Exploring Gender Differential Item Functioning (DIF) in Eighth Grade Mathematics Items for the United States and Taiwan

Ming-Chih Lan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2014

Reading Committee:
Min Li, Chair
Catherine S. Taylor
Robert D. Abbott

Program Authorized to Offer Degree:
College of Education

University of Washington

Abstract

Exploring Gender Differential Item Functioning (DIF) in Eighth Grade Mathematics Items for the United States and Taiwan

Ming-Chih Lan

Chair of the Supervisory Committee:
Associate Professor Min Li, Ph.D.
College of Education

Gender differences in mathematics performance have drawn much attention from researchers. Prior to making gender comparisons, a test with unbiased items should be used. Unbiased items are expected to have equal item correct response rates for males and females matched in their abilities. Differential Item Functioning (DIF) techniques have been developed to detect items that exhibit unequal success rates.

The focus of this secondary study with the international large-scale TIMSS mathematics assessment was to investigate gender DIF items for eighth grade populations in the U.S. and Taiwan based on ordinal logistic regression and Poly-SIBTest. For DIF items flagged, the study examined the gender DIF patterns across DIF methods, cognitive demand levels, and countries. The study concluded that the total amount of DIF items was low, and the number of items identified as DIF favoring each gender was approximately equal. This gender DIF pattern was consistent across items differing in cognitive demand levels, and in both the U.S. and Taiwan samples. That is, in general, neither males nor females were favored by test items differing in cognitive demand levels. The two DIF techniques were consistent in their abilities to identify DIF and non-DIF items, but the magnitude of DIF identified by the ordinal logistic regression DIF tended to be smaller than the magnitudes identified by the Poly-SIBTest. Although the gender DIF patterns were consistent across the focal countries, the actual items identified as DIF

and non-DIF differed because gender DIF items were country-dependent. This has important implications to deal with country-specific gender DIF items in any international assessment programs.

Table of Contents

List of Figures

List of Tables

Dedication

To my parents. My dear mom, always cares for me like a child even though I have already grown up to be a man with my own family. And especially to my dad, who passed away unexpectedly, just a week before my oral defense took place. This was an incident I believe I will never forget for the rest of my life. My dad frequently wore the jersey with the "UW Dad" logo I bought for him. I knew he was really proud of me, but I never realized how important he was to me until he passed away. Yes, my father never dies in my mind. He just fades away.

Acknowledgements

First, I would like to express my appreciation to my advisor, Dr. Min Li, for her inspiration, suggestions, and guidance in every phase I have experienced in the doctoral program at UW, and for helping me to reach the end successfully.

Second, I would like to thank my committee members, Dr. Catherine Taylor, Dr. Robert Abbott, Dr. Leonard Clark Johnson, and Dr. Craig Scott for their constructive feedback and comments, which were vital to improve my dissertation. I am very grateful to my fellow Dr. Maria Ilich, Ph.D. candidates, Mr. Phonraphee Thummaphan, Ms. Ting Wang, and Ms. Yuan-Ling Liaw, who selflessly shared their knowledge of my topic, helping to strengthen my dissertation and remind me of what I might have missed.

Most of all, my thanks go to my sisters and brothers in my country, for caring for my parents so that I was able to focus on my academic work without worry. In addition, thanks to my wife, Fandy Tsai. Without her support, financially, emotionally and physically, it would have been a much longer journey to fulfill my Ph.D. degree.

Last but not least, thanks to Mr. Matt Davidson. He showed up at the late stage of my project and accompanied me until I completed my work toward a Ph.D. Because of his help in proofreading and editing the dissertation chapters, my dissertation work can be published and shared in English.

Chapter I: Introduction

1.1 Background to the Study

1.1.1 The Importance of Tests without Biased Items

Researchers, specifically measurement specialists, are concerned with the fairness of tests, and in particular the possibility that tests may be biased against male or female groups. Unbiased tests will ensure that the scores from assessments are fair to all groups of examinees. Using unbiased tests allows further investigation into the internal or external factors contributing to performance differences between gender groups. Bias is defined as a phenomenon where "…examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose" (Zumbo, 1999, p.12). For example, a mathematics test with only multiple-choice items may result in higher scores for ability-matched males than females, if males are better at guessing on multiple-choice items. On the other hand, a mathematics test full of open-ended items may result in better test scores for ability-matched females than males, if females are better at writing, which helps them respond to items with open-ended formats. When guessing and writing abilities are not what a mathematics test is intended to measure, item type (multiple-choice or open-ended) becomes a source of test bias. In other words, test bias may occur when performance on a test requires sources of knowledge different from the knowledge the test items were intended to measure, resulting in less valid score interpretations for a particular group (Camilli & Shepard, 1994). Although no test can perfectly measure the targeted ability, it is expected that the impact of measurement bias on test scores due to group membership is either reduced or equally balanced.

An unbiased test ensures that the scores are valid for all test-takers. In the late 1960s and early 1970s, psychometricians began to respond to test bias issues. They defined the term "bias," developed rigorous methods to develop tests that are not biased, and empirically investigated biases in tests. Since then, there has been concerning what the term "bias" means. In the mid-80s, a more neutral and general term was proposed for detecting potentially biased test items, called Differential Item Functioning (Zhang, 2001). Differential Item Functioning (DIF) outlines the way to identify test items that affect the differential performance of certain groups on the test. In Angoff's (1993) words, DIF refers to the "…simple observation that an item displays different statistical properties in different group settings after controlling for differences in the abilities intended to be measured." The DIF approach focuses on the fact that different groups of examinees matched in ability may have different correct answer rates for the same test question.

1.1.2 DIF Shows Unequal Success Rates between Ability-matched Examinees

DIF refers to the differences in estimated item parameters after groups of examinees are matched by abilities or traits that the test is intended to measure (Dorans & Holland, 1993). DIF items in a test are those that have unequal rates of correct response for examinees with equal ability level, but with different group membership. In other words, for these DIF items, groups of individuals matched in their ability do not have equal probabilities of correct response rates on an item (Zumbo, 1999). A DIF item may be a *biased* item, but it is not necessarily the case. As Zumbo (1999) highlighted, biased items refer to the condition where test takers respond to items differently because of unrelated to traits or constructs. Thus, DIF is required, but not sufficient to claim that an item is biased, until the DIF for the item is proven to be unrelated to what the test is developed to measure (Zumbo, 1999). If DIF items are proven to be biased after a substantial

investigation following the identification of DIF items, test scores should be adjusted to correct for the resulting DIF effect in the test scores.

There are two kinds of DIF items, uniform and non-uniform (Metcalf, 2002). Uniform DIF items are those for which the probability of answering the item correctly in one group is greater or lower than the other group across all levels of abilities. In contrast, non-uniform DIF occurs when the probability of answering the item correctly is higher for one group at certain ability level, but lower for the same group at another ability level. On item characteristic curves (ICCs), which show the relationship between ability levels of examinees and the correct response rate on a test item, uniform DIF is represented by two parallel curved lines, whereas two non-parallel, or crossing, curve lines indicate non-uniform DIF (Burkes, 2009). These two types of DIF items are shown in Figure 1. The item on the left side indicates non-uniform DIF (i.e., crossing DIF), whereas the item on the right side indicates uniform DIF (i.e., unidirectional DIF; Metcalf, 2002).

DIF research aims at identifying potentially biased items, and provides test developers with guidelines for item development intended to create quality test items without bias. However, the statistical model used to detect DIF items does not necessarily tell researchers the source of the DIF.[1] To overcome the issue, Roussos and Stout developed a multidimensionality-based DIF analysis paradigm to identify the underlying causes of DIF (Gierl, 2005; Gierl, Bisanz, Bisanz, & Boughton, 2003).

Identifying the cause connected to the DIF items contribute to the development of more valid assessments, ensuring that any recommendations and conclusions made from that assessment are equally valid for the different sub-groups of the test-takers involved.

[1] In a related DIF study, Wang and Lane (1996) pointed out that it was relatively difficult to determine the actual factors that contributed to a significant DIF statistic after DIF items were identified.

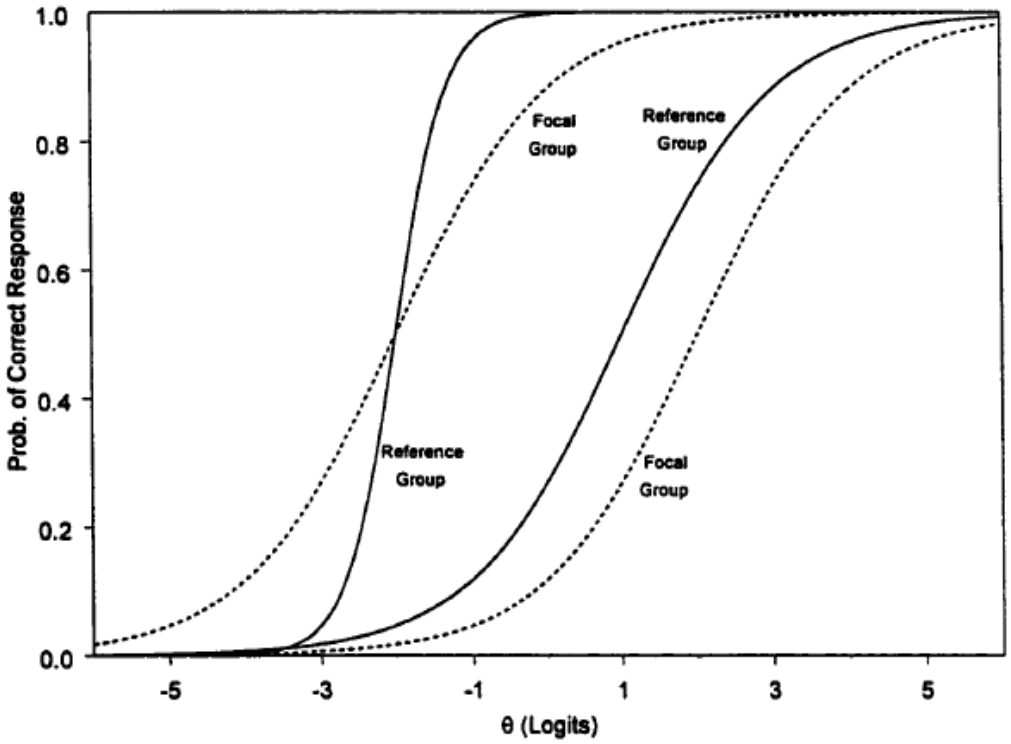


Figure 1. Item characteristic curves of two types of DIF items.

Different DIF detection methods have been developed over the past three decades to detect potentially biased items. Among them, the logistic regression DIF method and the simultaneous Item Bias Test (SIBTest) DIF method are frequently used. Both methods were selected to use in this dissertation. In this study, the SIBTest results was compared to the logistic regression results, to explore how similar or different their abilities to detect DIF and non-DIF items are.

1.1.3 DIF on Genders in Mathematics Items

Examination of items with potential DIF involves a comparison of performance on individual test items between two groups of test takers (e.g., males vs. females; Asian-American vs. African-American) who are assumed to have equal ability. If the probability of answering an item correctly for one group of examinees is significantly higher or lower than that of the other

group, the item is flagged as a DIF item, or a potentially biased item. The literature has identified several test item characteristics that potentially contribute to DIF items. These item characteristics are reviewed below.

Item characteristics refer to ways of grouping or classifying test items. These characteristics may include, but are not limited to, item type (e.g., true or false, multiple-choice, and open-ended items), content domain (e.g., number, algebra, geometry, and measurement items), cognitive demand (e.g., knowing facts, applying rules, and problem-solving items), item difficulty (e.g., easy, medium, and difficult items), and item context (e.g., visual, spatial, and application oriented items).

Studies to identify DIF items have shown that *item type* is one of the sources of gender DIF in mathematics items (Bolt, 2000; Burton, 1996; Garner & Engelhard, 1999; Henderson, 2001; Taylor & Lee, 2012; Zenisky, Hambleton, & Robin, 2004). For example, Burton (1996) and Henderson (2001) concluded that males outperformed females on multiple-choice items, whereas females outperformed males on open-ended items when both males and females were matched on mathematics ability. One possible explanation was that females exhibited better verbal ability (Beller & Gafni, 2000). In addition, different test-taking strategies used by each gender may account for the interaction between gender and item type. For example, as Ben-Shakhar and Sinai (1991) concluded, there was a tendency for males to guess when they did not know the correct answer on multiple-choice items, whereas females tended to skip those items. Also, Bennett (1993) concluded that the item type may affect the interpretation of test scores, because it limited the content and processes that can be measured. For example, multiple-choice items are less likely to measure productive or creative thinking, and do not elicit higher levels of mental skills to solve problems (Martinez, 1999).

Previous research also identified the *content domain* of test items as contributing to gender DIF in mathematics items. Langenfeld (1997) found that females performed better than males on algebra items, while males performed better on geometry items. However, Doolittle's (1989) study did not support the presence of gender DIF on algebra items. The results from Doolittle (1989) only supported the finding that females performed less well on geometry items than males. Doolittle urged test developers to try to understand the role that content area might play in gender DIF effects. The mixed results of previous gender DIF studies on content areas may result from an interaction effect with item type. In other words, the algebra items involved in the DIF study used by Langenfeld or Doolittle may have differences in item type that have an effect on gender DIF outcomes.

The findings from meta-analysis studies are also mixed. Hyde, Fennema, and Lamon (1990) reviewed 100 studies, and that computation items showed an advantage for elementary and middle school females, whereas problem-solving items showed no gender differences for elementary or middle school students. This result, however, was based on a comparison of mean scores on all test items combined, not on individual items with males and females matched in ability. To detect DIF items, the comparison should be based on individual test items with groups of students matched on observed test scores or estimated ability level. Thus, the issue of whether gender differences resulted from a true difference in abilities, or from DIF items favoring either group, should be examined further.

Freidman (1989) reviewed 98 studies on gender differences on mathematics tests published between 1974 and 1987. Freidman coded test items as either problem-solving oriented or not, and concluded that males outperformed females on the problem-solving items. However, this study also did not follow the requirement of detecting potentially biased items. The finding

of males outperforming females on problem-solving items in Friedman's (1989) study should be re-examined using the standard procedures for DIF item detection.

Cognitive demand (also known as cognitive skill, cognitive complexity, or cognitive domain) is another source of gender DIF for mathematics items discussed in the literature. Different mathematics items require different kinds or levels of cognitive skills. Cognitive demand is defined as the mental skills or abilities elicited to answer certain test items correctly (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). Different types of items need different kinds of mental skills to solve them. For example, the cognitive demand level of *knowing* refers to the mental abilities needed to understand the facts, procedures, and concepts; the cognitive demand level of *applying* refers to the mental ability to apply facts, knowledge, and conceptual understanding to solve familiar or routine problems; the level of *reasoning* refers to the ability to solve problems in unfamiliar situations or complex contexts, such as non-textbook practice items and multi-step problems (Mullis et al., 2009).

Research suggests that mathematics items classified as having higher levels of cognitive demand generally favor males over females (Engelhard, 1990). For example, males outperformed females on items with higher cognitive demand levels such as multi-step problems and real-life application problems, whereas females outperformed males on items with lower cognitive demand levels, like items requiring computation or tapping basic math knowledge) (Engelhard, 1990; Harris & Carlton, 1993). A similar finding was also reported by Harris and Carlton (1993), who examined gender DIF on multiple-choice items from the Scholastic Aptitude Test (SAT) mathematics test, administered to high school juniors and seniors for whom English was self-reported to be their first language. Harris and Carlton (1993) concluded that males outperformed females on items requiring a higher level of cognitive demand, whereas

females performed relatively better than males on items requiring a lower level of cognitive demand. Using content analysis, Taylor and Lee (2012) pointed out that many DIF items with higher levels of cognitive demand were constructed-response items. This finding suggests that the gender DIF found for higher cognitive demand items may be moderated by item characteristics like item type.

Unlike item type and content domain, gender DIF as a result of cognitive demands in mathematics items from large-scale datasets (e.g., statewide, national, and international assessment projects) has not drawn much attention from researchers. Specially, in 32 thesis and dissertation related to DIF studies conducted in the country Taiwan, no one study focused on cognitive demand. In addition, the findings from the limited research into this topic have been rather mixed (Bielinski, 1999; McKenzie, 2009; Yan, 2005; Zhang, 2001). This dissertation is motivated by these two concerns; this dissertation study will help us understand the relationship between gender DIF and the cognitive demand level of mathematics items, while also suggesting some productive directions to resolve the inconsistent research findings.

1.2 Purposes of This Study

For any assessment, in addition to gathering evidence about whether the test scores are reliable and valid, test developers must make sure that scores are free of potential bias, such as DIF. Although many DIF studies in the past thirty years have been conducted to explore gender DIF associated with item characteristics in different types of tests (Zhang, 2001), gender DIF studies investigating cognitive demand have not drawn as much attention as those involving item type and content domain. The first purpose of this dissertation is to fill in this particular research gap by focusing on gender DIF associated with the cognitive demands of mathematics test items. Specifically, this study will examine the influence of test items grouped by three cognitive

demand levels (i.e., knowing, applying, and reasoning) on the mathematics performance of eighth graders. This study will investigate whether items with different cognitive demand levels function differently across male and female groups of examinees matched in their abilities.

The second purpose of this dissertation is to compare two DIF detection methods (i.e., logistic regression and SIBTest), and demonstrate that the particular DIF detection method that a researcher adopts should come up with the same results. Thus, the ability of the logistic regression DIF approach is expected to be the same as the SIBTest DIF approach.

The third purpose of this study is to explore the similarities and differences of the gender DIF patterns across two countries. This serves as an extension of the first purpose of this study. This cross-cultural exploration of patterns and relationships between item-level gender DIF and cognitive demand levels of test items includes Taiwan and the U.S.

Taiwan and the U.S. were selected for this investigation based on the following ideas. First, compared to the U.S., Taiwan is less culturally diverse, as it does not have a large amount of immigrants from other countries. Although in recent years immigration has increased in Taiwan, to address the lack of labor needed to boost economic growth, the number of immigrants is still limited compared to the entire population in Taiwan.

Second, high school education in the U.S. is generally free and guaranteed, without entrance examinations for students who graduate from middle schools. Unlike the U.S., entrance examinations are required to screen students who wish to have high school education in Taiwan. However, Taiwan is experiencing a turning point. A reformed sort of K-12 education system is currently being developed. In response to public demand for educational reforms, the education system in Taiwan has undergone significant changes over the last several years. The current 9-year compulsory education policy has been implemented in schools since 1968. Extending the

education system begun in 2000 and will further expand to senior high schools with students up to the age of 17 years old. Schools will be guaranteed and free, without entrance examinations, as in the U.S., so long as students are willing to attend.

Third, Taiwan is primarily a nationally centralized educational system with a unified educational philosophy and national curriculum standards, whereas the U.S. is regionally centralized with higher responsibilities to manage educational practices assigned to the state, city, or school district level. Fourth, Taiwan students performed significantly higher than the U.S. students on the mathematics portion of the international large-scale Trends in International Mathematics and Science Study (TIMSS) assessment program (Mullis, Martin, Foy, & Arora, 2013).

1.3 Research Questions

This study explores the following three questions:

1. Are the gender DIF items identified by the logistic regression DIF method consistent with the items identified by the SIBTest method?
2. Do the gender DIF patterns for mathematics items differ in levels of cognitive demand?
3. Is the gender DIF pattern of items differing in cognitive demand levels similar across countries?

This dissertation study is a secondary analysis of data extracted from the Trend in the International Mathematics and Science Study 2011 (Foy, Arora, & Stanco, 2013), which included approximately 60 participating countries, including the U.S. and Taiwan (Mullis et al., 2013). These two countries were used to address the research questions.

The gender DIF items and patterns investigated for each research question were identified using both the logistic regression DIF and SIBTest methods. The cognitive demand levels of the mathematics items were defined according to the TIMSS 2011 mathematics assessment framework, developed by the TIMSS program.

1.4 Significance of the Study

The findings from this dissertation have important implications for interpreting gender differences in mathematics assessment. First, the study examines whether the cognitive demand serves as a source of gender DIF in mathematics performance. That finding contributes to the unsettled debate discussed above – whether test items with a high level of cognitive demand favor male examinees, and whether test items with a low level of cognitive demand favor female examinees.

Second, DIF analyses can help bias and sensitivity committees, formed for large-scale assessment projects, to investigate potential item bias and test unfairness for different levels of cognitive demands. DIF analyses are a routine part of large-scale assessment programs (Taylor & Lee, 2012; Zenisky, Hambleton, & Robin, 2004). To ensure fair test score interpretations regardless of race, ethnicity, gender, region, religion, special population, and socio-economic status, bias and sensitivity reviews by a group of committee members from a variety of backgrounds are a required step to develop bias-free tests prior to publishing. These committees review items for (1) negative or stereotypical representations of any group, (2) over- or under-representation of any group, (3) unfamiliar language or terms that may be confusing to students based on region, culture, socio-economic status, etc., and (4) controversial issues and topics that may affect some groups more than others. They do not look at cognitive demand as a potential source of item bias. In addition, the items identified as biased during such reviews may not be

the same items identified by DIF analyses. Thus, DIF analysis can serve as an auxiliary tool to help the bias and sensitivity committees locate potentially biased items.

Third, DIF analyses might also help to identify strengths or weakness in an educational program or assessment project by identifying items that function differently, or in an unexpected manner, for different groups of the same ability. For example, Taylor and Lee (2012), exploring gender DIF on mathematics items from state criterion-referenced tests for grades 4, 7, and 10, concluded that multiple-choice items generally favored males, while constructed-response items generally favored females. They speculated that the instruction rather than test development might explain this gender difference. Teaching students how to demonstrate solution strategies may decrease the number of constructed-response DIF items without threatening the validity of test scores (Taylor & Lee, 2012).

Fourth, student samples vary in many ways, such as social economic status, school district, poverty level, ethnicity, and culture. These differences may moderate the gender DIF patterns found by a DIF study. In particular, most of the DIF studies in literature were from the U.S.; so it is not possible to generalize from these studies to other countries. In order to understand the complex nature of gender DIF in mathematics items, and to develop interventions that focus on multiple factors, cross-country gender DIF analysis including more than just the U.S., is conducted in this dissertation.

1.5 Organization of the Dissertation

This dissertation is divided into five chapters. Chapter I – Introduction, describes the background and importance of the research questions raised in this study. Chapter II – Literature Review, summarizes studies on gender DIF related to mathematics tests, and then provides an overview of the DIF detection methods used in the studies, as well as alternatives from other

studies. In Chapter III – Methods, I describe the samples, datasets, mathematics items, gender DIF analysis procedures, and statistical tests that I used in this dissertation. In Chapter IV – Findings, the results of the gender DIF analysis are presented in the order of the research questions asked. In Chapter V – Discussions and Conclusion, conclusions from the findings are presented and discussed, along with implications from the findings, and suggestions for further study.

Chapter II: Literature Review

This chapter has two sections. The first section starts with an overview of the issue of gender differences in mathematics performance, which helps understand the ongoing debate about the existence of gender differences in mathematics achievement at the mean score level. Then, the focus is switched from the mean score level to the item level, and a detailed review of item characteristics associated with gender DIF for mathematics items. The second section describes the methods developed to detect gender DIF, and their properties.

2.1 Gender Differences in Mathematics Performance

2.1.1 Performance Differences between Genders at Mean Score Level

2.1.1.1 United States

Test results are widely used and profoundly affect us in career choice, opportunities in college education, and so on. The existence and nature of gender differences in performance on mathematics tests remains contentious (Bielinski, 1999). One widely known social stereotype is that females perform worse in mathematics than males do. For example, tests were intentionally described as producing differential performances between genders, females performed considerably worse than equally able males did (Spencer, Steele, & Quinn, 1999). Test results from various national and international large-scale assessments for school students may strengthen this viewpoint (Harris & Carlton, 1993).

According to the national large-scale mathematics tests included in the National Assessment of Educational Progress (NAEP), which were administered to 9-, 13-, and 17-year-old students every 2 to 5 years in the U.S., the overall mean score differences for the 38 year span from 1973-2008 between males and females at those ages were 0.17, 1.83, and 4.58, respectively over . The pattern of mean differences showed an increasing trend from ages 9 to

age 17. For 17-year-old students, the mean score difference ranged from 3 to 8 points across twelve testing cycles, indicating that male students performed relatively better than their female counterparts of the same high school age (Armstrong, 1981; Rampey, Dion, & Donahue, 2009). According to the U.S. mathematics scores on the 2003 TIMSS, an international, large-scale, long-term assessment project administered to 4th, 8th, and 12th grade students, eighth grade males had significantly higher overall mean math scores than their female counterparts (Mullis, Martin, Gonzalez, & Chrostowski, 2004).

In addition to the gender gap in K-12 mathematics achievement, differences between males and females at the mean score level have been found at postsecondary level, with a wider gap. For example, examination of mean score difference on the SAT mathematics test for the past four decades revealed that males' scores were consistently better than females' by more than 30 points (Zhang & French, 2010). Hyde et al.'s meta-analysis (1990) concluded that gender differences became increasingly larger when study samples were more selective such as undergraduate students, graduate students, and cognitively gifted students, compared with pre-K to 12 students. In addition, when older students were sampled, the magnitude of gender differences increased.

In sum, the issue of gender differences in mathematics performance has been and remains a focus of investigation for many researchers. Researchers and educators have been examining and discussing potential internal and external factors (Fennema & Tartre, 1985) that may contribute to the performance difference between genders. A number of explanations have been proposed, such as cognition, self-perception, attitude and affect, anxiety, stereotype threat, socio-cultural factors, course-taking patterns, approaches to learning, opportunities to learn, strategies to learn, praise from teachers, self-regulation, biological differences, and item bias (Zhang &

French, 2010). This dissertation focuses on potential item bias. That is, the focus of my study is the lack of measurement invariance between groups of examinees that have the same abilities (Zhang & French, 2010). This will be further explained in Section 2.1.2. The various perspectives of these research orientations highlight the nature-versus-nurture debate on the research of gender differences (Mckenzie, 2009).

