal',  0.5,0.20) ;
b4=Rand('normal',  1.5,0.20) ;
if b2 <= b1 then b2 = b1 + .2;
if b3 <= b2 then b3 = b2 + .2;
if b4 <= b3 then b4 = b3 + .2;
output ;
end ;
run ;


data yunique ;
do item = 1 to 35 ;
a = &range_a*(ranuni(&seedy))+ &lowest_a ;
b=rannor(&seedy) ;
```

```
call streaminit(&seedy) ;
c = Rand('Beta',8,32) ;
output ;
end ;
run ;


data yuniquecr ;
do item = 36 to 40 ;
a = &range_a*(ranuni(&seedy))+ &lowest_a ;
call streaminit(&seedy) ;
b1=Rand('normal', -1.5,0.20) ;
b2=Rand('normal', -0.5,0.20) ;
b3=Rand('normal',  0.5,0.20) ;
b4=Rand('normal',  1.5,0.20) ;
if b2 <= b1 then b2 = b1 + .2;
if b3 <= b2 then b3 = b2 + .2;
if b4 <= b3 then b4 = b3 + .2;
output ;
end ;
run ;


data anchormc ;
do item = 41 to &anchormc ;
a = &range_a*(ranuni(&seedc))+ &lowest_a ;
b=rannor(&seedc) ;
call streaminit(&seedc) ;
c = Rand('Beta',8,32) ;
output ;
end ;
run ;


data anchorcr ;
if &anchor >= 2 then do;
do item = &anchorcr to 50 ;
a = &range_a*(ranuni(&seedc))+ &lowest_a ;
call streaminit(&seedc) ;
b1=Rand('normal', -1.5,0.20) ;
b2=Rand('normal', -0.5,0.20) ;
b3=Rand('normal',  0.5,0.20) ;
b4=Rand('normal',  1.5,0.20) ;
if b2 <= b1 then b2 = b1 + .2;
if b3 <= b2 then b3 = b2 + .2;
if b4 <= b3 then b4 = b3 + .2;
output ;
end ;
end ;
run ;



data xtheta ;
do i = 1 to 3000 ;
theta = rannor(&seedtx) ;
output ;
```

```
end ;
drop i ;
run ;

data ytheta ;
do i = 1 to 3000 ;
theta = rannor(&seedty)+&groupdiff ;
output ;
end ;
drop i ;
run ;


proc iml ;

call randseed(&seedx) ;

use xtheta ;
read all var _num_ into theta ;
close xtheta ;

use xunique ;
read all var {a b c} into param ;
close xunique ;

resp = j(3000,35,0) ;

do i = 1 to 3000 ;
do j = 1 to 35 ;
a = param[j,1] ;
b = param[j,2] ;
c = param[j,3] ;
prob = c + (1-c)*(exp(&d*a*(theta[i]-b))/(1+exp(&d*a*(theta[i]-b)))) ;
u=ranuni(&seed) ;
if u<prob then resp[i,j]=1 ;
end ;
end ;

use xtheta ;
read all var _num_ into theta ;
close xtheta ;

use anchormc ;
read all var {a b c} into parameter ;
close anchormc ;

response = j(3000,&commonmc,0) ;

do i = 1 to 3000 ;
do j = 1 to &commonmc ;
a = parameter[j,1] ;
b = parameter[j,2] ;
c = parameter[j,3] ;
prob = c + (1-c)*(exp(&d*a*(theta[i]-b))/(1+exp(&d*a*(theta[i]-b)))) ;
u=ranuni(&seed) ;
if u<prob then response[i,j]=1 ;
end ;
```

```
end ;


if &anchor>=2 then do ;

use xtheta ;
read all var _num_ into theta ;
close xtheta ;

use anchorcr ;
read all var {a} into pa ;
close anchorcr ;

use anchorcr ;
read all var {b1 b2 b3 b4} into crpa ;
close crpa ;

pijkx = j(3000,&commoncr,0) ;
do i = 1 to 3000 ;
pjk5 = j(&commoncr,4,0) ;
do j = 1 to &commoncr ;
do k = 1 to 4 ;

prob = 1.7*pa[j]*(theta[i] - crpa[j,k]) ;
pjk5[j,k] = exp(prob)/(1+exp(prob)) ;
end ;
pt5 = j(&commoncr,5,0) ;
pt5[,1] = 1-pjk5[,1] ;
pt5[,2] = pjk5[,1] - pjk5[,2] ;
pt5[,3] = pjk5[,2] - pjk5[,3] ;
pt5[,4] = pjk5[,3] - pjk5[,4] ;
pt5[,5] = pjk5[,4] - 0 ;
cum5 = j(&commoncr,5,0) ;
cum5[,1] = pt5[,1] ;
cum5[,2] = cum5[,1] + pt5[,2] ;
cum5[,3] = cum5[,2] + pt5[,3] ;
cum5[,4] = cum5[,3] + pt5[,4] ;
cum5[,5] = cum5[,4] + pt5[,5] ;
u = uniform(0) ;
if (cum5[,1] >= u) then pijkx[i,j] = 0 ;
else if (cum5[,2] >= u) then pijkx[i,j] = 1 ;
else if (cum5[,3] >= u) then pijkx[i,j] = 2 ;
else if (cum5[,4] >= u) then pijkx[i,j] = 3 ;
else if (cum5[,5] >= u) then pijkx[i,j] = 4 ;
end ;
end ;
end ;


use xtheta ;
read all var _num_ into theta ;
close xtheta ;

use xuniquecr ;
read all var {a} into a ;
close xuniquecr ;
```

```
use xuniquecr ;
read all var {b1 b2 b3 b4} into crparam ;
close xuniquecr ;

pijk5 = j(3000,5,0) ;
do i = 1 to 3000 ;
pjk5 = j(5,4,0) ;
do j = 1 to 5 ;
do k = 1 to 4 ;

prob = 1.7*a[j]*(theta[i] - crparam[j,k]) ;
pjk5[j,k] = exp(prob)/(1+exp(prob)) ;
end ;
pt5 = j(5,5,0) ;
pt5[,1] = 1-pjk5[,1] ;
pt5[,2] = pjk5[,1] - pjk5[,2] ;
pt5[,3] = pjk5[,2] - pjk5[,3] ;
pt5[,4] = pjk5[,3] - pjk5[,4] ;
pt5[,5] = pjk5[,4] - 0 ;
cum5 = j(5,5,0) ;
cum5[,1] = pt5[,1] ;
cum5[,2] = cum5[,1] + pt5[,2] ;
cum5[,3] = cum5[,2] + pt5[,3] ;
cum5[,4] = cum5[,3] + pt5[,4] ;
cum5[,5] = cum5[,4] + pt5[,5] ;
u = uniform(0) ;
if (cum5[,1] >= u) then pijk5[i,j] = 0 ;
else if (cum5[,2] >= u) then pijk5[i,j] = 1 ;
else if (cum5[,3] >= u) then pijk5[i,j] = 2 ;
else if (cum5[,4] >= u) then pijk5[i,j] = 3 ;
else if (cum5[,5] >= u) then pijk5[i,j] = 4 ;
end ;
end ;

combresp3 = resp || pijk5 || response||pijkx ;
create Xcond2 from combresp3
[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
i11 i12 i13 i14 i15 i16 i17 i18 i19 i20
i21 i22 i23 i24 i25 i26 i27 i28 i29 i30
i31 i32 i33 i34 i35 i36 i37 i38 i39 i40
i41 i42 i43 i44 i45 i46 i47 i48 i49 i50}] ;
append from combresp3 ;

quit ;

data _null_ ; set Xcond2 ;
file "&workdir\x.dat" ;
put (i1-i50) (1.) ;
run ;


proc iml ;

call randseed(&seedy) ;

use ytheta ;
read all var _num_ into theta ;
```

```
close ytheta ;

use yunique ;
read all var {a b c} into param ;
close yunique ;

resp = j(3000,35,0) ;

do i = 1 to 3000 ;
do j = 1 to 35 ;
a = param[j,1] ;
b = param[j,2] ;
c = param[j,3] ;
prob = c + (1-c)*(exp(&d*a*(theta[i]-b))/(1+exp(&d*a*(theta[i]-b)))) ;
u=ranuni(&seed) ;
if u<prob then resp[i,j]=1 ;
end ;
end ;

use ytheta ;
read all var _num_ into theta ;
close xtheta ;

use anchormc ;
read all var {a b c} into parameter ;
close anchormc ;

response = j(3000,&commonmc,0) ;

do i = 1 to 3000 ;
do j = 1 to &commonmc ;
a = parameter[j,1] ;
b = parameter[j,2] ;
c = parameter[j,3] ;
prob = c + (1-c)*(exp(&d*a*(theta[i]-b))/(1+exp(&d*a*(theta[i]-b)))) ;
u=ranuni(&seed) ;
if u<prob then response[i,j]=1 ;
end ;
end ;

if &anchor>=2 then do ;

use ytheta ;
read all var _num_ into theta ;
close ytheta ;

use anchorcr ;
read all var {a} into pa ;
close anchorcr ;

use anchorcr ;
read all var {b1 b2 b3 b4} into crpa ;
close crpa ;

pijkx = j(3000,&commoncr,0) ;
do i = 1 to 3000 ;
pjk5 = j(&commoncr,4,0) ;
```

```
do j = 1 to &commoncr ;
do k = 1 to 4 ;

prob = 1.7*pa[j]*(theta[i] - crpa[j,k]) ;
pjk5[j,k] = exp(prob)/(1+exp(prob)) ;
end ;
pt5 = j(&commoncr,5,0) ;
pt5[,1] = 1-pjk5[,1] ;
pt5[,2] = pjk5[,1] - pjk5[,2] ;
pt5[,3] = pjk5[,2] - pjk5[,3] ;
pt5[,4] = pjk5[,3] - pjk5[,4] ;
pt5[,5] = pjk5[,4] - 0 ;
cum5 = j(&commoncr,5,0) ;
cum5[,1] = pt5[,1] ;
cum5[,2] = cum5[,1] + pt5[,2] ;
cum5[,3] = cum5[,2] + pt5[,3] ;
cum5[,4] = cum5[,3] + pt5[,4] ;
cum5[,5] = cum5[,4] + pt5[,5] ;
u = uniform(0) ;
if (cum5[,1] >= u) then pijkx[i,j] = 0 ;
else if (cum5[,2] >= u) then pijkx[i,j] = 1 ;
else if (cum5[,3] >= u) then pijkx[i,j] = 2 ;
else if (cum5[,4] >= u) then pijkx[i,j] = 3 ;
else if (cum5[,5] >= u) then pijkx[i,j] = 4 ;
end ;
end ;
end ;


use ytheta ;
read all var _num_ into theta ;
close xtheta ;

use yuniquecr ;
read all var {a} into a ;
close xuniquecr ;

use yuniquecr ;
read all var {b1 b2 b3 b4} into crparam ;
close crparam ;

pijk5 = j(3000,5,0) ;
do i = 1 to 3000 ;
pjk5 = j(5,4,0) ;
do j = 1 to 5 ;
do k = 1 to 4 ;

prob = 1.7*a[j]*(theta[i] - crparam[j,k]) ;
pjk5[j,k] = exp(prob)/(1+exp(prob)) ;
end ;
pt5 = j(5,5,0) ;
pt5[,1] = 1-pjk5[,1] ;
pt5[,2] = pjk5[,1] - pjk5[,2] ;
pt5[,3] = pjk5[,2] - pjk5[,3] ;
pt5[,4] = pjk5[,3] - pjk5[,4] ;
pt5[,5] = pjk5[,4] - 0 ;
cum5 = j(5,5,0) ;
```

```
cum5[,1] = pt5[,1] ;
cum5[,2] = cum5[,1] + pt5[,2] ;
cum5[,3] = cum5[,2] + pt5[,3] ;
cum5[,4] = cum5[,3] + pt5[,4] ;
cum5[,5] = cum5[,4] + pt5[,5] ;
u = uniform(0) ;
if (cum5[,1] >= u) then pijk5[i,j] = 0 ;
else if (cum5[,2] >= u) then pijk5[i,j] = 1 ;
else if (cum5[,3] >= u) then pijk5[i,j] = 2 ;
else if (cum5[,4] >= u) then pijk5[i,j] = 3 ;
else if (cum5[,5] >= u) then pijk5[i,j] = 4 ;
end ;
end ;


combresp = resp || pijk5 || response||pijkx ;
create Ycond2 from combresp
[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
i11 i12 i13 i14 i15 i16 i17 i18 i19 i20
i21 i22 i23 i24 i25 i26 i27 i28 i29 i30
i31 i32 i33 i34 i35 i36 i37 i38 i39 i40
i41 i42 i43 i44 i45 i46 i47 i48 i49 i50}] ;
append from combresp ;

quit ;


/*Multidimensional Condition*/


data xunique;
do item = 1 to 35;
ag = &range_a*(ranuni(&seedx))+ &lowest_a;
amc = &range_as*(ranuni(&seedx))+ &lowest_as;
acr = 0;
b=rannor(&seedx);
call streaminit(&seedx);
c = Rand('Beta',8,32);
d = -b*SQRT(ag**2+amc**2+acr**2);
output;
end;
run;

