The Effect of Differential Item Functioning on Population Invariance of Item Response Theory True Score Equating

Anne C. Huggins
University of Miami, a.huggins@umiami.edu

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation
https://scholarlyrepository.miami.edu/oa_dissertations/724

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
THE EFFECT OF DIFFERENTIAL ITEM FUNCTIONING ON POPULATION INVARIANCE OF ITEM RESPONSE THEORY TRUE SCORE EQUATING

By

Anne Corinne Huggins

A DISSERTATION

Submitted to the Faculty of the University of Miami in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Coral Gables, Florida

May 2012
A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

THE EFFECT OF DIFFERENTIAL ITEM FUNCTIONING ON POPULATION
INVARIANCE OF ITEM RESPONSE THEORY TRUE SCORE EQUATING

Anne Corinne Huggins

Approved:

Randall D. Penfield, Ph.D.                     Terri A. Scandura, Ph.D.
Professor of Educational and                  Dean of the Graduate School
Psychological Studies

Nicholas D. Myers, Ph.D.                     Soyeon Ahn, Ph.D.
Professor of Educational and                 Professor of Educational and
Psychological Studies                         Psychological Studies

Rebecca J. Shearer, Ph.D.
Professor of Psychology
Population invariance in equating exists when the relationship between two scales is the same for two or more subpopulations of examinees and hence the function used to equate the scales is not dependent on subpopulations. A lack of equating invariance (i.e., equating dependence) leads to a situation whereby examinees having the same score on one scale, but belonging to different subpopulations, have different expected test scores on the corresponding equated scale. This situation results in an expected advantage for one or more subpopulations of examinees and hence is a concern for fairness in assessment and disaggregated accountability. Little is known about the causes of equating dependence, and the purpose of this study is to locate a source of this problem. It is hypothesized that differential item functioning manifested in the anchor items of an assessment will have an effect on population invariance of equating. Findings show that when differential item functioning varies across forms in a differential manner across subpopulations, population invariance of equating can be compromised. Under these conditions, an increase in equating dependence is associated with increases in magnitudes of the differential item functioning and, to a lesser degree, increases in the frequency of anchor items with differential item functioning. These effects can be problematic in conditions of both unidirectional and bidirectional differential item functioning, and can pose problems for subpopulations that have equal or different mean ability levels.
For my family and Tripp

-Corinne
ACKNOWLEDGEMENTS

This dissertation would not have been completed in its entirety without the support of many people. To Dr. Michael Chajewski, I am thankful for the teaching, guidance, time, and patience offered in support of my computer coding tasks. To my dissertation committee members, Dr. Nicholas Myers, Dr. Soyeon Ahn, and Dr. Rebecca Shearer, I am grateful for the multiple reviews of my proposal and dissertation, and all the thoughtful suggestions for improvement to this study. I am also thankful to Dr. Isaac Prilletensky, Dr. Guerda Nicolas, and Dr. Etiony Aldarondo for the graduate assistant support while writing my dissertation. For my fellow graduate students and colleagues, I am in debt for the hours spent reading my drafts and listening to my struggles. This includes Dr. Karina Gattmorta, Dr. Jaime Maerten-Rivera, Allison Ames, Ying Jin, and Stefano Rainelli. Finally, to my mentor, Dr. Randall Penfield, I am thankful for his assistance in refining the purpose, method, discussion and organization of this study, and for the continued support throughout my pursuit of a doctoral degree. This was only possible with his guidance.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF ACRONYMS</td>
<td>viii</td>
</tr>
<tr>
<td><strong>Chapter</strong></td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2 LITERATURE REVIEW</td>
<td>4</td>
</tr>
<tr>
<td>3 OBJECTIVES AND RESEARCH QUESTIONS</td>
<td>30</td>
</tr>
<tr>
<td>4 METHOD</td>
<td>32</td>
</tr>
<tr>
<td>5 RESULTS</td>
<td>40</td>
</tr>
<tr>
<td>6 CONCLUSIONS AND DISCUSSION</td>
<td>52</td>
</tr>
<tr>
<td>WORKS CITED</td>
<td>59</td>
</tr>
<tr>
<td>TABLES</td>
<td>67</td>
</tr>
<tr>
<td>FIGURES</td>
<td>72</td>
</tr>
<tr>
<td>APPENDIX: R SYNTAX</td>
<td>88</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1. Methods for Assessing Invariance According to Two Dimensions 67
Table 2. Simulated Conditions of Differential Item Functioning 68
Table 3. Observed Anchor Item Parameters on a 2009 Statewide Assessment 70
Table 4. Pearson’s Correlations Assessing the Relationship between Form 1 Anchor DTF and Unconditional Equating Dependence 71
LIST OF FIGURES

Figure 1. Graphical Depiction of the Non-Equivalent Anchor Test Equating Design…...72
Figure 2. Conditional Equating Invariance Results for Condition 1 (Null Condition)....73
Figure 3. Conditional Equating Invariance Results for Condition 2 (Null Condition)....74
Figure 4. RMSD(x) Results for Conditions with No Differential Form DIF, No Group Mean Differences in Ability, and Unidirectional DIF..............................75
Figure 5. RMSD(x) Results for Conditions with No Differential Form DIF, No Group Mean Differences in Ability, and Bidirectional DIF...............................76
Figure 6. RMSD(x) Results for Conditions with No Differential Form DIF, Group Mean Differences in Ability, and Unidirectional DIF..............................77
Figure 7. RMSD(x) Results for Conditions with No Differential Form DIF, Group Mean Differences in Ability, and Bidirectional DIF...............................78
Figure 8. RMSD(x) Results for Conditions with Differential Form DIF, No Group Mean Differences in Ability, and Unidirectional DIF..............................79
Figure 9. RMSD(x) Results for Conditions with Differential Form DIF, No Group Mean Differences in Ability, and Bidirectional DIF...............................80
Figure 10. RMSD(x) Results for Conditions with Differential Form DIF, Group Mean Differences in Ability, and Unidirectional DIF..............................81
Figure 11. RMSD(x) Results for Conditions with Differential Form DIF, Group Mean Differences in Ability, and Bidirectional DIF...............................82
Figure 12. Group-to-overall Conditional Equating Dependence Results for Selected Conditions.........................................................................................83
Figure 13. Graphical Displays of REMSD Results: Interactions of Differential Form DIF with all other Factors in the Study..................................................84
Figure 14. Graphical Displays of RESD_R Results: Interactions of Differential Form DIF with all other Factors in the Study....................................................85
Figure 15. Graphical Displays of RESD_F Results: Interactions of Differential Form DIF with all other Factors in the Study....................................................86
Figure 16. Scatter Plots Displaying the Relationships between Unconditional Equating Dependence and True DTF in Form 1 Only..........................................87
**LIST OF ACRONYMS**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>AYP</td>
<td>Annual Yearly Progress</td>
</tr>
<tr>
<td>CTT</td>
<td>Classical Test Theory</td>
</tr>
<tr>
<td>DIF</td>
<td>Differential Item Functioning</td>
</tr>
<tr>
<td>DFIT</td>
<td>Differential Functioning of Items and Tests</td>
</tr>
<tr>
<td>DTF</td>
<td>Differential Test Functioning</td>
</tr>
<tr>
<td>DTM</td>
<td>Difference that Matters</td>
</tr>
<tr>
<td>IPD</td>
<td>Item Parameter Drift</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>NEAT</td>
<td>Non-equivalent Anchor Test Design</td>
</tr>
<tr>
<td>NCEST</td>
<td>National Council on Education and Statistics Testing</td>
</tr>
<tr>
<td>NCLB</td>
<td>No Child Left Behind</td>
</tr>
<tr>
<td>REMSD</td>
<td>Root Expected Mean Square Difference</td>
</tr>
<tr>
<td>RESD</td>
<td>Root Expected Squared Difference</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Square Difference</td>
</tr>
<tr>
<td>RSD</td>
<td>Root Squared Difference</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

Population invariance in equating exists when the relationship between two scales is the same for two or more subpopulations of examinees and hence the function used to equate the scales is not dependent on subpopulations (Angoff, 1971; Dorans & Holland, 2000; Flanagan, 1951). A lack of equating invariance leads to a situation whereby examinees having the same score on one scale, but belonging to different subpopulations, have different expected test scores on the corresponding equated scale. This situation results in an expected advantage for one or more subpopulations of students and hence a lack of invariance in equating is a concern for fairness and equity in assessment (Dorans, 2004, 2008).

While equating invariance is a concern of fairness and equity at the reported (i.e., equated) test score level, differential item functioning (DIF) concerns fairness with respect to statistical bias at the item level (Camilli & Shepard, 1994). DIF refers to the instance in which examinees of the same ability level, but belonging to different subpopulations, have different estimated probabilities of success on a test item (Camilli & Shepard, 1994; Penfield & Camilli, 2007). Differential test functioning (DTF) is an aggregate of DIF across all the items on a test. Therefore, DTF concerns statistical bias at the raw test score level. Equating invariance, DTF, and DIF are all concerned with invariant relationships, or lack thereof, across subpopulations of examinees; equating dependence is a lack of invariance at the reported test score level, DTF is a lack of invariance at the raw test score level, and DIF is a lack of invariance at the item level.

Invariant relationships in the internal mechanisms of estimating achievement scores on educational tests serve as the basis for concluding that a particular test is fair.
with respect to statistical bias concerns (AERA, APA, & NCME, 1999; AERA, APA, & NCME, 2011). The current culture surrounding American K-12 education is one of accountability, which is evaluated largely through results of standardized achievement test scores. Two facets of this accountability have implications for the importance of understanding reported score level invariance: high-stakes and disaggregated accountability. Ensuring that test scores used to make high-stakes decisions display adequate fairness with respect to a lack of statistical bias requires a firm understanding of the causes of a lack of statistical bias. If a lack of invariance occurring at the item level results in a lack of invariance at the equated score level, then the nature and causes of the latter can be understood in relation to the nature and causes of the former. The disaggregated nature of these high-stakes accountability systems further necessitates the need for this understanding; if data is to be reported at the subpopulation level, it is critical to understand how a lack of item level invariance and a lack of reported score level invariance dually arise and affect score equity across subpopulations.

Empirical studies on DIF and DTF can be found extensively in assessment literature and empirical studies on equating invariance have increased dramatically over the past decade. However, the relationship between equating invariance, DTF, and DIF is relatively unexplored. Few empirical studies have been conducted for the purpose of comparing these distinct, yet interrelated, facets of fairness as a lack of statistical bias, and those that have been conducted call for more attention to this area. A currently unanswered question is, “Under what conditions does DIF in anchor items of a test and DTF of an anchor test result in a problematic level of group dependency in an equating
function?” This study is a computer simulation that examines the effect of various conditions of DIF in anchor items, and the resultant DTF in an anchor test, on the presence and magnitude of population dependence in equating.
Chapter 2: Literature Review

Historically, the need to better understand invariant and variant relationships across subpopulations of examinees was prompted by two factors related to both the measurement community and American society in general. First, concerns with fairness in assessment have increased dramatically since the 1960’s and this mindset has created the desire and need to continuously examine the internal mechanisms of tests to ensure that they operate consistently across examinees from different subpopulations. Second, educational policy has shifted dramatically toward a system that bases high-stakes decisions on standardized test scores, further prompting the need to ensure fairness in test scores that have meaningful impacts on entire subpopulations of examinees and the school systems that serve them. These two factors deserve in depth treatment for introducing this study.

Fairness

The interest in DIF and DTF over the past few decades and the more recent interest in equating invariance stem from societal concerns with fairness, equity, and anti-discrimination that became prominent in America in the 1960’s. The Civil Rights Movement and legal sanctions that occurred in conjunction with the Movement brought to light the issues of discrimination and disparate impact at the group level rather than solely at the individual level (Camilli, 2006). While the prominent focus was on racial and ethnic groups, laws were established in relation to any group of person’s perceived to be generally affected by discrimination or adverse impacts in society (Camilli, 2006). These social and political shifts in issues of equity were mirrored in policies and practices of the psychometric community (Osterlind & Everson, 2009).
In 1966 the first edition of the *Standards for Educational and Psychological Tests and Manuals* was published (APA, AERA, & NCME, 1966) and since this time three other editions have been published and a new draft is currently being finalized, each producing a more nuanced definition of fairness in assessment. The most recent published version of the *Standards* (AERA, APA, & NCME, 1999) includes specific guidelines on fairness in assessment and these guidelines were altered in part to address the varying and vague definitions of fairness that existed in the measurement community (Eignor, 2001). In the *Standards* (1999) fairness is defined within four contexts: (a) fairness as lack of bias, (b) fairness as equitable treatment in the testing process, (c) fairness as opportunity to learn, and (d) fairness as equality in outcomes of testing.

The current revision draft of the standards, *Draft Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2011), differs from the 1999 version in the definition of fairness. The first three aspects of fairness in the 1999 standards (i.e., lack of bias, equitable treatment in the testing process, and opportunity to learn) are mirrored in the *Draft Standards* (2011) version, with fairness in treatment during the testing process no longer specifically requiring *equitable* treatment. Also, fairness as equality of outcomes is excluded from the definition of fairness as group differences in outcomes do not directly indicate that a test is bias or unfair. While there is a general consensus that tests producing very disparate results across groups are not desired, it is technically not an issue of bias or unfairness in testing (Huebert & Hauser, 1998). The *Draft Standards* (2011) excludes this portion of fairness but adds a new one, that of fairness in access to the construct(s) being measured, as a lack of access to the construct being measured can result in test bias or a lack of fairness. In this latest draft of the
standards, all four aspects of fairness are directly related to one important resulting issue; failure to meet any one of the four standards can result in scores from tests not reflecting the same construct and not possessing the same meaning for all individuals and groups of individuals in the test taking population (AERA, APA, & NCME, 2011; Huebert & Hauser, 1998).

Fairness in treatment during the testing process requires, among other things, that students are given equal opportunity to display their knowledge. The Draft Standards (2011) discuss at length that while standardization of the testing process is important (i.e., literal equitable treatment), there are instances in which it poses challenges for particular students being able to display their knowledge. When students cannot display their knowledge, scores across individuals are not comparable. Fairness as opportunity to learn relates to the conditions of instruction and curriculum provided to students before an assessment occurs. According to the Draft Standards (2011), when students or subgroups of students are exposed to instruction or curriculum that is not aligned with the test content, the test is unable to produce test scores that can be interpreted in the same manner as test scores of other students or groups of students who have been exposed to instruction and curriculum that is aligned with the content of the test. This is particularly problematic when test scores are used for high-stakes decisions (AERA, APA, & NCME, 2011). Fairness in access to the construct(s) being measured is directly related to fairness in the treatment of the testing process. However, it is given special treatment in the Draft Standards (2011) to call attention to the importance of accessibility for individuals with age, cultural, background, disability, and/or English language proficiency attributes that may require accommodations.
Fairness as lack of bias is directly related to the purpose of this study. Two types of bias are of concern: bias in measurement and bias in predictive relations (AERA, APA, & NCME, 2011). Predictive bias is defined as the differential prediction for one or more groups of examinees test scores on an outside criterion or measure, but examinations of this type of bias are often practically difficult to conduct and are also dependent on which outside criterion is utilized for prediction (Penfield & Camilli, 2007). Bias in measurement, in contrast, can be examined within the test scores and item responses of the test in question. It refers to the instance in which characteristics of a test result in different meanings of the scores earned by different subpopulations of test takers (AERA, APA, & NCME, 2011). While the other aspects of fairness (i.e., fairness in treatment during the testing process, fairness in opportunity to learn, and fairness as access to the construct) can result in different meanings of test scores for different individuals, lack of bias is the result of these fairness issues that can be observed in the test scores and item responses. Examinations of equating invariance, DIF (AERA, APA, & NCME, 2011; Camilli, 2006), and DTF (AERA, APA, & NCME, 2011) are directly related to this definition of fairness or lack thereof. Invariant relationships in reported (i.e., equated) test scores, raw test scores, and item responses are a prerequisite for establishing that fairness as lack of bias has been achieved.

There are other competing views of fairness than those defined by various versions of the Standards. Dorans (2004, 2008) conceptualizes fairness in relation to the statistical test that can be used to empirically assess fairness. However, this is a more limited view of fairness and it closely aligns with fairness as lack of bias in the Standards. Huebert and Hauser (1998) define fairness in three parts: psychometric
adequacy, appropriate opportunity to learn, and educational benefit (Penfield, 2010b). The *Standards* stress that the importance of ensuring lack of bias in test scores is necessitated by the increased use of high-stakes decisions based on test scores (AERA, APA, & NCME, 2011), while Huebert and Hauser’s (1998) inclusion of educational benefit stresses the importance of considering the effect that high-stakes decisions have on educational experiences of examinees.

**High-Stakes Testing**

A dramatic increase in the use of standardized tests was seen in the 1990s within elementary and secondary grade levels, and the practice of attaching high-stakes to test scores also increased. The need for attaching high-stakes to tests arose out of the concern of quality in the American public education system, particularly the lack of standards for what students should be learning and a system for holding schools accountable to those standards (National Council on Education and Statistics Testing [NCEST] Act, 1991). The NCEST Act (1991) implemented voluntary requirements of accountability and high-stakes decisions based on test scores, and despite the voluntary nature of the requirements, several states indeed chose to enact testing systems with high-stakes for schools and school districts (Koretz & Hamilton, 2006). A prominent system of attaching high-stakes to test scores was that of implementing accountability systems that provided rewards and sanctions to schools and districts based on aggregate test scores (Koretz & Hamilton, 2006).

In 2001, the No Child Left Behind (NCLB) Act was passed as federal law, which implemented sweeping changes to federal requirements not only on the amount of testing that has to be done in K-12 public institutions but also on the decisions that would be
attached to test scores (Koretz & Hamilton, 2006). These requirements were now compulsory. These high-stakes have prompted changes in the *Standards* (Eignor, 2001) and in the practices of measurement experts (Koretz & Hamilton, 2006). These changes have coincided with a shift in the way that measurement experts define and examine validity; validity in the context of test score use and stakes of those uses has become a central concern in the assessment community (Kane, 2006; Lissitz & Samuelsen, 2007).

In addition, the policies requiring test scores to be used for high-stakes accountability purposes have created a shift in testing from thinking about fairness at the individual level to thinking about fairness in reporting aggregated scores, both across entire schools and districts as well as disaggregated across different subpopulations that exist in those schools and districts.

NCLB (No Child Left Behind Act of 2001: State Plans) requires that schools and districts estimate and report Annual Yearly Progress (AYP) with sanctions in place for schools and districts that do not display growth in AYP over each school year. AYP is calculated as the percentage of students in a given grade level and subject area who are deemed as meeting or exceeding proficiency on a standardized measure of ability (No Child Left Behind Act of 2011: State Plans), and proficiency is defined at a particular cut-score on the score scale of the standardized measure of ability. Due to this definition of AYP, the different test forms given in different years of a particular measure of ability must be equated in order to ensure that the scale on which the cut-score is derived is consistent across years. This equating is intended to result in scores that are on the same scale and therefore interchangeable from year to year. Without equating, the scale of the measure would differ across years due to small but meaningful differences in difficulty.
across the test forms as well as differences in the non-equivalent test taking populations
across years. Changes from year to year in AYP estimates based on scales that are not
equated would reflect not only the increase or decrease in the number of students who are
proficient but also a change in the score scale. Equating must be utilized in order for
changes in AYP to accurately indicate changes in percentages of students who are
proficient.