2.1.1.2 Cross-country Comparisons

Gender differences in mathematics performance are also reported for students in countries, besides the U.S. in the TIMSS assessment projects. Previously known as the Third International Mathematics and Science Study (TIMSS)[2], TIMSS is an international educational assessment project that began in 1995. It provides comparative information about educational achievement across voluntarily participating countries, and mainly focuses on improving teaching and learning in mathematics and science for fourth, eighth, and twelfth graders. TIMSS is conducted on a regular 4-year cycle: 1995, 1999, 2003, 2007, 2011, etc. The latest one, TIMSS 2011, involved 63 participating countries[3] (Foy et al., 2013).

In the TIMSS 2011 report, gender differences in eighth grade mathematics achievement indicated that girls had, on average and in general, higher achievement than boys (i.e., scores 469 vs. 465) across the participating 42 countries. However, the differences varied from country to country. There was no difference in 22 of the 42 countries[4]. There was a difference favoring boys in seven countries, and a difference favoring girls in the remaining 13 countries (Mullis et al., 2013).

---

[2] The First International Mathematics Study (FIMS) was in 1964, and the Second International Mathematics Study (SIMS) was in 1980-1982 (Grouws, 1992).

[3] Fifty-two countries selected to participate in the fourth grade mathematics assessment, and forty-five countries participated in the eighth grade mathematics assessment.

[4] Three countries including Botswana, South Africa, and Honduras participated in the ninth grade mathematics assessment.

The U.S. and Taiwan samples were among the 22 countries in which no difference in mathematics performance between males and females was observed (i.e., scores 511 vs. 508 for the U.S sample and 606 vs. 613 for the Taiwan sample). In terms of mathematics content domains, the U.S. male students statistically significantly outperformed female students in *number* (i.e., 520 vs. 508). No statistically significant difference was found in *algebra* (i.e., 510 vs. 513), *geometry* (i.e., 487 vs. 482), or *data and chance* (i.e., 530 vs. 525). For the eighth grade Taiwanese population, male students significantly performed worse than female students in *algebra* (i.e., 621 vs. 636), but no difference in *number* (i.e., 599 vs. 597), *geometry* (i.e., 621 vs. 629), or *data and chance* (i.e., 583 vs. 585; Mullis et al., 2013).

2.1.2 Performance Difference between Ability-matched Genders at Individual Item Level

Research has indicated that mean-level differences in mathematics performance between females and males can be distorted by comparing mean-level scores alone, without paying attention to the item-level score which shows differential correct response rates (Harris & Carlton, 1993; Ryan & Fan, 1996). Using item-level scores allows researchers to identify potentially problematic items and then remove them from the test, to make sure that the remaining test items lead to valid mean-level score comparisons and interpretations. Thus, item-level score analysis can help assessments appropriately reflect examinees' abilities in mathematics. Otherwise, any test that contains biased items can result in test scores and score interpretations of gender differences that will be misleading and questionable.

In measurement theory, Differential Item Functioning (DIF) is a technique developed to identify potentially biased items by comparing item-level correct response rates between males and females with the same ability. Different from gender difference analysis at the mean score level, analysis of gender differences at the item-level by DIF techniques *requires males and*

females to be matched in the particular abilities a test is intended to measure. Similarly to gender difference analysis of mean test scores, analysis of gender differences at the item level involves test item characteristics, including item type (e.g., multiple-choice and free response items; Becker, 1990; Bolger & Kellaghan, 1990; Harris & Carlton, 1993), content domain (e.g., computation, algebra, measurement, and geometry; Becker, 1990; Engelhard, 1990; Harris & Carlton, 1993), item difficulty level (e.g., low, moderate, and high; Bielinski, 1999; Bielinski & Davison, 1998), the role of differential guessing and item serial position (Becker, 1990; Ben-Shakhar & Sinai, 1991), and cognitive demand (e.g., low, moderate, and high cognitive levels; Engelhard, 1990; Harris & Carlton, 1993), to name a few. The findings around each source of gender DIF are reviewed next.

Test item characteristics refer to ways grouping test items as defined in Chapter I (see Section 1.1.3). Item characteristics are used by researchers in order to identify the sources of potentially biased test items by using DIF techniques. However, definitions of item characteristics, and what/how many elements are included in a specific item characteristic, vary from study to study. For example, in the TIMSS 2003 project, all developed test items for eighth grade test-takers were grouped into five content domains: *number*, *algebra*, *measurement*, *geometry*, and *data* (Martin, Mullis, & Chrostowski, 2004). In 2007, only four content areas – *number*, *algebra*, *geometry*, and *data and chance* – were used; *measurement* was removed, and *data and chance* replaced the 2003 title *data* (Martin et al., 2004). In addition, while the TIMSS 2003 used four cognitive demands - *knowing facts and procedures*, *using concepts*, *solving routine problems*, and *reasoning* (Martin et al., 2004; Mullis et al., 2004), TIMSS 2007 used only three cognitive demands: *knowing*, *applying*, and *reasoning* (Olson, Martin, & Mullis, 2008). As a result, in 2007, *knowing* replaced the combined *knowing facts and procedures* and *using*

concepts in 2003. Also, *solving routine problems* from 2003 was renamed *applying* for 2007 renamed. On the other hand, NAEP used low, moderate, and high as labels to describe the mental skills required for answering math items correctly, instead of using the TIMSS framework to define cognitive demand. As another example, problem-solving items can be classified according to item context, content domain, or cognitive demand (Mendes-Barnett & Ercikan, 2006). These inconsistencies, and slight differences in definitions and classifications used in the research, make research outcomes incomparable.

2.1.2.1 Gender Differential Item Functioning in Item type

In terms of item type, also known as item format, there are two types: multiple-choice (MC) and open-ended (OE) items. OE items can also be labeled as constructed-response (CR) or free response items.

Multiple-choice items refer to a list of answer options for examinees to choose from in response to an item. There are different format of multiple-choice items such as simple multiple-choice, complex multiple-choice, true-false, multiple true-false, and so on. The simple multiple-choice format is used more frequently than other formats available for developing multiple-choice items. In general, multiple-choice items have been the dominant item type in test development for many years, because they are relatively inexpensive to develop, efficient to administer, objective in scoring procedures, and have a broad coverage of content. The multiple-choice item type is criticized for supporting the assessment of isolated facts and limiting the evaluation of higher order thinking and problem-solving skills (Zhang, 2001). Multiple-choice items are typically dichotomously scored – either as correct or incorrect (Yan, 2005).

On the other hand, OE items refer to the answer in response to a test item must be generated by the examinees. The answer may contain giving short answers from a single word or

number to a short sentence, providing a process, or justifying a chosen answer. OE items allow students to demonstrate higher-level thinking, problem solving, and reasoning (Zhang, 2001). Although there has been an increasing use of OE items in test development, multiple-choice items still play a dominant role.

Several studies have suggested that item type may result in gender DIF, and multiple-choice items tended to favor males, whereas open-ended items tended to favor females (Bolt, 2000; Burton, 1996; Garner & Engelhard, 1999; Henderson, 2001; Taylor & Lee, 2012; Zenisky, Hambleton, & Robin, 2004).

Burton (1996) had males and females matched in their performances on SAT to examine whether item type as a source favoring one gender group. The Mantel-Haenszel method was used to examine gender DIF effects results from genders. The statistics evidence showed males scored higher in multiple-choice items and females scored higher in open-ended items, indicating multiple-choice items favored males, whereas open-ended ones favored females. This finding was supported by two later studies.

In Henderson's (2001) examination of item type from the Alberta Education Diploma Examination composed of both multiple-choice and open-ended items. The generalized Mantel-Haenszel, Poly-SIBTest, and their counterparts, the Mantel-Haenszel and SIBTest DIF methods were selected to identify gender DIF patterns. The finding suggested that there may be a gender and item type interaction where males performed better on multiple-choice items and females performed better on open-ended items

Taylor and Lee (2012) explored gender DIF for graders 4, 7, and 10 on mathematics items composed of multiple-choice and open-ended formats using Poly-SIBTest and IRT-based

Rasch DIF methods. They concluded that multiple-choice items generally favored males while open-ended items generally favored females.

The gender-based DIF studies of Bolt (2000) and Garner and Engelhard (1999) also support males are favored by multiple-choice items and open-ended items favored females. Bolt (2000), who examined gender DIF due to item type using data from the mathematics sections of the Scholastic Aptitude Test (SAT-M). Forty mathematic items and the adjusted SIBTest DIF method were involved in the analysis. He reported 5 items favored males for the multiple-choice format and the magnitude of DIF, less than .05 for all five items, which were quite small. He concluded a small but statistically significant effect that favored male test-takers when the open-ended items were converted into the multiple-choice format.

Garner and Engelhard (1999) conducted a gender-based DIF study using 3,952 eleventh graders who took the mathematics portion of the 1994 Georgia High School Graduation Test (GHSGT), and found that open-ended items tended to favor females.

Moreover, in a study by Zenisky, Hambleton, and Robin (2004) examined gender DIF from approximately 360,000 students of elementary, middle, and high school students participating in a large-scale science assessment program. At each school level, data were collected by two forms of tests with each form containing 32 – 42 items. The tests at each level across forms were developed based on the same specifications. Both multiple-choice and open-ended items were used on each test form. They concluded that multiple-choice items tended to favor males while open-ended items tended to favor females. The gender DIF study in science items may reveal mathematics and science items share the gender DIF effect that multiple-choice items favoring and open-end items favoring females.

However, the findings in Lane, Wang and Magone's (1996) and Yan's (2005) do not support such a conclusion. In the study of Lane et al. (1996), data were collected by the QUASAR Cognitive Assessment Instrument from a total of 3,946 middle school students. Thirty-six open-ended items that assessed student's mathematical problem-solving strategies, conceptual understanding, reasoning, and communication ability were examined. The logistic discriminant function analysis (LDFA) was used in this study (Miller & Spray, 1993). The finding indicated open-ended items may favor either gender depending on other item feature.

In Yan's (2005) study, two item types and five content areas were investigated as sources of gender DIF. A sample of 1,132 eighth grade U.S. students, who took TIMSS-R 1999 mathematics test, were included in the analysis. Forty-five mixed multiple-choice and open-ended mathematics items were selected from multiple booklets for the DIF study. Items were tested as a group for gender DIF using the multifaceted Rasch model, an Item Response Theory-(IRT) based DIF method. Item types were not found to contribute to differential correct response rates between males and females.

As Taylor and Lee pointed out (2012), most of earlier studies that investigate gender differences in mathematics performance were not DIF studies because they did not follow the requirement of DIF studies – comparing two groups of subjects with equal abilities. However, researchers to from a hypothesis to explore gender DIF have cited those studies (Beller & Gafni, 2000; DeMars, 2000; O'Neil & Brown, 1998; Pomplun & Capps, 1999; Wester & Henriksson, 2000).

For example, Beller and Gafni (2000) investigated the influence of item type on gender DIF in mathematics using data from another international assessment, the International Assessment of Educational Progress (IAEP) in 1988 and 1991. The students were approximately

1,000 13-year-old middle school students from each of six countries in 1988 and approximately 1,650 9- to 13-year-olds for each of age group from some 20 countries. The open-ended items in these two studies asked for either a short answer or a single-number or word response. The results indicated that both multiple-choice and open-ended items favored males. The results of their study contradicted the assertion that females outperformed males on open-ended items.

DeMars (2000) examined the potential interaction between item type and gender DIF. The data of student performance on the mathematics section of Michigan's Higher School Proficiency Test (HSPT) were collected. Thirty-two multiple-choice and six open-ended items were involved in the DIF study. The finding showed there was an interaction between item type and gender DIF indicating multiple-choice favored males and open-ended favored females.

Wester and Henriksson (2000) examined the interaction between item type and gender DIF in mathematics performance using data from the TIMSS assessment program. The student samples were 9,779 6th, 7th, and 8th graders in Sweden with 4, 999 males and 4,780 females. Based on the results from the performance of students on the TIMSS test, they found females performed better than males on both multiple-choice and open-ended items. When hey changed multiple-choice items into comparable open-ended format items. As a result, there was no statistical evidence open-ended items would favor females.

The non-DIF studies mentioned (e.g., Beller & Gafni, 2000; DeMars, 2000; Wester & Henriksson, 2000) generally examined gender differences of performance in a mean score level instead an item score level. Additionally, in their studies, the DIF methods used to examine gender difference were not mentioned.

In general, empirical studies tend to support item type as a potential source of gender DIF in mathematics performance. Multiple-choice items tended to favor males, whereas open-ended

items tended to favor females. However, mixed results still exists. Identifying potentially biased items is the final goal of a DIF study. In order to identify such items, it is required that groups of subjects involved in a DIF study are matched in their abilities. In addition, either multiple-choice or open-ended items may occur in different content domains. Short answer, single-number, single-word, or extended free responses can all be considered as open-ended items.

Do males outperform females in all kinds of multiple-choice items? Do females perform better than males in all kinds of open-ended items or is there an interaction between gender DIF and item type when considering other item features? More research is necessary to address this mystery.

2.1.2.2 Gender Differential Item Functioning in Item Difficulty Level

Like item type and content domain, item difficulty level may serve as a source of gender DIF. Engelhard (1990) examined data gathered from students in Thailand and the U.S., extracted from the Second International Mathematics Study (SIMS). He investigated whether gender differences on mathematics items were associated with item difficulty. The results suggested a significant main effect from item difficulty levels, indicating that ability-matched males performed better than females on the more difficult items, whereas females outperformed males on the relatively easier items.

Consistent with what Engelhard (1990) concluded, Becker's (1990) study using SAT-M gathered data from the Study of Mathematically Precocious Youth (SMPY) during the years of 1973 through 1976. There were 2,382 talented students in grades 7, 8 and 9 participating in the study. Becker concluded easier items favored females, whereas harder items favored males.

Bielinski and Davison (1998) tested the hypothesis that gender DIF depend on item difficulty levels using data gathered from 8th graders taking a statewide mathematics. They plotted and examined the joint distribution between the difference in item difficulty (b_{male}-b_{female}) in the Y axis, and item difficulty estimated for the entire sample combining males and females in the X axis. Three item difficulty measures were used to examine mathematics items in nine multiple test forms. A linear correlation was shown between the gender difference in item difficulty and the overall item difficulty parameter. The findings suggested easier items favored females and more difficult items favored males.

Bielinski (1999) tried to replicate the findings of Bielinski and Davison (1998) to nationally representative samples. They explored the relationship between gender-related DIF and item difficulty level in three different age levels of the TIMSS 1995 for the U.S. students. A 3-parameter IRT-based model was used to estimate the item difficulty index for male, female, and total student samples on 124 multiple-choice items. The item difficulty index, equivalent to the item correct response rate, was estimated by the BILOG-MG DIF method (Du Toit, 2003). Then, item difficulty differences were computed between males and females, and finally, Pearson correlation was used to test the significance level of the correlation coefficients between item difficulty differences and total item difficulty. The study concluded that the easiest items tended to be easier for females than males, and the harder items tended to favor males across the three different age levels.

The literature on mathematics gender differences, as explored by DIF approaches, supported the idea, although not definitive, that easier items favor females and harder items favor males (Becker, 1990; Engelhard, 1990; Harris & Carlton, 1993; Lu & Dunbar, 1997).

2.1.2.3 Gender Differential Item Functioning in Content Domain

Content domain, also known as content focus, content area, or content knowledge, is defined as the content of a particular knowledge field, such as the subjects taught in school. Among the variety of subjects available as content domain in mathematics, algebra and geometry are discussed next.

Some studies fully supported the conclusion that geometry favors males and algebra favor females (Garner & Engelhard, 1999; Harris and Carlton, 1993; O'Neil & McPeek; 1993; Ryan & Fan, 1996). For example, in the study of Garner and Engelhard (1999) using the Multifaceted Rasch DIF method to examine the performance of 3,952 11th genders on the Georgia High School Graduation Test. They found that geometry items favors males and abstract algebra items favored females.

Harris and Carlton (1993) classified SAT mathematics items in content domains: arithmetic, algebra, geometry, and miscellaneous (e.g., structure of number systems, number sets, etc.). The Mantel-Haenszel method was used to detect gender DIF effect. They concluded that high school male students outperformed ability-matched females on geometry items, and that females outperformed males on algebra items.

O'Neil and McPeek (1993) reviewed a number of large sample DIF studies using Mantel-Haenszel, standardization, or IRT-oriented DIF methods on items from the American College Testing assessment (ACT), Graduate Management Admission Test (GMAT), Graduate Record Examination (GRE) General Test, the NTE core Battery (NTE), and SAT, in hopes of obtaining more reliable and consistent information on gender DIF outcomes. The review concluded that males performed better on geometry and mathematics problem-solving items than ability-matched females, while females performed better on algebra items than ability-matched males.

Ryan and Fan (1996) used a confirmatory differential bundle functioning (DBF) method on items from the SIMS mathematics assessment taken by eighth-grade students. DBF examined overall gender DIF effect for items grouped as a bundle within content areas (Gouglas, Roussos, & Stout, 1996). They found that geometry items favored males, and algebra favored females.

On the other hand, studies did not fully support the idea that geometry favors males whereas algebra favors females and even contradict such a tendency. For example, Doolittle and Cleary (1987) investigated gender DIF based on the performance of high school senior students on the ACT Mathematics Usage Test (ACTM). All items were in the multiple-choice format. A modification of an index based on a three-parameter logistic IRT model was used to identify DIF. They found that gender differences in performance were associated with content domain. Male students tended to perform better on geometry items. However, for algebra, either males or females may perform better depending on what item features connected to algebra. However, Becker (1990) concluded that males significantly outperformed females on algebra items.

Consistent with the finding in Becker's study, McKenzie (2009) explored gender DIF effects in TIMSS 2003 mathematics items using both the SIBTest and DBF methods on the U.S. sample group. Only a portion of the mathematics items was used as samples in the DIF study. For fourth graders, two out of fourteen final booklets (booklets 2 and 3) were selected, with around 743 students assigned to each booklet. For eighth graders, three out of twelve final booklets (booklets 1, 2, and 6) were selected, with around 819 students assigned to each booklet. Gender DIF items were identified in the first round of DIF analysis, and then those identified items were grouped based on six item characteristics, in order to explore the sources of the gender DIF effect. As a result, algebra was found to favoring male eighth graders, but not fourth graders.

Yan (2005) examined the following content areas as a potential source of gender DIF: fraction/number sense, measurement, data representation/ analysis/probability, geometry, and algebra. She concluded that females showed a statistically significant advantage over males on algebra items. However, geometry items did not favor either gender. In the studies of Boughton, Gierl, and Khaliq (2000) and Mendes-Barnett and Ercikan (2006), geometry was not found as a source of gender DIF as well.

In general, studies of gender DIF regarding content domains have yielded inconclusive findings, as Feng (2008) concluded. For algebra, mixed results were found in favor of both males and females. Geometry appears to favor males over females. However, more research is needed to clarify the picture. Grade levels may serve to moderate the gender effect. A content-domain-specific and grade-level-specific combination, which narrows down the scope of investigation, may be a better way to detect the direction and magnitude of the gender DIF effect due to content domain.

2.1.2.4 Gender Differential Item Functioning in Cognitive Demand

2.1.2.4.1 Models of Cognitive Demand

Cognitive demand is defined as cognitive or mental skills required in order to responding to items correctly (Mullis et al., 2009). In the past few decades, educators have attempted to develop systematic models to understand cognitive demand in mathematic learning (Kaplan & Plake, 1982), because it is closely associated with the design of instruction and assessment. Table 1 summarizes several models researchers have developed to describe cognitive abilities associated with learning, especially in mathematics.

Table 1

Models Describing Cognitive Demand Skills Required in Mathematics Learning

Developer	Content of Cognitive Demand Level
Bloom (1956)	6 levels: • Knowledge • Comprehension • Application • Analysis • Synthesis • Evaluation
National Assessment of Educational Progress (NAEP) (2006)	3 levels: • Low complexity • Moderate complexity • High complexity
Andrew Porter (2006)	5 levels: • Memorize • Perform procedure • Demonstrate understanding • Conjecture, generalize, and prove • Solve non-routine problems, and make connections
Norman Webb's Depth of Knowledge (DOK) (2006))	4 levels: • Recall • Basic application in skills and concepts • Strategic thinking • Extended thinking
Trends of International Mathematics and Science Study (TIMSS 2011)	3 levels: • Knowing: Recall facts and concepts; perform procedures and knowledge. • Applying: Apply knowledge and conceptual understanding to solve familiar, textbook-related problems. • Reasoning: Apply knowledge and conceptual understanding to solve unfamiliar, real world, and multi-step problems.
Harris and Carlton (1993)	6 levels: • Recall factual knowledge • Perform mathematical manipulations • Solve routine problems • Demonstrate comprehension of mathematical ideas and concepts • Solve non-routine problems requiring insight or ingenuity • Apply "higher" mental process to mathematics
Mendes-Barnett and Ercikan (2006)	3 levels: • Recognize a correct answer after performing a single computation, substituting values into a formula, or identifying specific characteristics of a graph or diagram • Form and solve equations, manipulate expressions, produce a graph or diagram, or interpret a graph or diagram • Analyze, synthesize, and evaluate

In the 1950's, Bloom and a group of researchers identified three aspects of educational objectives, which were *cognitive* – mental skills; *affective* – feelings or emotion; and *psychomotor* – manual or physical skills (Bloom, 1956). Cognitive objectives involve the development of intellectual skills for learning knowledge. Bloom identified six levels of cognitive mental skills: *knowledge*, *comprehension*, *application*, *analysis*, *synthesis, and evaluation*. *Knowledge* is classified as the lowest level of cognitive demand skills, and *evaluation* is classified as the highest level of mental skills. In addition, verbs were generated to represent the cognitively mental activity involved in each level of cognitive demand. For example, Level I, *knowledge*, includes test items that ask students to define, describe, duplicate, label, list, and so on. Level III, *application*, consists of verbs such as apply, choose, demonstrate, discover, and so on. Level IV, *synthesis*, involves verbs like arrange, collect, combine, create, design, and develop, and so on (Bloom, 1956).

Many others, following in Bloom's steps, have developed different models to describe the cognitive skills demanded in a variety of learning settings. In mathematics, for instance, the NAEP used *low*, *moderate*, and *high* levels of cognitive demands to describe the mental skills required for answering mathematics items correctly. A *low* level of cognitive demand relies heavily on the abilities of recall and recognition of previously learned fact, concepts, principles, and procedures. A *moderate* level of cognitive demand involves more flexible thinking and decision-making among alternatives. A *high* level of cognitive demand involves mental skills in planning, analysis, judgment, creative thought, and abstract reasoning (Hess, 2006).

The cognitive demand model developed by Porter involves five hierarchical levels of cognitive demand: (1) *memorize*: the ability to recall basic mathematics facts, etc.; (2) *perform procedures*: the ability to perform computational procedures or algorithms, etc.; (3) *demonstrate*

understanding: the ability to communicate mathematical ideas, using representations to model ideas, etc.; (4) *conjecture, generalize, and prove*: the ability to evaluate the truth of a mathematical pattern or proposition, write a formal or informal proof, etc.; and (5) *solve non-routine problems, and make connections*: the ability to apply and adapt proper strategies to solve mathematics problems, and so on (Hess, 2006).

Norman Webb's Depth of Knowledge (DOK) Levels is another cognitive demand model used to describe the mental abilities required to master learning in mathematics. Four levels are identified including (1) *recall*: recall or recognition of facts, concepts, information, and procedures; (2) *basic application in skills and concepts*: using information and concepts, following or choosing correct procedures, organizing or displaying data, etc.; (3) *strategic thinking*: reasoning, developing a plan or procedures to solve problems, decision making and justification, developing more than one possible answer, etc.; and (4) *extended thinking*: investigating or applying to real world situations, processing multiple conditions of problems or tasks, taking time to research and think, non-routine manipulating ability, and so on (Hess, 2006).

The TIMSS assessment program developed its own cognitive demand model to describe levels of mental skill elicited in learning mathematics (see Table 1). That model specifies three levels of cognitive demand: *knowing*, *applying*, and *reasoning* (Mullis et al., 2009). The first demand, *knowing*, is associated with the items testing "the facts, procedures, and concepts students need to know" (p. 40). The second demand, *applying*, includes items related to "the ability of students to apply knowledge and conceptual understanding to solve problems or answer questions" (p. 40). The third domain, *reasoning*, includes items that go "beyond the solution of routine problems to encompass unfamiliar situations, complex contexts, and multi-

step problems" (p. 40). Each content domain[5] includes items developed to address all the three levels of cognitive demand.

The cognitive demand skill *knowing* provides a knowledge base, including facts, procedures, and concepts that enables students to establish the foundations of mathematics. Without such a fundamental knowledge base, students would find mathematical learning difficult. Facts consist of the basic language of mathematics, essential mathematical facts, and the properties that shape the foundations of mathematical thought. Procedures connect basic knowledge so that students can use that connected knowledge for problem solving. Concepts allow students to make connections among elements of knowledge they have learned, and further allow them to judge, evaluate, and analyze mathematics statements and methods. As defined by the TIMSS mathematics framework, this level of cognitive demand skill covers the following six sub-skills: *recall*, *recognize*, *compute*, *retrieve*, *measure*, *and classify/order* (Mullis et al., 2009).

The cognitive demand skill *applying* enables students to apply mathematics knowledge of facts, procedures, and concepts (learned by cognitive demand of *knowing*) to create mathematics representations and solve problems. *Applying* focuses on problem solving that is more routine than that included in the next cognitive demand level of *reasoning*, which is described in the next paragraph. The routine problems are typically mathematics exercises that teachers teach more frequently in the classroom. They are designed to provide practice in particular methods or techniques. In addition, these types of "textbook" problems are expected to be familiar to students who can use learned facts, procedures, and knowledge of concepts to solve those problems (Mullis et al., 2009). As described in the TIMSS 2011 mathematics framework, this

[5] There are four content domains, including *Number*, *Algebra*, *Geometry*, and *Data and Chance*, in TIMSS 2011 mathematics test for Grade 8 (Mullis et al., 2009).

cognitive demand level covers the following five sub-skills: *select*, *represent*, *model*, *implement*, and *solve routine problems*.