/* create 5 unique CR items for form x*/
data xuniquecr;
do item = 36 to 40;
ag = &range_a*(ranuni(&seedx))+ &lowest_a;
acr = &range_as*(ranuni(&seedx))+ &lowest_as;
amc = 0;
call streaminit(&seedx);
b1=Rand('normal', -1.5,0.20);
b2=Rand('normal', -0.5,0.20);
b3=Rand('normal',  0.5,0.20);
b4=Rand('normal',  1.5,0.20);
if b2 <= b1 then b2 = b1 + .2;
if b3 <= b2 then b3 = b2 + .2;
if b4 <= b3 then b4 = b3 + .2;
```

```
d1 = -b1*SQRT(ag**2+acr**2+amc**2);
d2 = -b2*SQRT(ag**2+acr**2+amc**2);
d3 = -b3*SQRT(ag**2+acr**2+amc**2);
d4 = -b4*SQRT(ag**2+acr**2+amc**2);
output;
end;
run;

data yunique;
do item = 1 to 35;
ag = &range_a*(ranuni(&seedy))+ &lowest_a;
amc = &range_as*(ranuni(&seedy))+ &lowest_as;
acr = 0;
b=rannor(&seedy);
call streaminit(&seedy);
c = Rand('Beta',8,32);
d = -b*SQRT(ag**2+amc**2+acr**2);
output;
end;
run;

data yuniquecr;
do item = 36 to 40;
ag = &range_a*(ranuni(&seedy))+ &lowest_a;
acr = &range_as*(ranuni(&seedy))+ &lowest_as;
amc = 0;
call streaminit(&seedy);
b1=Rand('normal', -1.5,0.20);
b2=Rand('normal', -0.5,0.20);
b3=Rand('normal',  0.5,0.20);
b4=Rand('normal',  1.5,0.20);
if b2 <= b1 then b2 = b1 + .2;
if b3 <= b2 then b3 = b2 + .2;
if b4 <= b3 then b4 = b3 + .2;
d1 = -b1*SQRT(ag**2+acr**2+amc**2);
d2 = -b2*SQRT(ag**2+acr**2+amc**2);
d3 = -b3*SQRT(ag**2+acr**2+amc**2);
d4 = -b4*SQRT(ag**2+acr**2+amc**2);
output;
end;
run;

data anchormc;
do item = 41 to &anchormc ;
ag = &range_a*(ranuni(&seedc))+ &lowest_a;
amc = &range_as*(ranuni(&seedc))+ &lowest_as;
acr = 0;
b=rannor(&seedc);
call streaminit(&seedc);
c = Rand('Beta',8,32);
d = -b*SQRT(ag**2+amc**2+acr**2);
output;
end;
run;


data anchorcr;
```

```
if &anchor >= 2 then do;
do item = &anchorcr to 50 ;
ag = &range_a*(ranuni(&seedy))+ &lowest_a;
acr = &range_as*(ranuni(&seedy))+ &lowest_as;
amc = 0;
call streaminit(&seedc);
b1=Rand('normal', -1.5,0.20);
b2=Rand('normal', -0.5,0.20);
b3=Rand('normal',  0.5,0.20);
b4=Rand('normal',  1.5,0.20);
if b2 <= b1 then b2 = b1 + .2;
if b3 <= b2 then b3 = b2 + .2;
if b4 <= b3 then b4 = b3 + .2;
d1 = -b1*SQRT(ag**2+acr**2+amc**2);
d2 = -b2*SQRT(ag**2+acr**2+amc**2);
d3 = -b3*SQRT(ag**2+acr**2+amc**2);
d4 = -b4*SQRT(ag**2+acr**2+amc**2);
output;
end;
end;
run;


proc iml;

call randseed(&seedx);

use xunique;
read all var {ag amc acr c d} into m1;

use xtheta;
read all var _num_ into ability;

pijk35 = j(3000,35,0);

do i = 1 to 3000;

thetag=ability[i,1];
thetamc=ability[i,2];
thetacr=ability[i,3];

do j = 1 to 35;

ag=m1[j,1];
amc=m1[j,2];
acr=m1[j,3];
c=m1[j,4];
d=m1[j,5];

prob = c + (1-
c)*(exp(1.7*(((ag*thetag)+(amc*thetamc)+(acr*thetacr))+d)))/(1+exp(1.7*(((ag*
thetag)+(amc*thetamc)+(acr*thetacr))+d)));

u=ranuni(&seed);
if u<prob then pijk35[i,j]=1;
end;
end;
```

```
use anchormc ;
read all var {ag amc acr c d} into parameter ;
close anchormc ;

response = j(3000,&commonmc,0) ;

do i = 1 to 3000 ;

thetag=ability[i,1];
thetamc=ability[i,2];
thetacr=ability[i,3];

do j = 1 to &commonmc ;

ag = parameter[j,1] ;
amc = parameter[j,2] ;
acr = parameter[j,3] ;
c = parameter [j,4] ;
d = parameter [j,5] ;


prob = c + (1-
c)*(exp(1.7*(((ag*thetag)+(amc*thetamc)+(acr*thetacr))+d)))/(1+exp(1.7*(((ag*
thetag)+(amc*thetamc)+(acr*thetacr))+d)));

u=ranuni(&seed) ;
if u<prob then response[i,j]=1 ;
end ;
end ;


if &anchor>=2 then do ;

use xtheta ;
read all var _num_ into ability ;
close xtheta ;

use anchorcr ;
read all var {ag amc acr} into m2 ;
close anchorcr ;

use anchorcr;
read all var {d1 d2 d3 d4} into m3;
close anchorcr ;


pijkx = j(3000,&commoncr,0);

do i = 1 to 3000;

thetag=ability[i,1];
thetamc=ability[i,2];
thetacr=ability[i,3];

do j = 1 to &commoncr;
```

```
ag=m2[j,1];
amc=m2[j,2];
acr=m2[j,3];

pjk5 = j(&commoncr,4,0);

do k = 1 to 4;

prob = 1.7*((ag*thetag)+(amc*thetamc)+(acr*thetacr)+ m3[j,k]);

pjk5[j,k] = exp(prob)/(1 + exp(prob));
end;

pt5 = j(&commoncr,5,0);
pt5[,1] = 1-pjk5[,1];
pt5[,2] = pjk5[,1] - pjk5[,2];
pt5[,3] = pjk5[,2] - pjk5[,3];
pt5[,4] = pjk5[,3] - pjk5[,4];
pt5[,5] = pjk5[,4] - 0;
cum5 = j(&commoncr,5,0);
cum5[,1] = pt5[,1];
cum5[,2] = cum5[,1] + pt5[,2];
cum5[,3] = cum5[,2] + pt5[,3];
cum5[,4] = cum5[,3] + pt5[,4];
cum5[,5] = cum5[,4] + pt5[,5];

u = uniform(&seed);
if (cum5[,1] >= u) then pijkx[i,j] = 0;
else if (cum5[,2] >= u) then pijkx[i,j] = 1;
else if (cum5[,3] >= u) then pijkx[i,j] = 2;
else if (cum5[,4] >= u) then pijkx[i,j] = 3;
else if (cum5[,5] >= u) then pijkx[i,j] = 4;

end;
end;
end;


use xtheta ;
read all var _num_ into ability ;
close xtheta ;

use xuniquecr;
read all var {ag amc acr} into m4;

use xuniquecr;
read all var {d1 d2 d3 d4} into m5;

pijk5 = j(3000,5,0);

do i=1 to 3000;

thetag=ability[i,1];
thetamc=ability[i,2];
thetacr=ability[i,3];

do j = 1 to 5;
```

```
ag=m4[j,1];
amc=m4[j,2];
acr=m4[j,3];

pjk = j(1,4,0);

do k = 1 to 4;

prob = 1.7*((ag*thetag)+(amc*thetamc)+(acr*thetacr)+ m5[j,k]);

pjk[1,k] = exp(prob)/(1 + exp(prob));
end;

pti5 = j(1,5,0);
pti5[1,1] = 1-pjk[1,1];
pti5[1,2] = pjk[1,1] - pjk[1,2];
pti5[1,3] = pjk[1,2] - pjk[1,3];
pti5[1,4] = pjk[1,3] - pjk[1,4];
pti5[1,5] = pjk[1,4] - 0;
cum5 = j(1,5,0);
cum5[1,1] = pti5[,1];
cum5[1,2] = cum5[1,1] + pti5[1,2];
cum5[1,3] = cum5[1,2] + pti5[1,3];
cum5[1,4] = cum5[1,3] + pti5[1,4];
cum5[1,5] = cum5[1,4] + pti5[1,5];

u = uniform(&seed);
if (cum5[1,1] >= u) then pijk5[i,j] = 0;
else if (cum5[1,2] >= u) then pijk5[i,j] = 1;
else if (cum5[1,3] >= u) then pijk5[i,j] = 2;
else if (cum5[1,4] >= u) then pijk5[i,j] = 3;
else if (cum5[1,5] >= u) then pijk5[i,j] = 4;

end;
end;


comrespx = pijk35 || pijk5 || response || pijkx;
create xcond2 from comrespx
[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
i11 i12 i13 i14 i15 i16 i17 i18 i19 i20
i21 i22 i23 i24 i25 i26 i27 i28 i29 i30
i31 i32 i33 i34 i35 i36 i37 i38 i39 i40
i41 i42 i43 i44 i45 i46 i47 i48 i49 i50}];
append from comrespx;

quit;


proc iml;

call randseed(&seedy);

use yunique;
read all var {ag amc acr c d} into m1;
```

```
use ytheta;
read all var _num_ into theta;

pijk35 = j(3000,35,0);

do i = 1 to 3000;

thetag=theta[i,1];
thetamc=theta[i,2];
thetacr=theta[i,3];

do j = 1 to 35;

ag=m1[j,1];
amc=m1[j,2];
acr=m1[j,3];
c=m1[j,4];
d=m1[j,5];

prob = c + (1-
c)*(exp(1.7*(((ag*thetag)+(amc*thetamc)+(acr*thetacr))+d)))/(1+exp(1.7*(((ag*
thetag)+(amc*thetamc)+(acr*thetacr))+d)));

u=ranuni(&seed);
if u<prob then pijk35[i,j]=1;
end;
end;

use anchormc ;
read all var {ag amc acr c d} into parameter ;
close anchormc ;

response = j(3000,&commonmc,0) ;

do i = 1 to 3000 ;

thetag=theta[i,1];
thetamc=theta[i,2];
thetacr=theta[i,3];

do j = 1 to &commonmc ;

ag = parameter[j,1] ;
amc = parameter[j,2] ;
acr = parameter[j,3] ;
c = parameter [j,4] ;
d = parameter [j,5] ;


prob = c + (1-
c)*(exp(1.7*(((ag*thetag)+(amc*thetamc)+(acr*thetacr))+d)))/(1+exp(1.7*(((ag*
thetag)+(amc*thetamc)+(acr*thetacr))+d)));

u=ranuni(&seed) ;
if u<prob then response[i,j]=1 ;
end ;
end ;
```

```
if &anchor>=2 then do ;

use ytheta ;
read all var _num_ into theta ;
close ytheta ;

use anchorcr ;
read all var {ag amc acr} into m2 ;
close anchorcr ;


use anchorcr;
read all var {d1 d2 d3 d4} into m3;
close anchorcr ;


pijkcy = j(3000,&commoncr,0);

do i = 1 to 3000;

thetag=theta[i,1];
thetamc=theta[i,2];
thetacr=theta[i,3];

do j = 1 to &commoncr;

ag=m2[j,1];
amc=m2[j,2];
acr=m2[j,3];

pijkx = j(&commoncr,4,0);

do k = 1 to 4;

prob = 1.7*((ag*thetag)+(amc*thetamc)+(acr*thetacr)+ m3[j,k]);

pijkx[j,k] = exp(prob)/(1 + exp(prob));
end;

pt5 = j(&commoncr,5,0);
pt5[,1] = 1-pijkx[,1];
pt5[,2] = pijkx[,1] - pijkx[,2];
pt5[,3] = pijkx[,2] - pijkx[,3];
pt5[,4] = pijkx[,3] - pijkx[,4];
pt5[,5] = pijkx[,4] - 0;
cum5 = j(&commoncr,5,0);
cum5[,1] = pt5[,1];
cum5[,2] = cum5[,1] + pt5[,2];
cum5[,3] = cum5[,2] + pt5[,3];
cum5[,4] = cum5[,3] + pt5[,4];
cum5[,5] = cum5[,4] + pt5[,5];

u = uniform(&seed);
if (cum5[,1] >= u) then pijkcy[i,j] = 0;
else if (cum5[,2] >= u) then pijkcy[i,j] = 1;
```