NCLB not only requires AYP to be reported at the school level but also requires
AYP estimates to be disaggregated across subpopulations of examinees, specifically
economically disadvantaged students, students from major ethnic and racial groups,
students with disabilities, and students with limited English proficiency (No Child Left
Behind Act of 2001: State Plans). While AYP estimation requires equating from year to
year, the disaggregation of AYP estimates has an additional requirement of the equating
procedure; the equating function(s) must be consistent across subpopulations in order to
result in interchangeable scores across the years at the subpopulation level. Hence,
fairness as lack of bias at the reported (i.e., equated) test score level requires that the
equating from year to year be subpopulation invariant, ensuring that disaggregated AYP
scores mean the same thing across the different subpopulations and the different years of
assessment. As educational policies continue to change, the current climate indicates that
measuring subpopulations from year to year with AYP-type indicators will be a
requirement of public schools and districts for years to come (Klein, 2011; Klein, 2012;

High-stakes testing for purposes of accountability further necessitates the
importance of ensuring that invariant relationships exist in reported test scores,
particularly with respect to the subpopulations reported under NCLB. Ensuring a lack of DIF in the items on an assessment is not as critical for high-stakes decisions as is ensuring that any DIF in the items of an assessment does not result in an unacceptable degree of equating dependence. Ensuring this is necessary for producing reported scores that are interchangeable across years, common in definition across subpopulations, and on a scale that can be used to develop a cut-score that remains consistent in scale location across years and subpopulations.

**Item Response Theory**

All methods utilized in this study are based on Item Response Theory (IRT). Hence, a brief overview of IRT theory and application is required. IRT is a model-based, item-level, latent trait measurement theory (Lord, 1980). In IRT, item characteristic curves are used to model the probability of correct response on an item given a person’s latent ability level. Test characteristic curves can be estimated from the item characteristic curves of all items on a test and they model expected true scores on a test given a person’s latent ability level.

There are multiple unidimensional IRT models available for dichotomous items. The three-parameter logistic model (3PL) (Lord, 1980) is the most widely-used in large-scale testing and is defined as

\[
P_{ij}(1|\theta_i) = c_j + (1 - c_j) \frac{\exp[1.7a_j(\theta_i - b_j)]}{1+\exp[1.7a_j(\theta_i - b_j)]},
\]

where \(P_{ij}(1|\theta_i)\) is the probability of a correct response (i.e., a score of 1) on item \(j\) for individual \(i\), \(\theta_i\) is the latent trait ability level of individual \(i\), \(a_j\) is the discrimination parameter of item \(j\), \(b_j\) is the difficulty parameter of item \(j\), and \(c_j\) is the pseudo-guessing parameter of item \(j\). The two-parameter logistic model (2PL) is a special case of the 3PL
in which $c_j$ is fixed to zero and the one-parameter logistic model (1PL) is a special case of the 2PL in which $a_j$ is set equal across items (Lord, 1980). The 3PL model is utilized in this study as it is the most widely used IRT model in large-scale educational testing where stakes and accountability requirements are attached to test scores. To maintain consistency in methodology, the equating methods and DIF/DTF detection methods utilized are all IRT-based.

**Equating Invariance**

Equating is the process of transforming scores from one scale to a different scale while maintaining the comparability of scores across those scales to a degree that the scores can be treated as interchangeable (Dorans & Holland, 2000; Holland, 2007; Kolen & Brennan, 2004). For example, large-scale testing companies must produce multiple forms of a particular assessment to ensure test security and an equating process is used to place scores from each of the different forms onto the same scale. Interchangeability of scores across the forms is required, as the form to which an examinee is exposed needs to be a matter of indifference in order to prevent advantaging a group who happened to be exposed to a slightly easier form.

In order to produce interchangeable scores from an equating process, the following five conditions must be met (Dorans & Holland, 2000): (a) the tests or forms being equated must measure the same construct, (b) the tests or forms being equated must have equal reliability, (c) the equating function used to transform scores from form 1 to form 0 must be the inverse of the equating function used to transform scores from form 0 to form 1, (d) the examinees must have the same expected score and expected distribution of performance on both of the scales being equated, and (e) the equating function must be
invariant to subpopulations of test takers. It is always possible to use a function to place scores from one scale onto another, but having that process result in scores that are interchangeable across the two scales is a matter of how well the test construction, equating design, and equating method meet the five said requirements. In the case of large-scale standardized tests used for high-stakes purposes, it is critical that equating of scores be confirmed to meet the five requirements and thus produce interchangeable scores.

This study focuses on the property of equating that requires population invariance in the equating function. Equating invariance refers to the instance in which the relationship between two scales (i.e., the scale of raw scores and the equated scale onto which raw scores are placed) is consistent across subpopulations of test takers (Dorans & Holland, 2000). Stated differently, a lack of equating invariance exists when examinees having the same raw score, but belonging to different subpopulations, have different expected scores on the equated scale. The lack of equating invariance, or equating dependence, is associated with a differential difficulty within the data (Cook & Petersen, 1987; Dorans, 2004; Dorans & Holland, 2000; Holland & Dorans, 2006; von Davier & Wilson, 2008). This differential difficulty can be manifested in several ways (e.g., differential mean difficulties across all items, or differential rank ordering of anchor item difficulties used to equate the tests), and the nature of the differential difficulty is directly related to the underlying lack of invariance in the equating function.

Lack of invariance in equating by definition produces an expected disadvantage for at least one subpopulation of test takers (Holland & Dorans, 2006; von Davier, 2007). For this reason examining the invariance of an equating procedure can be used to assess
equity in reported scores. Termed as score equity analysis (Dorans, 2004, 2008), equating invariance examinations are recommended for all large-scale assessment companies (Dorans, 2004). Yet score equity analysis has only recently been defined and hence in practice there are still many assessments that have not been examined for equating invariance. In addition, score equity analysis methods have only recently been established and the relationship of each of the methods to the different realities of tests, testing procedures, equating methods, DIF in items, DTF in test forms, and diverse populations of test takers is still in the process of being empirically explored and supported.

**The multidimensional explanation for a lack of equating invariance.** In the case where a unidimensional model is assumed (i.e., a measurement model that assumes there is only one underlying trait being measured) but a secondary factor is present, standard observed score equating methods may be population dependent (van der Linden, 2000). A secondary factor refers to an unintended additional construct being measured, and it can be problematic for equating invariance when the content of the secondary factor interacts with population ability distributions (Kolen, 2004). Population invariance of equating is not expected to hold when the distributions of ability on a secondary factor differ by subpopulation, and this is believed to affect not only observed score equating but true score equating as well (Kolen, 2004).

Most studies on multidimensionality and equating invariance have focused on secondary dimensions that do not interact with the subpopulations of interest (i.e., the subpopulations did not differ in their ability distributions on that secondary factor) and found that equating results were robust to this type of multidimensionality (Cook, Dorans, Eignor, & Petersen, 1985; Dorans & Kingston, 1985; Yen, 1984). Studies that
focused on secondary factors with differential ability distributions across groups showed more equating dependence (Angoff & Cowell, 1985), although not always a level of dependence that was deemed problematic (De Champlain, 1996; Snieckus & Camilli, 1993 [as cited by De Champlain, 1996, p. 2]). These studies took place before the introduction of many current methods for detecting equating dependence, so more research is needed on how secondary factors that interact with subpopulation ability distributions affect equating invariance as measured by these recently developed methods.

**Equating designs and methods.** Equating invariance must be considered in relation to the equating design and the construction of the tests or test forms being equated. The equating design can have substantial implications for ensuring and examining equating invariance. There are many equating designs available for use, including the random groups design (Kolen & Brennan, 2004), the single group design (Kolen & Brennan, 2004), and the non-equivalent anchor test design (NEAT) (otherwise known as the common-item non-equivalent groups design [Kolen & Brennan, 2004]), the last of which is utilized in this study. This design indicates that anchor items are being used to establish the equating function between two or more test forms. Figure 1 displays a picture of a NEAT equating design, for which the purpose is to equate scores on form 1 to the scale of scores on form 0. As can be seen in this figure, there are two test forms each with fifty items that are intended to measure the same construct across forms (e.g., science ability). Anchor items refer to a subset of items within forms that is identical across forms. In Figure 1, the two test forms contain fifty items, with forty of the items in a form being unique to that form and ten of the items in the forms being common to both
forms. Under the NEAT design, these anchor items are used to establish the equating relationship between the scale of scores on form 1 and the scale of scores on form 0.

Utilizing anchor items for equating requires careful construction and empirical examination of anchor items if equating invariance is expected to hold. Longer anchor forms (Cook, 2007) and anchor forms that well represent the content of the entire test (Dorans, Liu, & Hammond, 2008; Petersen, Marco, & Stewart, 1982) and correlate well with the entire test (von Davier, Holland, & Thayer, 2004) are expected to result in a higher degree of equating invariance than shorter, non-representative, and poorly correlated anchors. The important consideration is that the appropriate equating method be chosen for any given equating design (Brennan, 2008; Dorans, Holland, Thayer, & Tateneni, 2002; Yang, Dorans, & Tateneni, 2003; Yin, Brennan, & Kolen, 2004), and that the tests and anchor portions of the test are assessed for the degree to which they satisfy the other properties of equating, such as the same construct requirement (Dorans and Holland, 2000; Kolen & Brennan, 2004; Liu & Holland, 2008; Petersen, 2008) and the equal reliability requirement (Brennan, 2008; Liu & Holland, 2008; Yang & Gao, 2008).

The choice of equating method also must be considered when examining equating invariance. There are two IRT-based equating methods available in the literature, IRT true score equating and IRT observed score equating, the former of which is more commonly used in practice and will therefore be utilized in this study. In unidimensional, dichotomous IRT models, true scores of individuals on any given test form are defined as the sum of the probabilities of correct response across all the items on the test form, or

\[ \tau_i = \sum_j P_{ij}(1|\theta_i), \]  

(2)
where \( \tau_i \) is the true score of individual \( i \), \( j \) is an item on a test form, and \( \theta_i \) is the latent ability of individual \( i \). IRT true score equating involves a three step process (Kolen & Brennan, 2004) in which a true score on one form is specified, the IRT ability estimate \( (\theta) \) that corresponds to that true score is located, and the true score on another form that corresponds to that \( \theta \) value is deemed as the equivalent true score. Determining equivalent true scores can be done using the item parameters themselves, and subsequently the true score relationship in the data of a representative sample can be utilized to establish an estimated true score relationship that can then be applied across the score range (Kolen & Brennan, 2004). This equating method assumes that true score theories can be applied to observed scores. IRT observed score equating does not make this assumption and estimates the equating relationship between two tests or test forms through the estimated distributions of number correct scores on said tests or forms. In IRT observed score equating, the estimated distributions of number correct scores are obtained through a compound binomial distribution integrated over a population distribution of \( \theta \) (Lord & Wingersky, 1984). Traditional equipercentile equating methods are then used to equate the number correct score distributions (Kolen & Brennan, 2004). The two IRT equating methods tend to produce similar results (Lord & Wingersky, 1984) and IRT true score equating is more commonly used in practice and in research studies (e.g., Bolt, 1999; De Champlain, 1996; Cook, 1983; Kingston & Dorans, 1984; Stocking, 1988; von Davier & Wilson, 2007).

IRT true score equating methods require that the latent ability scores on each test or test form first be placed on the same scale through a scaling procedure. There are multiple ways to scale ability estimates, one of which is the mean-sigma method (Kolen...
& Brennan, 2004). The mean-sigma method is utilized in this study as it defines scaling constants through the relationships of \( b \) parameters (which are the parameters manipulated for DIF in this study). Under the mean-sigma method, latent ability estimates are transformed from form 1 to form 0 such that

\[
\theta_{iF_0} = A\theta_{iF_1} + B, \tag{3}
\]

where \( \theta_{iF_1} \) is the latent ability level of individual \( i \) on test form 1, \( \theta_{iF_0} \) is the scaled latent ability level of individual \( i \) on the scale of form 0 scores, and \( A \) and \( B \) are scaling constants. The relationship between item parameters on the two scales of form 1 and form 0 is defined as

\[
a_{jF_0} = \frac{a_{jF_1}}{A}, \tag{4}
\]

\[
b_{jF_0} = Ab_{jF_1} + B, \tag{5}
\]

and

\[
c_{jF_0} = c_{jF_1}, \tag{6}
\]

where \( a_{jF_1} \), \( b_{jF_1} \), and \( c_{jF_1} \) are anchor item \( j \)’s discrimination, difficulty, and lower asymptote parameters (respectively) on form 1; and \( a_{jF_0} \), \( b_{jF_0} \), and \( c_{jF_0} \) are item \( j \)’s scaled discrimination, difficulty, and lower asymptote parameters on the scale of form 0. To scale the ability estimates and item parameter estimates using Equations 3 through 6, the scaling constants \( A \) and \( B \) first need to be established. Under the mean-sigma method, scaling constants \( A \) and \( B \) are defined as

\[
A = \frac{\sigma(b_{F_0})}{\sigma(b_{F_1})}, \tag{7}
\]

and

\[
B = \mu(b_{F_0}) - A\mu(b_{F_1}), \tag{8}
\]
where $\sigma(b_{F0})$ refers to the standard deviation of $b$ parameters across form 0 anchor items, $\sigma(b_{F1})$ refers to the standard deviation of $b$ parameters across form 1 anchor items, $\mu(b_{F0})$ refers to the mean of $b$ parameters across form 0 anchor items, and $\mu(b_{F1})$ refers to the mean of the $b$ parameters across form 1 anchor items. Once the scaling is complete, IRT true score equating can be implemented, and the scores resulting from these transformations can be assessed for equating invariance.

**Methods for evaluating equating invariance.** There are multiple methods available for assessing equating invariance. The methods are best introduced in accordance with two dimensions (Huggins & Penfield, 2012): (a) a dimension defined by which of the multiple equating functions used in an invariance analysis are being compared (i.e., omnibus methods or group-to-overall methods), and (b) a dimension defined by whether or not a method assesses invariance at individual score levels of a test (i.e., conditional or unconditional on score level). Table 1 displays and describes each of the invariance methods used in this study in relation to the two dimensions.

**Omnibus Methods.** Omnibus methods produce indices that compare an equating function based on an overall (entire) test taking population (i.e., the overall equating function) *simultaneously* to multiple equating functions based on each of the subpopulations of interest (Huggins & Penfield, 2012). For example, if interested in assessing the degree of invariance in an equating function with respect to gender subpopulations, three equating functions would be estimated (i.e., one for the overall population, one for males only, and one for females only) and an omnibus method would compare both the male and female equating functions *simultaneously* to the overall
equating function. Omnibus methods produce one value that represents the equating invariance, or lack thereof (i.e., equating dependence), of all subpopulations as compared to the overall population.

Omnibus methods can be conditional on score level or unconditional on score level, as is true with all equating invariance methods (Huggins & Penfield, 2012). An omnibus method conditional on score level measures the degree of invariance between the equating functions of all subpopulations and the overall population at each score level of a test. An omnibus method unconditional on score level measures the degree of invariance between the equating functions of all subpopulations and the overall population across all score levels of a test.

Dorans and Holland (2000) introduced two omnibus methods, one conditional on score level and the other unconditional on score level. In the case of the single or equivalent group equating design and a linear equating function, the Root Mean Square Difference, or $RMSD(x)$, is defined as

$$RMSD(x) = \frac{\sqrt{\sum_k w_k (d_k(x))^2}}{\sigma_Q},$$

(9)

where $x$ represents a particular score level of test form 1, $d_k(x)$ represents the difference between an equated score $y$ based on subpopulation $k$’s equating function and an equated score $y$ based on the overall equating function at score level $x$, $w_k$ represents a weighting for each subpopulation determined by the proportion of test takers classified into that subpopulation, $\sum_k$ represents a summation across subpopulations, and $\sigma_Q$ represents the unconditional standard deviation of scores in population $Q$ (under the single or random groups equating designs, there is only one population of concern). The resulting value represents the typical distance between the subpopulation equating functions and the
overall equating function at one score level \( x \). An invariance study produces a unique \( \text{RMSD}(x) \) value for each of the \( m \) score levels on test form 1, with score levels defined as \( x = 0, 1, \ldots, m \), where \( m \) is the maximum number of score levels on a test form. As with most of the invariance indices discussed in this literature review, the \( \text{RMSD}(x) \) can be adapted to other equating designs and methods (von Davier, Holland, & Thayer, 2004; von Davier & Wilson, 2008). Also, the \( \text{RMSD}(x) \) can be reported as an unstandardized value by removing the denominator component of Equation 9. This is true of all the methods for estimating equating invariance in this literature review, and the unstandardized version of these methods will be implemented in this study.

The Root Expected Mean Square Difference (\( \text{REMSD} \)) (Dorans & Holland, 2000) is the unconditional version of the \( \text{RMSD}(x) \). An invariance study produces one \( \text{REMSD} \) value for any one definition of subpopulations and this value represents the typical distance between the subpopulation equating functions and the overall equating function across all score levels of test form 1. The \( \text{REMSD} \) is defined as

\[
\text{REMSD} = \frac{\sqrt{\sum_{x=1}^{p_x} \sum_w w_k[d_k(x)]^2}}{\sigma_q},
\]

(10)

where \( p_x \) represents the proportion of test takers at score level \( x \) and \( \sum_{x=1}^{p_x} \) is the expectation of \( \{\sum_k w_k[d_k(x)]^2\} \) across all score levels of test form 1.

**Group-to-overall methods.** Group-to-overall invariance methods assess the degree of similarity or difference between one subpopulation’s equating function and the overall equating function, ignoring all other subpopulations’ equating functions (Huggins & Penfield, 2012). One unconditional, group-to-overall method is the Root Expected Squared Difference (\( \text{RESD}_k \)) (Yang, 2004), defined as
\[ RESD_k = \frac{\sqrt{\sum_{x=1}^{X} P_x [((d_k(x)))]^2}}{\sigma_Q}. \]  

(11)

where \( k \) represents a single subpopulation (e.g., females). The \( RESD_k \) assesses invariance between one subpopulation’s equating function and the overall equating function across all score levels on test form 1. A conditional version of this method (Huggins & Penfield, 2012) is defined as

\[ RSD_k(x) = \frac{|d_k(x)|}{\sigma_Q}. \]  

(12)

which measures the standardized distance between one subpopulation’s equating function and the overall equating function at one score level of test form 1.

**Standard errors of equating.** All equating functions are defined by statistical estimates of parameters, and hence there is statistical error in the equating functions. This error needs to be examined because differences in equating functions that fall within the confidence interval for the true equating function can be considered negligible in terms of equating dependence (Harris & Kolen, 1986; Holland & Dorans, 2006; Kolen, 2004; Yi, Harris, & Gao, 2008), and because large standard errors should call attention to the adequacy of the equating function in general. With respect to invariance, the concern is with differences in true, and unknown, parameters of the equating functions rather than differences that can be attributed to statistical error in the specification of the equating functions.

Analytic methods for estimating standard errors of equating are available for IRT equating methods. In particular, Lord (1982) and Ogasawara (2001) developed a delta method for deriving standard errors of IRT true score equating and Ogasawara (2003) developed a delta method for deriving standard errors of IRT observed score equating.
However, these analytic methods are computationally intensive. Bootstrapping techniques can be utilized as an alternative method for estimating standard errors of equating (Kolen & Brennan, 2004). These techniques repeatedly sample person scores with replacement and estimate the equating function for each sample, resulting in a distribution of equating estimates for which the standard deviation is the standard error of equating.

**Differential Item Functioning and Differential Test Functioning**

DIF, as a measure of group dependence at the item level, and DTF, as an aggregate of DIF across items on a test, can be manifested in several different ways. Across a test, DIF can exist in a manner that favors one subpopulation across all items with DIF (i.e., unidirectional DIF), or in a manner that favors one subpopulation on some items with DIF and another subpopulation on other items with DIF (i.e., bidirectional DIF). In addition, the magnitude of DIF and subpopulation favored by DIF on a single item can be uniform across proficiency levels of examinees or non-uniform (Osterlind & Everson, 2009). Uniform DIF exists when the group differences in odds ratios of correct response are independent of ability level (Hanson, 1998; Penfield, 2010a). Non-uniform DIF exists when these group differences in odds ratios of correct response are a function of ability level (Hanson, 1998; Penfield, 2010a). Another way of defining the type of DIF in items is with respect to group differences in the $a$, $b$, and/or $c$ parameters of items modeled under IRT. Group differences in $b$ parameters, or the item difficulty parameters, are the most common in educational testing. Therefore, $b$ parameter DIF and the resultant DTF of $b$ parameter DIF is the focus of this study.
The multidimensional explanation of DIF. The multidimensional explanation of DIF is relevant to this study as dimensionality issues are seen as an underlying cause of both DIF and equating invariance. This common cause of these two interrelated types of statistical bias has informed the implied hypothesis that DIF in items has an effect on equating invariance.