The cognitive demand skill *reasoning* enables students to use logical and systematic thinking and to solve non-routine problems. Non-routine problems are questions unfamiliar to students, compared to familiar problems required for the cognitive demand skill *applying*. Non-routine problems require a higher level of cognitive demand from students. Students need the mental ability to transfer skills and knowledge they have learned to new, creative, or complex situations. Problem solving is included not only in the *applying* cognitive demand, with emphasis on the more familiar and routine tasks, but also in the *reasoning* domain, which highlights solving novel or complex problems. According to the TIMSS 2011 mathematics framework, this level of cognitive demand covers the following five sub-skills: *analyze*, *generalize*, *synthesize/integrate*, *justify*, *and solve non-routine problems*. More details about the sub-skills that make up the cognitive demand model were provided in the TIMSS 2011 assessment framework (Mullis et al., 2009).

Harris and Carlton (1993) developed a 6-point hierarchy of cognitive demand. The model contains six levels of cognitive demand ranging from the lowest cognitive demand to the highest cognitive demand. The coding categories for assessing levels of cognitive demand for a given test item are summarized in Table 1.

Mendes-Barnett and Ercikan (2006) developed a three level model of cognitive demand based on a modified version of Bloom's Taxonomy of Educational Objectives (See Table 1). The three levels of understanding and their applications to test items are shown in Table 1. The items at the *lower* cognitive demand level required students to recognize a correct answer after performing a single computation, substituting values into a formula, or identifying specific

characteristics of a graph or diagram. The items at the *middle* cognitive demand level required students to form and solve equations, manipulate expressions, produce a graph or diagram, or interpret a graph or diagram. The items identified with the *higher* cognitive demand level required students to analyze, synthesize, and evaluate.

2.1.2.4.2 Gender DIF Studies Related to Cognitive Demand

The main purpose of a gender DIF analysis around cognitive demand levels is to understand whether cognitive demand affects gender DIF effects. Research findings on the DIF patterns attributable to cognitive demand are inconsistent or questionable partially because the cognitive demand models used to describe cognitive or mental skills elicited to solve mathematics items have differed.

For example, Harris and Carlton (1993), using a 6-level cognitive demand model (see Table 1), concluded that males outperformed females on items at higher levels of cognitive demand, whereas females outperformed males on items requiring lower levels of cognitive demand. In their investigation, 181,228 male and 198,668 female students high school students, for whom English was self-reported to be the best language, took each of six Scholastic Aptitude Test (SAT) mathematics tests. Each form consisted of 60 items: 40 problem-solving items and 20 quantitative comparison items, with 360 items in total involved in the DIF research. The Mantel-Haenszel procedure (Holland & Thayer, 1988) was used to investigate gender DIF items. The researchers estimated delta values, which are commonly used in the Mantel-Haenszel DIF method, not only for individual items but also for items grouped in each of the six cognitive demand levels.

The mean delta values estimated for each of six cognitive demand levels were 0.66, 0.31, 0.13, -0.01, -0.09, and -0.14. The mean delta values showed the direction and magnitude of DIF

in favor of either males or females. A positive mean delta indicated items favoring females, and a negative measure indicated test items to be in favor of males. Subsequently, a one-way analysis of variance (ANOVA) was used to compare those estimated mean delta values for items classified into the six cognitive demand levels. The ANOVA test was statistically significant, and a multiple comparison procedures revealed that this effect was due to the difference between levels 1 and 6 (i.e., 0.66 vs. -0.14). The study concluded that there is a consistent and significant shift in relative performance as cognitive demand level increases. In addition, the study concluded that female students outperformed males on mathematics items requiring lower mental processing, whereas males outperformed females on items requiring higher levels of mental processing.

However, this conclusion does not seem to be adequately supported by the results of the statistical analysis conducted in the research. The ANOVA involving six levels of cognitive demand may not allow us to draw such a conclusion. Actually, Instead of an omnibus test, the gender performance patterns should be evaluated according to the DIF effect sizes that in this case reached statistical significance. This dissertation re-examined the six mean gender DIF magnitudes, and found that none reached statistical significance. Thus, this study should have concluded that items differing in cognitive demand level did not favor males or females. It should be noted that the cognitive coding system in the study represented cognitive demand levels increasing from one to six. The information supporting the claim that the cognitive demand abilities differed, were ordinal, and were increasingly complex was not available. Such information should be provided for readers to better understand the coding system and use it when needed in further research.

In the gender DIF study by Mendes-Barnett and Ercikan (2006), the DBF module of the SIBTest DIF method was used to investigate the source of gender DIF in mathematics items. Forty-five mathematics items were grouped by lower, middle, and higher levels of cognitive demand of (see Tables 1 and 4). All items were multiple-choice and dichotomously scored. The study only focused on uniform DIF items. In the study, the samples were grade 12 students, including 5,069 males and 4,335 females, with no ELL students involved in the study. Although they initially found gender DIF only in the highest cognitive level of items that favored males, they refined their conclusion in the follow-up investigation in which they took a closer look at the content domain. They reported that items in the higher cognitive level that related to algebra, arithmetic, and computation exhibited no DIF, whereas DIF with an effect favoring males were only found with items related to geometry, measurement, and the context of spatial ability or spatial visualization. The results indicated the gender DIF effect of higher level items might vary depending on the mathematics content area.

Two studies (Gallagher & DeLisi, 1994; O'Neil & McPeek, 1993) that explored the relationship between gender DIF and how items were represented are introduced next. It is possible to apply a cognitive demand model to the items analyzed in these studies, to see what the results might suggest about gender DIF due to cognitive demand even though neither study included the cognitive demand model in approaching the DIF patterns.

O'Neil and McPeek (1993) reviewed a number of large samples DIF studies that used several DIF methods on the ACT, GMAT, GRE, NTE and SAT tests. Their review concluded that males generally did better on word problems that are situated in actual contexts, while females did better on abstract, pure mathematics items that focused on formula, equation, or theory. The key point is that these word problem items were real-world relevant, usually not taught in

classrooms, not ordinarily found in textbooks, and may have involved insightful or novel solutions. Based on the description of the word problem items, it is likely such items require the higher level ability of *reasoning* as defined by the TIMSS 2011 cognitive demand framework.

Gallagher and DeLisi (1994) investigated gender DIF using the Mantel-Haenszel DIF method on the SAT mathematics tests of junior and senior high school students who scored at or above 670. The findings indicated that males performed better than their ability-matched female counterparts on unconventional problems, whereas females performed better on conventional problems. In the study, conventional problems tended to be routine items typically found in textbooks. In contrast, unconventional problems were items that required an unusual use of a familiar algorithm, an atypical solution strategy, or some type of estimation or insight. According to the cognitive demand model developed by the TIMSS 2011 assessment program, the middle level cognitive demand skill of *applying* was likely elicited to solve conventional problems, and the higher level cognitive demand ability of *reasoning* was likely required to solve unconventional problems, reported by Gallagher and DeLisi (1994) that favored males.

According to the literature review, the cognitive demand models used in gender DIF analyses have differed. At minimum, the number and label of cognitive demand levels are inconsistent among developed cognitive models. Fortunately, however, the models implicitly reveal something in common. The tendency is that items that require individuals to recall or recognize learned concepts, facts, fundamental knowledge, procedures, and principles tend to occupy the *lower* levels in a cognitive demand model. Test items that require individuals to apply what they have learned with concepts, facts, knowledge, procedures, and principles to solve so-called familiar problems tend to occupy the *moderate* level in a cognitive demand model. Finally, items that tend to be in the *higher* level of a cognitive demand model require individuals to use

what they have learned in facts, procedures, and knowledge to solve more complicated, unfamiliar, non-routine, or novel items, compared to those items requiring moderate cognitive demand ability (see Table 1).

Among the gender DIF studies involving cognitive demand models, studies differed in DIF detection methods, the source and the size of student samples, and cognitive demand models used (Harris & Carlton, 1993; Mendes-Barnett & Ercikan, 2006). According to the literature review, it seems reasonable to conclude that mathematics items do not favor either males or females across cognitive demand levels. If there is a gender DIF effect, the effect is expected to be favoring males on items in higher cognitive demand levels.

In addition, the literature remains limited and unclear about the gender DIF effect when additional item characteristics, such as item type and content domain, are added to the cognitive demand models. A number of important gaps in the literature remain. First, additional item characteristics might moderate the effect of cognitive demand levels on gender DIF. Second, most of the studies looking at gender DIF differences used multiple-choice items only, so there may be an interaction between item type and cognitive demand. Third, most of the gender difference studies were based on U.S. samples. Fourth, findings were inconclusive, due to inconsistent research methods (e.g., different cognitive demand model and DIF detection methods) being used. This dissertation is intended to fill part of these particular research gaps.

2.2 Detecting Differential Item Functioning (DIF)

2.2.1 Overview of Differential Item Functioning

Researchers have utilized many different statistical methods to investigate gender differences in mathematics tests. In general, some studies compare the average total scores between genders, while others compare average performances of males and females on particular

subsets of scores. In contrast to using average total scores, another possibility for exploring gender differences is the use of item-level analyses. According to Linn and Hyde (1989), mean differences in individual items may vary because mean-level gender differences are not homogeneous across test items.

Traditionally, two types of item-level analyses can be studied: *impact* and DIF. The comparison of performance differences between groups of examinees on individual items based on unequal abilities between genders is called an impact study. In such a study, the comparison does not require groups of examinees to be matched in their ability. On the other hand, comparing performance differences on individual items based on equal abilities between genders is called a DIF study.

Item impact occurs when test takers in different gender groups show differential correct response rates and there are true differences between groups in the ability being measured by the item. However, DIF occurs when test takers between groups show differing correct response rates and there are *no true* differences in ability between groups. In other words, DIF violates what we assume: there are equal correct response rates on an item for groups of examinees ability-matched.

Angoff (1993) referred to item impact as the "true" difference between the compared groups, while an item with DIF is considered an "artifactual" difference between groups. "Artifactual" highlighted that a difference between groups results from problematic items involved in a test. A DIF study compares item-level performance differences after matching test-takers on the ability the test intended to measure. In other words, forming groups of examinees with equal abilities is required for a DIF analysis.

However, not all identified DIF items are inappropriate or problematic. It should be noted that items with DIF might not be biased items. When DIF is apparent, its presence alone is not sufficient to declare item bias; rather, "…one would have to apply follow-up item bias analyses (e.g., content analysis and empirical evaluation) to determine the presence of item bias" (Zumbo, 1999, p. 12).

As described above, DIF occurs when examinees with the same ability level, but different group membership, have a different probability of responding to an item correctly. DIF studies are usually used to identify potentially biased items in two groups of examinees. One group is labeled as a *focal group*, which refers to the group of examinees the researchers are interested in, and the other as a *reference group*, which refers to the group of examinees selected by the researcher as a reference for the comparison.

The procedure of "ability matching" is crucial to match examinees and form a pair of equal groups for a DIF study. DIF analyses can use the total test score, or model-based strategy, to estimate the so-called true or latent ability level for test-takers. When groups of examinees are matched on their achieved total scores, the tests are assumed to be homogeneous and unidimensional – that is, measuring a single ability (Nandakumar & Stout, 1993; Roussos & Stout, 1996; Stout, 1990). However, researchers have suggested that the assumption that tests are unidimensional might not be tenable. If unidimensionality is not assumed, tests area treated as multidimensional scales and different procedures to detect DIF items are used (Gierl et al., 2003; Mazor, Hambleton, & Clauser, 1998; Shealy & Stout, 1993b).

For example, if a mathematics test requires only computational ability to respond to test items correctly, the test would be unidimensional. If a mathematics test required one or more abilities (e.g., reading comprehension ability and computational ability) to respond items

correctly, the test would be multidimensional. Item Response Theory assumes that tests are unidimensional. However, more experts are supporting the idea that most achievement and aptitude tests are actually multidimensional (Shealy & Stout, 1993b). Using unidimensional DIF procedures on multidimensional tests violates the unidimensionality assumption, resulting in more DIF items identified (Clauser, Nungester, Mazor, & Ripkey, 1996). Thus, if a test requires more than one ability to respond to items correctly, test developers should use a scoring model that reflects all dimensions of the ability score so that false positive DIF items can be reduced (Taylor & Lee, 2012).

For any DIF investigation, the first step is to identify the groups of interest. Gender, ethnic, or cultural groups can be selected as a focus (Zhang, 2001). Then, the reference and focal groups are matched on the measured trait. Typically, one item at a time will be examined for DIF item detection, until the whole test is analyzed.

For large-scale assessment projects, DIF evaluation is a routine part of standard procedure (Taylor & Lee, 2012). Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985) highlight the importance of operational use of a test to investigate potentially biased items and identify the source(s) of the bias. This statement implies that after the source of DIF is identified, guidelines to create quality items will be developed, leading to a quality test that can provide a better estimate of the true abilities of test-takers. However, the techniques for DIF studies are still developing, and there is no consensus on a rigorous statistical method to identify DIF due to the inconsistency among the numerous DIF methods.

2.2.2 Methods for Detecting Items with Differential Item Functioning

Concern about DIF items has led measurement professionals to develop various methods for investigating such occurrences. DIF analyses are typically conducted at the individual item

level.[6] It is assumed that the absence of DIF items will lead to an unbiased test. Statistical methods proposed for identifying DIF can be classified into two groups. One group employs an IRT-oriented approach, which is based on model-based procedures or true-score approaches. The other group employs non-IRT theory, which tends to rely on observed score approaches. Both IRT- and non-IRT-based DIF methods first create ability-matched groups of examinees, and then determine whether examinees in each group have equal success rates for an item, at all ability levels.[7]

One apparent difference between these two approaches is that in IRT-based methods the item parameters, *a* (discriminant index), *b* (difficulty index), and/or *c* (guessing index), must be estimated, whereas item parameter estimation is not required for the non-IRT based DIF methods. IRT-based approaches can be further classified based on the following techniques: (1) Lord chi-square test, (2) IRF (Item Response Function)/ICC area comparison method, (3) likelihood ratio test (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). For each method, researchers can select the number of parameters (i.e., 1PL, 2PL, or 3PL) involved in the DIF item detection process. For example, the SIBTest method is an IRT-based approach with 1PL involved in the DIF analysis.

Non-IRT based approaches include the following: (1) Mantel-Haenszel method (Holland & Thayer, 1988; Mantel-Haenszel, 1959), (2) Logistic Regression method (Swaminathan & Rogers, 1990), (3) Standardization method (Dorans & Kulick, 1986), and (4) Delta Plot/Transformed Item-difficulty method (Angoff, 1982a, 1982b).

---

[6] There are methods developed to examine bias at the level with a group of items, such as differential test functioning (DTF) (Camilli & Penfield, 1997; Shealy & Stout, 1993a).

[7] Another way to label the techniques developed to detect DIF items is based on parametric or nonparametric methods. Parametric methods assume there is an underlying item response model (e.g., based on the item response theory of measurement) to support the developed DIF detection method, while the nonparametric methods do not make such an assumption.

Whether IRT or non-IRT based approaches are used, each method has its own advantages and weaknesses. In addition, each method has corresponding statistical models and software packages to perform the analysis.

The next section will describe four commonly used methods in gender DIF item detection: Mantel-Haenszel, logistic regression, SIBTest, and multilevel DIF. I will discuss the strengths and weaknesses of these four methods.

2.2.2.1 Mantel-Haenszel

Mantel-Haenszel (MH; Mantel-Haenszel, 1959) uses chi-square statistics in which a full range of total test scores from examinees are divided into equal score intervals (e.g., high, medium, and low level of proficiency). Test takers between reference and focal groups are compared at each proficiency level for each item by means of an odds ratio. Typically, the reference group represents examinees suspected to be favored by test items, whereas the focal group represents examinees suspected to be at a disadvantage (Roussos & Stout, 1996). Multiple 2 x n tables are created (where n is the number of score points possible for the item, with $n = 2$ for dichotomous items and $n > 2$ for polytomous items) – one table for each proficiency level – and chi-square statistics (α_{MH}) are calculated for the tables across all proficiency levels for a given item (Smith, 2009).

Often five ability levels are created, but there can be as many proficiency levels (k) as the researchers expect. The reference and focal groups are matched by their proficiency levels, and each performance level has its own contingency table. Multiple (k) contingency tables for a given dichotomous item can be viewed in Table 2, where k is the number of subgroups according to the number of proficiency levels determined by the researchers; T_k is the total number of examinees at a specific proficiency level; n_{Rk} *and* n_{Fk} are the number of examinees in reference

and focal groups, respectively; m_{1k} and m_{0k} are the number of examinees that answered correctly or incorrectly to the item, respectively; A_k and C_k are the number of examines that answered the item correctly for the reference and focal groups, respectively; and B_k and D_k are the number of examines that answered the item incorrectly for the reference and focal groups, respectively (Smith, 2009).

Table 2

Example of a Contingency Table for k Proficiency Levels for a Given Item with Mantel-Haenszel Method

	Item Score		
Group	Right (1)	Wrong (0)	Totals
Reference Group (*R*)	A_k	B_k	n_{Rk}
Focal Group (*F*)	C_k	D_k	n_{Fk}
Totals	m_{1k}	m_{0k}	T_k

The null hypothesis for the Mantel-Haenszel test states that the odds of getting an item correct across all levels of the matching variable are the same for all subgroups of the focal and reference groups, expressed as either

$$H_0: \frac{A_k}{B_k} = \frac{C_k}{D_k} \quad \text{k} = 1, \ldots, \text{K} \tag{2.01}$$

or

$$\alpha_{MH} = \frac{\sum_k \frac{A_k D_k}{T_k}}{\sum_k \frac{B_k C_k}{T_k}} = 1 \tag{2.02}$$

The parameter α_{MH} is referred to as the common odds ratio or contrast odds ratio, which yields a measure of the effect size for evaluating the magnitude of DIF. One formula,

developed by Mantel and Haenszel (1959), to detect the presence of DIF for α_{MH} is expressed as:

$$\chi^{2}{}_{MH} = \frac{\left[\left|\sum_{k} A_k - \sum_{k} E(A_k)\right| - 0.5\right]^2}{\sum_{k} Var(A_k)}, \qquad (2.03)$$

where $E(A_k)$ is the expected value of A_k, and $Var(A_k)$ is the variance of A_k. 0.5 is the Yates' correction for continuity (Yates, 1934), to prevent overestimation of statistical significance for small data. This formula is chiefly used when at least one cell of the table has an expected count smaller than 5.The chi-square estimate, x^2_{MH}, is tested based on one degree of freedom (for dichotomous items) for the chi-square distribution. If no DIF or bias is present in the item, the resulting chi-square statistic will not be significant. Items with significant chi-square results will be flagged for the presence of DIF (Angoff, 1993; Dorans & Holland, 1993).

The chi-square estimate, x^2_{MH}, is sensitive to the sample sizes. Small x^2_{MH}, which describes a trivial difference between focal and reference groups, may reach statistical significance because of the large sample size often used in DIF studies. In order to solve the sample size problem, α_{MH} is often transformed to $\hat{\Delta}_{MH}$ using the formula $\hat{\Delta}_{MH} = -2.35 \ln(\alpha_{MH})$, to enhance the interpretability of the results and reduce the effect of the large sample size (Holland & Thayer, 1988). $\hat{\Delta}_{MH}$, called MH D-DIF (Holland & Thayer, 1988), was developed by Educational Testing Service (ETS) to described three categories of DIF magnitude ranging from negligible, moderate, to large, and labeled A, B, and C. The positive or negative sign indicates either the reference or focus group being favored by the DIF item, with the aim of obtaining a symmetrical scale in which a zero value of $\hat{\Delta}_{MH}$ indicates an absence of DIF.

Based on this transformation, the following interpretation guideline is used to evaluate the DIF effect size (Dorans & Holland, 1993; Zieky, 1993; Zwick & Ercikan, 1989; Zwick, Thayer, Lewis, 1999), taking both x^2_{MH} (with an α = .05 significant level) and the absolute value of $\hat{\Delta}_{MH}$ into consideration:

• Negligible or A-level DIF: $|\hat{\Delta}_{MH}| \leq 1.0$, or x^2_{MH} test is not statistically significant,

• Moderate or B-level DIF: $1< |\hat{\Delta}_{MH}| < 1.5$, and x^2_{MH} test is statistically significant,

• Large or C-level DIF: $|\hat{\Delta}_{MH}| \geq 1.5$, and x^2_{MH} test is statistically significant.

Zwick and Ercikan (1989) suggested that B-level (moderate DIF) items could be used in the test if there are no replacement items available, and that C-level, or large DIF items could be selected if test specifications cannot be achieved without them. The criteria ETS developed to evaluate what constitutes a DIF item are summarized in Table 3.

Table 3

Levels of DIF Magnitude Developed by Educational Testing Service

	x^2_{MH} significant ($p \leq .05$)	x^2_{MH} not significant ($p > .05$)
$\|\hat{\Delta}_{MH}\| \geq 1.5$	C	A
$1< \|\hat{\Delta}_{MH}\| < 1.5$	B	A
$\|\hat{\Delta}_{MH}\| \leq 1.0$	A	A

The Mantel-Haenszel method works well for small sample sizes. The Mantel-Haenszel method for dichotomous items can be connected to either one-, two- or three-parameter IRT-

based logistic function (Donoghue, Holland, & Thayer, 1993; Roussos, Schnipke, & Pashley, 1999; Spray & Miller, 1992; Zwick et al., 1999; Zwick, Thayer, & Wingersky, 1994). In addition, it can be extended to detect DIF for polytomous items (Smith, 2009). This extension is referred to as the Generalized Mantel-Haenszel (GMH) method (Allen & Donoghue, 1996). The analysis examines a 2 (sub-populations) × *j* (scoring categories) × *k* (categories in matching criterion) contingency table similar to Table 2, but with more than two (e.g., correct and incorrect) scoring categories.

The claim that the Mantel-Haenszel method is insensitive to non-uniform DIF may be exaggerated (Rogers, 1989). Non-uniform DIF items can be identified by the Mantel-Haenszel method. Items that are easy or difficult and display non-uniform DIF are fairly and consistently identified by the Mantel-Haenszel method. Items that are of medium difficulty and exhibit non-uniform DIF are more likely to be missed, because the item characteristic curves between compared groups cross each other close to the middle of the ability range. The DIF effects are cancelled out as a result of the statistical models developed to perform the Mantel-Haenszel method. A possible solution is to use an adapted Mantel-Haenszel method, developed by Mazor, Clauser, and Hambleton (1994), that splits a dataset into two datasets by breaking the full sample at approximately the middle of the test score distribution and reanalyzing with both the low-performing sample and the high-performing sample. Mazor et al. (1994) demonstrated that the adjusted Mantel-Haenszel method was useful and improved the identification rates of non-uniform DIF items without increasing the Type I error rate. In addition, when using such an adjusted method, items with larger difficulty and discrimination parameters are more likely to be identified as non-uniform DIF items (Hidalgo & Lopez-Pina, 2004).

A number of simulation studies have shown that Mantel-Haenszel statistics are sometimes overestimated or underestimated due to factors such as the choice of matching variables, the amount of DIF items, guessing, and sample size (Brennan, 2006). Swaminathan and Rogers' simulation study (1900) indicated that small sample sizes result in inadequate recovery of DIF and underpowered statistics. This finding was confirmed by Paek and Guo (2011) that the Mantel-Haenszel DIF method gave adequate results with large sample sizes when the compared groups were asymmetrically unbalanced in sizes.

Unlike the logistic regression DIF and SIBTest DIF methods, the Mantel-Haenszel method requires ability to be split into a number of levels. Donoghue and Allen (1993) concluded that *thin matching*, where each total test score is used as an ability level, may lead to poor results when compared to *thick matching*, where ability is spilt into a number of levels specified by researchers. Zilberberg, Phan, Socha, Kong and Keng (2011) explored how Type I error rates and statistical power were affected by matching type, total sample size, and sample size ratio between groups in a total of 48 (4 x 3 x 4) simulated conditions. The simulation included 62 multiple-choice items, with the item difficulty parameters developed according to a real state high-school math test. Ten DIF items, varying in magnitude of uniform DIF effects and all favoring one group of subjects, were developed, and equal abilities between groups were specified with a mean of 0 and standard deviation of 1. For each of the 48 conditions, 100 replications were generated.

This study concluded that the quartile matching type, with abilities split into four levels, performed best in the recovery of DIF and non-DIF items than other three types (i.e., deciles, two-adjacent, and four-adjacent matching). As expected, larger total sample size led to better classification rates across all matching types. Moderate magnitudes of DIF, equivalent to the B-

level in the ETS DIF classification system, are the most difficult to detect across sample sizes and matching types. This study noted that numerous factors, including test length, item difficulty level, percentage of items with DIF, proficiency level between ability-matched groups, the use of purification techniques, percentage of items favoring either group, or the normality of abilities for either group, potentially contribute to the Mantel-Haenszel DIF technique's ability to detect DIF, and merit further investigation (Zilberberg et al., 2011).

2.2.2.2 Logistic Regression

Swaminathan and Rogers (1990) introduced a method to identify DIF for dichotomous items based on logistic regression (LR). Logistic regression belongs to a broad class of mathematical models called generalized linear models (GLM). In logistic regression, the item response is taken as a random Bernoulli variable Y*i*, scored dichotomously as 0 or 1, for *i* individuals. The logistic regression DIF method has become one of the most commonly used DIF detection methods (Clauser & Mazor, 1998; Paek, 2012). The general logistic regression model for DIF detection is expressed as (Gierl, Jodoin, & Ackerman, 2000):

$$E(Y) = p = \frac{e^{\eta}}{1+e^{\eta}} \quad \text{or} \quad E(Y) = p = \frac{1}{1+e^{-\eta}} \tag{2.04}$$

where $E(Y)$ or p is the conditional probability of obtaining a correct answer given a set of independent variables denoted by η. The η is the function that defines the linear combination of the predictor variables and expressed as (Zumbo, 1999):

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(total\_score) + \beta_2(group) + \beta_3(total\_score \text{ x } group) \tag{2.05}$$

In a logistic regression DIF analysis, the dependent variable is the item score, 0 or1, and the independent variables are the observed ability (*total_score*) as a matching criterion, the group variable (*group*), used to signify reference and focal groups, and the group and ability interaction (*total_score* x *group*). Four parameters are estimated: β_0 is the intercept, β_1 is the ability

regression coefficient, β_2 is the regression coefficient for the group variable, and β_3 is the coefficient for the interaction term (Zumbo, 1999).