```
else if (cum5[,3] >= u) then pijkcy[i,j] = 2;
else if (cum5[,4] >= u) then pijkcy[i,j] = 3;
else if (cum5[,5] >= u) then pijkcy[i,j] = 4;

end;
end;
end;

use ytheta ;
read all var _num_ into theta ;
close ytheta ;

use yuniquecr;
read all var {ag amc acr} into m4;

use yuniquecr;
read all var {d1 d2 d3 d4} into m5;

pijk = j(3000,5,0);

do i=1 to 3000;

thetag=theta[i,1];
thetamc=theta[i,2];
thetacr=theta[i,3];

do j = 1 to 5;

ag=m4[j,1];
amc=m4[j,2];
acr=m4[j,3];

pjk = j(1,4,0);

do k = 1 to 4;

prob = 1.7*((ag*thetag)+(amc*thetamc)+(acr*thetacr)+ m5[j,k]);

pjk[1,k] = exp(prob)/(1 + exp(prob));
end;

pti5 = j(1,5,0);
pti5[1,1] = 1-pjk[1,1];
pti5[1,2] = pjk[1,1] - pjk[1,2];
pti5[1,3] = pjk[1,2] - pjk[1,3];
pti5[1,4] = pjk[1,3] - pjk[1,4];
pti5[1,5] = pjk[1,4] - 0;
cum5 = j(1,5,0);
cum5[1,1] = pti5[,1];
cum5[1,2] = cum5[1,1] + pti5[1,2];
cum5[1,3] = cum5[1,2] + pti5[1,3];
cum5[1,4] = cum5[1,3] + pti5[1,4];
cum5[1,5] = cum5[1,4] + pti5[1,5];

u = uniform(&seed);
if (cum5[1,1] >= u) then pijk[i,j] = 0;
else if (cum5[1,2] >= u) then pijk[i,j] = 1;
```

```
else if (cum5[1,3] >= u) then pijk[i,j] = 2;
else if (cum5[1,4] >= u) then pijk[i,j] = 3;
else if (cum5[1,5] >= u) then pijk[i,j] = 4;

end;
end;

comrespy = pijk35 || pijk || response || pijkcy;
create ycond2 from comrespy
[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
i11 i12 i13 i14 i15 i16 i17 i18 i19 i20
i21 i22 i23 i24 i25 i26 i27 i28 i29 i30
i31 i32 i33 i34 i35 i36 i37 i38 i39 i40
i41 i42 i43 i44 i45 i46 i47 i48 i49 i50}];
append from comrespy;

quit;
*/




Data TestX ;
set Xcond2 ;
total = sum(of i1-i50) ;
totalx = sum(of i1-i40) ;
totala = sum(of i41-i50) ;
run ;

data testx1 ;
set testx (keep = totalx totala) ;
run ;

data testy ;
set Ycond2 ;
total = sum(of i1-i50) ;
totaly = sum(of i1-i40) ;
totala = sum(of i41-i50) ;
run ;

data testy1 ;
set testy (keep = totaly totala) ;
run ;

data _null_ ; set testx1 ;
file "&workdir\xtotal.dat" ;
put @1 totalx @4 totala ;
run ;

data _null_ ; set testy1 ;
file "&workdir\ytotal.dat" ;
put @1 totaly @4 totala ;
run ;


%if &anchor = 1 %then %do ;
```

```
x '"c:\program files\r\r-2.14.0\bin\x64\rscript.exe"
c:\raw59\equipercentile1.R'  ;
%end ;

%if &anchor = 2 %then %do ;
x '"c:\program files\r\r-2.14.0\bin\x64\rscript.exe"
c:\raw59\equipercentile2.R'  ;
%end ;

%if &anchor = 3 %then %do ;
x '"c:\program files\r\r-2.14.0\bin\x64\rscript.exe"
c:\raw59\equipercentile3.R'  ;
%end ;

data testdata ;
set Xcond2 end=lastobs ;
id = _N_ ;
group = 1 ;
put id 4. +1 (i1-i50) (1.) @51 group 1. ;
run ;

data testdata1 ;
set Ycond2 end=lastobs ;
id = _N_ ;
group = 2 ;
put id 4. +1 (i1-i50) (1.) @51 group 1. ;
run ;

data groupdata ;
set testdata testdata1 ;
run ;

proc sort data=groupdata ;
by group ;
run;

data _null_ ;
set groupdata ;
file "&workdir\itemresp.dat" ;
if group=1 then
put id 4. +1 'A'  +1 (i1-i40)(1.) (i41-i50) (1.)
'99999999999999999999999999999999999999' ;
else
put id 4. +1 'B'  +1 '99999999999999999999999999999999999999' (i41-i50)(1.)
(i1-i40) (1.) ;
run ;



%if &anchor = 1 %then %do ;

x 'cd C:\raw59' ;
x 'psl0.exe condition1a' ;
x 'psl1.exe condition1a' ;
x 'psl2.exe condition1a' ;
```

```
data par_cond1 ;
infile "&workdir\condition1a.ph2" truncover ;
input string $128.  ;
if (string = '| ITEM |BLOCK|  SLOPE  |   S.E.  |LOCATION |   S.E.  |GUESSING
|   S.E.  |') then do ;
do i = 1 to 90 ;
input ;
input item $2-6 block $9-12 slope $15-22 se $25-32 location $35-42 sel $45-52
guessing $55-62 seg $65-72 ;
output ;
end ;
end ;
drop string i ;
run ;

data par_cond1b ;
set par_cond1 ;
if (location = 'NaN' | location = '0.000') then delete ;
run ;

data cond1 ;
set par_cond1b;
nid=_N_;
keep nid slope location guessing ;
run ;

data par_cond1x;
set cond1;
if nid <= 50 then output;
run;

data par_cond1y;
set cond1;
if nid > 50 then output;
run;
data par_cond11y;
set par_cond1y;
newid=_N_;
keep newid slope location guessing;
run;
data par_cond111y;
set par_cond11y;
if newid <= 10 then output;
run;
data par_cond1111y;
set par_cond11y;
if newid > 10 then output;
run;
data ycombined;
set par_cond1111y par_cond111y;
drop newid;
nid=_N_;
run;


data test1 ;
infile "&workdir\condition1a.ph2" truncover ;
```

```
input string $128.  ;
if (string = '[GROUP:   1  OLD     ]') then do ;
do i = 1 to 90 ;
do j = 1 to 1 ;
input ;
end;
do k = 1 to 3;
input type $ b1 $27-34 b2 $38-44 b3 $49-54 b4 $59-65  ;
output ;
end ;
end ;
end;
drop string i j k;
run ;

data test2_1 ;
set test1 ;
if type ~= 'CATEGORY' then delete ;
if b1 = 0.000 then delete;
run ;

data test3_1 ;
set test2_1 ;
id=_N_;
if id = 1 then nid=36 ;
else if id = 2 then nid=37 ;
else if id = 3 then nid=38 ;
else if id = 4 then nid=39 ;
else if id = 5 then nid=40 ;
drop id type ;
run ;

data formx ;
merge par_cond1x test3_1 ;
by nid ;
run ;

data formx1 ;
set formx ;
if (nid >= 36 & nid <= 40) then do ;
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
drop b1 b2 b3 b4 ;
run ;

data formx11 ;
set formx1 ;
if (nid >= 36 & nid <= 40) then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
drop parb3 parb4 ;
```

```
run ;

data  _null_ ; set formx11 ;
file "&workdir\par_cond1.dat" ;
put @1 nid  @10 slope @20 location @30 guessing @40 parb1 @50 parb2 ;
run ;



data test1y ;
infile "&workdir\condition1a.ph2" truncover ;
input string $128.  ;
if (string = '[GROUP:   2  NEW     ]') then do ;
do i = 1 to 90 ;
do j = 1 to 1 ;
input ;
end;
do k = 1 to 3;
input type $ b1 $27-34 b2 $38-44 b3 $49-54 b4 $59-65  ;
output ;
end ;
end ;
end;
drop string i j k;
run ;

data test2_1y ;
set test1y ;
if type ~= 'CATEGORY' then delete ;
if b1 = 0.000 then delete;
run ;

data test3_1y ;
set test2_1y ;
id=_N_;
if (id >= 1 & id <= 5) then nid = id + 35;
drop id type ;
run ;

data formy ;
merge ycombined test3_1y ;
by nid ;
run ;

data formy1 ;
set formy ;
if (nid >= 36 & nid <= 40) then do ;
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
drop b1 b2 b3 b4 ;
run ;

data formy11 ;
set formy1 ;
```

```
if (nid >= 36 & nid <= 40) then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
drop parb3 parb4 ;
run ;

data  _null_ ; set formy11 ;
file "&workdir\par_cond1y.dat" ;
put @1 nid  @10 slope @20 location @30 guessing @40 parb1 @50 parb2 ;
run ;


data readpar ;
infile "&workdir\par_cond1.dat" ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_junk524 ;
merge readpar junk524 ;
by newid ;
run ;


data _null_ ;
set param_junk524 ;
file "&workdir\testrun" ;
put newid type numresp ;
if numresp = 2 then do ;
put c1 c2 ;
put '1.7 ' slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put '1.7 ' slope location guessing parb1 parb2 ;
end ;
run ;


data readpary ;
infile "&workdir\par_cond1y.dat" ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_junk524y ;
merge readpary junk524 ;
by newid ;
run ;

data _null_ ;
set param_junk524y ;
file "&workdir\testruny" ;
put newid type numresp ;
if numresp = 2 then do ;
put c1 c2 ;
```

```
put '1.7 ' slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put '1.7 ' slope location guessing parb1 parb2 ;
end ;
run ;

%end ;

%if &anchor = 2 %then %do ;


x 'cd C:\raw59' ;
x 'psl0.exe condition2' ;
x 'psl1.exe condition2' ;
x 'psl2.exe condition2' ;



data par_cond2 ;
infile "&workdir\condition2.ph2" truncover ;
input string $128.  ;
if (string = '| ITEM |BLOCK|  SLOPE  |   S.E.  |LOCATION |   S.E.  |GUESSING
|   S.E.  |') then do ;
do i = 1 to 90 ;
input ;
input item $2-6 block $9-12 slope $15-22 se $25-32 location $35-42 sel $45-52
guessing $55-62 seg $65-72 ;
output ;
end ;
end ;
drop string i ;
run ;

data par_cond2b ;
set par_cond2 ;
if (location = 'NaN' | location = '0.000') then delete ;
run ;

data cond2 ;
set par_cond2b;
nid=_N_;
keep nid slope location guessing ;
run ;

data par_cond2x ;
set cond2 ;
if nid <= 50 then output ;
run ;

data par_cond2y ;
set cond2 ;
if nid > 50 then output ;
run ;
data par_cond22y;
set par_cond2y;
```

```
newid=_N_;
keep newid slope location guessing;
run;
data par_cond222y;
set par_cond22y;
if newid <= 10 then output;
run;
data par_cond2222y;
set par_cond22y;
if newid > 10 then output;
run;
data ycombined2;
set par_cond2222y par_cond222y;
drop newid;
nid=_N_;
run;

data test2 ;
infile "&workdir\condition2.ph2" truncover ;
input string $128.  ;
if (string = '[GROUP:   1  OLD     ]') then do ;
do i = 1 to 90 ;
do j = 1 to 1 ;
input ;
end;
do k = 1 to 3;
input type $ b1 $27-34 b2 $38-44 b3 $49-54 b4 $59-65  ;
output ;
end ;
end ;
end;
drop string i j k;
run ;

data test2_2 ;
set test2 ;
if type ~= 'CATEGORY' then delete ;
if b1=0.000 then delete;
run ;

data test3_2 ;
set test2_2 ;
id=_N_;
if (id >= 1 & id <= 5) then nid = id + 35 ;
else if id = 6 then nid = 50;
drop id type ;
run ;


data formx2 ;
merge par_cond2x test3_2 ;
by nid ;
run ;

data formx22 ;
set formx2 ;
if (nid >= 36 & nid <= 40) then do ;
```

```
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
if nid = 50 then do ;
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
drop b1 b2 b3 b4 ;
run ;

data formx222 ;
set formx22 ;
if (nid >= 36 & nid <= 40) then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
if nid = 50 then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
drop parb3 parb4 ;
run ;

data  _null_ ; set formx222 ;
file "&workdir\par_cond2.dat" ;
put @1 nid  @10 slope @20 location @30 guessing @40 parb1 @50 parb2 ;
run ;


data test2y ;
infile "&workdir\condition2.ph2" truncover ;
input string $128.  ;
if (string = '[GROUP:   2  NEW     ]') then do ;
do i = 1 to 90 ;
do j = 1 to 1 ;
input ;
end;
do k = 1 to 3;
input type $ b1 $27-34 b2 $38-44 b3 $49-54 b4 $59-65  ;
output ;
end ;
end ;
end;
drop string i j k;
run ;

data test2_2y ;
set test2y ;
if type ~= 'CATEGORY' then delete ;
```

```
if b1=0.000 then delete;
run ;

data test3_2y ;
set test2_2y ;
id=_N_;
if id = 1 then nid=50 ;
else if (id >= 2 & id <= 6) then nid = id + 34 ;
drop id type ;
run ;

proc sort data = test3_2y;
BY nid;
run;

data formy2 ;
merge ycombined2 test3_2y ;
by nid ;
run ;

data formy3 ;
set formy2 ;
if (nid >= 36 & nid <= 40) then do ;
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
if nid = 50 then do ;
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
drop b1 b2 b3 b4 ;
run ;

data formy22 ;
set formy3 ;
if (nid >= 36 & nid <= 40) then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
if nid = 50 then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
drop parb3 parb4 ;
run ;

data  _null_ ; set formy22 ;
file "&workdir\par_cond2y.dat" ;
put @1 nid  @10 slope @20 location @30 guessing @40 parb1 @50 parb2 ;
```