The presence of a secondary factor(s), or multidimensionality, in a test item that is assumed to be unidimensional can result in DIF (Camilli & Shepard, 1994; Dorans, 2004; Osterlind & Everson, 2009), as all DIF effects are the result of a secondary factor assessed by an item. However, not all items measuring multiple dimensions display DIF. Whether or not a secondary factor is manifested as DIF is determined by whether or not the subpopulations of interest have a differential distribution on the secondary factor (Ackerman, 1992; Camilli, 1992). If a secondary factor is present but there is no differential distribution of the secondary factor, then item impact may be present but item bias is not. It is the differential distribution of the secondary factor that defines item bias (Ackerman, 1992). Item impact represents true differences in group abilities that are independent of ability level, while item bias refers to group differences in item performance for examinees of a comparable ability level (Camilli, 1992). The presence of item bias resulting from the differential distribution of a secondary factor means that item performance is dependent on group membership (Penfield & Camilli, 2007) and it is a violation of the standard of testing for fairness as lack of bias. It indicates that the items are not testing a valid, unidimensional construct for at least one group of examinees (Ackerman, 1992) and as a result the scores cannot be interpreted the same across groups.
As mentioned in a previous section, there are some reasons in the literature to believe that multidimensionality can impact equating invariance (Angoff & Cowell, 1985; Kolen, 2004; van der Linden, 2000). This study is intended to test for a particular type of multidimensionality that may have an effect on equating dependence; the type of multidimensionality that results in DIF may also result in equating dependence. If multidimensionality is manifested as differential distributions between groups on a secondary factor assessed by an item, then DIF will be present in that item (Ackerman, 1992; Camilli, 1992), and this study tests if that DIF and the resultant DTF is related to the presence or magnitude of equating dependence.

**Methods for detecting DIF and DTF.** There are numerous methods available for detecting DIF effects. This study focuses on IRT methodology in measurement, and hence this literature review focuses on the IRT methods for DIF detection.

There are many methods for IRT-based DIF effect estimation, including area measures (Cohen, Kim, & Baker, 1993; Kim & Cohen, 1991; Raju, 1988), Lord’s chi-square (Lord, 1980), the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988), and methods that follow the differential functioning of items and tests (DFIT) framework (Oshima & Morris, 2008; Oshima, Raju, & Nanda, 2006; Raju, van der Linden, & Fleer, 1995). This study makes use of the DFITS framework to derive estimates of true DTF.

When the true item parameters of two groups of examinees are known, true DIF can be derived by calculating the differences between subpopulations in true item characteristic curves and true DTF can be derived by calculating the differences between subpopulations in true test characteristic curves. As true item parameters are known in
this study, these methods are the focus of this literature review. True DIF can be defined as

\[ DIF = E_F[d_j(\theta_s)^2], \]  

(13)

where

\[ d_j(\theta_s) = P_{jF}(\theta_s) - P_{jR}(\theta_s), \]  

(14)

and \( P_{jF}(\theta_s) \) is the probability of correct response on item \( j \) at score level \( \theta_s \) for examinees in a focal population \( (F) \), \( P_{jR}(\theta_s) \) is the probability of correct response on item \( j \) at score level \( \theta_s \) for examinees in a reference population \( (R) \), and \( E_F \) is an expectation across the focal distribution of \( \theta \) values (Oshima & Morris, 2008; Penfield & Camilli, 2007; Raju et al., 1995). Stated simply, differences in true probabilities of correct response across two groups can be derived at multiple quadrature points along the scale of \( \theta \), and the differences can then be squared, weighted, and summed across the \( \theta \) scale to obtain a measure of true DIF in each item.

A similar process can be completed with true test characteristic curves to obtain true DTF values across items on a test. Specifically, true DTF can be defined as

\[ DTF = E_F[D(\theta_s)^2], \]  

(15)

where

\[ D(\theta_s) = T_{F}(\theta_s) - T_{R}(\theta_s), \]  

(16)

and \( T_{F}(\theta_s) \) is the true score for examinees in a focal population at score level \( \theta_s \), and \( T_{R}(\theta_s) \) is the true score of examinees in a reference population at score level \( \theta_s \) (Oshima & Morris, 2008; Raju et al., 1995). These true scores are obtained with the true item parameters of each item. Under this method, true DTF is defined as an aggregate of the squared differences in true scores between the two groups across the scale of \( \theta \).
**Relationship between Equating Invariance and Differential Item Functioning**

The relationship between equating invariance and DIF is critical in developing a sense of how the multiple concerns for fairness as lack of bias are interrelated, yet the current understanding of this relationship has little empirical support. Dorans (2004) briefly discusses how particular types of multidimensionality are expected to affect both equating invariance and DIF and how it is expected that a lack of DIF is a prerequisite for achieving invariant equating relationships, yet these statements were not empirically tested in the study. von Davier and Wilson (2007) explored DIF in core and anchor items before performing an equating procedure, acknowledging that the presence of DIF violates the unidimensionality assumption of IRT which could pose issues for equating latent scores. However, there was little evidence of DIF or multidimensionality in the items for that study.

Two recent studies (Han, 2008; Li, 2008) have empirically assessed the effect of item parameter drift (IPD) in linking and equating designs. IPD refers to the instance in which the parameters of an item (e.g., the $b$, or difficulty, parameter) change, or drift, over time, and this can occur in any NEAT equating design where the groups exposed to the forms to be equated are non-equivalent. These studies examined the effect of IPD on item parameter estimates, scaling coefficients, and proficiency classifications, but not on estimates of population invariance of equating. Another distinction between these studies and the current study is that IPD is a special case of DIF in which subpopulations are defined by the time of assessment, whereas the DIF examined in this study is across subpopulations of examinees within one assessment occasion. IPD effects relate to changes in item parameters across test administrations, whereas the DIF in this study
focuses on the fairness issue of differential expectations of performance on items across examinees with equal estimated ability levels but from different subpopulations. The nature of the effect of multiple DIF conditions on population invariance in equating has yet to be empirically examined, and the findings of such a study have different theoretical implications than do studies on IPD and equating.

Research in classical test theory (CTT) equating has shown that a lack of equating invariance is associated with a differential difficulty across tests or test forms (Cook & Petersen, 1987; Dorans, 2004; Dorans & Holland, 2000; Holland & Dorans, 2006; von Davier & Wilson, 2008) and the nature of the differential difficulty is directly related to the underlying lack of invariance in the equating function. From this information it is reasonable to hypothesize what differential difficulty across forms would look like under IRT methodology; when all items on a test are free of DIF there should be no differential difficulty in the data and therefore equating invariance should be satisfied. It is also reasonable to hypothesize that the presence of large DIF in many items may be associated with violations of population invariance in equating. While these hypotheses seem reasonable and are often assumed to be true by professionals in the testing community (Huggins & Penfield, 2012), they are too general to be helpful in understanding the true theoretical nature of the relationship between DIF and population invariance. For example, researchers state that having DIF-free items is probably a prerequisite to establishing equating invariance, but there may be some instances of DIF that do not result in differential difficulty across the subpopulations and forms, and therefore those instances would potentially not be problematic for establishing equating invariance.
The specific relationship between DIF and equating invariance is unknown. In order to inform IRT equating theory, it is necessary to know what magnitudes, frequencies, and directions of DIF in items are associated with a problematic level of population dependence in equating. It is also necessary to know how the differences in DIF and DTF across forms as well as the differences in mean ability levels of subpopulations affects the relationship between item level DIF, test level DTF, and reported score level equating dependence. This study is designed to inform these unknown relationships between anchor item DIF conditions, anchor test DTF conditions, and population invariance of equating.
Chapter 3: Objectives and Research Questions

Fairness as lack of bias can be detected at the item level, raw score level, and equated score level of a test. Bias at each level can be examined through DIF methods, DTF methods, and equating invariance methods, respectively. The manifestations of bias at these three levels are interrelated, yet the exact form of that relationship is unknown. Specifically, the frequency, magnitude, and direction of DIF that is manifested at the item level may or may not be manifested at the reported (i.e., equated) score level. Also, differences in DIF across forms as well as mean ability differences of subpopulations may or may not affect the relationship of DIF, DTF, and equating invariance. The equated score level is particularly relevant to fairness as lack of bias because high-stakes decisions are based on these reported, equated scores. Violations of population invariance in equating compromise the interchangeability of reported scores from year to year for at least one subpopulation of test takers, making the definition of AYP and other similar high-stakes cut-points unreliable across years for all groups of examinees in the test taking population.

As high-stakes decisions are based on reported scores, and equating dependence violates assumptions of equity in those reported scores, it is critical to have a better theoretical understanding of the underlying causes of equating dependence. This simulation study examines the effect of various conditions of DIF in anchor items on population invariance of equating. To increase generalizability to operational testing conditions for which high-stakes decisions are based on reported scores, IRT methods are used throughout the analyses, the commonly used NEAT equating design is simulated,
and the simulated test form distributions mimic those of an operational standardized K-12 assessment. This simulation will address the following research questions:

1. What is the effect of differential form DIF in anchor items on population invariance of IRT true score equating?

2. What is the effect of the magnitude of DIF in anchor items on population invariance of IRT true score equating?

3. What is the effect of the frequency of anchor items displaying DIF on population invariance of IRT true score equating?

4. What is the effect of the direction of DIF in anchor items on population invariance of IRT true score?

5. Do mean ability differences between groups impact the relationship between DIF in anchor items and population invariance of IRT true score equating?

6. What is the relationship between true DTF in anchor tests and population invariance of IRT true score equating?
Chapter 4: Method

A computer simulation was conducted to examine the effects of various conditions of anchor item DIF and anchor test DTF on multiple measures of population invariance in IRT true score equating. The R language and environment (R Development Core Team, 2005) was utilized for all portions of the simulation and analysis, and BILOG-MG (Zimowski, Muraki, & Mislevy, 2003) was batched through R for purposes of item calibration.

Two test forms, form 0 and form 1, were simulated to mimic the NEAT design in which anchor items were imbedded in the forms. Figure 1 displays a picture of the forms created in this study. The anchor items were utilized to establish the equating relationship between form 0 and form 1, and subsequently all scores on form 1 were equated to the scale of form 0 based on that equating relationship. Each form consisted of 50 dichotomously scored items with 20% of the items common to both forms (i.e., 10 anchor items). All anchor items were treated as internal, meaning the scores of anchor items counted toward examinee test scores and the total possible raw test score range was 0 to 50. As subpopulations used for disaggregated reporting purposes are often small, all analyses were simulated so that one subpopulation (i.e., the focal group) comprised 25% of the examinee sample while the other subpopulation (i.e., the reference group) comprised 75% of the examinee sample. It is difficult to manipulate the subpopulation sizes in a manner that generalizes to all subpopulations for which disaggregated scores are reported; these subpopulations can vary widely in the percentage they comprise of the examinee test taking population. Simulating the focal group as 25% of the population is
meant to approximate the common situation in which subpopulations for which scores are reported for high-stakes purposes comprise substantially less than 50% of the examinee population.

Conditions of DIF were simulated in the anchor items only (i.e., items unique to one form only were all simulated to have no true DIF). Table 2 displays the 74 conditions of DIF that were manipulated in the anchor items, with 100 trials being completed under each of the conditions. Because the focus of the study was on effect size rather than rejection rates, 100 trials was deemed adequate to yield sufficient stability in the mean effect size across trials. All simulated DIF was $b$ parameter DIF, meaning that the true $b$ parameters of the items exhibited group differences but the true $a$ and $c$ parameters were invariant across groups. DIF in $b$ parameters was chosen for this study as it is often the primary concern of problematic DIF in operational settings and because it helps to isolate the problematic effects more easily than can be done when a combination of $a$, $b$, and/or $c$ parameter DIF is simultaneously occurring in the items.

Conditions 1 and 2 are null conditions in which the items were simulated to have no true DIF. Condition 1 is the null for subpopulations that have no mean differences in ability level, and condition 2 is the null for subpopulations that display mean differences in ability level. The remaining 72 conditions represent the manipulation of five fully crossed factors (i.e., $3 \times 3 \times 2 \times 2 \times 2$): the factor of magnitude of DIF with three levels (small [$b_{F} - b_{R} = 0.30$], medium [$b_{F} - b_{R} = 0.60$], and large [$b_{F} - b_{R} = 0.90$]), where $b_{F} - b_{R}$ is the difference in $b$ parameter estimates for the focal and reference subpopulations (i.e., the amount of true $b$ parameter DIF); the factor of frequency of items in the anchor test with DIF with three levels (i.e., 2, 4, and 6 items, which comprise 20%, 40%, and 60% of
the anchor test, respectively); the factor of directionality of DIF with two levels (unidirectional and bidirectional); the factor of mean differences in subpopulation ability levels with two levels (no mean differences and mean differences); and the factor of differential form DIF with two levels (no differential form DIF and DIF in form 1 only).

The choices of these five factors and the levels within each factor were not arbitrary. First, the magnitudes of DIF were chosen based on the Educational Testing Service’s classification of small, medium, and large DIF effects (Dorans & Holland, 1993). Second, the frequency of DIF was chosen to simulate DIF in small, medium, and large percentages of items in the total anchor test. Third, the direction of DIF is known to have an impact on DTF and therefore potentially have an impact on equating dependence, and was simulated such that unidirectional DIF always favored the reference group. Fourth, mean differences in the ability distribution were simulated with the focal group scoring on average one standard deviation unit below the reference group, as many of the subpopulations for which disaggregated scores are reported have lower mean achievement than the overall population.

Finally, the differential form DIF factor was introduced to examine the possible differences in anchor item DIF that can occur when forms are administered to non-equivalent groups. Specifically, one level of this factor was associated with DIF being equivalent across the test forms. For example, if anchor item 2 in form 0 had DIF at a magnitude of 0.30 favoring the focal group, then anchor item 2 in form 1 displayed the same type of DIF. The other level of this factor was associated with DIF being present in form 1 only. Under this level of the factor, form 0 anchor items were always simulated to have no true DIF while form 1 anchor items were simulated to have true DIF according to
the other factors (i.e., magnitude, frequency, and direction). Differential form DIF in anchor items can occur in a variety of situations when groups are non-equivalent across forms. One specific situation would be that in which IPD is differential across groups. For example, if an item’s $b$ parameter drifts over time (i.e., across two different administrations of a test) differently for the reference group as compared to the focal group, then that item may not have had DIF in an initial administration of form 0 but would display DIF in a subsequent administration of form 1. This factor was used in the study because it most directly corresponds with the differential difficulty found to be problematic for equating dependence in CTT equating studies.

Item parameters of the 3PL model for the non-anchor items were randomly generated under each condition and for each trial. The anchor item parameters (i.e., $a$, $b$, and $c$ parameters) on each form were selected to mimic the distribution of 31 observed anchor item parameter estimates from a 2009 statewide 5th grade science assessment. In other words, the means and standard deviations of $a$, $b$, and $c$ parameters from this assessment were used to develop distributions of item parameters from which items for this study could be randomly drawn. This was done to ensure that the study applies to operational testing situations, and the chosen assessment was one that was operationally equated under a very common equating design that could be mimicked in this study (i.e., the state department of this assessment maintains an item bank for which the items are used as anchors on different forms of the test given to non-equivalent groups of examinees). Table 3 displays the item parameters of these 31 anchor items. As the anchor
form must be a miniature version of the full form for equating purposes (von Davier et al., 2004), these distributional properties were also utilized to generate the anchor item parameters on each form.

The true values of $b$ parameter anchor item DIF were known in each condition of the study, and the resultant true DTF was calculated with Equations 15 and 16 for each anchor test in each condition of the study. The $b$ parameter DIF and the resultant true anchor test DTF are displayed in Table 2.

Ability values ($\theta$) distributions were drawn from a normal distribution with a mean of 0 and a standard deviation of 1. In conditions with mean differences, the focal group examinee $\theta$ values were generated from a normal distribution with a mean of -1 and a standard deviation of 1. A total of 4,500 simulated examinees were assigned to each form in each trial, with 3,000 belonging to the reference group and 1,500 belonging to the focal group. Using the simulated $\theta$ values and item parameters, dichotomous responses to the 50 items on each form were generated by: (a) obtaining the probability of correct response on each item for each individual with Equation 1; (b) selecting a random value from a uniform distribution ranging from 0 to 1; (c) determining if the probability of correct response was equal to or greater than the randomly selected value; and (d) assigning a response of 1 (i.e., correct) if the probability of correct response was equal to or greater than the randomly selected value, or assigning a response of 0 (i.e., incorrect) if the probability of correct response was less than the randomly selected value. Through this process, all simulated examinees had a pattern of correct and incorrect responses on all 50 items of one of the test forms.
Once the data was generated for each trial, the two forms within each trial were scaled to the same metric via the mean-sigma method. Subsequent to scaling, IRT true score equating was conducted. Specifically, the scaled item parameters were utilized to determine equivalent true scores across forms and an estimated true score equating relationship was established based on the sample data. The estimated true score equating relationship was then used to develop conversion tables, such that each possible integer true score (i.e., scores 0 to 50) on form 1 was associated with an equated score on the scale of form 0. Equated scores that fell outside of the number correct score range from 0 to 50 were truncated (e.g., an equated score of 50.12 was changed to 50.00). These scaling and equating procedures were completed for each condition in the study three times; once using the data of the overall population of examinees on the forms, once using the data from the reference group only, and once using the data from the focal group only.

From these results, population invariance of equating was estimated via the unstandardized $REMSD$, $RESD_k$, $RMSD(x)$, and $RSD_k(x)$ indices. In total, 100 trials were run within each condition, resulting in distributions of $REMSD$, $RESD_F$, $RESD_R$, $RMSD(x)$, $RSD_F(x)$, and $RSD_R(x)$ estimates for each condition ($n = 100$). The means of these distributions serve as an estimate of the population level of equating dependence for each condition, and the standard deviation was considered the standard error of the equating dependence indices.

Complete equating invariance would be confirmed with the equating indices producing values of zero (i.e., showing no differences in equated scores across the overall and subpopulation equating). However, the equated scores being compared across the
different populations are estimates, and hence it is expected that they will differ by more than zero due to measurement and equating error (Braun & Holland, 1982; Dorans, 2004; Dorans & Liu, 2009; Lord, 1980; Lord & Wingersky, 1984). Therefore, a difference that matters (DTM) needs to be defined to assess the magnitude of the equating dependence. For this study, two DTMs were used as benchmarks of problematic levels of equating dependence. The first is a conservative DTM set at 0.50 score points. This benchmark acknowledges that any differences in equated scores that are larger than half of a score point (0.50) would be problematic as the scores may be rounded to different reported values. This benchmark was applied in this study as many researchers have previously used DTMs equal to half of a score point in equating invariance studies (e.g., Dorans & Feigenbaum, 1994; Yang & Gao, 2008; Yi et al., 2008), however it can be misleading. There are many instances in which equated scores that differ by less than 0.50 would be rounded to different reported scores, and many instances in which equated scores that differ by more than 0.50 would be rounded to the same reported score (Huggins & Penfield, 2012). Therefore, a second DTM equal to 1.00 score points was applied in this study. Equated scores that differ by more than 1.00 score points will definitely be rounded to different reported scores and can therefore be deemed problematic levels of equated score differences. The DTM of 1.00 is more liberal than 0.50, but it can also be more informative about the nature of problematic levels of equating dependence.