The total score, defined group memberships, and the interaction terms are the variables used in a regression analysis for predicting the probability of a correct response to each item for a given individual. Testing for the statistical significance of DIF is based on a chi-square test conducted after sequentially adding the independent variables into Equation 2.05. That is, the DIF evaluation is carried out by statistically evaluating the incremental contribution of each successive model in terms of model fit (Zumbo, 1999).

For example, in step 1, a likelihood ratio chi-square is performed for the equation with the observed ability variable *total_score* only. Next, in step 2, a *group* variable is added to the equation. The presence of uniform DIF is determined through a chi-square test following a distribution at $df = 1$. A statistically significant chi-square test indicates the presence of uniform DIF, with the significant variable *group*. Similarly, in step 3, the interaction variable *total_score* x *group* is added, and the equation is tested against step 2, which only had the ability and group variables. The presence of non-uniform DIF is determined with a chi-square test also following a distribution at $df = 1$. A significant chi-square test indicates the presence of non-uniform DIF, with a significant variable *total_score* x *group*. In sum, if the group effect (*group*) is statistically significant and the effect of the interaction (*total_score* x *group*) is not, then the tested item has uniform DIF. On the other hand, if the interaction (*total_score* x *group*) is statistically significant, then the tested item has non-uniform DIF (Zumbo, 1999). A uniform DIF shows the plotted regression lines between two groups of subjects as parallel and indicates that one group of subjects exhibited global higher or lower correct response rates than the other group of subjects, across the ability spectrum. Non-uniform DIF shows the plotted regression lines between two

groups may cross somewhere. Non-uniform DIF items favor one group at higher end of the ability spectrum and the other at the lower end. However, favoring both groups simultaneously is not necessary.

A polytomous item is when the range of item responses available is ordinal (as opposed to nominal), rank-ordered, and includes more than two possible score points. An open-ended item, with partial credits assigned to test-takers, is a typical example of a polytomous item (Ilich, 2013; Zumbo, 1999). Swaminathan and Rogers' (1990) logistic regression approach for DIF detection of dichotomous items can be extended to handle polytomously scored items. Extending the analysis to polytomous items requires the use of link functions and dichotomizes the item response category (French & Miller, 1996; Miller & Spray, 1993).

French and Miller (1996) summarized three dichotomization methods for the logistic regression models (O'Connell, 2006; Wilson, 1993): (1) *cumulative odds*, denoted as $P(Y \leq j)$ (Agresti, 1996), (2) *continuation ratio*, denoted as $P(Y < j \mid Y \geq j)$ (Greenland, 1994), and (3) *adjacent categories*, denoted as $P(Y = j + 1 \mid Y = j)$ (Agresti, 1989). These three all closely follow the procedures and strategies of both binary logistic and ordinary least squares regression analysis. Logistic regression DIF analysis for polytomous items uses the same link Equation 2.05 described above but must be analyzed multiple times, based on how ordinal response items are dichotomized within the number of score levels available. For example, for an item that can be scored as 0, 1, 2, 3, or 4 points, four sets (i.e., the number of levels minus 1) of multiple logistic regression analyses would be required to detect DIF effects no matter which logistic regression model selected for handling polytomous items. Wilson, Spray, and Miller (1993) compared these three dichotomized models in their abilities of detecting non-uniform DIF items.

Detecting DIF in polytomous items is more complex than in dichotomous items, as the DIF tests potentially need to be performed multiple times according to the number of response category involved in the items (Kristjansson, Aylesworth, Mcdowell, & Zumbo, 2005). Making an overall decision on how best to identify DIF items based on a number of separate regressions tends to be difficult (French & Miller, 1996; Wilson, 1993). Thus, the ordinal logistic regression DIF model, developed by Zumbo (1999) for polytomous items, that is based on constrained cumulative odds models, offers an easier option to use without running multiple times of regressions.

Actually, there are two types of cumulative odds logistic regression models: *constrained* (e.g., the Zumbo's model mentioned above) and *unconstrained*. Constrained cumulative odds models assume equal slopes across multiple instances of binary logistic regression for a polytomous item, so that a consistency of effect across all response categories is assumed, and one omnibus test can be performed with a single parsimonious logistic regression model for the dataset (O'Connell, 2006). The constrained cumulative odds model is referred to as a proportional odds model, which Zumbo used as a base to develop this ordinal logistic regression DIF detection method (Ilich, 2013). The constrained cumulative odds logistic regression model is the most frequently used ordinal logistic regression model (Hosmer & Lemeshow, 2000) because it simplifies the analysis.

Unconstrained cumulated models needs to be used when slopes among repeated dichotomized logistic regressions for an independent ordinal variable are unequal, such as continuation ration, adjacent categories, partial proportional odds, and multinomial models as alternatives to relax the proportional odds assumption (O'Connell, 2006). For example, because the equal slopes assumption was violated in 75% of the data sets they generated for the

simulation study, Kristjansson et al. (2005) used the unconstrained cumulative logits ordinal logistic regression to detect DIF in polytomous items.

Several criteria have been suggested by researchers to evaluate the magnitude of DIF effects when DIF is detected. Zumbo and Thomas (1997) proposed $\hat{\Delta} R^2$ to quantify the magnitude of uniform and non-uniform DIF effect sizes.

Measuring the magnitude of DIF in logistic regression methods follows the same procedure used to test the presence of uniform and non-uniform DIF items. Without an examination of effect size, ignorable DIF effects could be statistically significant when the analysis is based on large sample size. In logistic regression DIF analysis, the magnitude of DIF, $\hat{\Delta} R^2$, is calculated by comparing the difference of the R^2 value between the steps involved in uniform and non-uniform DIF detection. For example, the comparison of R^2 values between step 1, the model with ability score only, and step 2, with ability and group membership variables, shows the DIF magnitude for a uniform item. The comparison of R^2 between steps 2 and 3, the models with and without the interaction of group membership and ability, reveals the magnitude of DIF effects for a non-uniform item.

Zumbo and Thomas (1997) suggested the following guidelines to evaluate the impact of DIF items:

• Negligible or A-level DIF: $\hat{\Delta} R^2 < .13$ and χ^2 test is statistically significant,

• Moderate or B-level DIF: $.13 \leq \hat{\Delta} R^2 \leq .26$ and χ^2 test is statistically significant,

• Large or C-level DIF: $\hat{\Delta} R^2 > .26$, and χ^2 test is statistically significant.

Jodoin and Gierl (2001) found that the thresholds proposed by Zumbo and Thomas were too conservative, resulting in under-estimation of DIF effects. For example, only 6.8% of items were classified as moderate DIF by Zumbo and Thomas' criteria, whereas 68% of items were

identified as moderate DIF by Jodoin and Gierl's guidelines (Hidalgo & Lopez-Pina, 2004). The alternative criteria Jodoin and Gierl (2001) suggested were based on the effect size measure developed for the SIBTest DIF method. The classification system for the magnitude of DIF they proposed is summarized as follows:

• Negligible or A-level DIF: $\hat{\Delta} R^2 < .035$, or χ^2 test is statistically significant,

• Moderate or B-level DIF: $.035 \leq \hat{\Delta} R^2 \leq .070$ and χ^2 test is statistically significant,

• Large or C-level DIF: $\hat{\Delta} R^2 > .070$ and χ^2 test is statistically significant.

Jodoin and Gierl (2001) also indicated the regression coefficients β_2 and β_3 can be used as measures to describe the magnitude of uniform and non-uniform DIF, respectively. The difference of regression coefficient β_1 between steps 1 and 2 can also be used to identify items with uniform DIF. Based on simulation studies, a 1% - 10% difference between β_1 coefficients was proposed as a practically meaningful effect of uniform DIF (Choi, Gibbons, & Crane, 2011).

Monahan, McHorney, Stump, and Perkins (2007) also proposed guidelines to evaluate the magnitude of DIF effect sizes. Accordingly, $e^{(\beta_2)}$ is the estimated odds ratio of getting item *i* correct for the logistic regression DIF methodology, and is denoted as $\hat{\alpha}_{LR}$. Subsequently, effect sizes of DIF items are converted to a delta scale based on the formula, $\hat{\Delta}_{LR} = -2.35 \ln(\hat{\alpha}_{LR})$, and the derived magnitudes are then categorized according to the Educational Testing Services (ETS) DIF classification system (Dorans & Holland, 1993). DIF items were placed into one of the following three classes:

• Negligible or A-level DIF: $|\hat{\Delta}_{RL}| < 1.0$, or χ^2 test is not statistically significant,

• Moderate or B-level DIF: $1 \leq |\hat{\Delta}_{LR}| \leq 1.5$, and χ^2 test is statistically significant,

• Large or C-level DIF: $|\hat{\Delta}_{LR}| > 1.5$, and χ^2 test is statistically significant.

Alternately, the formula above can be converted to a predicted probability for each member of either the focus or reference group as follows:

$$\Pr = \frac{1}{1 + e^{-(\beta_0 + \beta_1(total\_score) + \beta_2(group) + \beta_3(total\_score \times group))}}, \qquad (2.06)$$

where *total score* refers to a total correct score or so-called observed ability; *group* refers to group membership (i.e., focus versus reference group); the interaction term, *total_score* × *group*, refers to a product of *total_score* and *group*; and the coefficient β_0, β_1, β_2, and β_3 represent the intercept, the weights for the ability, the weights for the group membership, and the weights for the interaction term, respectively. Furthermore, comparing the computed R^2 values between pairs of steps measures the magnitude (i.e., effect size) of the DIF effects.

The logistic regression method has been found to be as powerful as the Mantel-Haenszel method in detecting uniform DIF (Swaminathan & Rogers, 1990). The logistic regression method yields DIF effect sizes similar to the magnitude of α_{MH} (see Equation 2.02) obtained by Mantel-Haenszel method. In addition, it allows examination of both uniform and non-uniform DIF items (Zumbo, 1999).

There are two important extension of applying logistic regression to the DIF analysis. First, logistic regression can be applied to a dataset with a nested (or hierarchical) structure, called multilevel logistic regression or hierarchical generalized linear modeling (HGLM) with a Bernoulli model. HGLM has a variety of sub-models, which can handle different types of datasets that involves continuous, binary, count, multi-nominal, or ordinal dependent variables. The Bernoulli model is one of HGLM's applications. The Bernoulli model is a binomial model with the number of trials, m_{ij}, equal to one (Raudenbush, Bryk, Cheng, Congdon, & du Toit, 2004). Second, the ordinal logistic regression DIF method has been advanced by incorporating

Item Response Theory. This hybrid ordinal logistic regression and IRT approach can be implemented by the *lordif* package. Either IRT-based latent ability or sum total scores can be used as ability-matching criteria to measure uniform and non-uniform DIF (Choi, Gibbons, & Crane, 2006).Thus, using *lordif* in *R*[8] is a good choice when latent ability, instead of observed total scores, needs to be used as a matching variable in a logistic regression DIF analysis.

2.2.2.3 Simultaneous Item Bias Test (SIBTest)

The simultaneous Item Bias Test (SIBTest), developed by Shealy and Stout (1993a, 1993b), is a statistical method to detects items with DIF implemented by a computer software program called DIFPACK.

Both the hypothesis testing statistics and estimators use the estimated beta, which represents the advantage in item correct response rate for one group of examinees over another group at the same ability (θ) level for individual item. The SIBTest first matches examinees in reference and focal groups on ability using true ability scores estimated from the total test score (Roussos & Stout, 1996). All test items are divided into two subsets; one is a matching/valid subtest, and the other a suspect/studied subtest. A matching/valid subtest, which is free of DIF items, provides an estimate of true ability (θ) level for each examinee. The members of focal and reference groups in the same true ability level are compared to each other on a suspect/studied item or a set of suspect/studied items. Because of the likelihood of a different ability distribution between focal and reference, a regression correction procedure is used. Without such a correction procedure, more items would be flagged as DIF, because the focal group tends to have more examinees on the lower end of the scoring range than the reference group (Metcalf, 2002).

[8] *R* is a programming language or statistical computing and graphics (http://www.r-project.org/).

For the SIBTest method, the null hypothesis that states no DIF exists in an item, and is expressed as (Gierl et al., 2000):

H_0: $B(T) = P_R(T) - P_F(T) = \beta = 0,$ (2.07)

where *B(T)* is the difference between the probability of a correct response on a given item for reference and focal groups matched on true scores *T*. $P_R(T)$ is the probability of a correct response on a given item for examinees in the reference group with true scores *T*. $P_F(T)$ is the probability of a correct response for examinees in the focal group with true scores *T* (Gierl et al., 2000). In other words, the SIBTest tests whether the DIF estimate, β, is significantly different from zero when comparing a reference with a focal group, including a regression correction method to ensure examinees are matched on ability (Bolt, 2000; Ilich, 2014).

The test statistic for measuring the null hypothesis is (Gierl et al., 2000):

$$B = \frac{\hat{\beta}_u}{\hat{\sigma}(\hat{\beta}_u)} \tag{2.08}$$

where,

$$\hat{\beta}_u = \sum_{k=0}^{n} \hat{p}_k (\overline{Y}^*_{Rk} - \overline{Y}^*_{Fk}) \tag{2.09}$$

and,

$$\hat{\sigma}(\hat{\beta}_u) = \left\{ \sum_{k=0}^{n} \hat{p}_k^2 [\frac{1}{N_{Rk}} \hat{\sigma}^2(Y \mid k, R) + \frac{1}{N_{Fk}} \hat{\sigma}^2(Y \mid k, F] \right\}^{1/2} \tag{2.10}$$

In these formulae, $\hat{p}_k$ is the proportion of examinees in the focal group who obtained a score *k*; $\overline{Y}^*_{Rk}$ *and* $\overline{Y}^*_{Fk}$ are the adjusted means for examinees in subgroup *k* after using a regression correction procedure; $\hat{\sigma}^2(Y \mid k, R)$ is the sample variance for examinees on the studied item for the reference group with a total score of *k* on the valid subtest; and $\hat{\sigma}^2(Y \mid k, F)$

is the sample variance for examinees on the studied item for the focal group with a total score of *k* on the valid subtest (Gierl et al., 2000; Ilich, 2013; Mckenzie, 2009).

The equations were originally developed to detect DIF for dichotomously scored test items. When evaluating potential DIF for polytomously scored items, the *n* in Equations 2.9 and 2.10 is replaced by *nh*, which indicates the number of points on the valid test rather than the number of items (Chang, Mazzeo, & Roussos, 1996; Taylor & Lee, 2012).

The DIF test statistic, *B*, should be greater than 1.96 or less than -1.96 at the $p < .05$ level of statistical significance for an item to be flagged as a DIF item. Once the item is identified as a DIF item, the magnitude of the DIF measured, $\hat{\beta}_u$, is calculated by taking the average performance difference between groups on an item, across all ability levels. Roussos and Stout (1996) proposed the following interpretation guidelines to evaluate the magnitude of DIF effect sizes (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; McKenzie, 2009):

• Negligible or A-level DIF: $\left|\hat{\beta}_u\right| < .059$, and null hypothesis is rejected,

• Moderate or B-level DIF: $.059 \leq \left|\hat{\beta}_u\right| < .088$, and null hypothesis is rejected,

• Large or C-level DIF: $\left|\hat{\beta}_u\right| \geq .088$, and null hypothesis is rejected.

The SIBTest has also been developed for polytomous (Chang, Mazzeo, & Roussos, 1996) and non-uniform DIF items (Li & Stout, 1996). The Poly-SIBTest is an extension of the basic SIBTest method, which can analyze open-ended items with more than two scoring points (Taylor & Lee, 2012).

The SIBTest method is suitable for large or small sample sizes because results can be sample independent (as in Mantel-Haenszel method), and it provides a statistically significant

test of the magnitude of DIF. In addition, it can identify non-uniform DIF, unlike the Mantel-Haenszel methods, which work better to detect uniform DIF items.

The ability of the SIBTest to identify DIF and non-DIF items has been examined. In a simulation study by Gierl et al. (2000), Type I error rates and power for DIF detection were investigated among the SIBTest, Mantel-Haenszel, and logistic regression DIF methods. Forty items were generated by using constrained and varying item difficulty parameters, varying discrimination parameters, and fixed guessing parameters across all items. Four factors that potentially contribute to Type I error rates and statistical power were manipulated, including the amount of DIF (2 levels), the magnitude of DIF (3 levels), examinee sample size (3 levels), and the ability difference between groups of examinees (2 levels). Only uniform items were generated in the study; half the DIF items favored one group, and half favored the other group. Each of the 36 conditions was replicated 100 times. An α level of .05 was used for all hypothesis testing.

Two major results were reported. First, a larger number of DIF items, up to 60%, did not adversely affect the Type I error rates. In general, Type I error rates were below the nominal level of .05. Overall, the SIBTest's ability to maintain Type I error rates is in line with that of the Mantel-Haenszel and logistic regression DIF methods. Second, statistical power to correctly identify items with DIF differed across the three methods. When there was a sample size of 250 per group, the power of the three methods was moderate across all conditions. When there was a sample size of 500 per group, power increased significantly, especially for the SIBTest. The study suggested that at least 500 subjects per group should be used to maintain Type I error and power rates, and concluded that the SIBTest exhibited the highest statistical power among the three methods.

2.2.2.4 Multilevel DIF (ML-DIF)

Several multilevel item response models can be found in the literature (Adams, Wilson, & Wu, 1997; Kamata, 1998, 1999, 2001, 2002). Kamata (1998, 1999, 2001) has proposed one of frequently used multilevel IRT-based DIF methods (ML-DIF). His approach features a two-level approach to the study of DIF that incorporates an IRT Rasch perspective, where items are the level-1 unit, and examinees are the level-2 unit. Kamata's methodology described the relationship between the hierarchical generalized linear model (HGLM) and IRT models for dichotomous items. This model is mathematically equivalent to the IRT Rasch model (Burke, 2009).

In the multilevel DIF method Kamata developed to detect DIF, the probability of the correct response on an item is shaped by both item difficulty and the ability of examinees. Due to the model's multilevel-oriented design, examinee-level and group-level measures are allowed to be added to the model. This accommodation allows one to examine the effect of the measures related to the examinees and groups on grand mean performance of all items.

Kamata and Binici (2003) used Kamata's ML-DIF method to detect uniform DIF items. Kamata's ML-DIF method has been extended to a three-level ML-DIF method. In addition, Kim (2003) extended the ability of Kamata's ML-DIF model to identify non-uniform DIF items. Other researchers have been adjusted Kamata's Ml-DIF method to detect DIF for polytomous items (Williams & Beretvas, 2006; Vaughn, 2006).

For a two-level multilevel DIF model from Kamata (1999), Level 1 is represented as follows:

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{(k-1)j}X_{(k-1)ij}, \qquad (2.11)$$

"where p_{ij} represents the probability of examinee *j* getting item *i* correct, and η_{ij} represents log odds of examinee *j* getting item *i* correct. The structural model of Level 1 linearly determines the probability of getting any item correct, except item *k*. The last item, *k*, is used as a reference item. The explicit exclusion of the reference item in Equation 2.11 ensures a full rank of the design matrix (Kamata, 2000). Accordingly, β_{0j} is the coefficient associated with reference item *k*; β_{1j}, β_{2j}, through $\beta_{(k-1)j}$ are coefficients associated with the rest of the items (i.e., items 1, 2, …, *k*-1, respectively); and X_{1ij}, X_{2ij}, through $X_{(k-1)ij}$ are dummy variables associated with items 1, 2, …, *k*-1, respectively. Dummy variables are coded as '-1' when items are being analyzed and '0' otherwise" (Burkes, 2009, p. 19; Kamata, 1999).

Each of the item coefficients, β_j, can be further modeled at Level 2. Level 2 of the two-level DIF model is specified as follows:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Student\_Level\_Measure)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(Student\_Level\_Measure)$$

$$\vdots$$

$$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}(Student\_Level\_Measure) \quad , \qquad (2.12)$$

where γ_{00} is the grand mean score of all items for all examinees; u_{0j} is the random error of β_{0j}. γ_{10}, γ_{20}, through $\gamma_{(k-1)0}$ are the mean performance on items 1, 2, …, *k*-1, respectively. γ_{11}, γ_{21}, through $\gamma_{(k-1)1}$ are the effect of the student-level measure (e.g., SES, gender, ethnicity, and living regions of students) on the mean performance of items 1, 2, …, *k*-1, respectively (Burke, 2009; Kamata, 1999).

The ML-DIF methodology allow us to examine higher-level measures (e.g., teacher, classroom, school, or school district) that are potential sources of DIF. the parameters in Level 2

can be further modeled with a set of level-3 equations. The additional Level 3 allows us to examine the effect of cross-level interactions between examinee-level and group-level measures on single item response. In a 3-level MD-DIF method, the equations of Level-1 remain the same (Burke, 2009; Kamata, 1999). The equation of Level 2 will be extended to contain a group-level variable (i.e., *j* for student-level measure). As a result, Equation (2.12) would be modified as follows:

$$\beta_{0jm} = \gamma_{00m} + \gamma_{01m}(Student\_Level\_Measure) + u_{0jm}$$

$$\beta_{1jm} = \gamma_{10m} + \gamma_{11m}(Student\_Level\_Measure)$$

$$\beta_{2jm} = \gamma_{20m} + \gamma_{21m}(Student\_Level\_Measure)$$

$$\vdots$$

$$\beta_{(k-1)jm} = \gamma_{(k-1)0m} + \gamma_{(k-1)1m}(Student\_Level\_Measure) \qquad (2.13)$$

In the equation, DIF coefficients γ_{11m}, γ_{21m}, through $\gamma_{(k-1)1m}$ are the random coefficients varying at the group level *m*. In order to model the group-level sources of DIF, the equation at Level 3 is presented as follows:

$$\gamma_{00m} = \pi_{000} + r_{00m}$$

$$\gamma_{01m} = \pi_{010} + \pi_{011}(Teacher\_Level\_Measure)$$

$$\gamma_{10m} = \pi_{100}$$

$$\gamma_{11m} = \pi_{110} + \pi_{111}(Teacher\_Level\_Measure)$$

$$\gamma_{20m} = \pi_{200}$$

$$\gamma_{21m} = \pi_{210} + \pi_{211}(Teacher\_Level\_Measure)$$

$$\vdots$$

$$\gamma_{(k-1)0m} = \pi_{(k-1)00}$$

$$\gamma_{(k-1)1m} = \pi_{(k-1)10} + \pi_{(k-1)11}(Teacher\_Level\_Measure), \quad (2.14)$$

where π_{000} is the grand mean score of all items for all examinees across all groups; r_{00m} is the random error of γ_{00m}. π_{100}, π_{200}, through $\pi_{(k-1)00}$ are the mean score of items 1, 2, …, *k*-1, respectively; π_{110}, π_{210}, through $\pi_{(k-1)10}$ are the effect of the individual-level measure on the mean performance of items 1, 2, …, *k*-1, respectively; and π_{111}, π_{211}, through $\pi_{(k-1)11}$ are the interaction between the student level measure and the teacher level measure on the mean performance of items 1, 2, …, *k*-1, respectively (Burke, 2009; Kamata, 1999).

To interpret identified DIF items, the magnitude of the DIF was quantified with the formula based on $\hat{\Delta}_{ML} = -2.35\ (\gamma_{i1})$. The derived magnitudes are then classified into three levels of DIF effect sizes according to the guidelines developed by Educational Testing Services (ETS) DIF classification system (Burke, 2009).

• Negligible or A-level DIF: $|\hat{\Delta}_{ML}| < 1.0$, or null hypothesis is not rejected,

• Moderate or B-level DIF: $1.0 \leq |\hat{\Delta}_{ML}| \leq 1.5$, and null hypothesis is rejected,

• Large or C-level DIF: $|\hat{\Delta}_{ML}| > 1.5$, and null hypothesis is rejected.

More commonly used DIF methods like Mantel-Haenszel and logistic regression DIF only detect the presence of items showing DIF and addition procedure or techniques is necessary to locate potential sources of DIF. Kamata's ML-DIF method not only has the ability to identify DIF items, but also to investigate potential sources of DIF. The ML-DIF method has been applied to large-scale assessment programs like NAEP, TIMSS, and Florida State mathematics assessment, implemented at state, national, and international levels for DIF item detection

(Chaimongkol, 2005). All data used in the ML-DIF studies featured hierarchical and nested structures of data.

2.2.2.5 Selection of Logistic Regression and SIBTest DIF methods

Since developed DIF methods employ different statistical models in a DIF analysis, they may identify a set of DIF items different from each other. Based on simulation studies, different methods performed well in certain conditions and have their own strengths and weaknesses (Gierl et al., 2000; Hidalgo & Lopez-Pina, 2004; Woods, 2011). To tackle the issue, researchers recommend the use of more than one DIF method to determine the presence of DIF (e.g., Camilli & Shepard, 1994).

Researchers can choose DIF methods depending on their research designs that best meet the conditions manipulated in a DIF simulation study. This dissertation compared the logistic regression method to the SIBTest method to examine the consistency of their abilities in identifying gender DIF items.

The advantages of choosing the logistic regression DIF method for this study are the following. First, it does not require categorizing the total scores or ability measures to have two groups of examinees matched in their ability as the Mantel-Haenszel method does. Second, it can identify both uniform and/or non-uniform DIF. Third, the magnitude of DIF effect size is computed the same way as the procedures used to perform the chi-square test for the statistical significance of DIF. Fourth, the ordinal logistic regression DIF method Zumbo (1999) developed is the same as regular logistic regression DIF method in their formats of reporting a DIF analysis result. Fifth, researchers may hesitate to include polytomous items in their DIF analyses because it is time-consuming and complicated in computation and interpretations (French & Miller, 1996;

Miller & Spray, 1993). The ordinal logistic regression DIF method developed by Zumbo (1999) can handle polytomous items or polytomous along with dichotomous items in a test.