```
run ;

data readpar2 ;
infile 'c:\raw59\par_cond2.dat' ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_junk524 ;
merge readpar2 fun524 ;
by newid ;
run ;

data _null_ ;
set param_junk524 ;
file "&workdir\testrun" ;
put newid type numresp ;
if numresp = 0 then do ;
end ;
if numresp = 2 then do ;
put c1 c2 ;
put "1.7 " slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put "1.7 " slope location guessing parb1 parb2 ;
end ;
run ;

data readpary2 ;
infile 'c:\raw59\par_cond2y.dat' ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_fun524y ;
merge readpary2 fun524 ;
by newid ;
run ;

data _null_ ;
set param_fun524y ;
file "&workdir\testruny" ;
put newid type numresp ;
if numresp = 0 then do ;
end ;
if numresp = 2 then do ;
put c1 c2 ;
put "1.7 " slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put "1.7 " slope location guessing parb1 parb2 ;
end ;
run ;


%end ;
```

```
%if &anchor = 3 %then %do ;

x 'cd C:\raw59' ;
x 'psl0.exe condition3' ;
x 'psl1.exe condition3' ;
x 'psl2.exe condition3' ;


data par_cond3 ;
infile "&workdir\condition3.ph2" truncover ;
input string $128.  ;
if (string = '| ITEM |BLOCK|  SLOPE  |   S.E.  |LOCATION |   S.E.  |GUESSING
|   S.E.  |') then do ;
do i = 1 to 90 ;
input ;
input item $2-6 block $9-12 slope $15-22 se $25-32 location $35-42 sel $45-52
guessing $55-62 seg $65-72 ;
output ;
end ;
end ;
drop string i ;
run ;

data par_cond3b ;
set par_cond3 ;
if (location = 'NaN' | location = '0.000') then delete ;
run ;

data cond3;
set par_cond3b;
nid=_N_;
keep nid slope location guessing ;
run ;

data par_cond3x;
set cond3;
if nid <= 50 then output;
run;

data par_cond3y;
set cond3;
if nid > 50 then output;
run;

data par_cond33y;
set par_cond3y;
newid=_N_;
keep newid slope location guessing;
run;

data par_cond333y;
set par_cond33y;
if newid <= 10 then output;
run;

data par_cond3333y;
set par_cond33y;
```

```
if newid > 10 then output;
run;

data ycombined3;
set par_cond3333y par_cond333y;
drop newid;
nid=_N_;
run;

data test3 ;
infile "&workdir\condition3.ph2" truncover ;
input string $128.  ;
if (string = '[GROUP:   1  OLD     ]') then do ;
do i = 1 to 90 ;
do j = 1 to 1 ;
input ;
end;
do k = 1 to 3;
input type $ b1 $27-34 b2 $38-44 b3 $49-54 b4 $59-65  ;
output ;
end ;
end ;
end;
drop string i j k;
run ;

data test3_3 ;
set test3 ;
if type ~= 'CATEGORY' then delete ;
if b1=0.000 then delete;
run ;

data test4_3 ;
set test3_3 ;
id=_N_;
if (id >= 1 & id <= 5) then nid = id + 35 ;
else if id = 6 then nid=49;
else if id = 7 then nid=50;
drop id type ;
run ;


data formx3 ;
merge par_cond3x test4_3 ;
by nid ;
run ;

data formx4 ;
set formx3 ;
if (nid >= 36 & nid <= 40) then do ;
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
if (nid = 49 | nid = 50) then do ;
parb1 = location - b1 ;
```

```
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
drop b1 b2 b3 b4 ;
run ;

data formx5 ;
set formx4 ;
if (nid >= 36 & nid <= 40) then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
if (nid = 49 | nid = 50) then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
drop parb3 parb4 ;
run ;

data  _null_ ; set formx5 ;
file "&workdir\par_cond3.dat" ;
put @1 nid  @10 slope @20 location @30 guessing @40 parb1 @50 parb2 ;
run ;


data test3y ;
infile "&workdir\condition3.ph2" truncover ;
input string $128.  ;
if (string = '[GROUP:   2  NEW     ]') then do ;
do i = 1 to 90 ;
do j = 1 to 1 ;
input ;
end;
do k = 1 to 3;
input type $ b1 $27-34 b2 $38-44 b3 $49-54 b4 $59-65  ;
output ;
end ;
end ;
end;
drop string i j k;
run ;

data test3_3y ;
set test3y ;
if type ~= 'CATEGORY' then delete ;
if b1=0.000 then delete;
run ;

data test4y ;
set test3_3y ;
id=_N_;
if id = 1 then nid=49;
```

```
else if id = 2 then nid=50 ;
else if (id >= 3 & id <= 7) then nid = id + 33 ;
drop id type ;
run ;

proc sort data = test4y;
By nid;
run;

data formy3 ;
merge ycombined3 test4y ;
by nid ;
run ;

data formy4 ;
set formy3 ;
if (nid >= 36 & nid <= 40) then do ;
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
if (nid = 49 | nid = 50) then do ;
parb1 = location - b1 ;
parb2 = location - b2 ;
parb3 = location - b3 ;
parb4 = location - b4 ;
end ;
drop b1 b2 b3 b4 ;
run ;

data formy5 ;
set formy4 ;
if (nid >= 36 & nid <= 40) then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
if (nid = 49 | nid = 50) then do ;
location = parb1 ;
guessing = parb2 ;
parb1 = parb3 ;
parb2 = parb4 ;
end ;
drop parb3 parb4 ;
run ;


data  _null_ ; set formy5 ;
file "&workdir\par_cond3y.dat" ;
put @1 nid  @10 slope @20 location @30 guessing @40 parb1 @50 parb2 ;
run ;

data readpar3 ;
infile "&workdir\par_cond3.dat" ;
input newid slope location guessing parb1 parb2 ;
```

```
run ;

data param_fun613 ;
merge readpar3 fun613 ;
by newid ;
run ;


data _null_ ;
set param_fun613 ;
file "&workdir\testrun" ;
put newid type numresp ;
if numresp = 0 then do ;
end ;
if numresp = 2 then do ;
put c1 c2 ;
put "1.7 " slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put "1.7 " slope location guessing parb1 parb2 ;
end ;
run ;

data readpary3 ;
infile "&workdir\par_cond3y.dat" ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_fun613y ;
merge readpary3 fun613 ;
by newid ;
run ;

data _null_ ;
set param_fun613y ;
file "&workdir\testruny" ;
put newid type numresp ;
if numresp = 0 then do ;
end ;
if numresp = 2 then do ;
put c1 c2 ;
put "1.7 " slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put "1.7 " slope location guessing parb1 parb2 ;
end ;
run ;


%end;

%if &anchor = 1 %then %do ;

data posteriorweights ;
infile "&workdir\condition1a.ph2" truncover ;
```

```
input string $128.  ;
if (string = 'QUADRATURE POINTS AND POSTERIOR WEIGHTS:') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
            end ;
      do j = 1 to 3 ;
      input type $ i1-i5 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data posteriorweights2 ;
infile "&workdir\condition1a.ph2" truncover ;
input string $128.  ;
if (string = 'GROUP   2   GROUP NAME: NEW ') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
      input;
            end ;
      do j = 1 to 3 ;
      input type $ i1-i5 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data postpoint ;
set posteriorweights ;
if type ~= 'POINT' then delete ;
run ;

proc transpose data = postpoint out=postgroup1 ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew1 ;
set postgroup1(rename=(col1=point))
postgroup1(rename=(col2=point))
postgroup1(rename=(col3=point))
postgroup1(rename=(col4=point))
postgroup1(rename=(col5=point))
postgroup1(rename=(col6=point)) ;
run ;

data postweight ;
set posteriorweights ;
if type ~= 'WEIGHT' then delete ;
run ;
```

```
proc transpose data = postweight out=postgroup1wt ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew1wt ;
set postgroup1wt(rename=(col1=weight))
postgroup1wt(rename=(col2=weight))
postgroup1wt(rename=(col3=weight))
postgroup1wt(rename=(col4=weight))
postgroup1wt(rename=(col5=weight))
postgroup1wt(rename=(col6=weight)) ;
run ;

data postpoints ;
set postgroupnew1 (keep = point) ;
run ;
data postweights ;
set postgroupnew1wt (keep = weight) ;
run ;
data combinedwts ;
merge postpoints postweights ;
if weight = . then delete;
run ;

options nonumber nodate nocenter ;
title 'qudrature points' ;

data _null_ ;
set combinedwts ;
file 'c:\raw59\qfunx1' print title ;
put point weight 12.10 ;
run ;


data postpoint2 ;
set posteriorweights2 ;
if type ~= 'POINT' then delete ;
run ;

proc transpose data = postpoint2 out=postgroup2 ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew2 ;
set postgroup2(rename=(col1=point))
postgroup2(rename=(col2=point))
postgroup2(rename=(col3=point))
postgroup2(rename=(col4=point))
postgroup2(rename=(col5=point))
postgroup2(rename=(col6=point)) ;
run ;

data postweight2 ;
set posteriorweights2 ;
if type ~= 'WEIGHT' then delete ;
```

```
run ;

proc transpose data = postweight2 out=postgroup2wt ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew2wt ;
set postgroup2wt(rename=(col1=weight))
postgroup2wt(rename=(col2=weight))
postgroup2wt(rename=(col3=weight))
postgroup2wt(rename=(col4=weight))
postgroup2wt(rename=(col5=weight))
postgroup2wt(rename=(col6=weight)) ;
run ;

data postpoints2 ;
set postgroupnew2 (keep = point) ;
run ;
data postweights2 ;
set postgroupnew2wt (keep = weight) ;
run ;
data combinedwts2 ;
merge postpoints2 postweights2 ;
if weight = . then delete;
run ;


options nonumber nodate nocenter ;
title 'qudrature points' ;
data _null_ ;
set combinedwts2 ;
file 'c:\raw59\qfuny1' print title ;
put point weight 12.10 ;
run ;

%end;

%if &anchor = 2 %then %do;

data posteriorweights21 ;
infile "&workdir\condition2.ph2" truncover ;
input string $128.  ;
if (string = 'QUADRATURE POINTS AND POSTERIOR WEIGHTS:') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
            end ;
      do j = 1 to 3 ;
      input type $ i1-i5 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;
```

```
data posteriorweights22 ;
infile "&workdir\condition2.ph2" truncover ;
input string $128.  ;
if (string = 'GROUP   2   GROUP NAME: NEW ') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
      input;
            end ;
      do j = 1 to 3 ;
      input type $ i1-i5 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data postpoint21 ;
set posteriorweights21 ;
if type ~= 'POINT' then delete ;
run ;

proc transpose data = postpoint21 out=postgroup21 ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew21 ;
set postgroup21(rename=(col1=point))
postgroup21(rename=(col2=point))
postgroup21(rename=(col3=point))
postgroup21(rename=(col4=point))
postgroup21(rename=(col5=point))
postgroup21(rename=(col6=point)) ;
run ;

data postweight21 ;
set posteriorweights21 ;
if type ~= 'WEIGHT' then delete ;
run ;

proc transpose data = postweight21 out=postgroup21wt ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew21wt ;
set postgroup21wt(rename=(col1=weight))
postgroup21wt(rename=(col2=weight))
postgroup21wt(rename=(col3=weight))
postgroup21wt(rename=(col4=weight))
postgroup21wt(rename=(col5=weight))
postgroup21wt(rename=(col6=weight)) ;
run ;

data postpoints21 ;
```

```
set postgroupnew21 (keep = point) ;
run ;
data postweights21 ;
set postgroupnew21wt (keep = weight) ;
run ;
data combinedwts21 ;
merge postpoints21 postweights21 ;
if weight = . then delete;
run ;

options nonumber nodate nocenter ;
title 'qudrature points' ;

data _null_ ;
set combinedwts21 ;
file 'c:\raw59\qfunx2' print title ;
put point weight 12.10 ;
run ;


data postpoint22 ;
set posteriorweights22 ;
if type ~= 'POINT' then delete ;
run ;

proc transpose data = postpoint22 out=postgroup22 ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew22 ;
set postgroup22(rename=(col1=point))
postgroup22(rename=(col2=point))
postgroup22(rename=(col3=point))
postgroup22(rename=(col4=point))
postgroup22(rename=(col5=point))
postgroup22(rename=(col6=point)) ;
run ;

data postweight22 ;
set posteriorweights22 ;
if type ~= 'WEIGHT' then delete ;
run ;

proc transpose data = postweight22 out=postgroup22wt ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew22wt ;
set postgroup22wt(rename=(col1=weight))
postgroup22wt(rename=(col2=weight))
postgroup22wt(rename=(col3=weight))
postgroup22wt(rename=(col4=weight))
postgroup22wt(rename=(col5=weight))
postgroup22wt(rename=(col6=weight)) ;
run ;
```

```
data postpoints22 ;
set postgroupnew22 (keep = point) ;
run ;
data postweights22 ;
set postgroupnew22wt (keep = weight) ;
run ;
data combinedwts22 ;
merge postpoints22 postweights22 ;
if weight = . then delete;
run ;


options nonumber nodate nocenter ;
title 'qudrature points' ;
data _null_ ;
set combinedwts22 ;
file 'c:\raw59\qfuny2' print title ;
put point weight 12.10 ;
run ;