The results across conditions were compared via the following analyses in order to address all research questions. A portion of the analyses were completed in Statistical Package for the Social Sciences (SPSS) version 19. First, omnibus conditional equating invariance results were displayed graphically and compared visually across conditions.
Second, multiple ANOVA procedures were utilized to determine the nature of the effect of each of the factors in the study (i.e., magnitude of DIF, frequency of items with DIF, direction of DIF, true mean ability differences in DIF, and differential form DIF) on equating dependence. The main effects and interactions tested in each ANOVA were based on the findings from the results of the omnibus conditional equating invariance analysis. Plots of significant interactions were examined graphically and post hoc tests were used on factors with more than two levels (i.e., magnitude of DIF and frequency of items with DIF). Third, Pearson’s correlations and scatter plots were used to examine the relationship between true anchor DTF and unconditional equating invariance in the conditions.
Chapter 5: Results

The results of this study are presented in four sections. The first section describes the equating dependence results for the two null conditions (i.e., condition 1 and 2). The second concerns the findings of the effects of anchor item DIF on conditional measures of equating dependence. The third presents the findings of the effects of anchor item DIF on unconditional measures of equating dependence. The final section concerns the relationship between true anchor DTF and multiple measures of unconditional equating dependence.

Equating Dependence in the Null Conditions

Two null conditions were simulated and analyzed in this study. Condition 1 simulated no true DIF in the anchor items and the samples of examinees were simulated to have the same true mean ability levels across the focal and reference groups. The conditional equating dependence results (i.e., $RMSD(x)$ and $RSD_k(x)$) for condition 1 are shown in Figure 2. This type of graph for the conditional dependence results will be used repeatedly in this study. It displays the $RMSD(x)$ and $RSD_k(x)$ results as a function of the form 1 unequated true scores. It also displays the two DTM$s$ used in this study. Conditional equating dependence values above the DTM$s$ would be considered problematic for reported score equity, while values below the DTM$s$ would not be considered problematic. The graphs for condition 1 results show that when no true DIF is present and when the groups have equal true mean ability levels, all conditional equating dependence is below the DTM of 1.00 and most are below the DTM of 0.50. The
conditional indices are slightly larger for the focal group as compared to the reference group, and all indices have larger values of dependence on the lower end of the score scale where sample sizes are smaller.

The $RMSD(x)$ and $RSD_k(x)$ results of condition 2 are shown in Figure 3. This is the null condition of DIF when the focal group had a true mean ability level one standard deviation unit lower than the reference group. While the shapes of the $RMSD(x)$ and $RSD_k(x)$ values across the score scale are slightly different from condition 1, it is still the case that all measures of $RMSD(x)$ and $RSD_k(x)$ are below the DTM of 1.00 and many are below the DTM of 0.50.

The two null conditions are similar in results of unconditional equating invariance indices (i.e., the $REMSD$, $RESD_R$ and $RESD_F$). In condition 1, $REMSD = 0.37$, $RESD_R = 0.29$, and $RESD_F = 0.51$. In condition 2, $REMSD = 0.48$, $RESD_R = 0.39$, and $RESD_F = 0.60$. The unconditional indices of equating dependence are larger when there are true mean differences in ability between the groups (i.e., in condition 2), but all equating dependence estimates in the null condition are below the DTM of 1.00 and most are below the DTM of 0.50. The $RESD_F$ and $RSD_F(x)$ indices comparing the focal group to the overall population, ignoring the reference group, have the largest estimates of equating dependence in the null conditions.

**Differential Item Functioning and Conditional Measures of Equating Dependence**

The differences in $RMSD(x)$ results across the various DIF conditions can be seen through visual inspections of graphical displays similar to those in Figures 2 and 3. Figures 4 and 5 show the $RMSD(x)$ results for all conditions with no differential form DIF and no true mean ability differences between the groups. Figure 4 is the
unidirectional DIF results and Figure 5 is the bidirectional DIF results. As can be seen in these figures, the $RMSD(x)$ values are very similar across these conditions, smaller than the DTM of 1.0, and often smaller than the DTM of 0.50. The magnitude, frequency, and direction of DIF do not have an effect on the presence or magnitude of $RMSD(x)$ values at each score level. Figures 6 and 7 are the same as Figures 4 and 5, except they show conditions in which the focal group was simulated to have a mean ability level one standard deviation unit below the reference group. Again, the $RMSD(x)$ results are very similar across these conditions and the magnitude, frequency, and direction of DIF do not have an effect on the presence or magnitude of equating dependence as estimated by the $RMSD(x)$. The manner in which the $RMSD(x)$ values fluctuate across the score scale varies between the conditions with no mean differences in ability and the conditions with mean differences in ability, in a very similar manner to how the two null conditions differ in $RMSD(x)$ fluctuation across the score scale. In sum, Figures 4 through 7 show the $RMSD(x)$ results for all conditions in this study in which there was no difference in DIF across the two forms, and the graphical results are indistinguishable from the results of the null conditions. The $RSD_F(x)$ and $RSD_R(x)$ results (not pictured for reasons of parsimony) showed the same findings as the $RMSD(x)$; the graphs were indistinguishable from the null condition $RSD_F(x)$ and $RSD_R(x)$ equating results shown in Figures 2 and 3.

The conditions in which differential form DIF is present show different results for the $RMSD(x)$ results of equating dependence. Figures 8 through 11 are laid out the same as Figures 4 through 7, except the former set of figures displays all of the conditions that have differential DIF across the forms (i.e., form 0 has no true DIF and form 1 has true DIF). Figure 8 displays conditions with differential form DIF, no group mean differences
in ability, and unidirectional DIF. In these conditions, frequency of items with unidirectional DIF had an effect on $RMSD(x)$ results of equating dependence, yet not in a consistent way. Some conditions with six items with unidirectional DIF (e.g., condition 53) have more problematic magnitudes of equating dependence than other conditions with four items with unidirectional DIF (e.g., condition 47). However, in other instances having four items with unidirectional DIF (e.g., condition 45) appears to sometimes be more problematic than having six items with unidirectional DIF (e.g., condition 51), with respect to $RMSD(x)$ results of equating dependence. The factor of magnitude of unidirectional DIF on the other hand had a consistent effect on the $RMSD(x)$ results. As the magnitude of unidirectional DIF increases, the magnitude of $RMSD(x)$ values increases for a majority of the score levels. In addition, the magnitude of $RMSD(x)$ values fluctuates across the score scale more in conditions with higher magnitudes of unidirectional DIF than it does in conditions with lower magnitudes of unidirectional DIF.

Figure 9 is the same as Figure 8 except that it concerns bidirectional DIF as opposed to unidirectional DIF. Bidirectional DIF appears to have a similar effect on the $RMSD(x)$ as does unidirectional DIF, but the effect on the magnitude of equating dependence in bidirectional DIF appears smaller. Frequency of items with bidirectional DIF has an effect on the $RMSD(x)$ results, albeit an inconsistent effect. Magnitude of bidirectional DIF has a consistent effect on the $RMSD(x)$ results; as magnitude of bidirectional DIF increases, $RMSD(x)$ results of equating dependence increase and the fluctuation of this type of equating dependence across the score scale increases.
Figures 10 and 11 show the results for conditions with differential form DIF and true group mean ability differences, with the former focusing on conditions of unidirectional DIF and the latter focusing on conditions of bidirectional DIF. For these results, frequency appears to have a small to negligible effect on $RMSD(x)$ results, with increases in frequency from 4 to 6 items resulting in either no increase in the magnitude of equating dependence or small increases in the magnitude of equating dependence. Magnitude of DIF has a larger effect on the presence and magnitude of $RMSD(x)$; as magnitude of DIF increases, $RMSD(x)$ results increase, the fluctuation of dependence across the score scale increases, and many $RMSD(x)$ values are above both DTMs used in this study. Both levels of the directionality factor (i.e., unidirectional and bidirectional) appear to have effects on the $RMSD(x)$ equating dependence results. True group mean differences appear to change the shape (i.e., fluctuation across the score scale) of the $RMSD(x)$ results, and the effects of frequency, magnitude, and direction of DIF appear to be slightly larger when there are true group mean ability differences.

The group-to-overall conditional equating dependence (i.e., the $RSD_F(x)$ and $RSD_R(x)$) results for all conditions with differential form DIF showed the same patterns of results as the $RMSD(x)$. For reasons of parsimony, only a select sample of the group-to-overall conditional equating dependence results are displayed in Figure 12. These two conditions show conditional equating dependence for the focal group and reference group separately, and the patterns shown in these conditions were similar to the patterns in other conditions with differential DIF across forms. However, for the group-to-overall results it was clear that the magnitudes of equating dependence were larger for the focal group as
compared to the reference group, and the fluctuation of equating dependence across the score scale was more dramatic for the focal group as compared to the reference group.

**Differential Item Functioning and Unconditional Measures of Equating Dependence**

ANOVA procedures were utilized to examine the effect of the DIF conditions on conditional measures of equating dependence (i.e., $REMSD$, $RESD_R$, and $RESD_F$). All distributions of $REMSD$, $RESD_R$, and $RESD_F$ within conditions were positively skewed. A natural logarithmic transformation was applied to each of the distributions before performing the ANOVAs in order to meet the assumption of normality. The homogeneity of variance assumption was not a concern as the sample sizes of equating dependence estimates within the levels of each factor were equal.

The first ANOVA focused on the $REMSD$ (i.e., omnibus unconditional invariance) as the outcome of interest. The interaction effects tested in the first ANOVA were based on the above results seen with the $RMSD(x)$. Differential form DIF was the factor that had the most prominent effect on changes in the $RMSD(x)$ results, in that conditions without differential form DIF showed little to no variation in $RMSD(x)$ results and conditions with differential form DIF varied widely in $RMSD(x)$ results. Hence, an ANOVA was run in which all factors interacted with the differential form DIF factor. All interaction terms were statistically significant and together they explained 59% of the variance in $REMSD$ values. The interaction of magnitude of DIF and differential form DIF explained 23% of the variance in $REMSD$ values ($\eta^2 = 0.23$, $F(4, 7186)= 539.25$, $p < .001$), controlling for all other interactions. The interaction of frequency of items with DIF and differential form DIF explained 7% of the variance in $REMSD$ values ($\eta^2 = 0.07$, $F(4, 7186)= 130.71$, $p < .001$), controlling for all other interactions. The interaction of
directionality of DIF and differential form DIF explained 7% of the variance in $REMSD$ values ($\eta^2 = 0.07, F(2, 7186)= 279.60, p < .001$), controlling for all other interactions. The interaction of mean ability differences and differential form DIF explained 15% of the variance in $REMSD$ values ($\eta^2 = 0.15, F(2, 7186)= 616.08, p < .001$), controlling for all other interactions.

Graphical depictions of these interactions are displayed in Figure 13. For each graph in Figure 13, two horizontal lines in the graphs are located at the two DTMs for this study as points of reference for the presence and magnitude of equating dependence results. For the magnitude of DIF, frequency of items with DIF, direction of DIF, and mean ability differences factors, the interactions with differential form DIF are the same; when DIF is equivalent in the anchor items across form 0 and form 1 the levels of the factors do not differ significantly in mean $REMSD$ values, but when DIF is present in form 1 only each of the levels of the factors has a different effect on the mean $REMSD$ values. When DIF is in form 1 only, an increase in the magnitude of DIF is associated with an increase in mean $REMSD$ values, an increase the frequency of items with DIF is associated with an increase in mean $REMSD$ values, unidirectional DIF is associated with larger mean $REMSD$ values than is bidirectional DIF, and samples with mean differences in true ability levels are associated with larger mean $REMSD$ values than are samples with no mean differences in true ability levels. As shown in the graphs in Figure 13, all mean $REMSD$ values for conditions in which there are no differences in DIF across the forms are close to and/or below the DTM of 0.50. In contrast, all mean $REMSD$ values for conditions in which there are differences in DIF across forms are above the DTMs of
1.00 and/or 0.50, meaning that the differences in these conditions are more problematic for reported scores as compared to conditions with no differences in DIF across forms.

Based on these interaction findings, a second ANOVA was run using only the conditions with differential form DIF (i.e., where DIF is in form 1 only). While there were some statistically significant interactions in this ANOVA, they explained negligible amounts of variance in the REMSD values. Therefore, the main effects of each of the four factors in this study were the main findings of this ANOVA. Magnitude of DIF had the largest effect on REMSD values, explaining 46% of the variance amongst the conditions with differential form DIF ($\eta^2 = 0.46, F(2, 3593)= 1515.10, p < .001$), controlling for the three other factors. Mean differences in true ability levels had the second largest effect on REMSD values, explaining 25% of the variance amongst the conditions with differential form DIF ($\eta^2 = 0.25, F(1, 3593)= 1219.11, p < .001$), controlling for the three other factors. Direction of DIF explained 18% of the variance in REMSD values ($\eta^2 = 0.18, F(1, 3593)= 764.46, p < .001$), controlling for the three other factors. Frequency of items with DIF had the smallest effect of the four factors, explaining 16% of the variance in REMSD values ($\eta^2 = 0.16, F(2, 3593)= 351.36, p < .001$), controlling for the three other factors.

Post hoc analyses were required for the magnitude and frequency factors as they each contained three levels, and a Bonferroni test was run for each of these factors. For magnitude of DIF, all three magnitude levels of DIF (i.e., small [$b_F-b_R = 0.30$], medium [$b_F-b_R = 0.60$], and large [$b_F-b_R = 0.90$]) were statistically significantly different from each other. For frequency of items with DIF, the levels of 4 items with DIF and 6 items
with DIF did not differ significantly from each other ($MD_{6-4} = 0.02, SE = 0.02, p = .752$).

The right side of each graph in Figure 13 displays the means that were compared for this ANOVA and the post hoc analyses.

The ANOVA analyses from this section were repeated for the group-to-overall unconditional equating dependence results (i.e., the $RESDF$ and $RESDR$). For the reference group, there were some slight differences in results of the ANOVAs on $RESDR$ values as compared to the results for the $REMSD$ analysis. In the first ANOVA testing the interactions of differential form DIF with all other factors, the interactions were all statistically significant, but they explained 40% of the variance in $RESDR$ values as compared to the 59% of variance explained in the $REMSD$. The interactions are graphed in Figure 14. From this figure, it is clear that the patterns of the interactions mirror those seen with the $REMSD$ interactions, but the magnitudes of the mean $RESDR$ values are lower across the different conditions and often lower than one or both of the DTMs.

When comparing the reference group’s equating results to the overall group’s equating results, ignoring the focal group, unconditional differences in the two equatings when DIF is in form 1 only are overall less problematic for reported scores than the omnibus (i.e., $REMSD$) unconditional differences.

In the second $RESDR$ ANOVA using only the conditions of differential form DIF and testing for the main effects of all other factors, the effect sizes of the main effects were more evenly dispersed across the four factors as compared to the $REMSD$ main effects in a similar model. Magnitude of DIF explained 15% of the variance in $RESDR$ values ($\eta^2 = 0.15, F(2, 3593)= 305.18, p < .001$), controlling for all other factors. Frequency of items with DIF explained 18% of the variance in $RESDR$ values ($\eta^2 = 0.18,$
Direction of DIF explained 12% of the variance in $RESD_R$ values ($\eta^2 = 0.12$, $F(1, 3593)= 484.58, p < .001$), controlling for all other factors. Mean differences in ability explained 13% of the variance in $RESD_R$ values ($\eta^2 = 0.13$, $F(1, 3593)= 549.30, p < .001$), controlling for all other factors. In post hoc follow up analyses using the Bonferroni test, all three conditions of magnitude of DIF were statistically different from each other and all three conditions of frequency of items with DIF were statistically different from each other.

For the focal group’s $RESD_F$ values, the two ANOVAs showed similar results to those from the $REMSD$ ANOVAs. All factors interacted with the differential form DIF factor and together explained 56% of the variance in $RESD_F$ values. Figure 15 displays the interactions of each factor with the differential form DIF factor. As can be seen in the graphs of this figure, the patterns of the interactions are consistent with the patterns seen in the $REMSD$ interactions. However, the magnitudes of the mean $RESD_F$ values for each factor are larger than the mean $REMSD$ and $RESD_R$ values. Mean $RESD_F$ values on the right side of each graph (i.e., those associated with conditions in which DIF was in form 1 only) are well above the two DTM for the study indicating problematic levels of dependence for score reporting. When comparing the focal group’s equating results to the overall group’s equating results, ignoring the reference group, unconditional differences in the two equatings when DIF is in form 1 only are overall more problematic for reported scores than the omnibus (i.e., $REMSD$) unconditional differences.

Within the conditions that displayed differential form DIF, the main effects of the other four factors on $RESD_F$ values were as follows: (a) magnitude of DIF had the largest main effect, explaining 50% of the variance in $RESD_F$ values ($\eta^2 = 0.50$, $F(2, 3593)= 385.72, p < .001$), controlling for all other factors.
1785.38, \( p < .001 \)), controlling for other factors; (b) mean differences in true ability levels had the second largest effect, explaining 17% of the variance in \( RESD_F \) values (\( \eta^2 = 0.17 \), \( F(1, 3593) = 746.19, p < .001 \)), controlling for other factors; (c) direction of DIF explained 15% of the variance in \( RESD_F \) values (\( \eta^2 = 0.15 \), \( F(1, 3593) = 610.23, p < .001 \)), controlling for other factors; (d) frequency of items with DIF had the smallest effect, explaining 8% of the variance in \( RESD_F \) values (\( \eta^2 = 0.08 \), \( F(2, 3593) = 163.52, p < .001 \)), controlling for other factors.

**Differential Test Functioning and Equating Dependence**

True DTF of the anchor test of form 1 was calculated with Equation 15 and 16. True DTF in form 0 was equal to true DTF in form 1 in conditions where true DIF was identical across forms 0 and 1, and true DTF in form 0 was equal to zero in conditions where true DIF was only simulated in form 1. Therefore, form 1 true DTF on the anchor test was the focus of this analysis in addressing research question 6.

The relationship between form 1 true DTF on the anchor test and unconditional invariance was assessed with bivariate Pearson’s correlations. Each condition in the study has a form 1 true DTF value associated with it, as well as a mean \( REMSD, RESD_F, \) and \( RESD_R \) value. Given the results of the above sections of this study, it was hypothesized that the correlations of form 1 true DTF and unconditional equating invariance values would differ across conditions with equivalent DIF in both the forms and conditions with DIF in form 1 only. Treating conditions as the level of analysis, six Pearson’s correlations were estimated to understand the relationship between anchor test DTF and equating dependence.
Table 4 displays all results of the Pearson’s correlations. True DTF in the anchor of form 1 is statistically significantly correlated with unconditional equating dependence values when the DTF is in form 1 only. Specifically, when DTF is in form 1 only it explains 64% of the variance in the $REMSD$ values, 67% of the variance in the $RESD_R$ values, and 58% of the variance in the $RESD_F$ values. However, when the true anchor DTF is equivalent across both forms the correlations with unconditional equating dependence are small and non-significant. Plots of the significant correlations are shown in Figure 16 with least squares regression lines added to the graphs. Figure 16 shows that the nature of the conditions in this study led to many values of DTF at or below 4.0, and only a few values of DTF above 4.0. The high level of heterogeneity of equating dependence estimates at the lower DTF values can be explained by the fact that these DTF values are based on numerous conditions that varied (e.g., magnitude, frequency, directionality, and mean differences) and are likely introducing the high level of variance in equating dependence estimates. This is a limitation in that the correlation of DTF and equating dependence is based on a limited number of DTF values of high magnitude as well as a large amount of variability in the equating dependence associated with DTF values of lower magnitude. Nonetheless, as shown in Table 4 and Figure 16, when DTF is differential across forms, increases in the magnitude of DTF in form 1 are associated with increases in the magnitude of unconditional equating dependence. The correlational relationships are large, positive, and statistically significant at the $p < .001$ level.
Chapter 6: Conclusions and Discussion

The purpose of this study was to examine the effects of various conditions of DIF in anchor items, and the resultant anchor test DTF, on the magnitude and presence of equating dependence in IRT true score equating. Research question 1 asked about the relationship between differential form DIF and equating dependence in IRT true score equating. This factor was manipulated in the study because researchers focusing on CTT equating methods have shown that differential difficulty across test forms and subpopulations can impact CTT equating dependence (Dorans, 2004; Dorans & Holland, 2000). The results of this study extended this understanding of differential difficulty to equating in the IRT framework. Results showed that when DIF is equivalent in anchor items across forms, equating dependence is not affected and it closely mirrors equating dependence findings in conditions with no DIF in any anchor items. While DIF in anchor items (or any items on an assessment) is a possible concern for validity of test scores, results of this study show that it does not impact equating dependence if it is identical across forms. However, when DIF in anchor items differs across forms to be equated, increases in the magnitude of equating dependence occur and potentially result in a level of equating dependence that is problematic for reported scores.