The SIBTest DIF method was selected in my DIF study is based on the following reasons. First, the SIBTest has been found to be more effective in detecting DIF than the Mantel-Haenszel and logistic regression methods (Bolt & Stout, 1996; Jiang & Stout, 1998). Second, the SIBTest uses non-parametric approach to design its DIF detection model, which is different from those DIF methods developed by the parametric approach. The parameter-oriented DIF method requires stronger assumptions that a dataset may fail to meet its assumption. Third, after items identified as DIF, they can be grouped as a bundle to examine the potential sources that contribute to the DIF effect. Fourth, instead of using observed test score, the SIBTest runs a regression to estimate the true score used to match students on ability, which results in an improved estimation of ability level where the examinees should stay.

Chapter III: Methods

Chapter III is split into four sections. First, I provide three sections of detailed information on the data source, participants, and test items involved in the analysis. The last section describes the two DIF methods, logistic regression and SIBTest, including the specified statistical model of each method and procedures for comparing the DIF results from these two methods.

3.1 Data Source

TIMSS is an international large-scale assessment project that is routinely conducted by the International Association for the Evaluation of Educational Achievement (IEA) on a regular 4-year cycle since 1995. It claims to be the largest and most comprehensive international study of mathematics and science education. One of TIMSS's goals is helping participating countries improve teaching and learning in mathematics and science. TIMSS 2011 represented the fifth round of IEA's study, following projects in 1995, 1999, 2003, and 2007. It was conducted primarily with fourth and eighth grade students from approximately 60 countries, to assess student performance differences in mathematics and science[9] (Foy et al., 2013). The formal definitions of the TIMSS target populations makes use of UNESCO's International Standard Classification of Education (ISCED) in identifying appropriate target grades. ISCED provides an international standard for describing levels of schooling from Level 0, the pre-primary, to Level 6, the second level of tertiary education (OECD, 1999). The eighth grade population within each country was defined as the students enrolled in the grade that represents 8 years of formal schooling, counting from the first year of ISCED Level 1, provided that the mean age at the time of testing is at least 13.5 years old (OECD, 1999).

[9] TIMSS 2011 were administered to the sixth and ninth grade students for several of the countries because the assessment was found too difficult for fourth and eighth grade students in these countries.

Nationally representative sample of approximately 4,000 eighth grade students[10] from 150-200 schools in each of 43 voluntarily participating counties participated in the TIMSS 2011 project, which resulted in more than 300,000 student participants (Mullis et al., 2013). TIMSS adopted a two-stage stratified cluster sampling, with schools as the first stage of selection, and intact classes within schools as the second stage of sample selection (Martin & Mullis, 2013; Olson et al., 2008). Therefore, the collected database includes student performance data and background questionnaire data containing three levels: students, teachers, and schools.

3.2 Participants

The present study used students in the eighth grade from the U.S. and Taiwan labeled as Chinese Taipei (Martin & Mullis, 2013), focusing on results from the TIMSS 2011 mathematics assessment. Among all countries with eighth grade participants, Taiwan was considered a high performing country, with an average national total score of 609. Taiwan performed similarly to the Republic of Korea (613) and Singapore (611), and had higher achievement than the rest of participating countries. The U.S. was included in the top ten high-achieving countries with an average of 509. U.S. performance was statistically significantly higher than the TIMSS scale center point of 500[11] (Mullis et al., 2013).

The sample size included 10,477 eighth grade students from the U.S. (5,180 males and 5,297 females) and 5,042 students from Taiwan (2,594 males and 2,448 females) (Mullis et al., 2013). The average age at the time of testing was 14.2 years old for both the U.S. and Taiwan samples (Mullis et al., 2013). Due to the TIMSS multiple booklet design, which will be

[10] The number of participating eighth grade students for three countries, including United Arab Emirates, United States, and South Africa, ranged from 10,500 to 14,000.

[11] The scale average of 500 was set to correspond to the mean of the overall achievement distribution for all participating countries, and 100 points on the scale was set to correspond to the standard deviation (Martin & Mullis, 2013).

described in detail in next section, I only used approximately half of each country's sample in the DIF analysis.

The number of student participants with complete data, including the number of males and females, for the DIF analysis is summarized in Table 4. Students with missing data[12] in their answer records were eliminated from the DIF analysis. There were 19 students removed from the U.S. sample, and two students removed from the Taiwan sample (see Appendix A for student records with missing data of the U.S. and Taiwan samples). Eliminating incomplete student data records resulted in a very small number of students being removed – a change that would not have an impact on the validity of the study.

Table 4

Selected Booklets and Valid Number of Males and Females Involved in Gender DIF Analyses

Final Booklet	Item Block	U.S. Sample			Taiwan Sample		
		Male	Female	Total	Male	Female	Total
Booklet 1	M1, M2	375	353	728	189	172	361
Booklet 3	M3, M4	383	355	738	190	171	361
Booklet 5	M5, M6	372	370	742	189	175	364
Booklet 7	M7, M8	407	338	745	182	176	358
Booklet 9	M9, M10	365	377	742	193	168	361
Booklet 11	M11, M12	358	402	760	185	174	359
Booklet 13	M13, M14	347	394	741	186	175	361
Odd Booklets	M1 - 14	2,607	2,589	5,196	1,314	1,211	2,525

3.3 Instruments

3.3.1 Booklet Design of the Mathematics Test

The mathematics assessment framework in TIMSS 2011 was developed along two dimensions: (1) a content dimension specifying the knowledge domain or subject matter, and (2)

[12] A student who did not respond at least one test item was considered a record containing missing value/data.

a cognitive dimension specifying the cognitive demand elicited to solve test items. The content and cognitive dimensions are the foundation of the TIMSS 2011 assessment (Mullis et al., 2009). All 217 items were developed based on these two dimensions. Approximately half of the items were multiple-choice and half were constructed-response items, with a total of 232 score points (Martin & Mullis, 2013).

TIMSS 2011 used a matrix-sampling strategy to assign the entire pool of 217 mathematics test items into multiple booklets. The core idea for matrix-sampling is that all test items are divided into multiple final test booklets, so that all sampled students for a given country can be split into groups of approximately equal size, and each group can be randomly assigned to one booklet. Such a design allows for the inclusion of more items in the test to represent broader contents of curriculum (Martin & Mullis, 2013). By having groups of students using multiple booklets, the performance distribution of a given country's student population can be estimated based on the aggregation of all test items to which a student responds.

For TIMSS 2011, all 217 eighth grade mathematics items were first grouped into 14 mutually exclusive test item blocks. Item blocks were assembled and balanced by taking content domain, cognitive demand, and item format into consideration. Then, each pair of adjacent item blocks was placed into one booklet, providing a bridge to connect the student responses together across multiple booklets. As a result, each of the final 14 booklets contained two item blocks, and each block was used twice (Mullis et al, 2009) as shown in Table 5. The time limit for students to respond to the TIMSS mathematics booklet was 45 minutes[13] (Mullis et al., 2009).

[13] Students were allowed two 45-minute sessions, one session for mathematics and one for science, to complete the test booklet. They were also given an additional 30 minutes to complete the student background questionnaire (Mullis et al., 2009).

In the study, seven odd numbers of booklets (i.e., booklets 1, 3, 5, 7, 9, 11, and 13) were selected as a data source for the DIF analysis. Theoretically, the studies of either odd or even booklets are expected to be the same, because the booklets were balanced by taking item characteristics into account and were randomly assigned to roughly equal number of students.

Table 5

TIMSS 2011 Item Block and Booklet Design for Mathematics Assessment

Item Block	Booklet Assigned	Item Released (Yes/No)	Number of Stand-alone Items	Number of Items with Shared Stimulus	Items in total
M1	BK14, BK1	Y	8	3	11
M2	BK1, BK2	Y	13	2	15
M3	BK2, BK3	Y	17	0	17
M4	BK3, BK4	N	15	0	15
M5	BK4, BK5	Y	14	0	14
M6	BK5, BK6	Y	9	8	17
M7	BK6, BK7	Y	12	3	15
M8	BK7, BK8	N	13	2	15
M9	BK8, BK9	N	14	4	18
M10	BK9, BK10	N	11	4	15
M11	BK10, BK11	N	12	5	17
M12	BK11, BK12	N	13	2	15
M13	BK12, BK13	N	14	3	17
M14	BK13, BK14	N	10	6	16
M1-14			175	42	217

It is important to note that 42 TIMSS items were developed based on shared stimulus, as shown in Table 5. Items with a shared stimulus refer to a common passage used to make up a set of test items. According to Beretvas and Walker (2012), shared stimulus may be a potential source contributing to DIF effects. In the present study, there were only approximately 20% of stimulus-shared items. The gender DIF analysis with a model of items nested within shared stimulus was not considered. Instead, all test items were treated as stand-alone items for the analysis.

The number of schools, classes, and students assigned to each TIMSS booklet is summarized in Table 6. For example, in the U.S. student sample, Booklet One was assigned to 472 schools with 490 classes for 731 students. On average, fewer than two students per school were tested with Booklet One (731/472=1.5). As a result of this booklet design in combination of using stratified sampling strategies, the nested number of students in either a school or a class was fewer than two per booklet in the U.S. sample and fewer than three per booklet in the Taiwan sample. For this reason, the multilevel-DIF approach described in Chapter II of this DIF study was not possible, because the data did not meet the requirement of minimum sample size, which could lead to an unreliable estimation of the DIF parameters (Burkes, 2009; Hox, 1998).

Table 6

Number of Schools, Classes, and Students Sampled in the U.S. and Taiwan for TIMSS 2011 Mathematics Assessment

Final Booklet	Item Block	U.S. Sample			Taiwan Sample		
		School	Class	Student	School	Class	Student
Booklet 1	M1, M2	472	490	731	150	152	362
Booklet 2	M2, M3	469	488	743	150	151	357
Booklet 3	M3, M4	477	494	740	150	152	361
Booklet 4	M4, M5	472	491	754	150	152	363
Booklet 5	M5, M6	480	499	743	150	152	364
Booklet 6	M6, M7	479	496	752	150	152	365
Booklet 7	M7, M8	468	486	746	150	152	358
Booklet 8	M8, M9	471	486	746	150	152	358
Booklet 9	M9, M10	473	491	745	150	152	361
Booklet 10	M10, M11	478	494	750	148	150	357
Booklet 11	M11, M12	477	499	765	150	152	359
Booklet 12	M12, M13	475	498	759	150	152	357
Booklet 13	M13, M14	467	490	745	149	151	362
Booklet 14	M14, M1	472	494	758	150	152	358
Booklet 1-14		501	557	10,477	150	152	5,042

3.3.2 Description of TIMSS Test Items

In this sub-section, I first provide information on the number of items for each dimension. Then I report the breakdown of item numbers by cognitive demand in relation to each of the other two dimensions. The TIMSS assessment framework is based on two dimensions: content domain and cognitive demand. In other words, all test items were developed to sufficiently represent these two dimensions in the test. I also discuss item type or item format, another item characteristic cited frequently in the gender DIF research.

Content domain. The content domain refers to the specific mathematics subject matter covered by mathematics items. There are four content domains used for the eighth grade population: *number, algebra, geometry*, and *data and chance*. The number of items for each content domain was: (a) Number (n = 61), (b) Algebra (n = 70), (c) Geometry (n = 43), and (d) Data and Chance (n = 43) (Martin & Mullis, 2013). For each of the four content domains, several topic areas were further identified. For example, the *number* content domain consists of understanding and skills related to four topic areas, including "whole numbers", "fractions and decimals", "number sentences with whole numbers", and "patterns and relationships." (Mullis et al., 2009).

Cognitive demand. Cognitive demand is the characteristic of interest in this dissertation study. Cognitive demand is defined as the cognitive skills or mental abilities required in order for students to respond to items correctly. As described in Chapter II, the TIMSS assessment framework defined three levels of cognitive demand: *knowing*, *applying*, and *reasoning* (Mullis et al., 2009). The number of items for each category according to their performance expectation was: (a) Knowing (n = 80), (b) Applying (n = 85), and (c) Reasoning (n = 52) (Martin & Mullis, 2013).

Item type. Item type is defined as the formats that a test item is presented in. The TIMSS uses two common types or formats: multiple-choice and constructed-response (equivalent to open-ended items). The TIMSS 2011 mathematics test includes 118 multiple-choice and 99 constructed-response items (Martin & Mullis, 2013). Multiple-choice items provided students with four response options to choose from, with only one correct option, and were worth one score point. Constructed-response items required students to show their work and explain how they solved the problem instead of selecting a response from a set of options. Constructed-response items were worth one or two points. If they were 1-point constructed-response items, they were scored as correct (1 score point) or incorrect (0 score points). If they were 2-point items, they were scored from 0 to 2 as fully correct (2 score points), partially correct (1 score point), or incorrect (0 score points) (Martin & Mullis, 2013). A summary of the 217 mathematics items for eighth graders in TIMSS 2011 is presented in Appendix B.

Cognitive demand in relation to content domain and item type. Table 7 shows the number of items differing in cognitive demand level in each content domain for the TIMSS 2011 eighth grade mathematics assessment. Each cognitive demand level includes a range of content domains. A chi-square test was used to examine whether the distribution patterns of the percentage of items in each content domain were consistent across cognitive demand levels. The statistic showed a two-way interaction (χ^2 (6) = 18.49, $p < .01$). The results indicated the distribution patterns of items in each content domain were inconsistent across cognitive demand levels.

Table 7

TIMSS 2011 Number of Mathematics Items by Cognitive Demand and Content Domain

Cognitive Demand	Content Domain				
	Number	Algebra	Geometry	Data & Chance	Total
Knowing	29 (36.3)	32 (40.0)	6 (7.5)	13 (16.3)	80 (100)
Applying	22 (25.9)	21 (24.7)	21 (24.7)	21 (24.7)	85 (100)
Reasoning	10 (19.2)	17 (32.7)	16 (30.8)	9 (17.3)	52 (100)
Total	61 (28.1)	70 (32.3)	43 (19.8)	43 (19.8)	217 (100)

Note. The percentage of cell items divided by the number of total items in rows is shown in parenthesis.

Items developed in the cognitive demand level *knowing* tended to be in the content domains of *number* and *algebra* (76.3%); items developed in the cognitive demand level *applying* were almost evenly distributed in each of the four content domains, with 25% of the items in each category; and items developed in the cognitive demand level *reasoning* tended to appear in the domains of *algebra* and *geometry* (63.5%). The inconsistency of item distribution in content areas across cognitive demand levels indicated the heterogeneity of items in each of the cognitive demand levels. This heterogeneity constrained the study's ability to draw casual conclusions, because the heterogeneity of items may be a factor the potentially contributes to the gender DIF result. In addition, previous research has pointed to the content domain as a potential source of gender DIF. To address this issue, content domain was added as a factor, along with cognitive demand, to explore the gender DIF results.

Table 8

TIMSS 2011 Number of Mathematics Items by Cognitive Demand and Item Type

Cognitive Demand	Item Type		Total
	Multiple-Choice	Constructed-Response	
Knowing	53 (66.3)	27 (33.8)	80 (100)
Applying	47 (55.3)	38 (44.7)	85 (100)
Reasoning	18 (34.6)	34 (65.4)	52 (100)
Total	118 (54.5)	99 (45.6)	217 (100)

Note. The percentage of cell items divided by the number of total items in rows is shown in parenthesis.

Table 8 shows the number of items differing in cognitive demand level within the multiple-choice and constructed-response formats. Each cognitive demand level includes both item types. A chi-square test examining whether the distribution patterns of the two item types were consistent across cognitive demand levels showed a two-way interaction (χ^2 (6) = 12.76, p < .01). The results indicated the distribution patterns varied depending on cognitive demand levels. Items developed in the cognitive demand level *knowing* tended to appear in the multiple-choice format (66.3%); items developed in the cognitive demand level *applying* were roughly half multiple-choice and half constructed-response; and items developed in the cognitive demand level *reasoning* tended to be constructed-response (65.4%). The association between the cognitive demand and the item type will result in difficulty drawing causal conclusions about cognitive demand as a factor in gender DIF findings. The causal relationship can be demonstrated only when the items are homogeneously distributed across cognitive demand levels. Thus, in this dissertation, item type was combined with cognitive demand to explore their interaction effects on gender DIF results.

3.4 Procedures for Data Analysis

3.4.1 Tested Models Involved

Two of the DIF detection methods described in Chapter II were selected for this dissertation: ordinal logistic regression DIF and Poly-SIBTest DIF. In this section, I first describe the specified models for both logistic regression and SIBTest DIF methods, using notation consistent with the models formulated in Chapter II. In addition, I introduce the procedure for restructuring data to meet the requirements of the gender DIF analysis.

Next, I describe the statistical methods used to examine (1) the consistency of the two methods in identifying DIF and non-DIF items, (2) the gender DIF pattern of items differing in cognitive demand levels, and (3) the consistency of the identified gender DIF patterns between countries.

3.4.1.1 Ordinal Logistic Regression DIF Analysis

I used the logistic regression DIF method as described in Chapter II. The equation for ordinal logistic regression DIF analysis for polytomous items is expressed as (Zumbo, 1999):

$$\text{logit}[P(Y \leq j)] = \propto_j + \beta_1(total\_score) + \beta_2(group) + \beta_3(total\_score \text{ x } group), \quad (3.01)$$

$$\text{logit}[\frac{P(Y \leq j)}{P(Y > j)}] = \propto_j + \beta_1(total\_score) + \beta_2(group) + \beta_3(total\_score \text{ x } group), \quad (3.02)$$

where $P(Y \leq j)$ denotes the cumulative probabilities that the actual item response Y falls in category j or lower. A logit is the natural logarithm of the ratio of two probabilities as seen in Equation 3.02; $j = 1, 2, \ldots, k\text{-}1$, where k is the number of score levels in the polytomous item. The variables *total_score*, *group*, and *total_score* x *group*, as well as the estimated parameters β_1, β_2, and β_3 are the same as described in the logistic regression DIF section of Chapter II. The

model ends up with a separate intercept parameter α_j for each cumulative probability. Based on the equal slope or the proportional odds assumption, only a single set of regression coefficients is estimated for all cumulative logits with varying intercepts (α_j) (O'Connell, 2006; Zumbo, 1999).

In order to test the significance of DIF for each polytomous item, three nested models are formed in a hierarchical structure as follows:

Model 1: $\text{logit}[P(Y \leq j)] = \propto_j + \beta_1(total\_score)$,

Model 2: $\text{logit}[P(Y \leq j)] = \propto_j + \beta_1(total\_score) + \beta_2(group)$,

Model 3: $\text{logit}[P(Y \leq j)] = \propto_j + \beta_1(total\_score) + \beta_2(group) + \beta_3(total\_score \text{ x } group)$.

(3.03)

Statistical hypothesis testing for the significance of DIF is based on the following procedure. In Model 1, the conditioning variable is entered first (i.e., the total score). In Model 2 the group variable is entered, and finally, in Model 3, the interaction term is entered into the equation. For each model, a chi-square value is computed. Comparing those values allows the researchers to detect uniform and/or non-uniform DIF.

Two procedures are required to identify whether the DIF status is uniform or non-uniform. Statistical significance is tested for the chi-square difference between Models 1 and 2, which provides evidence for uniform DIF status. Furthermore, statistical significance testing of the chi-square difference between Models 2 and 3 provides evidence to confirm a DIF item exhibits a non-uniform pattern (Zumbo, 1999).

The estimation for the magnitude of DIF effect follows the same procedures as above. The chi-square measure for each model is used to compute their corresponding R^2 measures. The difference in R^2 between a pair of models reveals the magnitude of DIF for an identified uniform

or non-uniform item. The magnitude of the R^2 *difference* (denoted as $\hat{\Delta} R^2$) is evaluated according to the guidelines suggested by Jodoin and Gierl (2001).

• Negligible or A-level DIF: $\hat{\Delta} R^2 < .035$, or χ^2 test is statistically significant,

• Moderate or B-level DIF: $.035 \leq \hat{\Delta} R^2 \leq .070$ and χ^2 test is statistically significant,

• Large or C-level DIF: $\hat{\Delta} R^2 > .070$ and χ^2 test is statistically significant.

3.4.1.2 Poly-SIBTest DIF Analysis

This study also used the equation and procedure of the SIBTest, described in Chapter II, to identify items with DIF. The SIBTest requires that two subtests are created: a suspect subtest, which consists of items that are tested for DIF, and a matching subtest, which is both free of DIF items and measures what the test is intended to measure. Both the reference and focal group examinees were matched on achievement, or observed ability, based on the matching subtest. In this dissertation, male students were set as a reference group and female students as a focus group. The null hypothesis states that the suspect items exhibit no DIF, and that both the reference and focal group examinees, matched on ability, are expected to have equivalent scores on the suspect item(s), within statistical error (Stout & Roussos, 1996). The alternative hypothesis is that one group (the focal female or reference male group) of test-takers, matched in their ability, will have a significantly higher score on the suspect item(s) (Ilich, 2013).

Four steps take place in order to calculate the beta estimates used to identify items with DIF. First, examinees from both the reference and focal groups are put into the same subgroups by matching scores. Second, subgroup members from the focal and reference groups are combined to form statistical calculation cells for each matching score. Third, for each matching score level, the average number of correct responses on the suspect items is computed for both the reference group, represented as $P_R(T)$, and the focal group, represented as $P_F(T)$. Fourth, the

difference of $P_R(T)$- $P_F(T)$ is calculated and represented by $B(T)$. If $B(T)$ equals zero, there is no DIF. The DIF is present when $B(T)$ is statistically different from zero. The matching scores may not be reliable, which results in a biased $B(T)$. As a result, the SIBTest includes a regression correction procedure to fix this issue. Then, the true matching scores are estimated and used to adjust $P_R(T)$and $P_F(T)$ values. A weighted sum of the $B(T)$ values (i.e., difference between adjust $P_R(T)$and $P_F(T)$) is used to calculate the beta estimates (Metcalf, 2002).

Because the TIMSS 2011 data includes both dichotomously and polytomously scored items, the Poly-SIBTest option in the DIFPACK v1.7 software application was selected to detect gender DIF items. The Poly-SIBTest requires the creation of separate ASCII (American Standard Code for Information Interchange) files for each gender group (males and females) within each booklet under each country sample. The DIFPACK software was configured under the following six conditions in order to correctly implement the gender DIF detection for this dissertation. First, males were selected as the reference group, while females were selected as the focal group. Second, the minimum cell size was specified as 2. Third, the *p*-value was type "E" - indicating that DIF could be found for either the male or female group. Fourth, the guessing parameter was specified as 0.2, and the option 'default' was selected for the weighting parameters. Fifth, all mathematics items on a given booklet were selected as suspect items. Sixth, each suspect item was selected to run separately, while all remaining items were used as the matching subtest (Smith, 2009).

The output files reported a variety of information. Test score summary statistics reported the mean scores and their standard deviations for both reference and focal groups. Basic item statistics included the ordered item number, mean item scores, and point-biserial correlations, which are correlations between an individual item's total score and overall test score. DIF results

for each item examined included: the beta estimate, the standard error for the beta, the significance level of the *p*-value for the beta, the direction of the *p*-value (indicating whether the item favored either the reference or focus group), the proportion of both groups not used in SIBTest calculations for any examined item, the MS/SSD measure (representing the standardized difference in mean observed scores between the reference and focal groups on the matching subtest), and the FLAG message indicating whether the DIF run was successfully implemented (Smith, 2009).

A gender DIF direction in favor of either the reference or focus group is determined by the size of the beta estimate ($\hat{\beta}_u$). Positive beta values indicate a DIF effect in favor of the reference group (in this study, male students) while negative beta values indicate a DIF effect in favor of the focus group (in this study, female students).

The magnitude of the DIF effect, $\hat{\beta}_u$, is evaluated according the guidelines developed by Roussos and Stout (1996):

• Negligible or A-level DIF: $\left|\hat{\beta}_u\right| < .059$, and null hypothesis is rejected,

• Moderate or B-level DIF: $.059 \leq \left|\hat{\beta}_u\right| < .088$, and null hypothesis is rejected,

• Large or C-level DIF: $\left|\hat{\beta}_u\right| \geq .088$, and null hypothesis is rejected.

3.4.2 Selection for Type I Error Rate and Purification Procedures

In this dissertation study, a nominal α equal to .01, which is commonly used, was selected in the hypothesis testing to control Type I error rate. In such a circumstance, 1% of the items would be falsely identified as DIF when they did not truly show DIF.

In DIF analyses, just like in any hypothesis testing, researchers are concerned about whether observed Type I error rate is exactly as it is nominated. A well-controlled Type I error

rate as expected reflects the quality of a DIF study to a certain extent. Factors potentially affecting the stability of Type I error rates can be sample-related or item-related (Finch & French, 2007; Gierl et al., 2000; Pei & Li, 2010; Woods, 2011; Zilberberg et al., 2011). For example, the simulation study of Zilberberg et al. (2011) found that (1) equal and balanced sample groups consistently exhibited expected Type I error rates as compared to unequal and unbalanced sample sizes, and (2) the observed Type I error rate was no more than the nominal 5% error rate when sample sizes were at least 250 for each group when running Mantel-Haenszel, SIBTest, and logistic regression DIF methods given the proportion of DIF items ranging from 20% to 60%. These findings supported the decision of 1% Type I error rate in the current study in order to control the potential inflation of Type I error rate for the Taiwan sample which was on average 180 students for each gender group per booklet.

Another decision for the DIF procedures was related to the use of purification when item scores are used to form groups of ability-matched examinees. Examinees are matched either on the total score of all items or a common measure based on purification procedures. The latter, referred to as *criterion refinement*, was proposed to address the contamination issues of the total score as a matching variable when a number of DIF items, especially a large number of DIF items, are identified (Clauser, Mazor, & Hambleton, 1993; Miller & Oshima, 1992).

In the current study, I chose the DIF analysis without purification procedure. This decision was supported by the research finding that the purification procedures may not provide improved DIF detection over the single-step DIF procedure without purification, especially when the percentage of DIF items was as small as 5% to 10% (Clauser, Mazor, & Hambleton, 1993; French & Maller, 2007; Ilich, 2013; Magis & Facon, 2012; Zenisky, Hambleton, & Robin, 2003).

3.4.3 Dataset Prepared for Logistic Regression and SIBTest DIF Analysis

The U.S. and Taiwan samples were analyzed separately using both the logistic regression DIF and Poly-SIBTest methods. To meet the dataset formats required to implement these methods, separate data files were generated containing student scores for each of the 217 mathematics items as well as total scores for the male and female subgroups.