%end;

%if &anchor = 3 %then %do;

data posteriorweights31 ;
infile "&workdir\condition3.ph2" truncover ;
input string $128.  ;
if (string = 'QUADRATURE POINTS AND POSTERIOR WEIGHTS:') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
             end ;
      do j = 1 to 3 ;
      input type $ i1-i5 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data posteriorweights32 ;
infile "&workdir\condition3.ph2" truncover ;
input string $128.  ;
if (string = 'GROUP   2   GROUP NAME: NEW ') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
      input;
            end ;
      do j = 1 to 3 ;
      input type $ i1-i5 ;
      output ;
      end ;
      end ;
   stop ;
end ;
```

```
drop i j k string  ;
run ;

data postpoint31 ;
set posteriorweights31 ;
if type ~= 'POINT' then delete ;
run ;

proc transpose data = postpoint31 out=postgroup31 ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew31 ;
set postgroup31(rename=(col1=point))
postgroup31(rename=(col2=point))
postgroup31(rename=(col3=point))
postgroup31(rename=(col4=point))
postgroup31(rename=(col5=point))
postgroup31(rename=(col6=point)) ;
run ;

data postweight31 ;
set posteriorweights31 ;
if type ~= 'WEIGHT' then delete ;
run ;

proc transpose data = postweight31 out=postgroup31wt ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew31wt ;
set postgroup31wt(rename=(col1=weight))
postgroup31wt(rename=(col2=weight))
postgroup31wt(rename=(col3=weight))
postgroup31wt(rename=(col4=weight))
postgroup31wt(rename=(col5=weight))
postgroup31wt(rename=(col6=weight)) ;
run ;

data postpoints31 ;
set postgroupnew31 (keep = point) ;
run ;
data postweights31 ;
set postgroupnew31wt (keep = weight) ;
run ;
data combinedwts31 ;
merge postpoints31 postweights31 ;
if weight = . then delete;
run ;

options nonumber nodate nocenter ;
title 'qudrature points' ;

data _null_ ;
set combinedwts31 ;
```

```
file 'c:\raw59\qfunx3' print title ;
put point weight 12.10 ;
run ;


data postpoint32 ;
set posteriorweights32 ;
if type ~= 'POINT' then delete ;
run ;

proc transpose data = postpoint32 out=postgroup32 ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew32 ;
set postgroup32(rename=(col1=point))
postgroup32(rename=(col2=point))
postgroup32(rename=(col3=point))
postgroup32(rename=(col4=point))
postgroup32(rename=(col5=point))
postgroup32(rename=(col6=point)) ;
run ;

data postweight32 ;
set posteriorweights32 ;
if type ~= 'WEIGHT' then delete ;
run ;

proc transpose data = postweight32 out=postgroup32wt ;
by type ;
var i1 i2 i3 i4 i5 ;
run ;

data postgroupnew32wt ;
set postgroup32wt(rename=(col1=weight))
postgroup32wt(rename=(col2=weight))
postgroup32wt(rename=(col3=weight))
postgroup32wt(rename=(col4=weight))
postgroup32wt(rename=(col5=weight))
postgroup32wt(rename=(col6=weight)) ;
run ;

data postpoints32 ;
set postgroupnew32 (keep = point) ;
run ;
data postweights32 ;
set postgroupnew32wt (keep = weight) ;
run ;
data combinedwts32 ;
merge postpoints32 postweights32 ;
if weight = . then delete;
run ;


options nonumber nodate nocenter ;
title 'qudrature points' ;
```

```
data _null_ ;
set combinedwts32 ;
file 'c:\raw59\qfuny3' print title ;
put point weight 12.10 ;
run ;

%end;


%if &anchor = 1 %then %do ;
x "cd &workdir" ;
x 'poly.bat' ;

data equatetrue ;
infile "&workdir\dissertation.out" truncover ;
input string $128.  ;
if (string = 'x-raw   theta       y-equiv     scale     rounded') then do ;
   do k = 1 to 11 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 6 ;
      input raw i1-i4 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data truequ ;
set equatetrue ;
i2r=round(i2) ;
difft1 = i2-raw ;
keep raw i2 i2r difft1;
run ;

data KStr;
set truequ;
group = 1;
keep raw group;
run;

data Kstr;
set Kstr;
rename raw = score;
run;

data KStr1;
set truequ;
group = 2;
keep i2 group;
run;

data Kstr1;
set Kstr1;
rename i2 = score;
```

```
run;


data Kstr4;
set Kstr Kstr1;
run;

proc npar1way data = KStr4 edf noprint ;
class group;
var score;
output out = KStr5 edf;
run;


data KS_testtr ;
set KStr5 (keep = _D_ P_KSA);
rename _d_ = kstr;
rename p_ksa = ptr;
run;

data equateobst ;
infile "&workdir\dissertation.out" truncover ;
input string $128.  ;
if (string = 'observed score equivalents') then do ;
input ;
input ;
   do k = 1 to 11 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 6 ;
      input raw i1-i3 ;
      output ;
      end ;
      end ;
   stop ;
end ; * if ;
drop i j k string  ;
run ;

data obsequ ;
set equateobst ;
i1r=round(i1) ;
diffob1 = i1-raw ;
keep raw i1 i1r diffob1 ;
run ;


data KS;
set obsequ;
group = 1;
keep raw group;
run;

data KS;
set KS;
rename raw = score1;
run;
```

```
data KS_1;
set obsequ;
group = 2;
keep i1 group;
run;

data KS_1;
set KS_1;
rename i1 = score1;
run;


data KS_com1;
set KS KS_1;
run;

proc npar1way data = KS_com1 edf noprint ;
class group;
var score1;
output out = KS_com2 edf;
run;

data KS_testobs ;
set KS_com2 (keep = _D_ P_KSA);
rename _d_ = ksob;
rename p_ksa = pob;
run;


data eqfunctionfr ;
infile "&workdir\concordfr1.dat" ;
input raw yequ ;
yequr=round(yequ) ;
diffr1 = yequ-raw ;
run ;


data KSfr;
set eqfunctionfr;
group = 1;
keep raw group;
run;

data KSfr;
set KSfr;
rename raw = score2;
run;

data KSfr1;
set eqfunctionfr;
group = 2;
keep yequ group;
run;

data KSfr1;
set KSfr1;
```

```
rename yequ = score2;
run;

data KSfr5;
set KSfr Ksfr1;
run;

proc npar1way data = KSfr5 edf noprint;
class group;
var score2;
output out = KSfr6 edf ;
run;

data KS_testfr ;
set KSfr6 (keep = _D_ P_KSA);
rename _d_ = ksfr;
rename p_ksa = pfr;
run;

data eqfunctionch ;
infile "&workdir\concordch1.dat" ;
input raw yequ2 ;
yequr=round(yequ2) ;
diffch1 = yequ2-raw ;
run ;

data KSch1;
set eqfunctionch;
group = 1;
keep raw group;
run;

data KSch1;
set KSch1;
rename raw = score3;
run;

data KSch11;
set eqfunctionch;
group = 2;
keep yequ2 group;
run;

data KSch11;
set KSch11;
rename yequ2 = score3;
run;

data KSch51;
set KSch1 KSch11;
run;

proc npar1way data = KSch51 edf noprint;
class group;
var score3;
output out = KSchn edf  ;
run;
```

```
data KS_testch ;
set KSchn (keep = _D_ P_KSA) ;
rename _d_ = ksch;
rename p_ksa = pch;
run;

%end;


%if &anchor = 2 %then %do;
x "cd &workdir" ;
x 'poly2.bat' ;

data equatetrue2 ;
infile 'c:\raw59\dissertation2.out' truncover ;
input string $128.  ;
if (string = "x-raw   theta       y-equiv     scale     rounded") then do ;
   do k = 1 to 23 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 3 ;
      input raw i1-i4 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data truequ2 ;
set equatetrue2 ;
i2r=round(i2) ;
difft1 = i2-raw ;
keep raw i2 i2r difft1;
run ;

data KStr2;
set truequ2;
group = 1;
keep raw group;
run;

data KStr2;
set KStr2;
rename raw = score5;
run;

data KStr22;
set truequ2;
group = 2;
keep i2 group;
run;

data KStr22;
set KStr22;
```

```
rename i2 = score5;
run;

data KStr42;
set KStr2 KStr22;
run;


proc npar1way data = KStr42 edf noprint;
class group;
var score5;
output out = KStr52 edf;
run;

data KS_testtr ;
set KStr52(keep = _D_ P_KSA);
rename _d_ = kstr;
rename p_ksa = ptr;
run;

data equateobst2 ;
infile 'c:\raw59\dissertation2.out' truncover ;
input string $128.  ;
if (string = "observed score equivalents") then do ;
input ;
input ;
   do k = 1 to 69 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 1 ;
      input raw i1-i3 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data obsequ2 ;
set equateobst2 ;
i1r=round(i1) ;
diffob1 = i1-raw ;
keep raw i1 i1r diffob1;
run ;

data KSobs2;
set obsequ2;
group = 1;
keep raw group;
run;

data KSobs2;
set Ksobs2;
rename raw = score6;
run;
```

```
data KSobs21;
set obsequ2;
group = 2;
keep i1 group;
run;

data KSobs21;
set Ksobs21;
rename i1 = score6;
run;

data KS23;
set Ksobs2 Ksobs21;
run;


proc npar1way data = KS23 edf noprint ;
class group;
var score6;
output out = KS_obs2 edf ;
run;

data KS_testobs ;
set KS_obs2 (keep = _D_ P_KSA);
rename _d_ = ksob;
rename p_ksa = pob;
run;

data eqfunctionfr2 ;
infile "&workdir\concordfr2.dat" ;
input raw yequ ;
yequr=round(yequ) ;
diffr1 = yequ-raw ;
run ;

data KSfr21;
set eqfunctionfr2;
group = 1;
keep raw group;
run;

data Ksfr21;
set Ksfr21;
rename raw = score7;
run;

data KSfr22;
set eqfunctionfr2;
group = 2;
keep yequ group;
run;

data Ksfr22;
set Ksfr22;
rename yequ = score7;
run;
```

```
data KSfr25;
set Ksfr21 Ksfr22;
run;

proc npar1way data = KSfr25 edf noprint ;
class group;
var score7;
output out = KS2fr edf;
run;

data KS_testfr ;
set KS2fr (keep = _D_ P_KSA);
rename _d_ = ksfr;
rename p_ksa = pfr;
run;

data eqfunctionch2 ;
infile "&workdir\concordch2.dat" ;
input raw yequ2 ;
yequr=round(yequ2) ;
diffch1 = yequ2-raw ;
run ;

data KSch21;
set eqfunctionch2;
group = 1;
keep raw group;
run;

Data KSch21;
set KSch21;
rename raw = score8;
run;

data KSch22;
set eqfunctionch2;
group = 2;
keep yequ2 group;
run;

data KSch22;
set KSch22;
rename yequ2 = score8;
run;

data Ksch25;
set Ksch21 Ksch22;
run;


proc npar1way data = KSch25 edf noprint ;
class group;
var score8;
output out = KSch26 edf;
run;

data KS_testch ;
```

```
set KSch26 (keep = _D_ P_KSA);
rename _d_ = ksch;
rename p_ksa = pch;
run;


%end;


%if &anchor = 3 %then %do;
x "cd &workdir" ;
x 'poly3.bat' ;

data equatetrue3 ;
infile 'c:\raw59\dissertation3.out' truncover ;
input string $128.  ;
if (string = "x-raw   theta       y-equiv     scale     rounded") then do ;
   do k = 1 to 12 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 6 ;
      input raw i1-i4 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data truequ3 ;
set equatetrue3 ;
i2r=round(i2) ;
difft1 = i2-raw ;
keep raw i2 i2r difft1 ;
run ;

data KStr3;
set truequ3;
group = 1;
keep raw group;
run;

data KStr3;
set KStr3;
rename raw = score9;
run;

data KStr31;
set truequ3;
group = 2;
keep i2 group;
run;