Researchers in educational assessment have increasingly become concerned with potential problematic effects of IPD on equating (e.g., Bock, Muraki, & Pfeiffenberger, 1988; DeMars, 2004; Donoghue & Isham, 1998; Han, 2008; Li, 2008; Wells, Subkoviak, & Serlin, 2002), but this study calls for a focus on differential IPD as one of several potential causes of equating dependence under the NEAT design. If the $b$ parameters of anchor items change over time in a different manner for different subpopulations,
differential form DIF will be present and can have an impact on equating dependence. Differential IPD is possible in any equating design in which non-equivalent groups are exposed to different forms of an assessment. However, differential IPD is not the only potential cause of differential form DIF in assessment. Any equating under the NEAT design has the potential to be affected by differential form DIF as the non-equivalent groups can show different values of DIF across forms.

Research questions 2, 3, 4, and 5 call for examinations of the effects of magnitude of anchor item DIF, frequency of anchor items with DIF, direction of anchor item DIF, and differences in group mean ability levels on equating dependence, respectively. The findings of this study show that these factors of DIF are related to equating dependence only in instances in which DIF in anchor items is differential across test forms. When anchor item DIF is differential across forms, magnitude of DIF has the largest impact on equating dependence of these four factors. Even if DIF is limited to a few items and is bidirectional, medium to large magnitudes of DIF in anchor items can have a substantial impact on equating dependence. Mean differences in ability levels across groups can also have an effect on equating dependence. In this study, mean differences in ability resulted in larger magnitudes of equating dependence when DIF was differential across forms. It is unclear if this will always be the case, but it has implications for understanding possible problems for reporting disaggregated scores across groups with varying levels of mean ability; differential IPD and, more generally, differential form DIF may be seen as more concerning for score equity when considering groups with significantly lower or higher mean ability levels.
Direction of true DIF in anchor items had a consistent effect on equating dependence when DIF was differential across forms; both unidirectional and bidirectional DIF can pose problems for equating dependence, but the former appears to be more problematic. Potentially, certain conditions of bidirectional DIF would have less of an effect on equating dependence than others. Specifically, if bidirectional DIF occurred across items that had the same location on the ability metric (i.e., the same $b$ parameter across the overall population), it could potentially cancel itself out in terms of its effect on equating dependence. However, in this study the items with bidirectional DIF were randomly selected and the different locations of the items on the ability metric appeared to affect the simulated examinees in a way that was inconsistent across the items, and therefore did not allow the equivalent magnitudes of the directionality of the DIF to fully cancel itself. In sum, the findings of this study show that unidirectional DIF across anchor items is problematic for equating dependence when the DIF is differential across forms, and bidirectional DIF across anchor items also has the potential of being problematic for equating dependence when the DIF is differential across forms.

The final DIF factor examined in this study was frequency of anchor items with DIF. While an increase in the number of anchor items with differential form DIF was generally associated with an increase in equating dependence, this was not always the case and the relationship was sometimes inconsistent and small. The number of anchor items with differential form DIF appears to be less influential on the magnitude of equating dependence than does the other three factors affecting equating dependence under differential form DIF conditions. The effect of frequency was difficult to see when visually comparing the conditional equating results, and for the unconditional equating
results there were instances in which conditions with 4 anchor items with DIF and 6 anchor items with DIF did not differ in equating dependence. Nonetheless, the effect of frequency of anchor items displaying differential form DIF was non-negligible in all other instances and therefore cannot be ruled out as a cause of equating dependence.

Research question 6 was posed to understand the nature of the relationship between true anchor item DIF conditions, the resultant true anchor test DTF of those conditions, and the magnitude of equating dependence found in those conditions. When anchor DTF is identical across forms, the amount of DTF is not related to the magnitude of equating dependence. Under these conditions equating invariance was confirmed within this study. When anchor DTF differed across forms, the relationship of these three types of group invariance concerns was clear; more extreme conditions of anchor item DIF resulted in larger magnitudes of DTF, which in turn were associated with larger magnitudes of equating dependence.

There are at least two additional findings in this study that were not specifically addressed through research questions, but are nonetheless important findings. First, all of the conditions for which DIF affected equating dependence did so differently across the two subpopulations of the study. Equating dependence for the reference (i.e., larger) group was consistently lower in magnitude than equating dependence for the focal (i.e., smaller) group. Restated, the anchor item DIF effects on equating dependence in this study posed more problems for groups constituting 25% of the overall population than they did for groups constituting 75% of the overall population. In instances in which equating dependence is present, it will always be more problematic for a smaller group in the dependence study as compared to a larger group. This is true because the larger group
(i.e., the reference group in this study) has more weight in defining the overall scaling
and equating, and will therefore always have equated scores that are more similar to the
overall equating as compared to a smaller group. This finding further promotes careful
attention to disaggregating scores across subpopulations of unequal sizes, and the
necessity of examining for differential form DIF across subpopulations of unequal sizes.

Second, while the factors in this study were able to explain large portions of the
variance in equating dependence indices, it is clear that other factors must have an effect
on equating dependence as well. This must be true as the true DIF conditions and DTF
values did not explain 100% of the variance in equating dependence results, and also
because they were not able to fully explain the extreme fluctuations of conditional
dependence across score levels. It appears as if sample size differences across score
levels as well as the location of anchor items with DIF on the ability scale have impacts
on equating dependence, and potentially that the effects of sample size and location are
magnified as the magnitude of differential form anchor DIF increases. Presumably,
measurement error and equating error also impact the presence and magnitude of
equating dependence. The conditions of this study were not systematically varied in a
manner that can inform the true nature of these extraneous effects (i.e., extraneous to this
study). For example, the conditions of this study cannot confirm the nature of the extreme
fluctuation of equating dependence across the scores scales, and this is a recommended
direction for future research. But it is clear that particular conditions of DIF in anchor
items and anchor test DTF have substantial effects on the presence and magnitude of
equating dependence, above and beyond the effects of sample size, measurement error,
and the like.
This study makes several contributions to understanding the nature of the relationship between anchor item DIF and equating dependence in the NEAT design. However, it is not without limitations. Some of the conditions in this study were realistic to instances of DIF in anchor items that may be seen in operational settings (e.g., conditions with two anchor items displaying small to moderate DIF), while other conditions of this study had extreme instances of DIF that may not be seen in operational testing situations (e.g., conditions with 6 out of 10 anchor items displaying large DIF, as well as null conditions in which no true DIF exists in any anchor items on the test). Additional details of the simulation may rarely be observed in operational assessment, including the lack of true DIF in all non-anchor items as well as the consistent $b$ parameter nature of the DIF manipulated throughout the study. Also, to appropriately contain the scope of the study, the simulation and analyses focused on one equating design (i.e., the NEAT design), one equating method (i.e., IRT true score equating), one scaling method (i.e., mean-sigma scaling), one set of sample sizes, and a series of singular decisions in assessment construction (e.g., number of items on the forms and in the anchor tests). These limitations of the study make it difficult to provide direct implications for practice (which is true of any equating study that does not mirror a specific operational test), but they do not limit the ability of the study to inform the theoretical understanding of equating invariance within the IRT framework.

The results of this study have multiple implications that contribute to furthering knowledge on the causes of IRT-based equating dependence in educational assessment. DIF in anchor items and DTF of anchor tests that differ across forms are direct causes of equating dependence. When differential DIF in anchor items is present, the effect that it
has on equating dependence is magnified by increases in the magnitude of the DIF and, to a lesser degree, increases in the frequency of items with DIF. These effects can be problematic in conditions of both unidirectional and bidirectional DIF, and can pose problems for subpopulations that have equal or different mean ability levels. Based on these results, it is recommended that practitioners performing an equating under the NEAT design monitor for anchor items that display differential DIF across forms being equated and for items that display differential IPD, as use of these items in establishing an equating relationship can lead to problematic levels of equating dependence. Since the issues that differential DIF across forms causes for equating dependence are most problematic for the smallest group of interest in the population, it is also recommended that practitioners assess equating dependence with group-to-overall methods in order to ensure that the differential DIF across forms did not impact the smallest group in the examinee population.
Works Cited


Klein, Alyson. “Clock Ticks on Senate Bill to Overhaul NCLB.” *Education Week* 2 Nov. 2011: 1,18-19. Print.


Table 1

*Methods for Assessing Invariance According to Two Dimensions*

<table>
<thead>
<tr>
<th>Method</th>
<th>Conditional Method</th>
<th>Unconditional Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omnibus Method</td>
<td>Description: The typical distance between each subpopulation’s linking function (simultaneously) and the overall linking function at one score level.</td>
<td>Description: The typical distance between each subpopulation’s linking function (simultaneously) and the overall linking function across all score levels.</td>
</tr>
<tr>
<td></td>
<td>Indices: ( \text{RMSD}(x) ) (Dorans &amp; Holland, 2000)</td>
<td>Indices: ( \text{REMSD} ) (Dorans &amp; Holland, 2000)</td>
</tr>
<tr>
<td>Group-to-overall Method</td>
<td>Description: The distance between one subpopulation’s linking function and the overall linking function at one score level.</td>
<td>Description: The typical distance between one subpopulation’s linking function and the overall linking function across all score levels.</td>
</tr>
<tr>
<td></td>
<td>Indices: ( \text{RSD}_k(x) ) (Huggins &amp; Penfield, 2012)</td>
<td>Indices: ( \text{RESD}_k ) (Yang, 2004)</td>
</tr>
</tbody>
</table>
Table 2

Simulated Conditions of Differential Item Functioning

<table>
<thead>
<tr>
<th>Condition</th>
<th>Differential Form DIF (No or Yes)</th>
<th>Group Mean Ability Difference ($M_F - M_R$)</th>
<th>Mean (# of Items with DIF)</th>
<th>Magnitude ($b_F - b_R$)</th>
<th>Direction (Unidirectional or Bidirectional)</th>
<th>Form 0 Anchor True DTF</th>
<th>Form 1 Anchor True DTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (null 1)</td>
<td>N</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>---</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 (null 2)</td>
<td>N</td>
<td>-1 SD</td>
<td>0</td>
<td>0.00</td>
<td>---</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>2</td>
<td>0.30</td>
<td>U</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>0</td>
<td>2</td>
<td>0.30</td>
<td>B</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>5</td>
<td>N</td>
<td>0</td>
<td>2</td>
<td>0.60</td>
<td>U</td>
<td>.37</td>
<td>.37</td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>0</td>
<td>2</td>
<td>0.60</td>
<td>B</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>2</td>
<td>0.90</td>
<td>U</td>
<td>.75</td>
<td>.75</td>
</tr>
<tr>
<td>8</td>
<td>N</td>
<td>0</td>
<td>2</td>
<td>0.90</td>
<td>B</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>4</td>
<td>0.30</td>
<td>U</td>
<td>.43</td>
<td>.43</td>
</tr>
<tr>
<td>10</td>
<td>N</td>
<td>0</td>
<td>4</td>
<td>0.30</td>
<td>B</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>11</td>
<td>N</td>
<td>0</td>
<td>4</td>
<td>0.60</td>
<td>U</td>
<td>1.63</td>
<td>1.63</td>
</tr>
<tr>
<td>12</td>
<td>N</td>
<td>0</td>
<td>4</td>
<td>0.60</td>
<td>B</td>
<td>.11</td>
<td>.11</td>
</tr>
<tr>
<td>13</td>
<td>N</td>
<td>0</td>
<td>4</td>
<td>0.90</td>
<td>U</td>
<td>3.39</td>
<td>3.39</td>
</tr>
<tr>
<td>14</td>
<td>N</td>
<td>0</td>
<td>4</td>
<td>0.90</td>
<td>B</td>
<td>.33</td>
<td>.33</td>
</tr>
<tr>
<td>15</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>0.30</td>
<td>U</td>
<td>.88</td>
<td>.88</td>
</tr>
<tr>
<td>16</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>0.30</td>
<td>B</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>17</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>0.60</td>
<td>U</td>
<td>3.37</td>
<td>3.37</td>
</tr>
<tr>
<td>18</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>0.60</td>
<td>B</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>19</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>0.90</td>
<td>U</td>
<td>7.10</td>
<td>7.10</td>
</tr>
<tr>
<td>20</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>0.90</td>
<td>B</td>
<td>.32</td>
<td>.32</td>
</tr>
<tr>
<td>21</td>
<td>N</td>
<td>-1 SD</td>
<td>2</td>
<td>0.30</td>
<td>U</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>22</td>
<td>N</td>
<td>-1 SD</td>
<td>2</td>
<td>0.30</td>
<td>B</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>23</td>
<td>N</td>
<td>-1 SD</td>
<td>2</td>
<td>0.60</td>
<td>U</td>
<td>.17</td>
<td>.17</td>
</tr>
<tr>
<td>24</td>
<td>N</td>
<td>-1 SD</td>
<td>2</td>
<td>0.60</td>
<td>B</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>25</td>
<td>N</td>
<td>-1 SD</td>
<td>2</td>
<td>0.90</td>
<td>U</td>
<td>.32</td>
<td>.32</td>
</tr>
<tr>
<td>26</td>
<td>N</td>
<td>-1 SD</td>
<td>2</td>
<td>0.90</td>
<td>B</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>27</td>
<td>N</td>
<td>-1 SD</td>
<td>4</td>
<td>0.30</td>
<td>U</td>
<td>.25</td>
<td>.25</td>
</tr>
<tr>
<td>28</td>
<td>N</td>
<td>-1 SD</td>
<td>4</td>
<td>0.30</td>
<td>B</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>29</td>
<td>N</td>
<td>-1 SD</td>
<td>4</td>
<td>0.60</td>
<td>U</td>
<td>.86</td>
<td>.86</td>
</tr>
<tr>
<td>30</td>
<td>N</td>
<td>-1 SD</td>
<td>4</td>
<td>0.60</td>
<td>B</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>31</td>
<td>N</td>
<td>-1 SD</td>
<td>4</td>
<td>0.90</td>
<td>U</td>
<td>1.65</td>
<td>1.65</td>
</tr>
<tr>
<td>32</td>
<td>N</td>
<td>-1 SD</td>
<td>4</td>
<td>0.90</td>
<td>B</td>
<td>.26</td>
<td>.26</td>
</tr>
<tr>
<td>33</td>
<td>N</td>
<td>-1 SD</td>
<td>6</td>
<td>0.30</td>
<td>U</td>
<td>.55</td>
<td>.55</td>
</tr>
<tr>
<td>34</td>
<td>N</td>
<td>-1 SD</td>
<td>6</td>
<td>0.30</td>
<td>B</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>35</td>
<td>N</td>
<td>-1 SD</td>
<td>6</td>
<td>0.60</td>
<td>U</td>
<td>1.94</td>
<td>1.94</td>
</tr>
<tr>
<td>36</td>
<td>N</td>
<td>-1 SD</td>
<td>6</td>
<td>0.60</td>
<td>B</td>
<td>.07</td>
<td>.07</td>
</tr>
<tr>
<td>37</td>
<td>N</td>
<td>-1 SD</td>
<td>6</td>
<td>0.90</td>
<td>U</td>
<td>3.79</td>
<td>3.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>N</td>
<td>-1 SD</td>
<td>6</td>
<td>0.90</td>
<td>B</td>
<td>.27</td>
<td>.27</td>
</tr>
<tr>
<td>39</td>
<td>Y</td>
<td>0</td>
<td>2</td>
<td>0.30</td>
<td>U</td>
<td>0</td>
<td>.10</td>
</tr>
<tr>
<td>40</td>
<td>Y</td>
<td>0</td>
<td>2</td>
<td>0.30</td>
<td>B</td>
<td>0</td>
<td>.00</td>
</tr>
<tr>
<td>41</td>
<td>Y</td>
<td>0</td>
<td>2</td>
<td>0.60</td>
<td>U</td>
<td>0</td>
<td>.82</td>
</tr>
<tr>
<td>42</td>
<td>Y</td>
<td>0</td>
<td>2</td>
<td>0.60</td>
<td>B</td>
<td>0</td>
<td>.03</td>
</tr>
<tr>
<td>43</td>
<td>Y</td>
<td>0</td>
<td>2</td>
<td>0.90</td>
<td>U</td>
<td>0</td>
<td>1.65</td>
</tr>
<tr>
<td>44</td>
<td>Y</td>
<td>0</td>
<td>2</td>
<td>0.90</td>
<td>B</td>
<td>0</td>
<td>.15</td>
</tr>
<tr>
<td>45</td>
<td>Y</td>
<td>0</td>
<td>4</td>
<td>0.30</td>
<td>U</td>
<td>0</td>
<td>.61</td>
</tr>
<tr>
<td>46</td>
<td>Y</td>
<td>0</td>
<td>4</td>
<td>0.30</td>
<td>B</td>
<td>0</td>
<td>.01</td>
</tr>
<tr>
<td>47</td>
<td>Y</td>
<td>0</td>
<td>4</td>
<td>0.60</td>
<td>U</td>
<td>0</td>
<td>1.63</td>
</tr>
<tr>
<td>48</td>
<td>Y</td>
<td>0</td>
<td>4</td>
<td>0.60</td>
<td>B</td>
<td>0</td>
<td>.00</td>
</tr>
<tr>
<td>49</td>
<td>Y</td>
<td>0</td>
<td>4</td>
<td>0.90</td>
<td>U</td>
<td>0</td>
<td>4.67</td>
</tr>
<tr>
<td>50</td>
<td>Y</td>
<td>0</td>
<td>4</td>
<td>0.90</td>
<td>B</td>
<td>0</td>
<td>.06</td>
</tr>
<tr>
<td>51</td>
<td>Y</td>
<td>0</td>
<td>6</td>
<td>0.30</td>
<td>U</td>
<td>0</td>
<td>.88</td>
</tr>
<tr>
<td>52</td>
<td>Y</td>
<td>0</td>
<td>6</td>
<td>0.30</td>
<td>B</td>
<td>0</td>
<td>.01</td>
</tr>
<tr>
<td>53</td>
<td>Y</td>
<td>0</td>
<td>6</td>
<td>0.60</td>
<td>U</td>
<td>0</td>
<td>3.37</td>
</tr>
<tr>
<td>54</td>
<td>Y</td>
<td>0</td>
<td>6</td>
<td>0.60</td>
<td>B</td>
<td>0</td>
<td>.08</td>
</tr>
<tr>
<td>55</td>
<td>Y</td>
<td>0</td>
<td>6</td>
<td>0.90</td>
<td>U</td>
<td>0</td>
<td>7.10</td>
</tr>
<tr>
<td>56</td>
<td>Y</td>
<td>0</td>
<td>6</td>
<td>0.90</td>
<td>B</td>
<td>0</td>
<td>.32</td>
</tr>
<tr>
<td>57</td>
<td>Y</td>
<td>-1 SD</td>
<td>2</td>
<td>0.30</td>
<td>U</td>
<td>0</td>
<td>.09</td>
</tr>
<tr>
<td>58</td>
<td>Y</td>
<td>-1 SD</td>
<td>2</td>
<td>0.30</td>
<td>B</td>
<td>0</td>
<td>.00</td>
</tr>
<tr>
<td>59</td>
<td>Y</td>
<td>-1 SD</td>
<td>2</td>
<td>0.60</td>
<td>U</td>
<td>0</td>
<td>.35</td>
</tr>
<tr>
<td>60</td>
<td>Y</td>
<td>-1 SD</td>
<td>2</td>
<td>0.60</td>
<td>B</td>
<td>0</td>
<td>.01</td>
</tr>
<tr>
<td>61</td>
<td>Y</td>
<td>-1 SD</td>
<td>2</td>
<td>0.90</td>
<td>U</td>
<td>0</td>
<td>.71</td>
</tr>
<tr>
<td>62</td>
<td>Y</td>
<td>-1 SD</td>
<td>2</td>
<td>0.90</td>
<td>B</td>
<td>0</td>
<td>.05</td>
</tr>
<tr>
<td>63</td>
<td>Y</td>
<td>-1 SD</td>
<td>4</td>
<td>0.30</td>
<td>U</td>
<td>0</td>
<td>.48</td>
</tr>
<tr>
<td>64</td>
<td>Y</td>
<td>-1 SD</td>
<td>4</td>
<td>0.30</td>
<td>B</td>
<td>0</td>
<td>.02</td>
</tr>
<tr>
<td>65</td>
<td>Y</td>
<td>-1 SD</td>
<td>4</td>
<td>0.60</td>
<td>U</td>
<td>0</td>
<td>1.82</td>
</tr>
<tr>
<td>66</td>
<td>Y</td>
<td>-1 SD</td>
<td>4</td>
<td>0.60</td>
<td>B</td>
<td>0</td>
<td>.14</td>
</tr>
<tr>
<td>67</td>
<td>Y</td>
<td>-1 SD</td>
<td>4</td>
<td>0.90</td>
<td>U</td>
<td>0</td>
<td>3.82</td>
</tr>
<tr>
<td>68</td>
<td>Y</td>
<td>-1 SD</td>
<td>4</td>
<td>0.90</td>
<td>B</td>
<td>0</td>
<td>.40</td>
</tr>
<tr>
<td>69</td>
<td>Y</td>
<td>-1 SD</td>
<td>6</td>
<td>0.30</td>
<td>U</td>
<td>0</td>
<td>.99</td>
</tr>
<tr>
<td>70</td>
<td>Y</td>
<td>-1 SD</td>
<td>6</td>
<td>0.30</td>
<td>B</td>
<td>0</td>
<td>.02</td>
</tr>
<tr>
<td>71</td>
<td>Y</td>
<td>-1 SD</td>
<td>6</td>
<td>0.60</td>
<td>U</td>
<td>0</td>
<td>3.72</td>
</tr>
<tr>
<td>72</td>
<td>Y</td>
<td>-1 SD</td>
<td>6</td>
<td>0.60</td>
<td>B</td>
<td>0</td>
<td>.14</td>
</tr>
<tr>
<td>73</td>
<td>Y</td>
<td>-1 SD</td>
<td>6</td>
<td>0.90</td>
<td>U</td>
<td>0</td>
<td>7.72</td>
</tr>
<tr>
<td>74</td>
<td>Y</td>
<td>-1 SD</td>
<td>6</td>
<td>0.90</td>
<td>B</td>
<td>0</td>
<td>.51</td>
</tr>
</tbody>
</table>
Table 3