For the logistic regression DIF analyses, I used the ordinal logistic regression DIF codes developed by Zumbo (1999; see Appendix C), which can be implemented by the statistical software application Statistical Package for the Social Sciences (SPSS). Fourteen SPSS files were generated: 7 files for the U.S. and 7 files for the Taiwan samples, each corresponding to the selected booklets from the publicly accessible master SPSS files created by the TIMSS 2011 project (http://www.timss.org/).

In each of these 14 files, each individual mathematics item was automatically scored by TIMSS 2011 programing codes, with multiple-choice items scored either a 1 or 0 (1 = correct and 0 = incorrect answers), and constructed-response items scored either a 2, 1, or 0 (2 = correct, 1 = partial correct, and 0 = incorrect responses). In addition to item scores, the files included a variable for gender with female = 1 and male = 2, and a variable for the total scores of each student. Variables irrelevant to the gender-based DIF analysis were eliminated from the SPSS files. In addition, students with missing data were eliminated from the SPSS file as well (see Appendix A). Finally, the original TIMSS 2011 mathematics item variable names were manually re-named in order to easily identify items in the logistic regression DIF analysis. A sample SPSS data file for Booklet One from the U.S. sample, developed for this DIF study, is shown in Appendix D.

Next, I modified the codes originally developed by Zumbo (1999) to conduct the gender-based logistic regression DIF analysis for each of the seven SPSS files generated for each country sample. A sample of the modified codes for the U.S. sample is shown in Appendix E.

For the Poly-SIBTest DIF analysis, I used the DIFPACK software application (see Appendix F) and selected the Poly-SIBTest option, one of three modules in the DIFPACK application. The Poly-SIBTest DIF analyses required separate files for males and females, therefore each of the original 14 SPSS files (i.e., 7 files for the U.S. sample and 7 files for the Taiwan sample) was split into two files, one for each gender. These 28 ASCII files contained only the individual mathematics item scores. The Poly-SIBTest estimated total scores for each student based on individual item-level score records. A sample ASCII file for Booklet One, developed for the U.S. male and female samples, is shown in Appendix G. A pair of male and corresponding female ASCII files were linked to estimate gender DIF effects for a given booklet, which resulted in seven gender DIF results for each country sample.

3.4.4 Statistical Methods Used to Answer Research Questions

To answer research question 1: *Are the gender DIF items identified by the logistic regression DIF method consistent with the items identified by the SIBTest method*, the chi-square test, Cohen's kappa, Bowker's test, and Pearson correlation were used.

The chi-square test of independence can be used on a contingency table of any size to examine the degree of the dependence association between two variables (Glass & Hopkins, 1984; Kiess, 2002). After a chi-square value is estimated, the *contingency coefficient* (McNemar, 1969), which incorporates the chi-square estimate, can measure the degree of association or correlation between categorical variables. In this dissertation, because each of the DIF methods contains three nominal DIF categories (favoring males, favoring females, and non-DIF), chi-

square and contingency coefficients were used to evaluate the association between DIF methods. Similarly, Cohen's kappa (GraphPad Software, 2014) was used to test the degree of agreement between the DIF classifications from the two methods, and to compare with the contingency coefficient.

The Bowker's test of symmetry (Marascuilo & McSweeney, 1977), which is an extension of the McNemar test, focuses on tests of axial symmetry in a two-way table. The Bowker's measure is estimated by a chi-squared approximation (PQStat Software, 2009). The method was used to examine whether one DIF method tends to generate higher DIF item rates than the other method.

The Pearson correlation coefficient was used to test whether there is a positive correlation between DIF methods of the magnitude of DIF estimates. When the correlation coefficient is statistically significant, it is possible to claim that the magnitude of DIF identified by one DIF method can effectively predict the magnitude identified by the other DIF method.

To answer research question 2: *Do gender DIF patterns in mathematics items differ in levels of cognitive demand*, chi-square and log-linear tests were used.

The chi-square test can also be used to test homogeneity. If applied to a single categorical variable from different groups (i.e., category levels x groups), it can be used to determine whether frequency counts are distributed identically across different populations. In this study, the categorical variable is gender DIF pattern, and the group is cognitive demand level. Thus, for research question 2, the chi-square tests a null hypothesis that the gender DIF patterns are homogeneous, or equal, across cognitive demand levels. An alternative hypothesis claims that they are not.

When either content domain or item type is added as factors to examine their interaction effects with the gender DIF pattern and cognitive demand levels, a log-linear test is used. The log-linear test is a version of the chi-square test that is extended to a 3-way contingency table. When a chi-square value is computed by the log-liner method, it is labeled as G^2. Since the G^2 distribution is approximately the same as chi-square, its probability can be estimated by referencing the sampling distribution of chi-square. G^2 values are usually close to the corresponding chi-square values computed using the regular procedure.

To answer research question 3: *Is the gender DIF pattern of items differing in cognitive demand levels similar across countries,* a log-linear analysis (Jeansonne, 2002) was used. For research question 3, the collected data was categorized by country sample, gender DIF pattern, and cognitive demand level to form a 3-way contingency table. As described in the previous paragraph, a log-linear test was used to examine if the gender DIF pattern of items differing in cognitive demand levels was moderated by country samples. The null hypothesis asserted that the gender DIF patterns were homogeneous across countries; the alternative hypothesis claims that they were not.

The consistency of items identified as DIF and non-DIF between countries was also examined with the chi-square test and Cohen's kappa. Next, I performed the McNemar test to examine whether one country generated a higher rate of gender DIF items than the other one.

Lastly, Fisher's exact probability test (Wikipedia, 2014) and it's extension (Soper, 2006; Lowry, 2001) for *i* (i.e., row) x *j* (i.e., column) contingency tables was used to examine the homogeneity of distribution patterns of DIF item percentages in cognitive demand, content domain, and item type between two country samples.

Chapter IV: Results

In this chapter, the findings of the study are ordered by the research question they investigate.

4.1 Consistency of Gender DIF Item Analyses

Research Question 1 is "*Are the gender DIF items identified by the logistic regression method consistent with the items identified by the SIBTest method*?" To answer this question, I used both logistic regression and SIBTest methods to analyze the TIMSS mathematics items. By doing so, I could statistically flag DIF items and examine the consistency of both methods' abilities to identify items as DIF and non-DIF for the U.S. and Taiwan student samples. In the next two sub-sections, I report the results of the analysis first for the U.S. sample and then for the Taiwan sample.

4.1.1 Consistency of Logistic Regression and SIBTest DIF Results for the U.S. Sample

In this sub-section, I summarize the items flagged as DIF and non-DIF by both detection methods on the U.S. sample. When items were identified as showing DIF, they were labeled with the TIMSS item ID (e.g., M032166). Item type for DIF items was reported as well. I performed the chi-square, contingency coefficient, Cohen's kappa, Bower's, and Pearson correlation coefficient test to compare the consistency of items identified as DIF and non-DIF between the DIF methods.

Both dichotomous and polytomous items were identified as exhibiting DIF if any of the differential step analysis for the 3-step hierarchical logistic regression analysis was statistically significant at an $\alpha = .01$ level. For the SIBTest DIF method, an item was identified as DIF at an $\alpha = .01$ level was used as well for the DIF detection.

Table 9 summarizes the items identified as showing DIF by both the logistic regression and SIBTest DIF methods. Table 9 is split into three tiers, with items identified as DIF by both DIF methods in the upper tier, items identified as DIF only by logistic regression DIF method in the middle tier, and items identified as DIF only by SIBTest method in the lower tier.

As shown in Table 9, the logistic regression DIF analysis flagged 20 items with DIF status: 10 items favoring male students, and 10 items favoring female students (see column 5 in Table 9; see Appendix H for the sample statistical reports of ordinal logistic regression gender DIF results in the U.S. sample). The percentage of the total number of items identified with DIF was 9.2% (i.e., 20/217). Of the 20 items with significant DIF, 15 items were uniform and 5 were non-uniform. In terms of content domain, 8 items belonged to *number*, 9 items belonged to *algebra*, only one item belonged to *geometry*, and 2 items belonged to *data and chance*. In terms of cognitive demands, 13 items were *knowing* items, 5 were *applying* items, and 2 were *reasoning* items. In terms of item type, 7 were multiple-choice, and the remaining 13 were constructed-response items.

As shown in column 6 of Table 9, 16 out of 20 items were classified as having a smaller DIF effect, with a range of the magnitudes from .010 to .034. Only 4 items had an effect size equal to or above .035, which is required to be classified as a moderate DIF effect (Jodoin & Gierl, 2001).

Table 9

Summary of Identified DIF Items by Logistic Regression and SIBTest Methods for the U.S. Sample

TIMSS 2011 Item ID	Content Domain	Cognitive Demand	Item Type	LR Coefficient	LR ($\hat{\Delta}$ R^2) Effect Size	SIBTest Beta & Effect Size	DIF Favoring
M032595	Number	Applying	MC	0.74*	0.028	**0.13***	Male
M052302	Algebra	Knowing	MC	-0.81*	0.027	**-0.08***	Female
M042226	Algebra	Knowing	CR	-0.66*	0.016	**-0.10***	Female
M042103	Algebra	Knowing	CR	-0.64*	0.019	**-0.10***	Female
M042086	Algebra	Applying	CR	0.78*	0.033	**0.13***	Male
M042229A	Algebra	Applying	CR	-0.73*	0.023	**-0.10***	Female
M052217	Number	Reasoning	CR	0.82*	0.031	**0.15***	Male
M052422B	Data & Chance	Applying	MC	-0.40*	0.010	**-0.09***	Female
M042194	Number	Knowing	CR	0.79*	0.028	**0.12***	Male
M042114A	Number	Knowing	CR	0.69*	0.018	**0.12***	Male
M042114B	Number	Applying	CR	0.68*	0.020	**0.12***	Male
M042050	Algebra	Knowing	CR	-0.64*	0.023	**-0.13***	Female
M042076	Algebra	Knowing	MC	0.69*	0.032	**0.14***	Male
M042302C	Number	Reasoning	CR	-0.61*	0.025	**-0.19***	Female
M042152	Geometry	Knowing	MC	0.55*	0.023	-0.05	Male
M042169B	Data & Chance	Knowing	CR	2.74*	**0.035**	-0.04	Male
M052130	Algebra	Knowing	MC	4.55*	0.034	-0.01	Male
M042081	Number	Knowing	CR	-5.44*	**0.035**	0.02	Female
M042049	Algebra	Knowing	MC	-4.56*	**0.041**	-0.01	Female
M052229	Number	Knowing	CR	-5.69*	**0.044**	-0.02	Female
M042245	Algebra	Applying	MC	-0.45	0.010	**-0.10***	Female
M042269	Data & Chance	Reasoning	MC	-0.31	0.004	**0.14***	Male
M052066	Algebra	Applying	MC	-0.47	0.011	**-0.11***	Female
M032681B	Data & Chance	Applying	CR	0.92	0.004	**0.10***	Male
M052073	Algebra	Knowing	MC	-0.28	0.005	**-0.08***	Female
M052502	Data & Chance	Applying	CR	-0.43	0.006	**-0.10***	Female
M042060	Number	Knowing	MC	0.49	0.017	**0.08***	Male
M042197	Number	Reasoning	CR	0.50	0.012	**0.10***	Male
M042224	Data & Chance	Knowing	CR	0.41	0.012	**0.10***	Male
M052021	Number	Reasoning	CR	0.46	0.009	**0.16***	Male
M052422A	Data & Chance	Applying	MC	-0.42	0.011	**-0.09***	Female
M052058A	Number	Applying	CR	0.44	0.013	**0.08***	Male
M052125	Number	Reasoning	MC	-0.43	0.008	**-0.09***	Female

Note. MC = Multiple-choice item; CR = Constructed-response items; LR = Logistic Regression. The significant DIF items with moderate to large magnitudes of DIF effects are in bold.

* $p < .01$.

On the other hand, the SIBTest DIF analysis flagged 27 DIF items (see column 7 in Table 9; see Appendix I for the sample statistical report of Poly-SIBTest gender DIF results in the U.S. sample), 7 more items than the logistic regression DIF analysis. Fourteen items favored males, and 13 items favored females. The percentage of total number of items identified with DIF was 12.4% (i.e., 27/217). All 27 DIF items were uniform. In terms of content domain, 11 items belonged to *number*, 10 items belonged to *algebra*, zero items belonged to *geometry*, and 6 items belonged to *data and chance*. In terms of cognitive demand, 10 items were *knowing* items, 11 were *applying* items, and 6 were *reasoning* items. In terms of item type, 11 were in multiple-choice formats, and 16 were in constructed-response formats (see Table 9).

For the SIBTest DIF analyses, the estimated beta was used to both identify DIF items and assess the magnitude of DIF for a given item (see column 7 in Table 9). The effect sizes ranged from .08 to .19 for all 27 items flagged with significant DIF. Four items had a magnitude of .08, which were considered a moderate magnitude of DIF effect. The remaining 23 items were classified as large DIF effects.

Table 10 reports the types of DIF items identified by each method. It shows three types of items: items flagged as DIF favoring males, items flagged as DIF favoring females, and items without DIF, arranged by each DIF detection method. The percentage of DIF items identified by logistic regression and SIBTest DIF methods was 9.2% (i.e., 20/217) and 12.4% (i.e., 27/217; see Table 10), respectively. The difference between those percentages was small, at 3.2%. Specifically, the two methods achieved 91.2% (i.e., 198/217) agreement; they mutually tagged 7 DIF items favoring males, 7 DIF items favoring females, and 184 non-DIF items. Please note that this high percentage of agreement is mostly due to the large number of mutually identified non-DIF items. Only 9% of the items (i.e., 19 items) were inconsistently flagged, with 6 items

identified as DIF only by the logistic regression DIF method, and 13 items identified as DIF only by the SIBTest DIF method. No items were identified as favoring one gender by one method and the other gender by the other method.

Table 10

Number of Items identified with DIF Favoring Males, favoring Females, or Non-DIF by Logistic Regression and SIBTest for the U.S. Sample

	SIBTest			
Logistic Regression	DIF (M)	DIF (F)	Non-DIF	Total
DIF (M)	7	0	3	10
DIF (F)	0	7	3	10
Non-DIF	7	6	184	197
Total	14	13	190	217

Note. Items identified as DIF by both LR and SIBTest methods are in boldface. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females.

A chi-square test of independence can test whether both methods were highly correlated in their ability to identify items as DIF and non-DIF. The chi-square test with Table 10 showed statistical significance (χ^2 (4) = 145.98, $p < .001$), indicating that both methods were consistent in identifying items as DIF and non-DIF status. The *contingency coefficient*, indicating the magnitude of association between two DIF methods, was 0.63 (C[14]). Cohen's kappa was also used to examine the degree of consistency between the two methods. The statistic evidence showed the kappa value is different from zero and reached statistical significance (*kappa* = .56; .39 to .74 with 95% confidence interval). Consistent with the chi-square test finding, the kappa indicates an agreement in identifying items as DIF and non-DIF between methods.

[14] $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$.

In addition, Bowker's test of symmetry was used to test whether the number of items identified by only one of the DIF methods was symmetric. The non-significant result (χ^2 (3) = 2.60, $p > .05$) indicated the symmetry for items that were not mutually identified. That is, for those items identified by only one method, neither of the methods showed a higher rate of identifying items as DIF.

In addition, the Pearson correlation coefficient was calculated for the DIF effect sizes between the logistic regression and SIBTest DIF methods. A coefficient of -.37 ($p < .05$) showed a significantly negative relationship of effect sizes between the methods. However, the scatterplot in Figure 2 indicates two clusters of effect sizes, one for uniform DIF items on the upper left of the graph, and one for non-uniform items on the lower right. The heterogeneity between uniform and non-uniform DIF suggested that a correlation test for each group would be more appropriate to describe the relationship between the methods. As a result, correlation tests for uniform and non-uniform DIF items were computed separately, but neither was statistically significant ($r = .320$, $p > .05$; $r = -.230$, $p > .05$). This finding suggests that the two DIF methods were inconsistent in identifying the magnitude of DIF. For example, a DIF item identified as large DIF effect size by one method was not as large when identified by the other method.

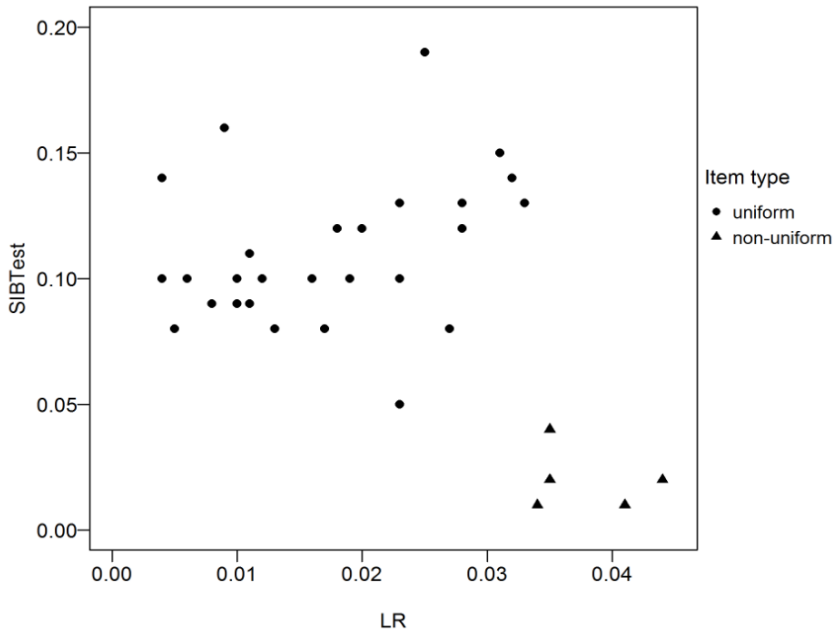


Figure 2. Scatterplot of DIF effect sizes by logistic regression DIF and SIBTest methods for 33 identified DIF items in the U.S. sample.

4.1.2 Consistency of Logistic Regression and SIBTest DIF Results for the Taiwan Sample

I used the same DIF methods to identify statistically significant DIF items and compare the consistency of results between them for the Taiwan sample. For this sample, both methods flagged relatively fewer DIF items compared to those identified in the U.S. sample. The items identified as DIF status by both the logistic regression and SIBTest DIF methods are summarized in Table 11.

Table 11

Summary of Identified DIF Items by Logistic Regression and SIBTest Methods for the Taiwan Sample

TIMSS 2011 Item ID	Content Domain	Cognitive Demand	Item Type	LR Coefficient	LR ($\hat{\Delta} R^2$) Effect Size	SIBTest Beta & Effect size	DIF Favoring
M042059	Number	Knowing	CR	0.78*	0.019	**0.18***	Male
M052126	Algebra	Applying	CR	-1.22*	**0.035**	**-0.13***	Female
M052174B	Number	Applying	CR	0.28*	**0.041**	**0.18***	Male
M042255	Data & Chance	Applying	MC	-0.90*	**0.037**	**-0.13***	Female
M052068	Algebra	Knowing	MC	-0.95*	0.032	**-0.14***	Female
M032692	Geometry	Reasoning	CR	10.60*	**0.073**	-0.07	Male
M042226	Algebra	Knowing	CR	-1.04*	**0.039**	-0.09	Female
M042152	Geometry	Knowing	MC	0.81*	**0.043**	0.11	Male
M052041	Geometry	Reasoning	CR	0.90*	0.023	0.07	Male
M052036	Geometry	Applying	CR	5.26*	**0.057**	-0.10	Male
M042243	Algebra	Knowing	MC	12.97*	**0.137**	-0.02	Male
M042194	Number	Knowing	CR	7.98*	**0.058**	0.05	Male
M042198C	Algebra	Reasoning	CR	-0.87	0.011	**-0.12***	Female
M052422B	Data & Chance	Applying	MC	-0.54	0.015	**-0.12***	Female

Note. MC = Multiple-choice item; CR = Constructed-response items; LR = Logistic Regression. The Significant DIF items with moderate to large magnitude of DIF effects are in boldface.
* $p < .01$.

As shown in column 5 of Table 11, the logistic regression DIF method flagged 12 items with DIF status (i.e., 5.5% as 12 out of 271 items; see Appendix J for the sample statistical reports of ordinal logistic regression gender DIF results in the Taiwan sample), 8 items favoring male students and 4 items favoring female students (see Table 11 on column5). Of the 12 DIF items, 8 were uniform and 4 were non-uniform. In terms of content domain, 3 items belonged to *number*, 4 items belonged to *algebra*, 4 items belonged to *geometry*, and only one item belonged to *data and chance*. In terms of cognitive demand, 6 were *knowing* items, 4 were *applying* items, and 2 were *reasoning* items. In terms of item type, 4 were multiple-choice items, and 8 were constructed-response items.

As shown in column 6 of Table 11, of the 12 items flagged with significant DIF by the logistic regression method, 3 were classified as negligible DIF with a range of effect sizes from .019 to .032. Seven items were classified as moderate DIF, ranging from .035 to .058. Only two items, with .073 and .137, were identified as having large DIF effect sizes (Jodoin & Gierl, 2001).

On the other hand, the SIBTest DIF analysis flagged 7 items with DIF status, 2 items favoring male and 5 items favoring female students (see column 7 in Table 11; see Appendix K for the sample statistical report of Poly-SIBTest gender DIF results in the Taiwan sample). The percentage of the total items identified as DIF items was 3.2% (i.e., 7/217). All of the 7 items were uniform. In terms of content domain, 3 items belonged to *number*, 3 items belonged to *algebra*, zero items belonged to *geometry*, and 2 items belonged to *data and chance*. In terms of cognitive demand, 2 items belonged to *knowing*, 4 were *applying*, and only 1 item was *reasoning*. In terms of item type, 3 were multiple-choice and 4 were constructed-response items.

The magnitudes of the DIF effects are shown in column 7 of Table 11. All 7 items flagged with DIF had large effect sizes, ranging from .12 to .18.

I compared the types of DIF status between the two methods to determine whether they produced comparable DIF results for the Taiwan sample, similar to how I analyzed the U.S. sample. Table 12 shows the number of three kinds of DIF items identified by each DIF method: items flagged with DIF in the favor of males, items flagged with DIF in the favor of females, and items without DIF.

Table 12

Number of Items identified with DIF Favoring Males and Females and Non-DIF by Logistic Regression and SIBTest for the Taiwan Sample

	SIBTest			
Logistic Regression	DIF (M)	DIF (F)	Non-DIF	Total
DIF (M)	2	0	6	8
DIF (F)	0	3	1	4
Non-DIF	0	2	203	205
Total	2	5	210	217

Note. Items mutually identified as DIF by both LR and SIBTest methods are in boldface. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females.

The percentage of DIF items identified by logistic regression and SIBTest DIF methods were 5.5% (i.e., 12/217) and 3.2% (i.e., 7/217; refer to Table 12) respectively, and the percent difference was only 2.3%. As shown in Table 12, the two methods achieved 95.9% (i.e., 208/217) of an agreement; they mutually flagged 2 DIF items favoring males, 3 DIF items favoring females, and 203 non-DIF items. This figure was higher than that for the U.S. sample, which found 91.2% agreement. Please note that the agreement is an overall effect from the agreement for DIF and non-DIF items, respectively. Only 4.1% of the items (i.e., 9 items) were inconsistently flagged between the DIF methods, with 7 items identified as DIF only by the logistic regression DIF method, and 2 items identified as DIF only by the SIBTest DIF method. Likewise, the items where there was disagreement only occurred as one method identified an item as DIF and another method identified as non-DIF, as in the U.S. sample. Consistent with what was observed in the U.S. sample, no items were identified as favoring one gender by one method and the other gender by the other method.

The consistency of item identification between the methods was examined with a chi-square test, which showed statistical significance (χ^2 (4) = 148.38, $p < .001$). This evidence indicates that both methods were consistent in identifying DIF and non-DIF items. The

magnitude of association between methods, investigated with the *contingency coefficient*, was .77 (C[15]). Cohen's kappa was also used to examine the degree of consistency between the two methods. It also showed an agreement between the methods (*kappa* = .51, .24 to .79 with 95% confidence interval).

In addition, Bowker's test, examining the symmetry of items disagreed between the two methods was systematic. The statistical evidence showed the gender DIF rates was equal between methods (χ^2 (3) = 6.3, p > .05). Neither method generated a higher rate of gender DIF items.

Pearson correlation coefficient was computed to determine whether there was an association of DIF effect sizes between the logistic regression and SIBTest DIF methods for the Taiwan sample. As in the U.S. sample, the scatterplot, shown in Figure 3 for the Taiwan sample, seems to indicate two clusters of effect sizes, one for uniform DIF items and one for non-uniform. Again, a correlation test conducted for each type of DIF item would be more appropriately describe the association of the effect sizes between methods, due to the heterogeneity of flagged DIF items. The correlation test for both uniform and non-uniform DIF items was not statistically significant (r = .040, p > .05; r = -.799, p > .05, respectively). The statistical evidence indicated the inconsistency between the two DIF methods in identifying the magnitude of DIF effects.

[15] $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$

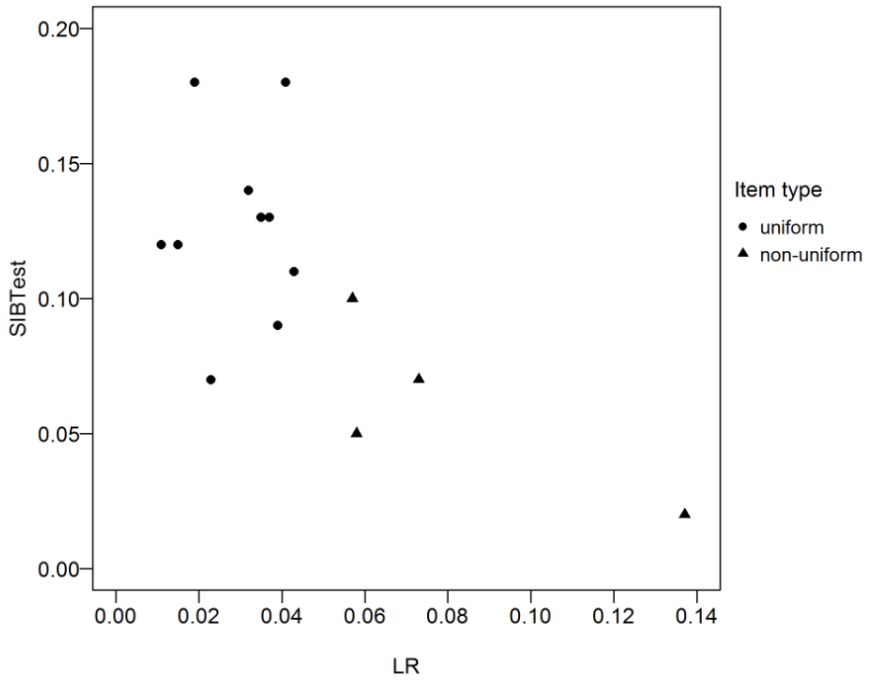


Figure 3. Scatterplot of DIF effect sizes by logistic regression DIF and SIBTest methods for 14 identified DIF items in the Taiwan sample.