Data KStr31;
set KStr31;
rename i2 = score9;
```

```
run;

data KStr34;
set KStr3 KStr31;
run;

proc npar1way data = KStr34 edf noprint ;
class group;
var score9;
output out = KStr53 edf;
run;

data KS_testtr ;
set KStr53 (keep = _D_ P_KSA);
rename _d_ = kstr;
rename p_ksa = ptr;
run;

data equateobst3 ;
infile 'c:\raw59\dissertation3.out' truncover ;
input string $128.  ;
if (string = "observed score equivalents") then do ;
input ;
input ;
   do k = 1 to 72 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 1 ;
      input raw i1-i3 ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;

data obsequ3 ;
set equateobst3 ;
i1r=round(i1) ;
diffob1 = i1-raw ;
keep raw i1 i1r diffob1;
 run ;

data KSobs31;
set obsequ3;
group = 1;
keep raw group;
run;

data KSobs31;
set KSobs31;
rename raw = score10;
run;

data KSobs32;
set obsequ3;
```

```
group = 2;
keep i1 group;
run;

data KSobs32;
set KSobs32;
rename i1 = score10;
run;

data KSobs34;
set KSobs31 KSobs32;
run;

proc npar1way data = KSobs34 edf noprint ;
class group;
var score10;
output out = KS_obs3 edf;
run;

data KS_testobs ;
set KS_obs3 (keep = _D_ P_KSA);
rename _d_ = ksob;
rename p_ksa = pob;
run;

data eqfunctionfr3 ;
infile "&workdir\concordfr3.dat" ;
input raw yequ ;
yequr=round(yequ) ;
diffr1 = yequ-raw ;
run ;

data KSfr31;
set eqfunctionfr3;
group = 1;
keep raw group;
run;

data KSfr31;
set KSfr31;
rename raw = score11;
run;

data KSfr32;
set eqfunctionfr3;
group = 2;
keep yequ group;
run;

data KSfr32;
set KSfr32;
rename yequ = score11;
run;

data KSfr35;
set KSfr31 KSfr32;
run;
```

```
proc npar1way data = KSfr35 edf  noprint ;
class group;
var score11;
output out = KSfr7 edf;
run;

data KS_testfr ;
set KSfr7 (keep = _D_ P_KSA);
rename _d_ = ksfr;
rename p_ksa = pfr;
run;

data eqfunctionch3 ;
infile "&workdir\concordch3.dat" ;
input raw yequ3 ;
yequr=round(yequ2) ;
diffch1 = yequ3-raw ;
run ;

data KSch31;
set eqfunctionch3;
group = 1;
keep raw group;
run;

data Ksch31;
set Ksch31;
rename raw = score12;
run;

data KSch32;
set eqfunctionch3;
group = 2;
keep yequ3 group;
run;

data Ksch32;
set KSch32;
rename yequ3 = score12;
run;

data KSch35;
set KSch31 KSch32;
run;

proc npar1way data = KSch35 edf noprint ;
class group;
var score12;
output out = KS_3ch edf;
run;

data KS_testch ;
set KS_3ch (keep = _D_ P_KSA);
rename _d_ = ksch;
rename p_ksa = pch;
run;
```

```
%end;


%if &anchor = 1 %then %do ;

data readpar ;
infile "&workdir\par_cond1.dat" ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_junk524 ;
merge readpar junk524 ;
by newid ;
run ;


data _null_ ;
set param_junk524 ;
file "&workdir\testrun" ;
put newid type numresp ;
if numresp = 2 then do ;
put c1 c2 ;
put '1.7 ' slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put '1.7 ' slope location guessing parb1 parb2 ;
end ;
run ;

data readpary ;
infile "&workdir\par_cond1y.dat" ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_junk524y ;
merge readpary junk524 ;
by newid ;
run ;

data _null_ ;
set param_junk524y ;
file "&workdir\testruny" ;
put newid type numresp ;
if numresp = 2 then do ;
put c1 c2 ;
put '1.7 ' slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put '1.7 ' slope location guessing parb1 parb2 ;
end ;
run ;


options nonumber nodate nocenter  PS = 100;
```

```
title 'Raw Score Equivalents Form X' ;

data _null_ ;
set rawscore_con1 ;
file "&workdir\rawscx1.rs" print title ;
put @1 r1 @5 r2 @10 r3 ;
run ;


options nonumber nodate nocenter ;
title 'quadrature points' ;

data _null_ ;
set combinedwts2 ;
file "&workdir\funx1.pst" print title ;
put point weight 12.10 ;
run ;


x 'cd C:\raw59' ;
x 'polycsx1.bat' ;


data eqfunctionfr ;
infile "&workdir\concordfr1.dat" ;
input raw yequ ;
yequr=round(yequ) ;
diffr1 = yequ-raw ;
run ;

data eqfunctionch ;
infile "&workdir\concordch1.dat" ;
input raw yequ2 ;
yequr=round(yequ2) ;
diffch1 = yequ2-raw ;
run ;

options nonumber nodate nocenter PS=100 ;
title 'Raw Score Equivalents' ;

data _null_ ;
set eqfunctionfr ;
file "&workdir\rawscy1.rs" print title ;
put raw yequ yequr ;
run ;

options nonumber nodate nocenter PS=100 ;
title 'Raw Score Equivalents' ;

data _null_ ;
set eqfunctionch ;
file "&workdir\rawscy2.rs" print title ;
put raw yequ2 yequr ;
run ;

options nonumber nodate nocenter PS=100 ;
title 'Raw Score Equivalents' ;
```

```
data _null_ ;
set truequ ;
file "&workdir\rawscy3.rs" print title ;
put raw i2 i2r ;
run ;

options nonumber nodate nocenter PS=100 ;
title 'Raw Score Equivalents' ;

data _null_ ;
set obsequ ;
file "&workdir\rawscy4.rs" print title ;
put raw i1 i1r ;
run ;

options nonumber nodate nocenter ;
title 'quadrature points' ;

data _null_ ;
set combinedwts2 ;
file "&workdir\funy1.pst" print title ;
put point weight 12.10 ;
run ;


x 'cd C:\raw59' ;
x 'polycsy1.bat' ;

x 'cd C:\raw59' ;
x 'polycsy2.bat' ;

x 'cd C:\raw59' ;
x 'polycsy3.bat' ;

x 'cd C:\raw59' ;
x 'polycsy4.bat' ;


%end ;


%if &anchor = 2 %then %do ;

data readpar2 ;
infile 'c:\raw59\par_cond2.dat' ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_junk524 ;
merge readpar2 fun524 ;
by newid ;
run ;

data _null_ ;
set param_junk524 ;
file "&workdir\testrun" ;
```

```
put newid type numresp ;
if numresp = 0 then do ;
end ;
if numresp = 2 then do ;
put c1 c2 ;
put "1.7 " slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put "1.7 " slope location guessing parb1 parb2 ;
end ;
run ;

data readpary2 ;
infile 'c:\raw59\par_cond2y.dat' ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_fun524y ;
merge readpary2 fun524 ;
by newid ;
run ;

data _null_ ;
set param_fun524y ;
file "&workdir\testruny" ;
put newid type numresp ;
if numresp = 0 then do ;
end ;
if numresp = 2 then do ;
put c1 c2 ;
put "1.7 " slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put "1.7 " slope location guessing parb1 parb2 ;
end ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents Form X" ;
data _null_ ;
set rawscore_con2 ;
file "&workdir\rawscx2.rs" print title ;
put @1 r1 @5 r2 @10 r3 ;
run ;

options nonumber nodate nocenter ;
title "quadrature points" ;
data _null_ ;
set combinedwts22 ;
file "&workdir\funx2.pst" print title ;
put point weight 12.10 ;
run ;

x 'cd "C:\raw59"' ;
x 'polycsx2.bat' ;
```

```
data eqfunctionfr2 ;
infile "&workdir\concordfr2.dat" ;
input raw yequ ;
yequr=round(yequ) ;
diffr1 = yequ-raw ;
run ;

data eqfunctionch2 ;
infile "&workdir\concordch2.dat" ;
input raw yequ2 ;
yequr=round(yequ2) ;
diffch1 = yequ2-raw ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents" ;
data _null_ ;
set eqfunctionfr2 ;
file "&workdir\rawscy12.rs" print title ;
put raw yequ yequr ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents" ;
data _null_ ;
set eqfunctionch2 ;
file "&workdir\rawscy22.rs" print title ;
put raw yequ2 yequr ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents" ;
data _null_ ;
set truequ2 ;
file "&workdir\rawscy32.rs" print title ;
put raw i2 i2r ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents" ;
data _null_ ;
set obsequ2 ;
file "&workdir\rawscy42.rs" print title ;
put raw i1 i1r ;
run ;

options nonumber nodate nocenter ;
title "quadrature points" ;
data _null_ ;
set combinedwts22 ;
file "&workdir\funy2.pst" print title ;
put point weight 12.10 ;
run ;

x 'cd "C:\raw59"' ;
x 'polycsy12.bat' ;
```

```
x 'cd "C:\raw59"' ;
x 'polycsy22.bat' ;

x 'cd "C:\raw59"' ;
x 'polycsy32.bat' ;

x 'cd "C:\raw59"' ;
x 'polycsy42.bat' ;

%end;

%if &anchor = 3 %then %do;

data readpar3 ;
infile "&workdir\par_cond3.dat" ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_fun613 ;
merge readpar3 fun613 ;
by newid ;
run ;


data _null_ ;
set param_fun613 ;
file "&workdir\testrun" ;
put newid type numresp ;
if numresp = 0 then do ;
end ;
if numresp = 2 then do ;
put c1 c2 ;
put "1.7 " slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put "1.7 " slope location guessing parb1 parb2 ;
end ;
run ;

data readpary3 ;
infile "&workdir\par_cond3y.dat" ;
input newid slope location guessing parb1 parb2 ;
run ;

data param_fun613y ;
merge readpary3 fun613 ;
by newid ;
run ;

data _null_ ;
set param_fun613y ;
file "&workdir\testruny" ;
put newid type numresp ;
if numresp = 0 then do ;
end ;
```

```
if numresp = 2 then do ;
put c1 c2 ;
put "1.7 " slope location guessing ;
end ;
if numresp = 5 then do ;
put c1 c2 c3 c4 c5 ;
put "1.7 " slope location guessing parb1 parb2 ;
end ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents Form X" ;
data _null_ ;
set rawscore_con3 ;
file "&workdir\rawscx3.rs" print title ;
put @1 r1 @5 r2 @10 r3 ;
run ;


options nonumber nodate nocenter ;
title "quadrature points" ;
data _null_ ;
set combinedwts32 ;
file "&workdir\funx3.pst" print title ;
put point weight 12.10 ;
run ;

x 'cd "C:\raw59"' ;
x 'polycsx3.bat' ;

data eqfunctionfr3 ;
infile "&workdir\concordfr3.dat" ;
input raw yequ ;
yequr=round(yequ) ;
diffr1 = yequ-raw ;
run ;


data eqfunctionch3 ;
infile "&workdir\concordch3.dat" ;
input raw yequ3 ;
yequr=round(yequ3) ;
diffch1 = yequ3-raw ;
run ;


options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents" ;
data _null_ ;
set eqfunctionfr3 ;
file "&workdir\rawscy13.rs" print title ;
put raw yequ yequr ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents" ;
data _null_ ;
```

```
set eqfunctionch3 ;
file "&workdir\rawscy23.rs" print title ;
put raw yequ3 yequr ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents" ;
data _null_ ;
set truequ3 ;
file "&workdir\rawscy33.rs" print title ;
put raw i2 i2r ;
run ;

options nonumber nodate nocenter PS=100 ;
title "Raw Score Equivalents" ;
data _null_ ;
set obsequ3 ;
file "&workdir\rawscy43.rs" print title ;
put raw i1 i1r ;
run ;


options nonumber nodate nocenter ;
title "quadrature points" ;
data _null_ ;
set combinedwts32 ;
file "&workdir\funy3.pst" print title ;
put point weight 12.10 ;
run ;

x 'cd "C:\raw59"'  ;
x 'polycsy13.bat'  ;

x 'cd "C:\raw59"'  ;
x 'polycsy23.bat'  ;

x 'cd "C:\raw59"' ;
x 'polycsy33.bat' ;

x 'cd "C:\raw59"' ;
x 'polycsy43.bat' ;

%end ;



data sdoldform ;
infile "&workdir\disx.out" truncover ;
input string $128.  ;
if (string = 'marginal results for raw scores') then do ;
input ;
input type $ 1-28 i1 31-40 i2 42-51 i3 54-62 i4 63-73 ;
output ;
keep i2 ;
end ;run ;
```

```
data resultsx ;
infile "&workdir\disx.out" truncover ;
input string $128.  ;
if (string = 'theta      exp raw    csem raw   exp sc     csem sc     exp rsc
csem rsc') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 3 ;
      input abil rawexpx rawcsemx scexpx sccsemx exprx csemrx ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;


data resultsyfr ;
infile "&workdir\disyfr.out" truncover ;
input string $128.  ;
if (string = 'theta      exp raw    csem raw   exp sc     csem sc     exp rsc
csem rsc') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 3 ;
      input abil rawexpy rawcsemy scexpy sccsemy expry csemry ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;


proc iml;

use sdoldform ;
read all var _num_ into sd;
close sdoldform ;

use resultsx ;
read all var {scexpx} into d1x;
close resultsx;

use resultsyfr ;
read all var {scexpy} into d1y;
close resultsyfr;

a = nrow(d1x) ;
Resd1 = j(a,1,0);
do i = 1 to a ;
```

```
Resd1 = sum(d1y-d1x)##2;
Resd1f = ((sqrt(Resd1))/sd);
end;

a = nrow(d1x) ;

Ref1 = j(a,1,0);
do i = 1 to a;

Ref1 = d1y-d1x ;
end;


D1_fr = Resd1f ;
create D1_fr1 from D1_fr
[colname = {D1_fr }] ;
append from D1_fr ;

fr_1 = Ref1 ;
create fr1e from fr_1
[colname = {fr1exp }] ;
append from fr_1 ;


use sdoldform ;
read all var _num_ into sd;
close sdoldform ;

use resultsx ;
read all var {sccsemx} into d2x;
close resultsx;

use resultsyfr ;
read all var {sccsemy} into d2y;
close resultsyfr;

a = nrow(d2x) ;
Resd2 = j(a,1,0);
do i = 1 to a ;