*Observed Anchor Item Parameters on a 2009 Statewide Assessment*

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.91</td>
<td>0.68</td>
<td>0.23</td>
</tr>
<tr>
<td>2</td>
<td>0.79</td>
<td>-1.53</td>
<td>0.11</td>
</tr>
<tr>
<td>3</td>
<td>0.61</td>
<td>-1.21</td>
<td>0.20</td>
</tr>
<tr>
<td>4</td>
<td>0.75</td>
<td>1.19</td>
<td>0.16</td>
</tr>
<tr>
<td>5</td>
<td>0.94</td>
<td>-0.58</td>
<td>0.20</td>
</tr>
<tr>
<td>6</td>
<td>0.92</td>
<td>-0.44</td>
<td>0.21</td>
</tr>
<tr>
<td>7</td>
<td>1.30</td>
<td>0.49</td>
<td>0.20</td>
</tr>
<tr>
<td>8</td>
<td>0.90</td>
<td>-0.97</td>
<td>0.20</td>
</tr>
<tr>
<td>9</td>
<td>0.55</td>
<td>-0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>0.68</td>
<td>-0.62</td>
<td>0.24</td>
</tr>
<tr>
<td>11</td>
<td>0.76</td>
<td>-1.39</td>
<td>0.10</td>
</tr>
<tr>
<td>12</td>
<td>1.06</td>
<td>-0.14</td>
<td>0.28</td>
</tr>
<tr>
<td>13</td>
<td>0.59</td>
<td>-0.04</td>
<td>0.31</td>
</tr>
<tr>
<td>14</td>
<td>0.99</td>
<td>-0.39</td>
<td>0.21</td>
</tr>
<tr>
<td>15</td>
<td>1.15</td>
<td>-0.98</td>
<td>0.25</td>
</tr>
<tr>
<td>16</td>
<td>0.99</td>
<td>0.39</td>
<td>0.06</td>
</tr>
<tr>
<td>17</td>
<td>1.28</td>
<td>0.52</td>
<td>0.17</td>
</tr>
<tr>
<td>18</td>
<td>0.99</td>
<td>-0.87</td>
<td>0.35</td>
</tr>
<tr>
<td>19</td>
<td>0.88</td>
<td>-0.64</td>
<td>0.31</td>
</tr>
<tr>
<td>20</td>
<td>1.36</td>
<td>0.21</td>
<td>0.18</td>
</tr>
<tr>
<td>21</td>
<td>0.78</td>
<td>0.30</td>
<td>0.24</td>
</tr>
<tr>
<td>22</td>
<td>0.90</td>
<td>0.34</td>
<td>0.18</td>
</tr>
<tr>
<td>23</td>
<td>0.81</td>
<td>0.31</td>
<td>0.43</td>
</tr>
<tr>
<td>24</td>
<td>1.20</td>
<td>-1.05</td>
<td>0.29</td>
</tr>
<tr>
<td>25</td>
<td>0.72</td>
<td>-0.54</td>
<td>0.13</td>
</tr>
<tr>
<td>26</td>
<td>0.64</td>
<td>-0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>27</td>
<td>0.66</td>
<td>-0.41</td>
<td>0.29</td>
</tr>
<tr>
<td>28</td>
<td>0.75</td>
<td>0.13</td>
<td>0.38</td>
</tr>
<tr>
<td>29</td>
<td>0.70</td>
<td>0.23</td>
<td>0.34</td>
</tr>
<tr>
<td>30</td>
<td>0.57</td>
<td>-0.41</td>
<td>0.17</td>
</tr>
<tr>
<td>31</td>
<td>0.57</td>
<td>0.19</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Table 4

*Pearson’s Correlations Assessing the Relationship between Form 1 Anchor DTF and Unconditional Equating Dependence*

<table>
<thead>
<tr>
<th>Conditions with Equivalent True DIF Across Forms</th>
<th>Form 1 DTF</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Remsd</td>
<td>$r$</td>
<td>0.28</td>
<td>0.08</td>
</tr>
<tr>
<td>$RESD_R$</td>
<td>$r^2$</td>
<td>0.25</td>
<td>0.06</td>
</tr>
<tr>
<td>$RESD_F$</td>
<td>$r^2$</td>
<td>0.26</td>
<td>0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conditions with True DIF in Form 1 Only</th>
<th>Form 1 DTF</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Remsd</td>
<td>$r$</td>
<td>0.80**</td>
<td>0.64**</td>
</tr>
<tr>
<td>$RESD_R$</td>
<td>$r^2$</td>
<td>0.82**</td>
<td>0.67**</td>
</tr>
<tr>
<td>$RESD_F$</td>
<td>$r^2$</td>
<td>0.76**</td>
<td>0.58**</td>
</tr>
</tbody>
</table>

*Correlation is significant at the $p < .05$ level

**Correlation is significant at the $p < .001$ level
Figure 1. Graphical Depiction of the Non-Equivalent Anchor Test Equating Design
Figure 2. Conditional Equating Invariance Results for Condition 1 (Null Condition)
Figure 3. Conditional Equating Invariance Results for Condition 2 (Null Condition)
Figure 4. RMSD(x) Results for Conditions with No Differential Form DIF, No Group Mean Differences in Ability, and Unidirectional DIF
<table>
<thead>
<tr>
<th>Frequency</th>
<th>Condition 4</th>
<th>Condition 6</th>
<th>Condition 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Magnitude = 0.30</td>
<td>Magnitude = 0.60</td>
<td>Magnitude = 0.90</td>
</tr>
<tr>
<td>Frequency</td>
<td>Condition 10</td>
<td>Condition 12</td>
<td>Condition 14</td>
</tr>
<tr>
<td>4</td>
<td>Magnitude = 0.30</td>
<td>Magnitude = 0.60</td>
<td>Magnitude = 0.90</td>
</tr>
<tr>
<td>Frequency</td>
<td>Condition 16</td>
<td>Condition 18</td>
<td>Condition 20</td>
</tr>
<tr>
<td>6</td>
<td>Magnitude = 0.30</td>
<td>Magnitude = 0.60</td>
<td>Magnitude = 0.90</td>
</tr>
</tbody>
</table>

Figure 5. RMSD(x) Results for Conditions with No Differential Form DIF, No Group Mean Differences in Ability, and Bidirectional DIF
Figure 6. RMSD(x) Results for Conditions with No Differential Form DIF, Group Mean Differences in Ability, and Unidirectional DIF
Figure 7. RMSD(x) Results for Conditions with No Differential Form DIF, Group Mean Differences in Ability, and Bidirectional DIF
Figure 8. RMSD(x) Results for Conditions with Differential Form DIF, No Group Mean Differences in Ability, and Unidirectional DIF
<table>
<thead>
<tr>
<th>Condition</th>
<th>Frequency</th>
<th>Magnitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>2</td>
<td>0.30</td>
</tr>
<tr>
<td>42</td>
<td>4</td>
<td>0.60</td>
</tr>
<tr>
<td>44</td>
<td>4</td>
<td>0.90</td>
</tr>
<tr>
<td>46</td>
<td>4</td>
<td>0.90</td>
</tr>
<tr>
<td>48</td>
<td>6</td>
<td>0.90</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>6</td>
<td>0.90</td>
</tr>
<tr>
<td>54</td>
<td>6</td>
<td>0.90</td>
</tr>
<tr>
<td>56</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 9. RMSD(x) Results for Conditions with Differential Form DIF, No Group Mean Differences in Ability, and Bidirectional DIF*
Figure 10. RMSD(x) Results for Conditions with Differential Form DIF, Group Mean Differences in Ability, and Unidirectional DIF
Figure 11. RMSD(x) Results for Conditions with Differential Form DIF, Group Mean Differences in Ability, and Bidirectional DIF
Condition 55
Differential Form DIF, No Mean Ability Differences,
Magnitude = .9, Frequency = 6, Unidirectional

Condition 73
Differential Form DIF, Mean Ability Differences,
Magnitude = .9, Frequency = 6, Unidirectional

Figure 12. Group-to-overall Conditional Equating Dependence Results for Selected Conditions
Figure 13. Graphical Displays of REMSD Results: Interactions of Differential Form DIF with all other Factors in the Study
Figure 14. Graphical Displays of RESD$_R$ Results: Interactions of Differential Form DIF with all other Factors in the Study
Figure 15. Graphical Displays of RESD$_F$ Results: Interactions of Differential Form DIF with all other Factors in the Study
Figure 16. Scatter Plots Displaying Relationships between Unconditional Equating Dependence and True DTF in Form 1 Only
Appendix: R Syntax

# Part 1: Data Generation

# Building Forms 0 and 1

# The remainder of this syntax is condition 3
# (i.e., frequency 2, magnitude .3, unidirectional, no group mean differences)

# Sampling anchor items

beta.anchor.ref <- rnorm(31, mean = -0.24677, sd = 0.65402)
alpha.anchor.ref <- runif(31, min = .55, max = 1.36)
gamma.anchor.ref <- runif(31, min = .06, max = .43)

beta.anchor.foc <- beta.anchor.ref
alpha.anchor.foc <- alpha.anchor.ref
gamma.anchor.foc <- gamma.anchor.ref

# Manipulating DIF according to condition
# (DIF was introduced here according to condition)
# (For conditions with differential form DIF, this portion of
# syntax comes after the creation of Form 0)


# Creating Form 0

beta0ref <- c(beta.anchor.ref[1:10], rnorm(40, mean = -0.24677, sd = 0.65402))
alpha0ref <- c(alpha.anchor.ref[1:10], runif(40, min = .55, max = 1.36))
gamma0ref <- c(gamma.anchor.ref[1:10], runif(40, min = .06, max = .43))

beta0foc <- c(beta.anchor.foc[1:10], beta0ref[11:50])
alpha0foc <- c(alpha.anchor.foc[1:10], alpha0ref[11:50])
gamma0foc <- c(gamma.anchor.foc[1:10], gamma0ref[11:50])

# Creating Form 1

beta1ref <- c(beta.anchor.ref[1:10], rnorm(40, mean = -0.24677, sd = 0.65402))
alpha1ref <- c(alpha.anchor.ref[1:10], runif(40, min = .55, max = 1.36))
gamma1ref <- c(gamma.anchor.ref[1:10], runif(40, min = .06, max = .43))
beta1foc <- c(beta.anchor.foc[1:10], beta1ref[11:50])
alpha1foc <- c(alpha.anchor.foc[1:10], alpha1ref[11:50])
gamma1foc <- c(gamma.anchor.foc[1:10], gamma1ref[11:50])

#Creating data frames for Form 0 and Form 1

itemparamF0ref <- data.frame("item"=c(1:50),
    "b"=beta0ref,
    "a"=alpha0ref,
    "c"=gamma0ref,
    "type"=c(rep("A",10),rep("C",40)))

itemparamF0foc <- data.frame("item"=c(1:50),
    "b"=beta0foc,
    "a"=alpha0foc,
    "c"=gamma0foc,
    "type"=c(rep("A",10),rep("C",40)))

itemparamF1ref <- data.frame("item"=c(1:10,51:90),
    "b"=beta1ref,
    "a"=alpha1ref,
    "c"=gamma1ref,
    "type"=c(rep("A",10),rep("C",40)))

itemparamF1foc <- data.frame("item"=c(1:10,51:90),
    "b"=beta1foc,
    "a"=alpha1foc,
    "c"=gamma1foc,
    "type"=c(rep("A",10),rep("C",40)))

#Saving true item parameters

write.csv(itemparamF0ref,
    file="C:/Users/corinne/Desktop/data_condition_3/
    itemparamF0ref.csv", row.names=FALSE)
write.csv(itemparamF0foc,
    file="C:/Users/corinne/Desktop/data_condition_3/
    itemparamF0foc.csv", row.names=FALSE)
write.csv(itemparamF1ref,
    file="C:/Users/corinne/Desktop/data_condition_3/
    itemparamF1ref.csv", row.names=FALSE)
write.csv(itemparamF1foc,
    file="C:/Users/corinne/Desktop/data_condition_3/
    itemparamF1foc.csv", row.names=FALSE)
#----------------------------------------#
# Simulating item response strings #
#----------------------------------------#

#Creating a loop of 100 trials

for(y in 1:100) {

#Sampling latent traits
#(mean differences were manipulated here according to condition)

thetaF0ref <- rnorm(3000, mean=0, sd=1)
thetaF0ref <- sort(thetaF0ref)
thetaF0foc <- rmom(1500, mean=0, sd=1)
thetaF0foc <- sort (thetaF0foc)
thetaF1ref <- rnorm(3000, mean=0, sd=1)
thetaF1ref <- sort(thetaF1ref)
thetaF1foc <- rmom(1500, mean=0, sd=1)
thetaF1foc <- sort (thetaF1foc)

#Creating empty matrices to hold future item responses

dataF0ref <- c()
dataF0foc <- c()
dataF1ref <- c()
dataF1foc <- c()

#Defining 3PL function

pix <- function(iparam, thp, i){
iparam[i,4]+(1-iparam[i,4])*(exp(1.7*iparam[i,3]*
    (thp-iparam[i,2]))/(1+exp(1.7*iparam[i,3]*(thp-iparam[i,2]))))
}

#Generating item responses

for(j in 1:50) {
dataF0ref <- cbind(dataF0ref,
    ifelse(pix(itemparamF0ref,thetaF0ref,j) >= runif(length(thetaF0ref)),1,0))
}
for(j in 1:50) {
dataF0foc <- cbind(dataF0foc,
    ifelse(pix(itemparamF0foc,thetaF0foc,j) >= runif(length(thetaF0foc)),1,0))
}
for(j in 1:50) {
  dataF1ref <- cbind(dataF1ref,
  ifelse(pix(itemparamF1ref,thetaF1ref,j) >= runif(length(thetaF1ref)),1,0))
}

for(j in 1:50) {
  dataF1foc <- cbind(dataF1foc,
  ifelse(pix(itemparamF1foc,thetaF1foc,j) >= runif(length(thetaF1foc)),1,0))
}

#Saving data for overall population

F0data1 <- rbind(dataF0ref,dataF0foc)
F0data <- cbind(c(10001:14500),c(rep("0",3000),rep("1",1500)),F0data1)
F1data1 <- rbind(dataF1ref,dataF1foc)
F1data <- cbind(c(14501:19000),c(rep("0",3000),rep("1",1500)),F1data1)

nextdataname0 <- paste("F0data","y",sep="")
nextdataname1 <- paste("F1data","y",sep="")

write.table(F0data,
  file=paste("C:/Users/corinne/Desktop/data_condition_3/",nextdataname0,
  "overall.dat",sep=""),col.names=FALSE,row.names=FALSE,quote=FALSE,
  sep="")
write.table(F1data,
  file=paste("C:/Users/corinne/Desktop/data_condition_3/",nextdataname1,
  "overall.dat",sep=""),col.names=FALSE,row.names=FALSE,quote=FALSE,
  sep="")

#Saving data for reference group

F0dataref <- cbind(c(10001:13000),c(rep("0",3000)),dataF0ref)
F1dataref <- cbind(c(14501:17500),c(rep("0",3000)),dataF1ref)

nextdatnameref0 <- paste("F0data","y",sep="")
nextdatnameref1 <- paste("F1data","y",sep="")

write.table(F0dataref,
  file=paste("C:/Users/corinne/Desktop/data_condition_3/",nextdatnameref0,
  "ref.dat",sep=""),col.names=FALSE,row.names=FALSE,quote=FALSE,
  sep="")
write.table(F1dataref,
  file=paste("C:/Users/corinne/Desktop/data_condition_3/",nextdatnameref1,
  "ref.dat",sep=""),col.names=FALSE,row.names=FALSE,quote=FALSE,
  sep="")
#Saving data for focal group

F0datafoc <- cbind(c(13001:14500),c(rep("1",1500)),dataF0foc)
F1datafoc <- cbind(c(17501:19000),c(rep("1",1500)),dataF1foc)

nextdatanamefoc0 <- paste("F0data",y,sep="")
nextdatanamefoc1 <- paste("F1data",y,sep="")

write.table(F0datafoc,
    file=paste("C:/Users/corinne/Desktop/data_condition_3/",nextdatanamefoc0,
    "foc.dat",sep=""),col.names=FALSE,row.names=FALSE,quote=FALSE, sep="")
write.table(F1datafoc,
    file=paste("C:/Users/corinne/Desktop/data_condition_3/",nextdatanamefoc1,
    "foc.dat",sep=""),col.names=FALSE,row.names=FALSE,quote=FALSE, sep="")

#Saving thetas

thetaF0refdf <- data.frame("theta"=thetaF0ref)
thetaF0focdf <- data.frame("theta"=thetaF0foc)
thetaF1refdf <- data.frame("theta"=thetaF1ref)
thetaF1focdf <- data.frame("theta"=thetaF1foc)