Summary of Findings on the Consistency of Gender DIF Item Analysis

The first research question asked whether the gender DIF items identified by the logistic regression method were consistent with those identified by the SIBTest method. The results showed that both DIF methods exhibited equal abilities to identify gender DIF and non-DIF items, regardless of country samples. First, both methods showed 91.2% agreement in identifying items with DIF and non-DIF in the U.S. sample, and 95.9% agreement for the Taiwan sample. Second, the difference between the methods in the percentage of total items identified as DIF was small: a 3.2% difference in the U.S. sample, and a 2.3% difference in the Taiwan sample. Third, according to the results of chi-square, kappa, and Bowker's tests, both methods showed comparable abilities in identifying items as DIF and non-DIF across country samples. Fourth, the effect sizes of DIF items identified by the logistic regression DIF method were smaller than those identified by SIBTest method. The correlation of DIF effect sizes between the

methods showed an insignificant association, which revealed the inconsistency of DIF effect sizes estimated by the methods. The scatterplot of effect sizes showed a cluster of items on the right lower corner, indicating a large difference between methods. Those items were identified as non-uniform by the logistic regression method, and failed to be identified as non-uniform by the SIBTest method.

Overall, the ability to identify items as DIF and non-DIF between the two methods was consistent. However, the two methods differed in classifying the magnitude of DIF effect sizes - the logistic regression method tended to identify DIF items with smaller effect sizes than the SIBTest method.

4.2 Gender DIF Pattern Analyses on Items differing in Cognitive Demand

Research Question 2 asks whether *the gender DIF patterns for mathematics items differ in levels of cognitive demand*. That is, do the gender DIF patterns tend to be the same for items differing in cognitive demand levels? Alternatively, do the gender DIF patterns depend on cognitive demand level? To answer this question, a chi-square test was performed to explore the consistency of gender DIF patterns between cognitive demand levels. This analysis focused mainly on the U.S. sample. The gender DIF pattern analysis for the Taiwan sample is reported in the cross-country comparison under research question 3.

4.2.1 Gender DIF Pattern Analysis in the U.S. Sample

Table 13, a 3 x 3 contingency table, shows the number of items identified as favoring males, favoring females, or non-DIF status for each cognitive demand level. For example, of 80 *knowing* items, 8 were flagged with DIF favoring males, another 8 were flagged with DIF favoring females, and the remaining 65 were flagged with non-DIF status. For the 73 *applying* items, 5 were flagged with DIF favoring males, 6 were flagged with DIF favoring females, and

the remaining 73 were flagged with non-DIF status. For the 52 *reasoning* items, 4 were flagged with DIF favoring males, 2 were favoring females, and the remaining 46 were non-DIF. These counts defined the gender DIF patterns for items differing in cognitive demand level. Please note the gender DIF pattern in Table 13 was developed from the results of gender DIF analysis of both the logistic regression and SIBTest DIF methods. The amount of gender DIF items for this pattern was larger than the amount of identified by each method, because a liberal selection criterion for defining DIF items was used. That is, the item was considered as DIF as long as one of the DIF methods identified the item as DIF.

Table 13

Identified Gender DIF Patterns for Items Differing in Cognitive Demand in the U.S. Sample

Cognitive Demand	DIF (M)	DIF (F)	Non-DIF	Total
Knowing	8(3/5)[a]	8(4/4)	65	80
Applying	5(3/2)	6(2/4)	73	85
Reasoning	4(1/3)	2(1/1)	46	52
Total	17(7/10)	16(7/9)	184	217

Note. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females. Non-DIF = Items not flagged as statistically significant DIF status.

[a] The left side of the figure in the parentheses is the number of DIF items flagged by both the logistic regression and SIBTest methods; the figure on the right side is the number of DIF items flagged by either method.

A chi-square test was used to examine whether the gender DIF patterns among cognitive demand levels were homogeneous. This test indicated the gender DIF patterns were consistent across demand levels, because the chi-square test was not statistically significant ($\chi^2(4)= 2.71$, $p > .05$). The number of items favoring males and females was approximately equal regardless of cognitive demand level as described in the previous paragraph. For example, within *knowing* items, there were 8 items favoring males and 8 items favoring females. Within *applying* items, there were 5 items favoring males and 6 items favoring females.

In the next gender DIF pattern analysis, content domain was added along with the factor of cognitive demand to explore the gender DIF patterns. Table 14, a 4 x 3 x 3 contingency table, reports the number of items identified with DIF favoring males, favoring females, and non-DIF status in each of the cognitive demand levels, grouped by the four content domains. For example, among a total of 61 mathematics items in the content domain *number*, 3 items were identified as DIF items favoring males, 2 items were favoring females, and 24 non-DIF items were identified for the cognitive demand level *knowing*. For 22 *applying*-level items, three items were DIF items favoring males, no items were DIF items favoring females, and 19 items were items without DIF. For the remaining 10 *reasoning*-level items, three were favoring males, two were favoring females, and five were non-DIF items. These three rows defined the gender DIF pattern for items grouped in the content domain *number*. Again, the next three groups of items defined the gender DIF patterns of items differing cognitive demand level for each content domain.

As shown in Table 14, adding content domain as a factor to explore the gender DIF patterns of items differing in cognitive demand level significantly reduced the sample sizes involved, ranging from 6 to 32 for each row of cognitive demand levels. In addition, the number of gender DIF item also decreased to zero to six. Given the reduced sample size for cells, content domain was not included as a factor in analysis of the gender DIF pattern of cognitive demand levels, considering these small sample sizes, the result of gender DIF pattern analyses may not be robust or convincing. In other words, statistical power tends to be lower for identifying a true difference between groups involved in the comparisons.

Table 14

Identified Gender DIF Patterns for Items Differing in Cognitive Demand by Content Domain in the U.S. Sample

Cognitive Demand	DIF (M)	DIF (F)	Non-DIF	Total
Number				
Knowing	3(2/1)	2(0/2)	24	29
Applying	3(2/1)	0(0/0)	19	22
Reasoning	3(1/2)	2(1/1)	5	10
Algebra				
Knowing	2(1/1)	6(4/2)	24	32
Applying	1(1/0)	3(1/2)	17	21
Reasoning	0(0/0)	0(0/0)	17	17
Geometry				
Knowing	1(0/1)	0(0/0)	5	6
Applying	0(0/0)	0(0/0)	21	21
Reasoning	0(0/0)	0(0/0)	16	16
Data & Chance				
Knowing	0(0/0)	0(0/0)	13	12
Applying	1(0/1)	3(1/2)	17	21
Reasoning	3(0/3)	0(0/0)	6	9
Total	17(7/10)	16(7/9)	184	217

Note. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females. Non-DIF = Items not flagged as statistically significant DIF status.

[a] The left side of the figure in the parentheses is the number of DIF items flagged by both the logistic regression and SIBTest methods; the figure on the right side is the number of DIF items flagged by either method.

In the next gender DIF pattern analysis, item type was added as a factor along cognitive demand to explore the gender DIF patterns. All 217 items were first split based on item type. This result in a 2 x 3 x 3 contingency table, as shown in Table 15, which presents the gender DIF pattern of items differing cognitive demand level, grouped by item type. For example, among a total 53 *knowing*-level items in the multiple-choice format, 4 items showed DIF in favor of males, 3 items showed DIF in favor of females, and 46 items were non-DIF. In *applying*-level items in

multiple-choice format, one item showed DIF in favor of males, 4 items showed DIF in favor of females, and 42 items were non-DIF. For the remaining 18 *reasoning*-level items, one item each was found to favor males or females, and 16 were not significant DIF items. The rows of items defined the gender DIF patterns for items classified as multiple-choice or constructed response formats.

Table 15

Identified Gender DIF Patterns for Items Differing in Cognitive Demand by Item Type in the U.S. Sample

Cognitive Demand	DIF (M)	DIF (F)	Non-DIF	Total
Multiple-choice				
Knowing	4(1/3)	3(1/2)	46	53
Applying	1(1/0)	4(1/3)	42	47
Reasoning	1(0/1)	1(0/1)	16	18
Constructed-response				
Knowing	4(2/2)	5(3/2)	18	27
Applying	4(2/2)	2(1/1)	32	38
Reasoning	3(1/2)	1(1/0)	30	30
Total	17(7/10)	16(7/9)	184	217

Note. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females. Non-DIF = Items not flagged as statistically significant DIF status.
[a] The left side of the figure in the parentheses is the number of DIF items flagged by both the logistic regression DIF and SIBTest methods; the figure on the right side is the number of DIF items flagged by either method.

As shown in Table 15, when adding item type as a factor to explore the gender DIF patterns of items differing in cognitive demand level, the sample sizes in the analysis ranged from 18 to 53 items in each level of cognitive demand. When there were DIF items present, the number of gender DIF items ranged from one to five. A log-linear model was used to test the consistency of gender DIF patterns across cognitive demand levels within item type. The statistics show neither a three-way interaction ($G^2(4) = 3.97$, $p = .41$) nor a two-way interaction

($G^2(4) = 1.81$, $p = .77$ for multiple-choice items; $G^2(4) = 6.59$, $p = .16$ for constructed-response items), indicating that item type did not change the gender DIF pattern of items differing in cognitive demand levels. However, conclusions based on this statistic evidence may not be robust, as there are both small sample sizes and a relatively scarce number of gender DIF items present, the same concern raised in the previous analysis involving content domain. Thus, in the gender DIF pattern analysis involving both cognitive demand and item type, the results did not support any conclusions until larger sample sizes and more gender DIF items are included in the analysis.

Summary of Findings on the Gender DIF Pattern Analysis in the U.S. Sample

The second research question asked what the gender DIF patterns of items differing in cognitive demand levels looked like in the U.S. sample. The statistical evidence indicated there is a common gender DIF pattern across items differing in cognitive demand level. That is, the pattern of gender DIF items did not change based on the cognitive demand level of the items. The shared gender DIF pattern was independent of cognitive demand level, and featured an equal amount of items favoring both genders.

It was not possible to add either content domain or item type as factors to examine the gender DIF pattern of items differing in cognitive demand, because of concerns about both small sample sizes and the relatively scarce presence of identified gender DIF items.

4.3 Gender DIF Pattern Analysis across Countries

Research Question 3 is "*Is the gender DIF pattern of items differing in cognitive demand levels similar across countries*?" The third research question asked whether the identified gender DIF pattern of items differing in levels of cognitive demand for the U.S. sample was consistent with other county samples. The research question called for a comparison of the identified

gender DIF patterns between countries. In this dissertation, Taiwan was chosen as a target country for the comparison. To answer this question, the gender DIF pattern of items differing in cognitive demand for the Taiwan sample was summarized and examined first, and then compared with the gender DIF pattern identified for the U.S. sample to examine whether the pattern was comparable across countries.

4.3.1 Gender DIF Item Analysis in the Taiwan Sample

Table 16, a 3 x 3 contingency table, reveals the gender DIF pattern of items differing in cognitive demand for the Taiwan sample. For example, the gender DIF pattern for items in the cognitive demand level *knowing* was 4 DIF items favoring males, 2 DIF items favoring females, and 74 non-DIF items. The gender DIF pattern for items in the cognitive demand level *applying* was 2 DIF items favoring males, 3 DIF items favoring females, and 80 non-DIF items. For items grouped in the cognitive level *reasoning*, the gender DIF pattern showed 2 DIF items favoring males, one DIF item favoring females, and 49 items without DIF.

Table 16

Identified Gender DIF Patterns for Items Differing in Cognitive Demand in the Taiwan Sample

Cognitive Demand	DIF (M)	DIF (F)	Non-DIF	Total
Knowing	4(1/3)	2(1/1)	74	80
Applying	2(1/1)	3(2/1)	80	85
Reasoning	2(0/2)	1(0/1)	49	52
Total	8(2/6)	6(3/3)	203	217

Note. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females. Non-DIF = Items not flagged as statistically significant DIF status.

[a] The left side of the figure in the parentheses is the number of DIF items flagged by both the logistic regression DIF and SIBTest methods; the figure on the right side is the number of DIF items flagged by either method.

A chi-square test was used to examine the consistency among those gender DIF patterns. The test showed there was no two-way interaction ($\chi^2(4)= 1.13$, $p > .05$), indicating that the

gender DIF pattern across cognitive demand levels in the Taiwan sample was consistent. When comparing the DIF items favoring males and females, the number of DIF items was equally distributed for both genders regardless of cognitive demand levels. For example, for *knowing* items, there were 4 items favoring males and the 2 items favoring females. For *applying* items, there were 2 items favoring males and 3 items favoring females. For items that tapped into *reasoning*, there were 2 items favoring males and one item favoring females.

In the next gender DIF pattern analysis, both content domain and item type were added as factors, along with cognitive demand, to explore the gender DIF patterns for the Taiwan sample. Tables 17 and 18 show the gender DIF patterns for content domain and item type, respectively. Both tables show that the number of items involved in the analysis decreased significantly. In addition, the number of gender DIF items also became relatively small, ranging from zero to two DIF items. Thus, I chose not to perform the statistical analyses of the gender DIF pattern including content domain and item type for the Taiwan sample.

Table 17

Identified Gender DIF Patterns for Items Differing in Cognitive Demand by Content Domain in the Taiwan Sample

Cognitive Demand	DIF (M)	DIF (F)	Non-DIF	Total
Number				
Knowing	2(1/1)	0(0/0)	27	29
Applying	1(1/0)	0(0/0)	21	22
Reasoning	0(0/0)	0(0/0)	10	10
Algebra				
Knowing	1(0/1)	2(1/1)	29	32
Applying	0(0/0)	1(1/0)	20	21
Reasoning	0(0/0)	1(0/1)	16	17
Geometry				
Knowing	1(0/1)	0(0/0)	5	6
Applying	1(0/1)	0(0/0)	20	21
Reasoning	2(0/2)	0(0/0)	14	16
Data & Chance				
Knowing	0(0/0)	0(0/0)	13	12
Applying	0(0/0)	2(1/1)	19	21
Reasoning	0(0/0)	0(0/0)	9	9
Total	8(2/6)	6(3/3)	203	217

Note. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females. Non-DIF = Items not flagged as statistically significant DIF status.

Table 18

Identified Gender DIF Patterns for Items Differing in Cognitive Demand by Item Type in the Taiwan Sample

Cognitive Demand	DIF (M)	DIF (F)	Non-DIF	Total
Multiple-choice				
Knowing	0(0/0)	0(0/0)	53	53
Applying	0(0/0)	2(1/1)	45	47
Reasoning	2(0/2)	1(1/0)	15	18
Constructed-response				
Knowing	2(1/1)	1(0/1)	24	27
Applying	2(1/1)	1(1/0)	35	38
Reasoning	2(0/2)	1(0/1)	31	34
Total	8(2/6)	6(3/3)	203	217

Note. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females. Non-DIF = Items not flagged as statistically significant DIF status.
[a] The left side of the figure in the parentheses is the number of DIF items flagged by both the logistic regression DIF and SIBTest methods; the figure on the right side is the number of DIF items flagged by either method.

4.3.2 Gender DIF Pattern Replication across the U.S. and Taiwan Samples

The gender DIF pattern analyses for both the U.S. and Taiwan sample showed a common and shared gender DIF pattern. Table 19 shows two panel datasets. The first two panel datasets, each with a 3 x 3 contingency table, describe the gender DIF pattern of items differing in cognitive demand for the U.S. and Taiwan samples, respectively. A log-linear model was used to test the consistency of the gender DIF patterns between countries. No statistical significant was found for a three-way interaction ($G^2(4) = .59$, $p = .96$) or for a two-way interaction ($G^2(12) = 12.97$, $p = .37$), indicating that the U.S. and Taiwan samples shared similar gender DIF patterns for items differing in cognitive demand levels.

Table 19

Identified Gender DIF Patterns for Items Differing in Cognitive Demand in the U.S. and Taiwan Samples

Cognitive Demand	DIF (M)	DIF (F)	Non-DIF	Total
U.S.				
Knowing	8(3/5)[a]	8(4/4)	65	80
Applying	5(3/2)	6(2/4)	73	85
Reasoning	4(1/3)	2(1/1)	46	52
Taiwan				
Knowing	4(1/3)	2(1/1)	74	80
Applying	2(1/1)	3(2/1)	80	85
Reasoning	2(0/2)	1(0/1)	49	52

Note. DIF (M) = flagged DIF items favoring males. DIF (F) = flagged DIF items favoring females. Non-DIF = Items not flagged as statistically significant DIF status.
[a] The left side of the figure in the parentheses is the number of DIF items flagged by both the logistic regression DIF and SIBTest methods; the figure on the right side is the number of DIF items flagged by either method.

Table 20 reports the distribution of items in both country samples identified as having DIF and non-DIF status. Of the 217 items, both country samples agreed on the DIF status of 178 items, and disagreed on 39 items. There were only 4 items mutually flagged with DIF status in the two country samples. This information was used for the next round of the DIF analysis. A chi-square test of independence showed that the items identified as DIF or non-DIF between countries may not be the same ($\chi^2(4) = 2.07$, $p > .05$). As shown in Table 20, the number of 184 mutually identified items between countries was not high enough to reach statistical significance. The finding does not allow us to claim both countries identify the same items as DIF or non-DIF, which was confirmed by the Cohen's kappa (*kappa* = .09; -.06 to .24 with 95% confidence interval).

Table 20

Number of Items Identified as gender DIF and Non-DIF between the U.S. and Taiwan Samples

		Taiwan		
		DIF	Non-DIF	Total
U.S.	DIF	4	29	33
	Non-DIF	10	174	184
	Total	14	203	217

The McNemar test examining the symmetry of items not mutually identified as DIF and non-DIF between countries, showed statistical significance ($\chi^2(1) = 9.26$, $p < .01$). It indicates that the U.S. sample showed a significantly higher rate of gender DIF items identified as 29 items in comparison to the 10 items identified in the Taiwan sample.

For those items identified as DIF in the U.S. and Taiwan samples, Table 21 reports the distribution patterns of items by cognitive demand, content domain, and item type. Fisher's exact probability test was used to examine the homogeneity of this distribution pattern between the countries. The results did not show significant differences between the patterns at an $\alpha = .05$ level ($p = .84$, .08, and .75 for cognitive demand, content domain, and item type, respectively), indicating a consistent distribution pattern of DIF items between countries, regardless of cognitive demand, content domain, and item type.

Table 21

Distribution of Gender DIF Items in Cognitive Demand, Content Domain, and Item Type within the U.S. and Taiwan Samples

	U.S.	Taiwan
Cognitive Demand		
Knowing	16	6
Applying	11	5
Reasoning	6	3
Content Domain		
Number	13	3
Algebra	12	5
Geometry	1	4
Data & Chance	7	2
Item Type		
Multiple-Choice	14	5
Open-ended	19	9

For the four mutually identified DIF items between countries, two favored males and two favored females in both country samples. Three items belonged to the cognitive demand level of *knowing*, and one item belonged to the cognitive demand level of *applying*. No mutually flagged items were found for the cognitive demand level of *reasoning*. The labels for these four mutually identified items in TIMSS 2011 assessment were M042152, M042194, M042226, and M052422B. Items M042152 and M042226 were released items, which are accessible online, while the other two were not released (see Table 5). Thus, I described these two released items in the following section.

Through content analysis, item features potentially contributing to the mutually identified gender DIF items might be generalized. Two items is too small of a sample to achieve this goal. However, these two items are briefly described for researchers to get a sense of what these items

look like. The item M042226, as shown in Figure 4, is a DIF item favoring females. It is an *algebra* item presented in constructed-response format. This question asks the answer for the variable P. To answer the question correctly, knowledge and basic computation skills related to equations, formulas or functions is required. According to the TIMSS 2011 assessment plan, this item was developed and classified in the cognitive demand level of *knowing*, which was defined as the facts, procedures, and concepts students need to know.

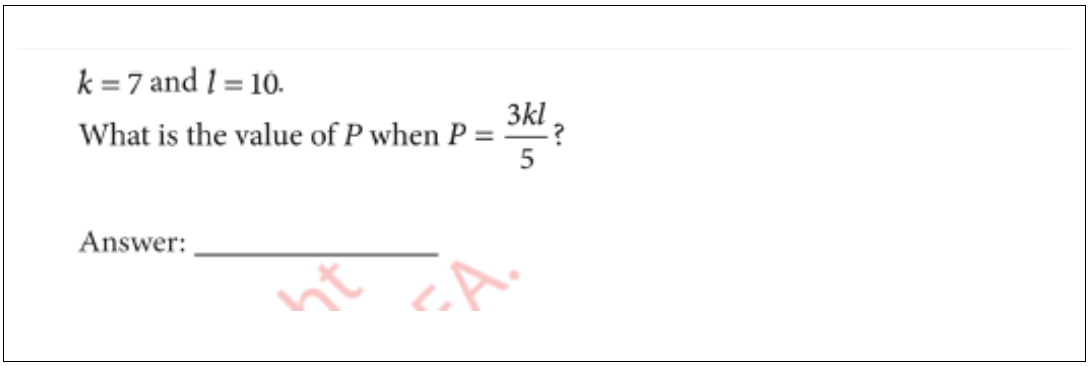


Figure 4. The content of test item ID M042226 in TIMSS 2011 assessment (Foy et al., 2013).

Shown in Figure 5, item M042152 is a DIF item favoring males that was mutually flagged in the U.S. and Taiwan samples. Unlike the item M042226, this is an item related to geometry and presented in the multiple-choice format. This question asks examinees to choose the right answer from four options after the figure in the item stem is turned clockwise 180 degrees. This question requires basic knowledge of location and movement. In the TIMSS 2011 assessment plan, this item was developed and classified as the same as the cognitive demand level of *knowing* required for item M042226 discussed above.

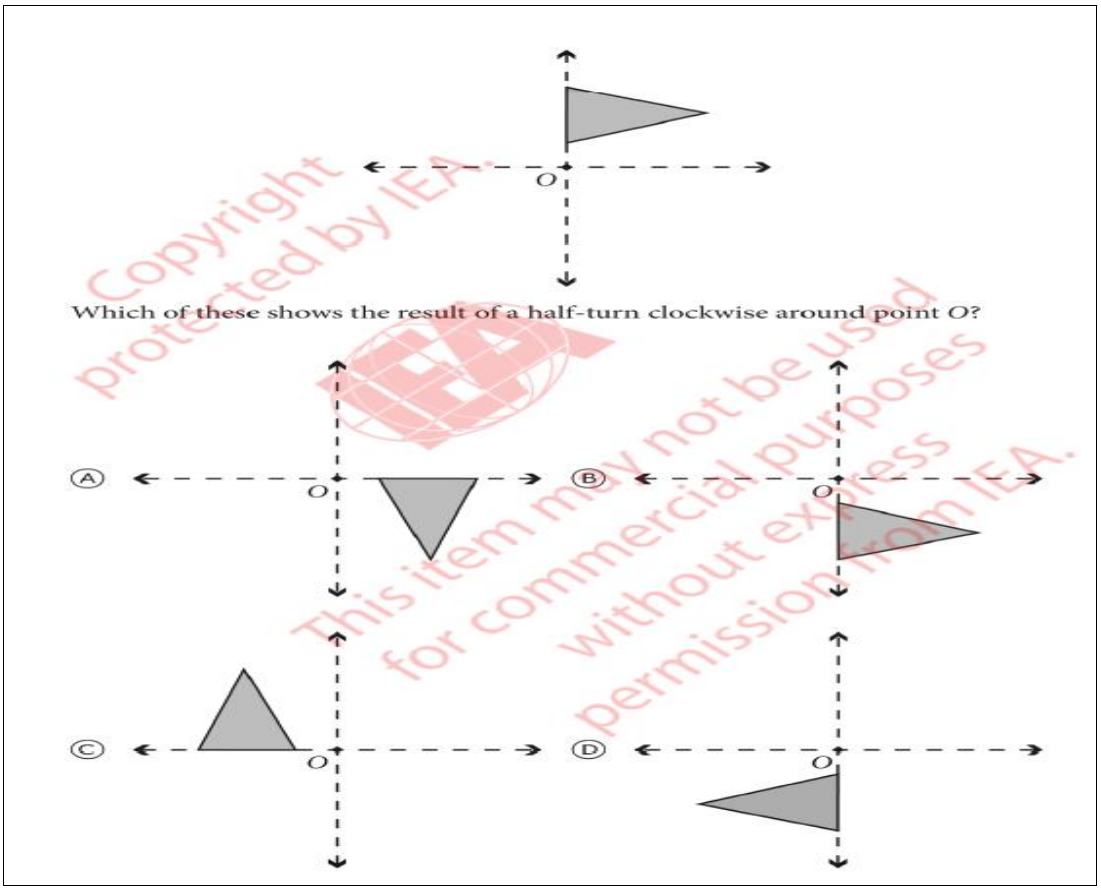


Figure 5. The content of test item ID M042152 in TIMSS 2011 assessment (Foy et al., 2013).