Resd2 = sum(d2y-d2x)##2;
Resd2f = ((sqrt(Resd2))/sd);
end;

a = nrow(d1x) ;
Refr = j(a,1,0) ;
do i = 1 to a;

Refr = d2y-d2x ;

end;

D2_fr = Resd2f ;
create D2_fr1 from D2_fr
[colname = {D2_fr }] ;
append from D2_fr ;
```

```
fr1cs = Refr ;
create fr1c from fr1cs
[colname = {fr1c }] ;
append from fr1cs ;

quit;


data resultsych ;
infile "&workdir\disych.out" truncover ;
input string $128.  ;
if (string = 'theta      exp raw    csem raw   exp sc     csem sc     exp rsc
csem rsc') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 3 ;
      input abil rawexpyc rawcsemyc scexpyc sccsemyc expryc csemryc ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;


proc iml;

use sdoldform ;
read all var _num_ into sd;
close sdoldform ;

use resultsx ;
read all var {scexpx} into d1x;
close resultsx;

use resultsych ;
read all var {scexpyc} into d1y;
close resultsych;

a = nrow(d1x) ;
Resd1 = j(a,1,0);
do i = 1 to a ;

Resd1 = sum(d1y-d1x)##2;
Resd1c = ((sqrt(Resd1))/sd);
end;

a = nrow(d1x) ;
Resch = j(a,1,0);
do i = 1 to a;

Resch = d1y-d1x ;

end;
```

```
D1_ch = Resd1c ;
create D1_ch1 from D1_ch
[colname = {D1_ch }] ;
append from D1_ch ;

ch1 = Resch ;
create ch1e from ch1
[colname = {ch1e }] ;
append from ch1 ;


use sdoldform ;
read all var _num_ into sd;
close sdoldform ;

use resultsx ;
read all var {sccsemx} into d2x;
close resultsx;

use resultsych ;
read all var {sccsemyc} into d2y;
close resultsych;

a = nrow(d2x) ;
Resd2c1 = j(a,1,0);
do i = 1 to a ;

Resd2c1 = sum(d2y-d2x)##2;
Resd2ch1 = ((sqrt(Resd2c1))/sd);
end;

a = nrow(d1x) ;
Rechc = j(a,1,0);
do i = 1 to a;

Rechc = d2y-d2x ;

end;

D2_ch1 = Resd2ch1 ;
create D2_ch1 from D2_ch1
[colname = {D2_Ch }] ;
append from D2_ch1 ;

ch1cs = Rechc;
create ch1c from ch1cs
[colname = {ch1c}] ;
append from ch1cs ;

quit;

data resultsytr ;
infile "&workdir\disytr.out" truncover ;
input string $128.  ;
if (string = 'theta      exp raw    csem raw   exp sc     csem sc     exp rsc
csem rsc') then do ;
```

```
   do k = 1 to 10 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 3 ;
      input abil rawexpy rawcsemy scexpy sccsemy expry csemry ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
run ;


proc iml;

use sdoldform ;
read all var _num_ into sd;
close sdoldform ;

use resultsx ;
read all var {scexpx} into d1x;
close resultsx;

use resultsytr ;
read all var {scexpy} into d1y;
close resultsytr;

a = nrow(d1x) ;
Resd1 = j(a,1,0);
do i = 1 to a ;


Resd1 = sum(d1y-d1x)##2;
Resd1tr = ((sqrt(Resd1))/sd);
end;

a = nrow(d1x) ;

Ret = j(a,1,0);
do i = 1 to a;

Ret = d1y-d1x ;

end;

D1_tr = Resd1tr ;
create D1_tr1 from D1_tr
[colname = {D1_tr }] ;
append from D1_tr ;

tr1 = Ret ;
create tr1e from tr1
[colname = {tr1e }] ;
append from tr1 ;
```

```
use sdoldform ;
read all var _num_ into sd;
close sdoldform ;

use resultsx ;
read all var {sccsemx} into d2x;
close resultsx;

use resultsytr ;
read all var {sccsemy} into d2y;
close resultsytr;

a = nrow(d2x) ;
Resd2 = j(a,1,0);
do i = 1 to a ;

Resd2 = sum(d2y-d2x)##2;
Resd2tr = ((sqrt(Resd2))/sd);
end;

a = nrow(d1x) ;
Retr = j(a,1,0);
do i = 1 to a;

Retr = d2y-d2x ;

end;

D2_tr = Resd2tr ;
create D2_tr1 from D2_tr
[colname = {D2_tr }] ;
append from D2_tr ;

tru1 = Retr ;
create tr1c from tru1
[colname = {tr1c }] ;
append from tru1 ;

quit;


data resultsyob ;
infile "&workdir\disyob.out" truncover ;
input string $128.  ;
if (string = 'theta      exp raw    csem raw   exp sc     csem sc     exp rsc
csem rsc') then do ;
   do k = 1 to 10 ;
   do i = 1 to 1 ;
                end ;
      do j = 1 to 3 ;
      input abil rawexpy rawcsemy scexpy sccsemy expry csemry ;
      output ;
      end ;
      end ;
   stop ;
end ;
drop i j k string  ;
```

```
run ;

proc iml ;

use sdoldform ;
read all var _num_ into sd ;
close sdoldform ;

use resultsx ;
read all var {scexpx} into d1x ;
close resultsx;

use resultsyob ;
read all var {scexpy} into d1y ;
close resultsyob ;

a = nrow(d1x) ;
Resd1 = j(a,1,0);
do i = 1 to a ;

Resd1 = sum(d1y-d1x)##2 ;
Resd1ob = ((sqrt(Resd1))/sd) ;
end;

a = nrow(d1x) ;
Re1 = j(a,1,0);
do i = 1 to a;

Re1 = d1y-d1x ;

end ;

D1_ob = Resd1ob ;
create D1_ob1 from D1_ob
[colname = {D1_ob }] ;
append from D1_ob ;

ob1 = Re1 ;
create ob1e from ob1
[colname = {ob1 }] ;
append from ob1 ;


use sdoldform ;
read all var _num_ into sd;
close sdoldform ;

use resultsx ;
read all var {sccsemx} into d2x ;
close resultsx ;

use resultsyob ;
read all var {sccsemy} into d2y ;
close resultsyob ;

a = nrow(d2x) ;
Resd2 = j(a,1,0);
```

```
do i = 1 to a ;

Resd2 = sum(d2y-d2x)##2 ;
Resd2ob = ((sqrt(Resd2))/sd) ;
end;

a = nrow(d1x) ;

Reoc = j(a,1,0);
do i = 1 to a;

Reoc = d2y-d2x ;

end ;

D2_ob = Resd2ob ;
create D2_ob1 from D2_ob
[colname = {D2_ob }] ;
append from D2_ob ;

ob1cs = Reoc ;
create ob1c from ob1cs
[colname = {ob1cse }] ;
append from ob1cs ;

quit;

data combinedd1d2 ;
set resetd1d2;
repnum = &replication ;
groupnum = &group ;
anchornum = &anchor ;
run;


data combinedd1d2 ;
merge D1_fr1 D2_fr1 D1_ch1 D2_ch1 D1_tr1 D2_tr1 D1_ob1 D2_ob1 ;
repnum = &replication ;
groupnum = &group ;
anchornum = &anchor ;
run ;

data _null_ ;
set combinedd1d2 ;
file "&workdir\resultsd1d2.dat" mod ;
put repnum groupnum anchornum D1_fr D1_ch D1_tr D1_ob D2_fr D2_ch D2_tr D2_ob
;
run ;


data combinedks;
set resetks;
groupnum = &group;
repnum = &replication;
anchornum = &anchor;
run;
```

```
data combinedks;
merge KS_testtr KS_testobs KS_testfr KS_testch  ;
groupnum = &group;
repnum = &replication;
anchornum = &anchor;
run;

data _null_ ;
set combinedks ;
file "&workdir\resultsks.dat" mod ;
put repnum groupnum anchornum KStr ptr KSob pob KSfr pfr KSch pch ;
run;


%if &anchor = 1 %then %do;

data sec_results ;
merge truequ obsequ eqfunctionfr eqfunctionch  ;
groupnum = &group ;
repnum = &replication ;
anchornum = &anchor ;
keep difft1 diffob1 diffr1 diffch1 groupnum repnum anchornum ;
output;
run;
%end;

%if &anchor = 2 %then %do;

data sec_results ;
merge truequ2 obsequ2 eqfunctionfr2 eqfunctionch2  ;
groupnum = &group ;
repnum = &replication ;
anchornum = &anchor ;
keep difft1 diffob1 diffr1 diffch1 groupnum repnum anchornum ;
output;
run;
%end;

%if &anchor = 3 %then %do;

data sec_results ;
merge truequ3 obsequ3 eqfunctionfr3 eqfunctionch3  ;
groupnum = &group ;
repnum = &replication ;
anchornum = &anchor ;
keep difft1 diffob1 diffr1 diffch1 groupnum repnum anchornum ;
output;
run;
%end;


data _null_ ;
set sec_results ;
file "&workdir\secrespart1.dat" mod ;
put repnum groupnum anchornum difft1 diffob1 diffr1 diffch1 ;
run;
```

```
data combinedpart2 ;
merge fr1e fr1c ch1e ch1c ob1e ob1c tr1e tr1c ;
repnum = &replication ;
groupnum = &group ;
anchornum = &anchor ;
run ;

data _null_ ;
set combinedpart2 ;
file "&workdir\secrespart2.dat" mod ;
put repnum groupnum anchornum fr1exp fr1c ch1e ch1c ob1 ob1cse tr1e tr1c ;
run ;

proc datasets lib=work nolist;
delete
Anchorcr
Anchormc
Ch1c
Ch1e
Combinedd1d2
Combinedks
Combinedpart2
Combinedwts
Combinedwts2
Combinedwts21
Combinedwts22
Combinedwts31
Combinedwts32
Cond1
Cond2
Cond3
D1_ch1
D1_fr1
D1_ob1
D1_tr1
D2_ch1
D2_fr1
D2_ob1
D2_tr1
Eqfunctionch
Eqfunctionch2
Eqfunctionch3
Eqfunctionfr
Eqfunctionfr2
Eqfunctionfr3
Equateobst
Equateobst2
Equateobst3
Equatetrue
Equatetrue2
Equatetrue3
Formx
Formx1
Formx11
Formx2
Formx22
```

```
Formx222
Formx3
Formx4
Formx5
Formy
Formy1
Formy11
Formy2
Formy22
Formy3
Formy4
Formy5
Fr1c
Fr1e
Groupdata
Ks
Ks23
Ks2fr
Ksch1
Ksch11
Ksch21
Ksch22
Ksch25
Ksch26
Ksch31
Ksch32
Ksch35
Ksch51
Kschn
Ksfr
Ksfr1
Ksfr21
Ksfr22
Ksfr25
Ksfr31
Ksfr32
Ksfr35
Ksfr5
Ksfr6
Ksfr7
Ksobs2
Ksobs21
Ksobs31
Ksobs32
Ksobs34
Kstr
Kstr1
Kstr2
Kstr22
Kstr3
Kstr31
Kstr34
Kstr4
Kstr42
Kstr5
Kstr52
Kstr53
```

Ks_1
Ks_3ch
Ks_com1
Ks_com2
Ks_obs2
Ks_obs3
Ks_testch
Ks_testfr
Ks_testobs
Ks_testtr
Ob1c
Ob1e
Obsequ
Obsequ2
Obsequ3
Param_fun524y
Param_fun613
Param_fun613y
Param_junk524
Param_junk524y
Param_population
Param_population2
Param_population3
Param_populationy
Param_populationy2
Param_populationy3
Par_cond1
Par_cond1111y
Par_cond111y
Par_cond11y
Par_cond1b
Par_cond1x
Par_cond1y
Par_cond2
Par_cond2222y
Par_cond222y
Par_cond22y
Par_cond2b
Par_cond2x
Par_cond2y
Par_cond3
Par_cond3333y
Par_cond333y
Par_cond33y
Par_cond3b
Par_cond3x
Par_cond3y
Populationpar
Populationpar1
Populationpar1y
Populationpar2
Populationpar2_1
Populationpar3
Populationpar3_1
Populationpary
Populationpary2
Populationpary2_1

```
Populationpary3
Populationpary3_1
Posteriorweights
Posteriorweights2
Posteriorweights21
Posteriorweights22
Posteriorweights31
Posteriorweights32
Postgroup1
Postgroup1wt
Postgroup2
Postgroup21
Postgroup21wt
Postgroup22
Postgroup22wt
Postgroup2wt
Postgroup31
Postgroup31wt
Postgroup32
Postgroup32wt
Postgroupnew1
Postgroupnew1wt
Postgroupnew2
Postgroupnew21
Postgroupnew21wt
Postgroupnew22
Postgroupnew22wt
Postgroupnew2wt
Postgroupnew31
Postgroupnew31wt
Postgroupnew32
Postgroupnew32wt
Postpoint
Postpoint2
Postpoint21
Postpoint22
Postpoint31
Postpoint32
Postpoints
Postpoints2
Postpoints21
Postpoints22
Postpoints31
Postpoints32
Postweight
Postweight2
Postweight21
Postweight22
Postweight31
Postweight32
Postweights
Postweights2
Postweights21
Postweights22
Postweights31
Postweights32
Readpar
```