F0theta1 <- rbind(thetaF0refdf,thetaF0focdf)
F0theta <- cbind("ID"=c(10001:14500),
    "group"=c(rep("0",3000),rep("1",1500)),F0theta1)
F1theta1 <- rbind(thetaF1refdf,thetaF1focdf)
F1theta <- cbind("ID"=c(14501:19000),
    "group"=c(rep("0",3000),rep("1",1500)),F1theta1)

nextthetanameF0 <- paste("F0theta",y,sep="")
nextthetanameF1 <- paste("F1theta",y,sep="")

write.csv(F0theta,
    file=paste("C:/Users/corinne/Desktop/data_condition_3/",nextthetanameF0,
    ".csv",sep=""),row.names=FALSE)
write.csv(F1theta,
    file=paste("C:/Users/corinne/Desktop/data_condition_3/",nextthetanameF1,
    ".csv",sep=""),row.names=FALSE)

#Closing loop

#--------------------------#
# Deriving True DTF   #
#--------------------------#

#Preparing item parameter files

anchoritemparamF0ref <- itemparamF0ref[1:10,1:4]
anchoritemparamF0foc <- itemparamF0foc[1:10,1:4]
anchoritemparamF1ref <- itemparamF1ref[1:10,1:4]
anchoritemparamF1foc <- itemparamF1foc[1:10,1:4]

#Creating quadrature points along theta continuum

thetaq <- c(-3.00, -2.85, -2.70, -2.55, -2.40, -2.25, -2.10, -1.95, -1.80,  
           -1.65, -1.50, -1.35, -1.20, -1.05, -0.90, -0.75, -0.60, -0.45,  
           -0.30, -0.15, 0.00, 0.15, 0.30, 0.45, 0.60, 0.75, 0.90, 1.05,  
           1.20, 1.35, 1.50, 1.65, 1.80, 1.95, 2.10, 2.25, 2.40, 2.55,  
           2.70, 2.85, 3.00)
quadpoints <- data.frame(point = c(1:41), thetaq = c(-3.00, -2.85, -2.70, -2.55,  
                                              -2.40, -2.25, -2.10, -1.95, -1.80, -1.65, -1.50, -1.35, -1.20,  
                                              -1.05, -0.90, -0.75, -0.60, -0.45, -0.30, -0.15, 0.00, 0.15,  
                                              0.30, 0.45, 0.60, 0.75, 0.90, 1.05, 1.20, 1.35, 1.50, 1.65, 1.80,  
                                              1.95, 2.10, 2.25, 2.40, 2.55, 2.70, 2.85, 3.00))

#Defining true score function

tsf <- function(thp, iparam) {
  sum(iparam[,4]+((1-iparam[,4]) /  
               (1 + exp(-1.7*iparam[,3]*(thp-iparam[,2]))))  
       )
}

#Calculating form 0 true scores

truesscoresF0refX <- sapply(quadpoints[,2], function(X)  
                           tsf(thp=X, iparam=anchoritemparamF0ref))
truesscoresF0focX <- sapply(quadpoints[,2], function(X)  
                           tsf(thp=X, iparam=anchoritemparamF0foc))

truesscoresF0ref <- data.frame(thetaq = quadpoints[,2], truescore=truesscoresF0refX)
truesscoresF0foc <- data.frame(thetaq = quadpoints[,2], truescore=truesscoresF0focX)
#Defining the proportion the focal population at each quadrature point
#(mean differences were accounted for here according to condition)

propF0focscores <- dnorm(thetaq, mean=0, sd=1)

#Calculating form 0 DTF

RajuDtest <- data.frame(D=truescoresF0foc[,2]-truescoresF0ref[,2])
RajuDtestsquaredF0 <- data.frame(thetaq=c(quadpoints[,2]),
   Dsquared = RajuDtest[,1]^2, propfoc=propF0focscores)
F0DTFX <- data.frame(Dtimesprop = RajuDtestsquaredF0[,2]*RajuDtestsquaredF0[,3])
anchorF0DTF <- sum(F0DTFX[,1])

#Calculating form 1 true scores

truescoresF1refX <- sapply(quadpoints[,2], function(X)
   {tsf(thp=X, iparam=anchoritemparamF1ref)})
truescoresF1focX <- sapply(quadpoints[,2], function(X)
   {tsf(thp=X, iparam=anchoritemparamF1foc)})

truescoresF1ref <- data.frame(thetaq = quadpoints[,2], truescore=truescoresF1refX)
truescoresF1foc <- data.frame(thetaq = quadpoints[,2], truescore=truescoresF1focX)

#Calculating form 1 DTF

RajuDtest <- data.frame(D=truescoresF1foc[,2]-truescoresF1ref[,2])
RajuDtestsquaredF1 <- data.frame(thetaq=c(quadpoints[,2]),
   Dsquared = RajuDtest[,1]^2, propfoc=propF1focscores)
F1DTFX <- data.frame(Dtimesprop = RajuDtestsquaredF1[,2]*RajuDtestsquaredF1[,3])
anchorF1DTF <- sum(F1DTFX[,1])

#Saving DTF Results

anchorDTF <- data.frame(Form0 = anchorF0DTF, Form1 = anchorF1DTF)

write.csv(anchorDTF, "C:/Users/corinne/Desktop/100_trial_complete/data_condition_3/
anchortrueDTF.csv", row.names=FALSE, quote=FALSE)

############################
#End Part 1     #
############################

#Setting directory routing path
#(folder B holds data from Part 1 according to condition)

setwd("C:/Users/corinne/Desktop/B/")
bpath <- "C:/Users/corinne/Desktop/B/
datpath <- "C:/Users/corinne/Desktop/B/"

#Creating loops

formcodes <- c("F0","F1")
trialcodes <- c("1","2","3","4","5","6","7","8","9","10",
                "11","12","13","14","15","16","17","18","19","20",
                "21","22","23","24","25","26","27","28","29","30",
                "31","32","33","34","35","36","37","38","39","40",
                "41","42","43","44","45","46","47","48","49","50",
                "51","52","53","54","55","56","57","58","59","60",
                "61","62","63","64","65","66","67","68","69","70",
                "71","72","73","74","75","76","77","78","79","80",
                "81","82","83","84","85","86","87","88","89","90",
                "91","92","93","94","95","96","97","98","99","100")

#Opening trial loop

for(tid in 1:length(trialcodes)) {

    #-------------------------------#
    # BILOG Calibration #
    #-------------------------------#

    #Opening form loop for overall population

    for(fid in 1:length(formcodes)) {

        #Creating BILOG syntax for overall population

        write.table(paste("Calibration for Overall Population",formcodes[fid],sep=""),
                    paste(bpath,formcodes[fid],"dat.BLM",sep=""),
                    append=FALSE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste("","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">COMMENT  Calibration of item parameters;","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">GLOBAL  DFNAME='",datpath,formcodes[fid],".data",trialcodes[tid],
        "overall.dat',NPARM=3,SAVE;","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">SAVE    PARM='",bpath,formcodes[fid],".data1overallcalib.PAR',"",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">LENGTH  NITEMS=50;","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">INPUT   NTOTAL=50, NALT=5, NIDCHAR=5;","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">ITEMS   INAMES=(i01(1)i50);","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">TEST1   TNAME='", formcodes[fid],".Form', INUMBER=(1(1)50);","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste("(5A1,T7,50A1)","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">CALIB   NQPT=31, CYCLES=25, NEWTON=10, CRIT=0.001;","",sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),
        append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)
write.table(paste(">
SCORE NOPRINT, RSCTYPE=4, INFO=2, POP;
",sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
append=TRUE, row.names=FALSE, col.names=FALSE, quote=FALSE)

#Running BILOG Calibration for overall population
system(paste("C:/Users/corinne/Desktop/BILOGMG/blm1",
paste(bpath,formcodes[fid],"dat",sep="")))
ystem(paste("C:/Users/corinne/Desktop/BILOGMG/blm2",
paste(bpath,formcodes[fid],"dat",sep="")))
system(paste("C:/Users/corinne/Desktop/BILOGMG/blm3",
paste(bpath,formcodes[fid],"dat",sep="")))

#Cleaning and formatting files for overall population
file.remove("IF.DAT");
file.remove("bmgicc_file")
file.remove("bmginfo_file")
file.remove("bmgscore_file")
file.remove("bmgtotinfo_file")
file.remove(paste(formcodes[fid],"dat.PH1",sep=""))
file.remove(paste(formcodes[fid],"dat.PH2",sep=""))
file.remove(paste(formcodes[fid],"dat.PH3",sep=""))
file.remove(paste(formcodes[fid],"dat.BLM",sep=""))

tempcalib <-
read.fwf(paste(formcodes[fid],"DATA1OVERALLCALIB.PAR",sep=""),
skip = 4, n=50, widths=c(-36,10,-10,10,-30,10))

assign(paste(formcodes[fid],"overallcalib1",sep=""),
data.frame(
  Item=c(1:50),
  b=tempcalib[,2],
  a=tempcalib[,1],
  c=tempcalib[,3],
  Use=c(rep("A",10),rep("C",40))))

rm(tempcalib)
file.remove(paste(formcodes[fid],"DATA1OVERALLCALIB.PAR",sep=""))

assign(paste(formcodes[fid],"overallscores1",sep=""),
read.fwf(paste(formcodes[fid],"SCORES1OVERALL.SCO",sep=""),
skip = 2, n=4500, buffersize=2,
widths=list(c(-5,6),c(-25,10,12,12)))
file.remove(paste(formcodes[fid],"SCORES1OVERALL.SCO",sep=""))

#Closing form loop for overall population
#
#
#Saving calibrated item parameters for overall population
write.csv(F0overallcalib1,  
    file=paste("F0overallcalib",trialcodes[tid],".csv",sep=""),  
    row.names=FALSE, quote = FALSE)
write.csv(F1overallcalib1,  
    file=paste("F1overallcalib",trialcodes[tid],".csv",sep=""),  
    row.names=FALSE, quote = FALSE)
write.csv(F0overallscores1,  
    file=paste("F0overallscores",trialcodes[tid],".csv",sep=""),  
    row.names=FALSE, quote = FALSE)
write.csv(F1overallscores1,  
    file=paste("F1overallscores",trialcodes[tid],".csv",sep=""),  
    row.names=FALSE, quote = FALSE)

#Opening form loop for reference population
for(fid in 1:length(formcodes)) {
    #Creating BILOG syntax for reference population
    write.table(paste(  
        "Calibration of Form for Reference Population",formcodes[fid]  
        ,sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),  
        append=FALSE, row.names=FALSE,col.names=FALSE, quote=FALSE)
    write.table(paste(  
        " "  
        ,sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),  
        append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
    write.table(paste(  
        ">COMMENT   Calibration of item parameters for test form;"  
        ,sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),  
        append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
    write.table(paste(  
        ">GLOBAL  DFNAME=',datpath,formcodes[fid],".dat",trialcodes[tid],  
        "ref.dat',NPARM=3,SAVE;"  
        ,sep=""),paste(bpath,formcodes[fid],".dat.BLM",sep=""),  
        append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  ">SAVE  PARM="bpath.formcodes[fid],"data1refcalib.PAR","  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  "SCORE="bpath.formcodes[fid],"scores1ref.SCO;"  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  ">LENGTH NITEMS=50;"  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  ">INPUT NTOTAL=50, NALT=5, NIDCHAR=5;"  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  ">ITEMS INAMES=(i01(1)i50);"  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  ">TEST1 TNAME="formcodes[fid],"Form', INUMBER=(1(1)50);"  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  "(5A1,T7,50A1)"  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  ">CALIB NQPT=31, CYCLES=25, NEWTON=10, CRIT=0.001;"  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(  ">SCORE NOPRINT, RSCTYPE=4, INFO=2, POP;"  
,sep=""),paste(bpath.formcodes[fid],"dat.BLM",sep=""),  
append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)

#Running BILOG Calibration for reference population

system(paste(""C:/Users/corinne/Desktop/BILOGMG/blm1"",  
paste(bpath.formcodes[fid],"dat",sep=""))))

system(paste(""C:/Users/corinne/Desktop/BILOGMG/blm2"",  
paste(bpath.formcodes[fid],"dat",sep=""))))

system(paste(""C:/Users/corinne/Desktop/BILOGMG/blm3"",  
paste(bpath.formcodes[fid],"dat",sep=""))))
#Cleaning and formatting files for reference population

```r
file.remove("IF.DAT")
file.remove("bmgicc_file")
file.remove("bmginfo_file")
file.remove("bmgscore_file")
file.remove("bmgscore_file")
file.remove(paste(formcodes[fid],"dat.PH1",sep=""))
file.remove(paste(formcodes[fid],"dat.PH2",sep=""))
file.remove(paste(formcodes[fid],"dat.PH3",sep=""))
file.remove(paste(formcodes[fid],"dat.BLM",sep=""))

tempcalib <-
    read.fwf(paste(formcodes[fid],"DATA1REFCALIB.PAR",sep=""),
          skip = 4, n=50,
          widths=c(-36,10,-10,10,-30,10))

assign(paste(formcodes[fid],"refcalib1",sep=""),
       data.frame(
           Item=c(1:50),
           b=tempcalib[,2],
           a=tempcalib[,1],
           c=tempcalib[,3],
           Use=c(rep("A",10),rep("C",40))))

rm(tempcalib)
file.remove(paste(formcodes[fid],"DATA1REFCALIB.PAR",sep=""))

assign(paste(formcodes[fid],"refscores1",sep=""),
       read.fwf(paste(formcodes[fid],"SCORES1REF.SCO",sep=""),
               skip = 2, n=3000, buffersize=2,
               widths=list(c(-5,6),c(-25,10,12,12))))

file.remove(paste(formcodes[fid],"SCORES1REF.SCO",sep=""))

#Closing form loop for reference population
```

#Saving calibrated item parameters for reference population

```r
write.csv(F0refcalib1,
        file=paste("F0refcalib",trialcodes[tid],".csv",sep=""),
        row.names=FALSE, quote = FALSE)
```
write.csv(F1refcalib1,
   file=paste("F1refcalib",trialcodes[tid],".csv",sep=""),
   row.names=FALSE, quote = FALSE)
write.csv(F0refscores1,
   file=paste("F0refscores",trialcodes[tid],".csv",sep=""),
   row.names=FALSE, quote = FALSE)
write.csv(F1refscores1,
   file=paste("F1refscores",trialcodes[tid],".csv",sep=""),
   row.names=FALSE, quote = FALSE)

#Opening form loop for focal population

for(fid in 1:length(formcodes)) {

 #Creating BILOG syntax for focal population
 write.table(paste(
   "Calibration of Form for Focal Population",formcodes[fid]
   ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
   append=FALSE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
   " "
   ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
   append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
   ">COMMENT   Calibration of item parameters for test form;"
   ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
   append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
   ">GLOBAL  DFNAME='",datpath,formcodes[fid],"data",trialcodes[tid],
   "foc.dat',NPARM=3,SAVE;"
   ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
   append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
   ">SAVE    PARM='",bpath,formcodes[fid],"data1foccalib.PAR','"
   ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
   append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
   ">LENGTH  NITEMS=50;"
   ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
   append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
    "">INPUT   NTOTAL=50, NALT=5, NIDCHAR=5;",
    ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
    append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
    ">
ITEMS   INAMES=(i01(1)i50);",
    ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
    append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
    "">TEST1   TNAME='', formcodes[fid],'Form', INUMBER=(1(1)50);",
    ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
    append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
    "(5A1,T7,50A1)
    ",sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
    append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
    ">CALIB   NQPT=31, CYCLES=25, NEWTON=10, CRIT=0.001;",
    ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
    append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)
write.table(paste(
    "">SCORE   NOPRINT, RSCTYPE=4, INFO=2, POP;",
    ,sep=""),paste(bpath,formcodes[fid],"dat.BLM",sep=""),
    append=TRUE, row.names=FALSE,col.names=FALSE, quote=FALSE)

#Running BILOG Calibration for focal population

system(paste("C:/Users/corinne/Desktop/BILOGMG/blm1",
    paste(bpath,formcodes[fid],"dat",sep="")))
system(paste("C:/Users/corinne/Desktop/BILOGMG/blm2",
    paste(bpath,formcodes[fid],"dat",sep="")))
system(paste("C:/Users/corinne/Desktop/BILOGMG/blm3",
    paste(bpath,formcodes[fid],"dat",sep="")))

#Cleaning and formatting files for focal population

file.remove("IF.DAT")
file.remove("bmgcicc_file")
file.remove("bmginfo_file")
file.remove("bmgscore_file")
file.remove("bmgtotinfo_file")
file.remove(paste(formcodes[fid],"dat.PH1",sep=""))
file.remove(paste(formcodes[fid],"dat.PH2",sep=""))
file.remove(paste(formcodes[fid],"dat.PH3",sep=""))
file.remove(paste(formcodes[fid],"dat.BLM",sep=""))
tempcalib <-
    read.fwf(paste(formcodes[fid],"DATA1FOCCALIB.PAR",sep=""),
    skip = 4, n=50,
    widths=c(-36,10,-10,10,-30,10))

assign(paste(formcodes[fid],"foccalib1",sep=""),
    data.frame(
        Item=c(1:50),
        b=tempcalib[,2],
        a=tempcalib[,1],
        c=tempcalib[,3],
        Use=c(rep("A",10),rep("C",40))))

rm(tempcalib)
file.remove(paste(formcodes[fid],"DATA1FOCCALIB.PAR",sep=""))

assign(paste(formcodes[fid],"focscores1",sep=""),
    read.fwf(paste(formcodes[fid],"SCORES1FOC.SCO",sep=""),
    skip = 2, n=1500, buffersize=2,
    widths=list(c(-5,6),c(-25,10,12,12))))

file.remove(paste(formcodes[fid],"SCORES1FOC.SCO",sep=""))

#Closing form loop for focal population

}

#Saving calibrated item parameters for focal population

write.csv(F0foccalib1,
    file=paste("F0foccalib",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote = FALSE)
write.csv(F1foccalib1,
    file=paste("F1foccalib",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote = FALSE)
write.csv(F0focscores1,
    file=paste("F0focscores",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote = FALSE)
write.csv(F1focscores1,
    file=paste("F1focscores",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote = FALSE)
#--------------------------#  
# Mean-Sigma Scaling #  
#--------------------------#

# Scaling item parameters and thetas for overall population

Aoverall <- sd(F0overallcalib1[1:10,2])/sd(F1overallcalib1[1:10,2])
Boverall <- mean(F0overallcalib1[1:10,2])-(Aoverall*mean(F1overallcalib1[1:10,2]))

scaledF1overallcalib1 <- data.frame(
    Item=F1overallcalib1[,1],
    b=(Aoverall*F1overallcalib1[,2])+Boverall,
    a=F1overallcalib1[,3]/Aoverall,
    c=F1overallcalib1[,4],
    Use=F1overallcalib1[,5])

write.csv(scaledF1overallcalib1,
    file=paste("scaledF1overallcalib",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)

scaledF1overallscores1 <- data.frame(
    Item=F1overallscores1[,1],
    ncorrect=F1overallscores1[,2],
    scaledtheta=(Aoverall*F1overallscores1[,3])+Boverall,
    se=F1overallscores1[,4])

write.csv(scaledF1overallscores1,
    file=paste("scaledF1overallscores",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)

# Scaling item parameters and thetas for reference population

Aref <- sd(F0refcalib1[1:10,2])/sd(F1refcalib1[1:10,2])
Bref <- mean(F0refcalib1[1:10,2])-(Aref*mean(F1refcalib1[1:10,2]))

scaledF1refcalib1 <- data.frame(
    Item=F1refcalib1[,1],
    b=(Aref*F1refcalib1[,2])+Bref,
    a=F1refcalib1[,3]/Aref,
    c=F1refcalib1[,4],
    Use=F1refcalib1[,5])

write.csv(scaledF1refcalib1,
    file=paste("scaledF1refcalib",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)
```r
scaledF1refscores1 <- data.frame(
  Item=F1refscores1[,1],
  ncorrect=F1refscores1[,2],
  scaledtheta=(Aref*F1refscores1[,3])+Bref,
  se=F1refscores1[,4])

write.csv(scaledF1refscores1,
  file=paste("scaledF1refscores",trialcodes[tid],".csv",sep=""),
  row.names=FALSE, quote=FALSE)

#Scaling item parameters and thetas for focal population

Afoc <- sd(F0foccalib1[1:10,2])/sd(F1foccalib1[1:10,2])
Bfoc <- mean(F0foccalib1[1:10,2])-(Afoc*mean(F1foccalib1[1:10,2]))

scaledF1foccalib1 <- data.frame(
  Item=F1foccalib1[,1],
  b=(Afoc*F1foccalib1[,2])+Bfoc,
  a=F1foccalib1[,3]/Afoc,
  c=F1foccalib1[,4],
  Use=F1foccalib1[,5])

write.csv(scaledF1foccalib1,
  file=paste("scaledF1foccalib",trialcodes[tid],".csv",sep=""),
  row.names=FALSE, quote=FALSE)

scaledF1focscores1 <- data.frame (
  Item=F1focscores1[,1],
  ncorrect=F1focscores1[,2],
  scaledtheta=(Afoc*F1focscores1[,3])+Bfoc,
  se=F1focscores1[,4])

write.csv(scaledF1focscores1,
  file=paste("scaledF1focscores",trialcodes[tid],".csv",sep=""),
  row.names=FALSE, quote=FALSE)
```

```
```
# IRT True Score Equating #

#Defining true score function

tsf <- function(thp, iparam) {
  sum(
    iparam[c(1:50),4] +
    ( (1-iparam[c(1:50),4]) /
      (1 + exp(-1.7*iparam[c(1:50),3]*
        (thp-iparam[c(1:50),2]))))))
}

#Equating true scores for overall population

F0overallts1 <- sapply(F0overallscores1[,3],
  function(X) {tsf(thp=X, iparam=F0overallcalib1)})
F1overallts1 <- sapply(scaledF1overallscores1[,3],
  function(X) {tsf(thp=X, iparam=scaledF1overallcalib1)})
equatedF1overallts1 <- sapply(scaledF1overallscores1[,3],
  function(X) {tsf(thp=X, iparam=F0overallcalib1)})

F0overalltruescores1 <- data.frame(
  ID = c(10001:14500),
  TS = c(F0overallts1))
F1overalltruescores1 <- data.frame(
  ID = c(14501:19000),
  TS = c(F1overallts1))
F1overalltruescoresequated1 <- data.frame(
  ID = c(14501:19000),
  equatedTS = c(equatedF1overallts1))

write.csv(F0overalltruescores1, 
  file=paste("F0overalltruescores",trialcodes[tid],".csv",sep=""),
  row.names=FALSE, quote=FALSE)
write.csv(F1overalltruescores1, 
  file=paste("F1overalltruescores",trialcodes[tid],".csv",sep=""),
  row.names=FALSE, quote=FALSE)
write.csv(F1overalltruescoresequated1, 
  file=paste("F1overalltruescoresequated",trialcodes[tid],".csv",sep=""),
  row.names=FALSE, quote=FALSE)

#Equating true scores for reference population

F0refts1 <- sapply(F0refscores1[,3],
  function(X) {tsf(thp=X, iparam=F0refcalib1)})
F1refts1 <- sapply(scaledF1refscores1[,3],
    function(X) {tsf(thp=X, iparam=scaledF1refcalib1)})
equatedF1refts1 <- sapply(scaledF1refscores1[,3],
    function(X) {tsf(thp=X, iparam=F0refcalib1)})
F0refttrue scores1 <- data.frame (ID = c(10001:13000),
    TS = c(F0refts1))
F1refttrue scores1 <- data.frame (ID = c(14501:17500),
    TS = c(F1refts1))
F1refttrue scoresequated1 <- data.frame (ID = c (14501:17500),
    equatedTS = c(equatedF1refts1))
write.csv(F0refttrue scores1,
    file=paste("F0refttrue scores",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)
write.csv(F1refttrue scores1,
    file=paste("F1refttrue scores",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)
write.csv(F1refttrue scoresequated1,
    file=paste("F1refttrue scoresequated",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)
#Equating true scores for focal population
F0focts1 <- sapply(F0focscores1[,3],
    function(X) {tsf(thp=X, iparam=F0foccalib1)})
F1focts1 <- sapply(scaledF1focscores1[,3],
    function(X) {tsf(thp=X, iparam=scaledF1foccalib1)})
equatedF1focts1 <- sapply(scaledF1focscores1[,3],
    function(X) {tsf(thp=X, iparam=F0foccalib1)})
F0foctrue scores1 <- data.frame (ID = c(13001:14500),
    TS = c(F0focts1))
F1foctrue scores1 <- data.frame (ID = c(17501:19000),
    TS = c(F1focts1))
F1foctrue scoresequated1 <- data.frame (ID = c (17501:19000),
    equatedTS = c(equatedF1focts1))
write.csv(F0foctrue scores1,
    file=paste("F0foctrue scores",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)
write.csv(F1foctrue scores1,
    file=paste("F1foctrue scores",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)
write.csv(F1foctrue scoresequated1,
    file=paste("F1foctrue scoresequated",trialcodes[tid],".csv",sep=""),
    row.names=FALSE, quote=FALSE)
write.csv(F0foctruescores1,  
    file=paste("F0foctruescores",trialcodes[tid],".csv",sep=""),  
    row.names=FALSE, quote=FALSE)
write.csv(F1foctruescores1,  
    file=paste("F1foctruescores",trialcodes[tid],".csv",sep=""),  
    row.names=FALSE, quote=FALSE)
write.csv(F1foctruescoresequated1,  
    file=paste("F1foctruescoresequated",trialcodes[tid],".csv",sep=""),  
    row.names=FALSE, quote=FALSE)

#------------------------#  
# Conversion Tables #  
#------------------------#

#Defining form 1 unequated scores for conversion table
possibletsv <- c(0:50)

#Estimating conversion table equivalents for overall population
equatedoverall <- F1overalltruescoresequated1[,2]
originaloverall <- F1overalltruescores1[,2]
dataoverall <- data.frame(equatedoverall,originaloverall)
polyfitoverall <- lm(equatedoverall~originaloverall+I(originaloverall^2),  
    data=dataoverall)
coefoverall <- coef(polyfitoverall)
overalltsr <- function(rts){(rts^2)*coefoverall[3]+(rts*coefoverall[2])+coefoverall[1]}
equatedoverallvalues <- overalltsr(possibletsv)

#Estimating conversion table equivalents for reference population
equatedref <- F1reftruescoresequated1[,2]
originalref <- F1reftruescores1[,2]
dataref <- data.frame(equatedref,originalref)
polyfitref <- lm(equatedref~originalref+I(originalref^2),  
    data=dataref)
coefref <- coef(polyfitref)
reftsr <- function(rts){(rts^2)*coefref[3]+(rts*coefref[2])+coefref[1]}
equatedrefvalues <- reftsr(possibletsv)

#Estimating conversion table equivalents for focal population

equatedfoc <- F1foctruescoresequated1[,2]
originalfoc <- F1foctruescores1[,2]
datafoc <- data.frame(equatedfoc,originalfoc)

polyfitfoc <- lm(equatedfoc~originalfoc+I(originalfoc^2),
data=datafoc)
coeffoc <- coef(polyfitfoc)

foctsr <- function(rts){(rts^2)*coeffoc[3]+
(rts*coeffoc[2])+coeffoc[1]}
equatedfocvalues <- foctsr(possibletsv)

#Truncating equated scores

equatedoverallvalues0s <- ifelse(equatedoverallvalues < 0,
0,equatedoverallvalues)
equatedrefvalues0s <- ifelse(equatedrefvalues < 0,
0,equatedrefvalues)
equatedfocvalues0s <- ifelse(equatedfocvalues < 0,
0,equatedfocvalues)

equatedoverallvalues50s <- ifelse(equatedoverallvalues0s > 50,
50,equatedoverallvalues0s)
equatedrefvalues50s <- ifelse(equatedrefvalues0s > 50,
50,equatedrefvalues0s)
equatedfocvalues50s <- ifelse(equatedfocvalues0s > 50,
50,equatedfocvalues0s)

equatedoverallvalues <- equatedoverallvalues50s
equatedrefvalues <- equatedrefvalues50s
equatedfocvalues <- equatedfocvalues50s

#Formatting and saving conversion table

conversiontable <- data.frame(
raw=possibletsv,
equatedoverall=equatedoverallvalues,
equatedref=equatedrefvalues,
equatedfoc=equatedfocvalues)
write.csv(conversiontable, file=paste("conversiontable", trialcodes[tid], ",".csv",sep=""), row.names=FALSE, quote=FALSE)

#--------------------------#
# Equating Invariance  #
#--------------------------#

#Defining DTM

DTMhalf <- .5
DTM1 <- 1

#Developing score level weights

proportion <- dnorm(possibletsv, mean=mean(F1overalltruescoresequated1[,2]), sd=sd(F1overalltruescoresequated1[,2]))

#Refining conversion table for calculations

raw <- possibletsv
overall <- equatedoverallvalues
groupr <- equatedrefvalues
groupf <- equatedfocvalues
frequency <- proportion

invariancedata <- data.frame(
  raw=possibletsv,
  overall= equatedoverallvalues,
  groupr=equatedrefvalues,
  groupf=equatedfocvalues,
  frequency=proportion)

#Calculating RMSD(x)

RMSD<-(sqrt((((groupf-overall)^2)*.25)+ (((groupr-overall)^2)*.75)))

#Calculating REMSD Code

REMSD<- (sqrt(sum((((groupf-overall)^2)*.25) + (((groupr-overall)^2)*.75))*frequency)))
#Calculating RESD\textsubscript{k}

\[
RESDR \leftarrow \left(\sqrt{\text{sum}((\text{groupr-overall})^2) * \text{frequency}}\right)
\]

\[
RESDF \leftarrow \left(\sqrt{\text{sum}((\text{groupf-overall})^2) * \text{frequency}}\right)
\]

#Calculating RSD\textsubscript{k}(x)

\[
RSDR \leftarrow \left(\sqrt{((\text{groupr-overall})^2)}\right)
\]

\[
RSDF \leftarrow \left(\sqrt{((\text{groupf-overall})^2)}\right)
\]

#Saving invariance results

conditionalinvariance <- data.frame(
  F1truescorevalues = possibletsv,
  RMSD = RMSD,
  RSDreference = RSDR,
  RSDfocal = RSDF)

unconditionalinvariance <- data.frame(
  REMSD = REMSD,
  RESDreference = RESDR,
  RESDfocal = RESDF)

write.csv(conditionalinvariance,
  file=paste("conditionalinvariance",trialcodes[tid],".csv",sep=""),
  row.names=FALSE, quote=FALSE)

write.csv(unconditionalinvariance,
  file=paste("unconditionalinvariance",trialcodes[tid],".csv",sep=""),
  row.names=FALSE, quote=FALSE)

#Plotting invariance results

jpeg(file=paste("omnibusinvariance",trialcodes[tid],".jpg",sep=""))
plot(possibletsv,RMSD,xlim=range(0,50),xlab="Form 1 True Score",
  ylim=range(0,5),ylab="Equated Score Differences",col="blue",
  pch=19, main="REMSD and RMSD(x)")
abline(DTMhalf,0,col="red")
abline(DTM1,0,col="red")
legend(x=40,y=5,c("RMSD","REMSD","DTM"),pch=c(19,19,19),
  col=c("blue","red","red"))
dev.off()
# Plotting Reference groups RESD and RSD
jpeg(file=paste("referenceinvariance",trialcodes[tid],".jpg",sep=""))
plot(possibletsv,RSDR,xlim=range(0,50),xlab="Form 1 True Score",
 ylim=range(0,5),ylab="Equated Score Differences",col="blue",
pch=19, main="RESD and RSD(x) for Reference Group")
abline(DTMhalf,0,col="red")
abline(DTM1,0,col="green")
legend(x=40,y=5,c("RSDr","RESD","DTM"),pch=c(19,19,19),
col=c("blue","red","red"))
dev.off()

# Plotting Focal groups RESD and RSD
jpeg(file=paste("focalinvariance",trialcodes[tid],".jpg",sep=""))
plot(possibletsv,RSDF,xlim=range(0,50),xlab="Form 1 True Score",
 ylim=range(0,5),ylab="Equated Score Differences",col="blue",
pch=19, main="RESD and RSD(x) for Focal Group")
abline(DTMhalf,0,col="red")
abline(DTM1,0,col="green")
legend(x=40,y=5,c("RSDf","RESD","DTM"),pch=c(19,19,19),
col=c("blue","red","red"))
dev.off()

# Closing trial loop
}

#-----------------------#
# Combining Trials  
#-----------------------#

# Creating population level invariance datasets
#(= code was shortened for parsimony)

unconditionalinvariancecombinedX <- rbind(unconditionalinvariance1,
unconditionalinvariance2, unconditionalinvariance3,
unconditionalinvariance4, ... , unconditionalinvariance98,
unconditionalinvariance99, unconditionalinvariance100)

unconditionalinvariancecombined <- data.frame(  
  trial = c(1:100),
  REMSD = unconditionalinvariancecombinedX[,1],
  RESDreference = unconditionalinvariancecombinedX[,2],
  RESDfocal = unconditionalinvariancecombinedX[,3])
RMSDcombinedX <- cbind(conditionalinvariance1[,2],
 conditionalinvariance2[,2], conditionalinvariance3[,2],
 conditionalinvariance4[,2], ..., conditionalinvariance98[,2],
 conditionalinvariance99[,2], conditionalinvariance100[,2])

RMSDcombined <- data.frame(
  RMSD1 = RMSDcombinedX[,1],
  RMSD2 = RMSDcombinedX[,2],
  RMSD3 = RMSDcombinedX[,3],
  RMSD4 = RMSDcombinedX[,4],
  ...,
  RMSD98 = RMSDcombinedX[,98],
  RMSD99 = RMSDcombinedX[,99],
  RMSD100 = RMSDcombinedX[,100])

RMSDmeans <- data.frame(
  F0truescore = possibletsv,
  RMSDmean = rowMeans(RMSDcombined))

RSDrefcombinedX <- cbind(conditionalinvariance1[,3],
 conditionalinvariance2[,3], conditionalinvariance3[,3],
 conditionalinvariance4[,3], ..., conditionalinvariance98[,3],
 conditionalinvariance99[,3], conditionalinvariance100[,3])

RSDrefcombined <- data.frame(
  RSDREF1 = RSDrefcombinedX[,1],
  RSDREF2 = RSDrefcombinedX[,2],
  RSDREF3 = RSDrefcombinedX[,3],
  RSDREF4 = RSDrefcombinedX[,4],
  ...,
  RSDREF98 = RSDrefcombinedX[,98],
  RSDREF99 = RSDrefcombinedX[,99],
  RSDREF100 = RSDrefcombinedX[,100])

RSDrefmeans <- data.frame(
  F0truescore = possibletsv,
  RSDref = rowMeans(RSDrefcombined))

RSDfoccombinedX <- cbind(conditionalinvariance1[,4],
 conditionalinvariance2[,4], conditionalinvariance3[,4],
 conditionalinvariance4[,4], ..., conditionalinvariance98[,4],
 conditionalinvariance99[,4], conditionalinvariance100[,4])

RSDfoccombined <- data.frame(
  RSDF1 = RSDfoccombinedX[,1],
  RSDF2 = RSDfoccombinedX[,2],
  RSDF3 = RSDfoccombinedX[,3],
  RSDF4 = RSDfoccombinedX[,4],
  ...,
  RSDF98 = RSDfoccombinedX[,98],
  RSDF99 = RSDfoccombinedX[,99],
  RSDF100 = RSDfoccombinedX[,100])

RSDfocmeans <- data.frame(
  F0truescore = possibletsv,
  RSDF = rowMeans(RSDfoccombined))
RSDfoccombined <- data.frame(  
    RSDFOC1 = RSDfoccombinedX[,1],  
    RSDFOC2 = RSDfoccombinedX[,2],  
    RSDFOC3 = RSDfoccombinedX[,3],  
    RSDFOC4 = RSDfoccombinedX[,4],  
    ...,  
    RSDFOC98 = RSDfoccombinedX[,98],  
    RSDFOC99 = RSDfoccombinedX[,99],  
    RSDFOC100 = RSDfoccombinedX[,100])

RSDfocmeans <- data.frame(  
    F0truescore = possibletsv,  
    RSDfoc = rowMeans (RSDfoccombined))

write.csv(unconditionalinvariancecombined,  
  "C:/Users/corinne/Desktop/B/combined_trials_results/  
  unconditionalinvariancecombined.csv",  
  row.names=FALSE, quote=FALSE)
write.csv(RMSDcombined,  
  "C:/Users/corinne/Desktop/B/combined_trials_results/  
  RMSDcombined.csv",  
  row.names=FALSE, quote=FALSE)
write.csv(RMSDmeans,  
  "C:/Users/corinne/Desktop/B/combined_trials_results/  
  RMSDmeans.csv",  
  row.names=FALSE, quote=FALSE)
write.csv(RSDrefcombined,  
  "C:/Users/corinne/Desktop/B/combined_trials_results/  
  RSDrefcombined.csv",  
  row.names=FALSE, quote=FALSE)
write.csv(RSDrefmeans,  
  "C:/Users/corinne/Desktop/B/combined_trials_results/  
  RSDrefmeans.csv",  
  row.names=FALSE, quote=FALSE)
write.csv(RSDfoccombined,  
  "C:/Users/corinne/Desktop/B/combined_trials_results/  
  RSDfoccombined.csv",  
  row.names=FALSE, quote=FALSE)
write.csv(RSDfocmeans,  
  "C:/Users/corinne/Desktop/B/combined_trials_results/  
  RSDfocmeans.csv",  
  row.names=FALSE, quote=FALSE)
# Plotting population level conditional results

```r
jpeg("C:/Users/corinne/Desktop/B/combined_trials_results/ RMSDmeans.jpg")
plot(possibletsv,RMSDmeans[,2],xlim=range(0,50),
    xlab="Form 1 True Score",ylab="Equated Score Differences",col="blue",pch=19,
    main="Average RMSD(x) Across Trials")
abline(DTM1,0,col="red",lty=2)
abline(DTMhalf,0,col="black")
legend(x=35,y=5,c("RMSD(x)","DTM(0.5)","DTM(1.0)"),
pch=c(19,-1,-1),lty=c(0,1,5),col=c("blue","black","red"))
dev.off()
```

```r
jpeg("C:/Users/corinne/Desktop/B/combined_trials_results/ RSDrefmeans.jpg")
plot(possibletsv,RSDrefmeans[,2],xlim=range(0,50),
    xlab="Form 1 True Score",ylab="Equated Score Differences",col="blue",pch=19,
    main="Average RSD(x) for Reference Group Across Trials")
abline(DTM1,0,col="red",lty=2)
abline(DTMhalf,0,col="black")
legend(x=35,y=5,c("RSDr(x)","DTM(0.5)","DTM(1.0)"),
pch=c(19,-1,-1),lty=c(0,1,5),col=c("blue","black","red"))
dev.off()
```

```r
jpeg("C:/Users/corinne/Desktop/B/combined_trials_results/ RSDfocmeans.jpg")
plot(possibletsv,RSDfocmeans[,2],xlim=range(0,50),
    xlab="Form 1 True Score",ylab="Equated Score Differences",col="blue",pch=19,
    main="Average RSD(x) for Focal Group Across Trials")
abline(DTM1,0,col="red",lty=2)
abline(DTMhalf,0,col="black")
legend(x=35,y=5,c("RSDf(x)","DTM(0.5)","DTM(1.0)"),
pch=c(19,-1,-1),lty=c(0,1,5),col=c("blue","black","red"))
dev.off()
```

```
#End Part 2
```