```
Readpar2
Readpar3
Readparm
Readparm2
Readparm3
Readparmy
Readparmy2
Readparmy3
Readpary
Readpary2
Readpary3
Resultsx
Resultsych
Resultsyfr
Resultsyob
Resultsytr
Sasmacr
Sdoldform
Sec_results
Test1
Test1y
Test2
Test2y
Test2_1
Test2_1y
Test2_2
Test2_2y
Test3
Test3y
Test3_1
Test3_1y
Test3_2
Test3_2y
Test3_3
Test3_3y
Test4y
Test4_3
Testdata
Testdata1
Testx
Testx1
Testy
Testy1
Tr1c
Tr1e
Truequ
Truequ2
Truequ3
Xcond2
Xtheta
Xunique
Xuniquecr
Ycombined
Ycombined2
Ycombined3
Ycond2
Ytheta
```

```
Yunique
Yuniquecr
;
run;

x "del &workdir\concord*.dat";
x "del &workdir\cond*.ph*";
x "del &workdir\par*.dat";
x "del &workdir\testrun*";
x "del &workdir\dissertation*.out";

x "del &workdir\*.pst";
x "del &workdir\*.out";
x "del &workdir\itemresp.dat";
x "del &workdir\x.dat";
x "del &workdir\xtotal.dat";
x "del &workdir\ytotal.dat";
x "del &workdir\y.dat";
x "del &workdir\qf*";

x "del &workdir\poparm*.dat";
x "del &workdir\populationparm*.itm";
x "del &workdir\rawscy*.rs";

%end ;
%end ;
%end ;


proc printto;
run;


quit;

%mend ;

%toomuchfun ;
```

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Albano. (2010). Equate.

Andrews, B. J. (2011). *Assessing First-And Second-Order Equating For the Common-Item Nonequivalence Groups Design using Multidimensional IRT*. Doctoral Dissertation. University of Iowa. Iowa City.

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.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*(1), 87-96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*(3), 239-251.

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

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale*. Unpublished doctoral dissertation. University of Massachusetts.

Béguin, A. A., & Hanson, B. A. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating (Measurement and Research Department Reports 2001-2)*. Paper presented at the National Council on Measurement in Education, Seattle,WA.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the National Council on Measurement in Education, New Orleans,LA.

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

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

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

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

Brennan, R. L. (2010). Assumptions About True-Scores and Populations in Equating. *Measurement: Interdisciplinary Research & Perspective, 8*(1), 1-3.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581-612.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of educational measurement, 32*(1), 79-96.

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common-item sets*. Unpublished doctoral dissertation. University of Maryland.

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

Cohen, J. (1988). *Statistical Power Analsysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conover, W. (1999). Statistics of the Kolmogorov-Smirnov type. *Practical nonparametric statistics*, 428-473.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of educational measurement, 25*(1), 31-45.

De Ayala, R. J. (2009). *The theory and practice of item response theory*: Guilford Press.

DeMars, C. E. (2006). Application of the Bi-Factor Multidimensional Item Response Theory Model to Testlet-Based Tests. *Journal of Educational Measurement, 43*(2), 145-168.

Desa, Z. N. D. M. (2012). *Bi-factor Multidimensional Item Response Theory Modeling for Subscores Estimation, Reliability, and Classification*. Unpublished doctoral dissertation. University of Kansas.

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

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

Duong, M. Q. (2011). *Evaluating Equating Results in the Nonequivalent Groups with Anchor Test Design using Equipercentile and Equity Criteria*. Unpublished doctoral dissertation. Michigan State University.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (Vol. 4): Lawrence Erlbaum.

Finch, H. (2006). Comparison of the performance of varimax and promax rotations: factor structure recovery for dichotomous items. *Journal of Educational Measurement, 43*(1), 39-52.

Gibbons, R. D., Bock, D. R., Hedeker, D. R., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*(1), 4-19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436.

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

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

Haertel, E. H., & Linn, R. L. (1996). *Comparability*. Washington,DC: National Center for Education Statistics.

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups*. Unpublished doctoral dissertation. University of Iowa.

Hagge, S. L., & Kolen, M. J. (2011). Equating Mixed-Format Tests with Format Representative and Non-Representative Common Items. In M. J. Kolen & W. Lee (Eds.), Mixed-format tests:Psychometric properties with a primary focus on equating (Vol. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Hagge, S. L., Liu, C. L., He, Y., Powers, S. J., Wang, M., & Kolen, M. J. (2011). A Comparison of IRT and Traditional Equipercentile Methods in Mixed-format Equating. In M. J. Kolen & W. Lee (Eds.), Mixed-format tests:Psychometric properties with a primary focus on equating (Vol. 1). Iowa City,IA:Center for Advanced Studies in Measurement and Assessment,The University of Iowa.

Haladyna, T. M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5*(1), 73-88.

Haladyna, T. M. (1997). *Writing Test Items To Evaluate Higher Order Thinking*. Needham Heights, MA: Allyn & Bacon.

Hambleton, R., Merenda, P., & Spielberger, C. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. *Adapting educational and psychological tests for cross-cultural assessment*, 3-38.

Han, T., Kolen, M. J., & Pohlmann, J. (1997). A comparison among IRT true-and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education, 10*(2), 105-121.

Hanson, B. A. (1994). An extension of the Lord-Wingersky algorithm to polytomous items.Unpublished research note.

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

Hanson, B. A., Zeng, L., & Colton, D. (1994). A Comparison of Presmoothing and Postsmoothing Methods in Equipercentile Equating. *ACT Research Report Series* (Vol. 94-4). Iowa City, IA: American College Testing.

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

He, Y. (2011). *Evaluating equating properties for mixed-format tests*. University of Iowa. Iowa City.

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

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

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*(1), 41-54.

Kamata, A., & Tate, R. L. (2005). The Performance of a Method for the Long-term Equating of Mixed-Format Assessment. *Journal of Educational Measurement, 42*(2), 193-213.

Kim, S. (2004). *Unidimensional IRT scale linking procedures and their robustness to multidimensionality*. Unpublished doctoral dissertation. University of Iowa.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

Kim, S., & Lee, W. C. (2006). An Extension of Four IRT Linking Methods for Mixed-Format Tests. *Journal of Educational Measurement, 43*(1), 53-76.

Kim, S., Walker, M. E., & McHale, F. (2008). Equating of mixed-format tests in large scale assessments (ETS Research Rep. No. RR-08-26). *Princeton, NJ: ETS*.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among Designs for Equating Mixed-Format Tests in Large-Scale Assessments. *Journal of Educational Measurement, 47*(1), 36-53.

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

Kirkpatrick, R. K. (2005). *The effects of item format in common item equating*. Unpublished doctoral dissertation. University of Iowa.

Kolen, M. J. (2004a). POLYCSEM.

Kolen, M. J. (2004b). POLYEQUATE. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

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

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*(4), 285-307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129-140.

Lane, S., & Stone, C. A. (2006). Performance Assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Lee, E., Lee, W., & Brennan, R. L. (2012). Exploring Equity Properties in Equating using AP Examinations (2012-4 ed.): College Board.

Lee, W., Hagge, S. L., He, Y., Kolen, M. J., & Wang, W. (2010). *Equating mixedformat tests using dichotomous anchor items.* Paper presented at the American Educational Research Association, Denver,CO.

Levine, R. (1955). *Equating the score scales of alternate forms adminsitered to samples of different ability*. Research Bulletin 55-23. Princeton,NJ:Educational Testing Service.

Li, Y. (2011). *EXPLORING THE FULL-INFORMATION BIFACTOR MODEL IN VERTICAL SCALING WITH CONSTRUCT SHIFT*. Unpublished doctoral dissertation. University of Maryland.

Li, Y., & Rupp, A. A. (2011). Performance of the S− χ2 Statistic for Full-Information Bifactor Models. *Educational and psychological measurement, 71*(6), 986-1005.

Linn, R. L. (1995). High-stakes uses of performance-based assessments: Rationale, examples, and problems of comparability *International perspectives on academic assessment* (pp. 49-73). Norwell, MA: Kluwer Academic Publishers.

Liu, C. L., & Kolen, M. J. (2011a). A Comparison Among IRT Equating Methods and Traditional Equating Methods for Mixed-format Tests. In M. Kolen & W. Lee (Eds.), Mixed-format tests:Psychometric properties with a primary focus on equating (Vol. 1). Iowa City, IA:Center for Advanced Studies in Measurement and Assessment,The University of Iowa.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 79-95.

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

Loyd, B. H., & Hoover, H. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. *Journal of Educational Measurement, 31*(3), 234-250.

Marco, G. L. (1977). ITEM CHARACTERISTIC CURVE SOLUTIONS TO THREE INTRACTABLE TESTING PROBLEMS1. *Journal of Educational Measurement, 14*(2), 139-160.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polytomous ordered data. *Applied Psychological Measurement, 18*, 245-256.

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

Muraki, E. (1998). RESGEN:Item response generator. [Program Manual]. Princeton, NJ: Educational Testing Service.

Muraki, E., & Bock, R. D. (1993). PARSCALE: Scientific Software International.

Muthén, L. K., & Muthén, B. O. Mplus. Los Angeles, CA.

Paek, I., & Kim, S. (2007). *Empirical Investigation of Alternatives for Assessing Scoring Consistency on Constructed Response Items in Mixed Format Tests*. Paper presented at the American Educational Research Association, Chicago, IL.

Perkhounkova, Y., & Dunbar, S. B. (1999). *Influences of Item Content and Format on the Dimensionality of Tests Combining Multiple-Choice and Open-Response Items: An Application of the Poly-DIMTEST Procedure*. Paper presented at the American Educational Research Association, Montreal, Canada.

Powers, S. J., Hagge, S. L., Wang, W., He, Y., Liu, C. L., & Kolen, M. J. (2011). Effects of Group Differences on Mixed-format Test Equating. In M. J. Kolen & W. Lee (Eds.), Mixed-format tests:Psychometric properties with a primary focus on equating (Vol. 1). Iowa City,IA:Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Powers, S. J., & Kolen, M. J. (2011). Evaluating Equating Accuracy and Assumptions for Groups that Differ in Performance. In M. J. Kolen & W. Lee (Eds.), Mixed-format tests:Psychometric properties with a primary focus on equating (Vol. 1). Iowa City, IA: Center for Advanaced Studies in Measurement and Assessment, The University of Iowa.

Proctor, T. P., Reshetar, R., & Patel, P. (2012). *An Investigation of Small Sample Equating Methods for Mixed-format Exams with Non-representative Anchors*. Paper presented at the National Council on Measurement in Education, Vancouver, Canada.

Samejima, F. (1997). Graded response model. *Handbook of modern item response theory*, 85-100.

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.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Sykes, R. C., Hou, L., Hanson, B. A., & Wang, Z. (2002). *Multidimensionality and the Equating of a Mixed-Format Math Examination*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Sykes, R. C., & Yen, W. M. (2000). The Scaling of Mixed-Item-Format Tests With the One-Parameter and Two-Parameter Partial Credit Models. *Journal of Educational Measurement, 37*(3), 221-244.

Tan, X., Kim, S., Paek, I., & Xiang, B. (2009). *An Alternative to the Trend Scoring Method for Adjusting Scoring Shifts in Mixed-Format Tests*.

Tang, K. L., & Eignor, D. R. (1997). Concurrent calibration of dichotomously and polytomously scored TOEFL items using IRT models. *TOEFL Technical Report* (Vol. 13). Princeton, NJ: Educational Testing Service.

Tate, R. L. (1999). A Cautionary Note on IRT-Based Linking of Tests With Polytomous Items. *Journal of Educational Measurement, 36*(4), 336-346.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*(4), 329-346.

Tate, R. L. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and psychological measurement, 63*(6), 893-914.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39-49.

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are Tests Comprising Both Multiple-Choice and Free-Response Items Necessarily Less Unidimensional Than Multiple-Choice Tests? An Analysis of Two Tests. *Journal of Educational Measurement, 31*(2), 113-123.

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

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

Von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The Chain and Post Stratification Methods for Observed Score Equating: Their Relationship to Population Invariance. *Journal of Educational Measurement, 41*(1), 15-32.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Walker, M. E., & Kim, S. (2009). *Linking Mixed-Format Tests Using Multiple Choice Anchors.* Paper presented at the National Council on Measurement in Education, San Diego,CA.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*(2), 141-162.

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

Wei, H., & Yi, Q. (2012). *Mixed-format test equating in the presence of multidimensionality and rater severity variations.* Paper presented at the National Council on Measurement in Education, Vancouver, Canada.

Wu, N., Huang, C., Huh, N., & Harris, D. (2009). *Robustness in using multiple-choice items as an external anchor for constructed-response test equating.* Paper presented at the National Council on Measurement in Education, San Diego,CA.

Yung, Y. F., Mcleod, L. D., & Thissen, D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 71*, 281-301.