Summary of Findings on the Gender DIF Pattern Replication across Countries

The third research question asked whether the gender DIF pattern of items differing in cognitive demand levels in the U.S. sample was replicable across samples. The findings were as follows. First, statistical evidence indicated that the gender DIF patterns of items differing in cognitive demand levels were consistent across countries (see Table 19). That is, the gender DIF pattern featured an equal amount of gender DIF items favoring both genders. Second, although the gender DIF pattern was shared across country samples, the specific items that shaped the pattern for each country were rather different. There were only 4 mutually identified gender DIF items among the 33 gender DIF items identified in U.S. sample and the 14 gender DIF items

identified in the Taiwan sample (see Table 20). In other words, an item identified with DIF status in the U.S. sample might appear as a non-DIF item in the Taiwan sample, and vice versa. Third, the number of gender DIF items in the U.S. sample was significantly higher than that in the Taiwan sample (i.e., 33 vs. 14). For those gender DIF items identified in each country sample, the distribution patterns of items according to cognitive demand, content domain, or item type was consistent between countries (see Table 21).

Chapter V: Discussions and Conclusions

In this chapter, I begin with a summary of the findings followed by discussion. I then report the limitations and directions for further research, and finally consider the implications and conclusions of this dissertation study.

5.1 Summary and Discussion of Findings

The purpose of this dissertation was to explore whether the TIMSS 2011 mathematics assessment contained gender-based DIF items, as identified by two selected DIF detection methods (i.e., logistic regression and SIBTest DIF), and if so, whether the identified patterns of gender DIF could be attributed to the cognitive demand level of the items. This study also explored whether the findings from the U.S. student sample was replicable for the Taiwan student sample.

5.1.1 The Presence of Gender DIF in Mathematics Items

The first research question aimed to investigate whether there are gender DIF items in the TIMSS 2011 mathematics test. If so, were the gender DIF items identified by the logistic regression DIF method consistent with the items identified by the SIBTest DIF method? With regard to this research question, there are three key findings.

The first key finding pertains to the presence of gender DIF items. Analysis indicated that gender DIF items do occur in the TIMSS mathematics test (see Tables 9 and 11). It is consistent with prior studies investigating gender DIF items for TIMSS mathematics assessments that examined different years of TIMSS administration. For example, Mckenzie (2009) examined gender DIF items in the TIMSS 2003 project and identified 15 items with DIF among 30 selected

mathematics items[16] for eighth grade U.S. student sample. Likewise, for fourth grade U.S. student sample, half of the 30 studied items were identified with DIF. In her study, the SIBTest method was used to flag items with DIF at a specified α-level of .05. In another study of eighth grade U.S. students with the TIMSS-R 1999, Yan (2005) grouped 45 mathematics items by content domain to detect the gender DIF items, using the multifaceted Rasch DIF method also at an α level of .05. Yan found that gender DIF existed in the content domains of *geometry*, *algebra*, *fraction* and *measurement*. The presence of gender DIF items in multiple TIMSS projects based on different DIF detection techniques, revealed that gender DIF does occur in such large-scale mathematics assessments (Mckenzie, 2009; Yan, 2005; Zhang, 2001; Zhang & French, 2010).

The percentage of DIF items identified in the present study was 15.2% (i.e. 33/217) for the U.S. sample, out of 217 total mathematics items (see Table 13) and 6.4% (i.e., 14/271) for the Taiwan sample (see Table 16). The gender DIF item rates reported in this dissertation were lower relative to two previous TIMSS-based gender DIF, which reported rates from 18.5% to 50% (Calvert, 2002; Mckenzie, 2009).[17] The varying amount of identified gender DIF items may result from the total number of items analyzed, the subject matter, and the DIF detection method the researchers selected for their studies.

5.1.2 Comparisons of DIF Methods in Identifying Items as DIF and Non-DIF

The second key finding was related to the ability of the two selected DIF methods to identify gender DIF and non-DIF items. Because of potential inconsistencies between DIF

[16] The TIMSS items were selected based on one of item attributes of: (1) algebra content, (2) visual stimuli presented in a question, (3) word problem format, (4) constructed-response format without visual stimuli, and (5) constructed-response format with visual stimuli.

[17] Researchers rarely provide the information on the rate of DIF items, for example, Bielinski (1999) and Yan (2005). Even in the study by McKenzie (2009), the DIF item rates were only inferred from the reported results instead of being explicitly reported.

methods in identifying the same items as DIF (Fidalgo, Ferreres, & Muñiz, 2004), two DIF detection methods, logistic regression and SIBTest, were used in this dissertation.

As reported in Tables 9 and 11, the two DIF methods were consistent in terms of identifying the same items as DIF and non-DIF, as well as in specifying the DIF direction. In the U.S. sample, there were 14 items mutually identified as DIF items, and 184 items mutually identified as non-DIF items by both methods (see Table 10); and in the Taiwan sample, 5 items were mutually identified as DIF items and 203 items as non-DIF items (see Table 12). The consistency between the two methods reached statistical significance for both country samples (for the U.S. sample, χ^2 (4) = 145.98, $p < .001$, $C = .63$, $p < .001$, and *kappa* = .56, which was also significantly different from zero; for the Taiwan sample, χ^2 (4) = 148.38, $p < .001$, $C = .77$, $p < .001$, and *kappa* = .51, which was also significantly different from zero). This statistical evidence supported the findings from prior studies that compared different DIF methods. For example, Ilich (2013) reported the consistency between the logistic regression DIF and SIBTest DIF methods on a statewide science test. In her study, English Language Learner (ELL) and non-ELL students were matched by science test scores. The study concluded a high agreement between these two DIF methods with regard to both which items were flagged as DIF and which group was favored by the flagged items. The average correlation of regression coefficients estimated by these two DIF methods was .90 across two testing years.

Because of the consistency of the selected DIF detection methods in their abilities of identifying DIF and non-DIF items, a conservative criterion to determine the amount of identified DIF items was not used in this dissertation. This conservative criterion was recommended by Fidalgo, Ferreres and Muniz (2004), as only keeping items identified as DIF items by more than one detection method when selecting DIF items. That is, items are not

considered as showing DIF unless they are identified by both of the selected DIF methods involved in a DIF analysis. The conservative criterion tends to minimize Type I error rates, minimize our ability to see DIF, and maximize Type II error, particularly when the number of DIF methods used in a DIF analysis increases. When applying this criterion to the present study, the amount of gender DIF items dramatically decreased, from 34 to 14 for the U.S. sample (see Table 10) and from 14 to 5 for the Taiwan sample (see Table 12).

Additionally, for those mutually identified DIF items in this study, the DIF direction (favoring either males or females) was identical between the two methods. That is, a DIF item that was flagged as favoring males by the logistic regression was also identified by the SIBTest as favoring males. This provided evidence to support the argument that the logistic regression and SIBTest DIF methods were consistency in identifying DIF items. Again, this finding is aligned with prior research (Ilich, 2013) comparing the consistency of direction for mutually identified DIF items between the same two methods used in this study. The agreement rate in Ilich's study was 100%.

However, I found that the two DIF methods were inconsistent with respect to the magnitude of DIF effects found. Most of the DIF items identified by the logistic regression exhibited small DIF effects in the U.S. sample and moderate effects in the Taiwan sample, while most of the DIF items identified by the SIBTest exhibited large DIF effects for both the U.S. and Taiwan samples (see Tables 9 and 11). Zumbo underscored the importance of including the magnitude of DIF effect when comparing the performance of selected DIF methods: "… This is necessary because small sample sizes can hide interesting statistical effects whereas large sample sizes (like the ones found in typical psychometric studies) can point to statistically significant findings where the effect is quite small and meaningless …" (1999, p. 26). The emphasis place a

threat to decrease the comparability between DIF methods in their ability to identify the same items as DIF and non-DIF. The findings in the present study provided empirical evidence that although the logistic regression and SIBTest methods were consistent in flagging the same items with DIF and non-DIF status, the magnitude of DIF identified by the SIBTest were larger than those identified by the logistic regression DIF method regardless of country samples.

This finding on smaller DIF magnitudes in a logistic regression DIF method is consistent with what Ilich (2013) concluded: DIF items identified by the SIBTest tended to show moderate to high DIF effects, in contrast to small DIF effects for items identified by the logistic regression method. Smaller DIF magnitudes were also found in a logistic regression DIF study conducted by Kelecioglu and Acar (2010) and Hauger and Sireci (2008). In the study of Hauger and Sireci (2008), students who more frequently spoke the language of tests with counterparts who less frequently spoke that language in three different countries were compared. They found that the number of identified DIF items ranged from eight to eleven for each country. All of the DIF items were classified as small without moderate or large magnitude of DIF effects detected in 48 multiple-choice science items from the 1999 TIMSS administration.

In the present study and other studies (Acar & Kelecioğlu, 2010; Hauger & Sireci, 2008; Ilich, 2013), small DIF items identified by logistic regression methods may show moderate to high magnitude of DIF effects as identified by other methods, like the SIBTest. Logistic regression DIF methods tend to have smaller DIF effect sizes, which may be a result of the conservative guidelines Zumbo (1999) developed to evaluate the magnitude of DIF effects. In this dissertation, the more liberal guidelines, developed by Jodoin and Gierl (2001), were used to assess the magnitude of DIF. However, logistic regression still found smaller magnitudes of DIF effect compared to the SIBTest methods. Thus, the guidelines set for logistic regression DIF

methods to recognize DIF effect sizes may be too conservative as French and Maller (2007) suggested that additional studies should re-visit the criterion of the DIF effect size in order to help significant tests to accurately identify DIF items in logistic regression DIF method.

Additionally, the scatterplots of the magnitude of DIF effects between the methods, which showed potential clustering, highlighted that the SIBTest failed to identify the non-uniform DIF items identified by the logistic regression method. Second, for the logistic regression method, non-uniform items tended to show larger DIF effects than those identified as uniform (see Figures 2 and 3). It seems that the magnitudes of DIF for non-uniform DIF items were over-estimated. This phenomenon might reveal an issue that is a result of the DIF detection methods being developed to handle both dichotomous and polytomous items. That is, the ordinal logistic regression DIF method, developed to accommodate polytomous items as well, may inflate the magnitude of DIF for non-uniform items.

5.1.3 Comparisons of DIF Methods in Identifying DIF Items as Uniform and Non-uniform

The inability of the Poly-SIBTest to identify non-uniform DIF items was highlighted as the third key finding for the first research question. This dissertation found, for the U.S. sample, that the ordinal logistic regression DIF method identified five additional non-uniform DIF items that the Poly-SIBTest failed to identify. The Taiwan sample showed the same result – four additional items were identified as non-uniform by the ordinal logistic regression than by the Poly-SIBTest method.[18] The failure to identify non-uniform DIF items by the Poly-SIBTest was

[18] If Poly-SIBTest had been developed to be capable of identifying non-uniform DIF items, the difference in the number of DIF items identified between logistic regression DIF and Poly-SIBTest methods would have been larger for the U.S. sample and relatively smaller for the Taiwan sample. Such a change would lead to potential inconsistency of the items identified with DIF and non-DIF between the two DIF methods for the U.S. sample but not for the Taiwan sample and, accordingly, would alter the current results found by this DIF study – the consistency between methods in their abilities to identify gender DIF across country samples.

also observed by Ilich (2013), who reported that the ordinal logistic regression DIF method identified all the non-uniform DIF items, whereas the Poly-SIBTest did not.

The Poly-SIBTest is one of the three applications included in the DIFPACK DIF software, along with the SIBTest and the crossing-SIBTest.[19] Although the Poly-SIBTest can handle polytomously scored items exclusively or polytomous mixed with dichotomous items, it is unable to identify non-uniform DIF items if analyzing polytomously scored items. To make the DIFPACK application complete and fully able to support the DIF method developed by Shealy and Stout (1993a), it is suggested to develop a function to identify non-uniform DIF for polytomously scored items. In such a circumstance, the outcome of a DIF analysis will more accurately reflect the gender DIF item and pattern that a test may have.

5.1.4 Gender DIF Patterns in Mathematics Items Differing in Cognitive Demand Level

The second research question in this dissertation explores the gender DIF patterns found in mathematics items differing in cognitive demand levels. The analysis for this question was mainly on the U.S. sample. The Taiwan sample is used later to further examine whether the gender DIF patterns found in the U.S. sample was replicable across countries.

The mathematics items in this study were grouped by three levels of cognitive demand: *knowing*, *applying*, and *reasoning*. The three levels implicitly represented mental complexity levels (low, moderate, and high, respectively) that were elicited to solve mathematics questions. The statistical evidence indicated that the gender DIF patterns for mathematics items differing in

[19] Poly-SIBTest, SIBTest and crossing-SIBTest were all developed based on Shealy and Stout's (1993a) unidimensional-based SIBTest model for DIF analysis. Later, the unidimensional concept was evolved into a multidimensional framework. The SIBTest is used to detect uniform DIF for dichotomously scored items. The crossing-SIBTest serves an extension of the SIBTest to detect non-uniform DIF for dichotomously scored items only. With the SIBTest and crossing-SIBTest together, both uniform and non-uniform DIF can be identified for dichotomously scored items. Unlike those two packages, the Poly-SIBTest can be considered an extension of SIBTest for identifying uniform DIF for polytomously scored items.

cognitive demand levels appeared to be similar. For example, as shown in Table 13 for the U.S. sample, the gender DIF pattern for 81 items grouped in the *knowing* level showed 8 DIF items favoring males and 8 favoring females. The gender DIF pattern for 84 items in the *applying* level showed 5 DIF items favoring males and 6 items favoring females, and the DIF pattern for 50 *reasoning* items showed 4 male-favoring and 2 female-favoring items. Because of the approximately equal amount of gender DIF items favoring each gender in each of cognitive demand levels, this study concluded that mathematics items in different cognitive demand levels do not favor either the male or female group.

The finding in this study is consistent with some previous research. First, as pointed out in the literature review section, as a result of re-visiting the gender DIF study by Harris and Carlton (1993), items differing in six levels of cognitive demand actually did not favor either gender. Second, Mendes-Barnett and Ercikan (2006) concluded that there was no gender DIF found favoring either gender on items in lower cognitive demand level. The identified gender DIF effect favoring males for items in the higher level of cognitive demand actually connected to certain content areas without overall DIF effects across all content areas. That is, their study reported that lower cognitive demand items did not favor either gender, and higher cognitive demand items favored males in the first round of an analysis. With further investigation, they found that the bias of higher cognitive demands against females was due to an interaction with content areas. In sum, according to the findings of this dissertation and previous studies, items differing in cognitive demand level do not seem to favor either males or females.

Although I initially planned to include item type and content domain as additional factors to further explore gender DIF patterns of items differing in cognitive demand level, I was unable to complete this analysis due to small number of total items and the small number of gender DIF

items flagged (see Tables 14 and 15). Thus, whether the overall gender DIF pattern can apply to both multiple-choice and constructed-response items or across content domains is still unknown. This dissertation suggests having a large sample size and more gender DIF items collected in order to draw more reliable and convincing conclusions.

5.1.5 Gender DIF Patterns between the U.S. and Taiwan Samples

The third research question in this dissertation asked whether the gender DIF patter on items differing in cognitive demand levels between countries was similar. First, the statistical evidence showed that the gender DIF pattern was replicable across the two country samples (see Table 19). The gender DIF pattern, shared by both samples, featured an equal amount of gender DIF items favoring males and favoring females. The finding is consistent with Calvert's (2002) conclusion, which showed a similar gender DIF pattern in different country samples. Please note that a different pair of countries and science used in her gender DIF study involving cognitive demand levels.

Second, taking a closer look at the DIF items that shaped the gender DIF pattern for each country revealed that the identified gender DIF items were quite different between countries. Of the 33 items identified with gender DIF for the U.S. sample and the 14 items identified with gender DIF for the Taiwan sample, only 4 items were identical (See Table 20). The finding was contradictory to Calvert's (2002). Calvert's DIF study indicated not only that the gender DIF pattern replicated across the U.S. and Spain, but also that the DIF and non-DIF items were identical across countries. Note that Calvert examined a different subject domain, the TIMSS science assessment, and compared different countries, the U.S. and Spain, than those analyzed in this dissertation. Whether or not country and subject matter serve as factors contributing to the inconsistent findings can guide the direction of future gender DIF research.

One potential reason that a different set of items were identified as DIF for each country is described as follows. As the literature pointed out, countries differ in many ways such as emphasized learning goals, expectations in the education system, curriculum arrangement, opportunities for students to learn, and content focus (Burkes, 2009; Klieme & Baumert, 2001). These factors may contribute not only to differential outcomes between genders in their overall mathematics performance but also to unequal correct response rates between genders for individual items identified with DIF status for a given country.

In addition, for those gender DIF items identified in the U.S. and Taiwan samples, the distribution patterns of DIF items according to cognitive demand, content domain, and item type did not differ between the countries (see Table 21).

5.2 Limitations and Recommendation

Findings from this study, and subsequent discussions and generalizations based on those findings, should be interpreted with caution due to some important limitations. In addition to presenting the limitations, this section offers recommendations for future research, including those that are within and beyond the scope of my study. Four limitations and three recommendations are described as follows.

The first limitation pertains to the sample size of each gender group involved in the logistic regression DIF analysis. Zumbo (1999) recommended that a minimum of 200 people per group is adequate for DIF analyses of binary items, with no missing data. Since the test items involved in the current study included both binary and polytomous items, the expected minimum number of examinees per group should be higher than the suggested 200, which was for dichotomously scored items only. After students with missing data were dropped from the current study, the sample size in the analysis ranged from 172 to 193 per group for the Taiwan

sample. Since this study did not adhere to the recommendation for logistic regression DIF studies, the resulting identification of gender DIF items and patterns may be biased (Zumbo, 1999). To investigate this bias, a simulation that varies sample size around the group size of 200 is suggested to find an adjustment or remedy that can tackle the issue of below ideal sample size.

The second limitation of my study deals with the dimensionality of test items. When a test meets the assumption of unidimensionality, the observed total score can be used as a reasonable matching criterion for a DIF analysis. The unidimensionality of a test ensures the homogeneity of all test items developed to measure a single trait or ability. Otherwise, matching on the observed total score may result in inflated Type I error rates, meaning that more items would be identified with DIF than are actually present. The assumption of unidimensionality was not explicitly examined in the present DIF study. Therefore, it is possible that the Type I error rate for the present study exceeded 1%, leading to more falsely identified DIF items. Future research should test the assumption of unidimensionality to ensure that the Type I error rate is nominal.

The third limitation deals with matching males and females to form ability-matched groups, which is required for DIF studies. Two types of measures, observed total score (i.e., the total score examinees obtain from a test) and latent ability, can be used as ability matching variables. IRT-based DIF methods tend to use the latent ability θ (Thompson, 2009), whereas non-IRT based DIF methods, like the two selected methods in this study, tend to use the observed total scores (Camilli, 2006). However, for this study, IRT-based ability scores were used for the logistic regression DIF method (Martin & Mullis, 2013), and observed total scores were used for the SIBTest. Identified gender DIF items and patterns may differ depending the matching variable. As a result, if the latent-ability scores estimated by TIMSS 2011 (Martin & Mullis,

2013) had been replaced by the observed scores for the logistic regression DIF method, the gender DIF items and pattern identified may have been altered. Thus, using the observed score as a matching variable for logistic regression DIF detection is suggested for further research, in order to explore whether the finding is still tenable. Additionally, an external measure of the same trait, from a test other than the TIMSS math test being used in the DIF analysis, could also be used to match males and females. If alternative measures had been available as replacement, the resulting gender DIF items and patterns might have been different from the ones found in present study. Thus, an external measure, serving as an indicator of student performance in mathematics, should be used in further research to test generalizability.

The fourth limitation is related to other item features that might affect the gender DIF pattern that this dissertation was unable to examine.[20] Items involved in each level of cognitive demand level may be different in some ways such as content domain, item type (see Tables 14 and 15), and item difficulty level. Without controlling those item features across cognitive demand level, the effects of cognitive demand on gender DIF patterns may be confounded with the effects of other variables. Particularly, as pointed out by Mendes-Barnett and Ercikan (2006), items in higher cognitive demand level initially found to favor males actually came from the specific content areas of *algebra*, *arithmetic*, and *computation*.

Informed by these limitations, further research should advance gender DIF research in the following three directions. First, gender DIF research should be guided by a cognitive demand model that is well supported by research in the learning sciences and learner cognition, and that

[20] This dissertation originally planned to analyze the confound effect by three *cognitive demands* either with two *item types* or four *content domains* data analysis. However, with only 14 or 33 out of 217 items identified DIF items in this study, a further investigation involving such interaction analysis was impossible. A small amount of DIF items identified as DIF and an insufficient sample size of total items resulted in under-power for the DIF detection.

is empirically validated for its psychometric quality in coding items. The following questions may provide some clues to evaluate the quality of such a cognitive demand model and the corresponding coding system developed to classify mathematics items: Does the coding system provide reliable and valid codes? How is student performance on items differing in cognitive demand level associated with student overall performance? Is the cognitive demand model helpful as a tool to develop learning objectives and goals?

Using with a well-supported cognitive model in gender DIF research not only strengthens the meaningfulness of the DIF findings, but also helps reconcile the research findings based on different cognitive models. For instance, four different cognitive models were used in the gender DIF literature (see Table 1; Harris & Carlton, 1993; Mendes-Barnett & Ercikan, 2006; Mullis et al., 2013). Although all of these models more or less conceptualize cognitive skills required to answer items from low to high, the number of levels vary greatly, ranging from 2 to 6; therefore, an item classified at the middle level in one cognitive demand system may belong to a higher level in another cognitive demand system. The inconsistency of DIF findings among studies may partially result from differences in the cognitive models used.

The second recommended direction is replication of this type of research in order to generalize the findings. For example, in the current study, only half of 14 possible booklets were selected. The other seven even-numbered booklets should be analyzed in order to evaluate the internal validity of this study as a cross-validation. Furthermore, it is unclear whether the results can be replicated with TIMSS items administered in other years, other countries, or the other TIMSS subject, science. Large-scale assessment programs share something in common, including the matrix-sampling strategy to design test items, multiple-stage clustered sampling to select subjects, and IRT-based theories to estimate student ability and item parameters.

Accordingly, the same research methods and research questions in the current study can be applied to large-scale assessment programs like NEAP and PISA to examine the generalizability of the results for large-scale assessment programs.

The third recommendation calls for incorporating demographic variables into gender DIF pattern investigations to examine potential interactions with the identified gender DIF items and patterns. Demographic variables include socio-economic status (SES), urbanization of school location, level of educational resources available, ethnic group membership, and so on. For example, are the identified gender DIF items and patterns consistent across ability-matched gender groups split into high, middle, and low socio-economic statuses? Are the gender DIF items and patterns consistent between gender groups studying in urban school districts versus rural school districts? Investigating these kinds of questions would provide broader ways to examine whether or not the equal correct response rates between ability-matched gender groups is tenable.

5.3 Implications and Conclusions

Differential item functioning (DIF) approaches allow researchers to examine the assumption that there are equal correct response rates among examinees in the same ability level, regardless of their demographic background. Whether a test item demonstrates differential correct response rates between males and females who are matched in their abilities has drawn much attention from researchers. An item that exhibits differential correct response rates between ability-matched males and females is referred to as a gender DIF item. In other words, when the difference of correct response rates between genders is statistically significant, an item will qualify as a gender DIF item. An item identified with DIF implies the item favors ability-matched males or females.

In this dissertation, I chose two DIF detection methods, the logistic regression DIF method and the SIBTest DIF method, to identify gender DIF and non-DIF items as well as the gender DIF pattern of items differing in cognitive demand levels, in TIMSS mathematics items administered in the U.S. and Taiwan. My findings directly contribute to existing literature by providing evidence of gender DIF patterns for mathematics items differing in cognitive demand levels. The implications and conclusions derived from the findings are described as follows.

First, ***gender DIF items were found on mathematics items.*** In this dissertation study, 33 items were identified as gender DIF items in the U.S. sample, equal to 15% of the total 217 items; for the Taiwan sample, 14 items were identified, equal to 6% of the total 217 items. This finding is consistent with other gender DIF analyses, which identified 18.5% to 50% of TIMSS related mathematics items as gender DIF items. The results of this study suggest that gender DIF items do exist in large-scale mathematics tests. In addition, the varying rates of flagged gender DIF items in different mathematics tests are expected because of the potential impacts of the detection method, grade level, the criteria to flag items as having DIF status, and other influential factors specified in each empirical gender DIF study. Regardless of these varying gender DIF identification rates, they clearly demonstrate the existence of gender DIF items in mathematics tests. A DIF item is a potentially biased test item that may contribute to the unfairness of a test. In testing practice, up to five percent of DIF items are considered tolerable because of sampling errors caused by the statistical analysis of gender DIF studies. In this dissertation, the amount of identified gender DIF items was between 6 and 15 percent, which is more than what is expected for a quality test. Therefore, minimizing the presence of items exhibiting unequal item correct response rates for examinees with equal abilities in mathematics tests continues to be a goal for test item developers.

It should be noted that the overall gender DIF effect was cancelled out due to the fact that the amount of DIF items identified favoring both males and females was close to each other. However, no matter what the overall gender DIF effect was, identified gender DIF items that showed significantly unequal response rates were present, and did challenge our fundamental assumption that examinees in same ability level should exhibit equal success rates on an item. Therefore, for each single item identified with DIF, how to adjust those items to minimize the unequal success rates between genders is still a crucial goal for test item developers to achieve.

Second, ***both the logistic regression DIF and the SIBTest DIF methods were consistent in their ability to identify items as DIF and non-DIF.*** This finding implies that, for practitioners of test development, either the ordinal logistic regression or the Poly-SIBTest DIF method can be selected for the DIF detection, if both methods are available. The ordinal logistic regression DIF method was able to detect both uniform and non-uniform DIF items, whereas the Poly-SIBTest method could only detect uniform DIF items. Thus, in testing practice, the ordinal logistic regression DIF method is suggested if the focus of a DIF analysis is to identify both uniform and non-uniform DIF items.

In addition, the logistic regression DIF method tended to classify DIF items as having small to moderate DIF effect sizes, but the Poly-SIBTest DIF method tended to treat those same items as notable DIF items with moderate to large effect sizes. This suggests that researchers focused on exploring the magnitude of DIF effects should expect different results from the logistic regression and SIBTest methods. The inconsistency of both DIF methods in evaluating the magnitude of DIF effects is worth exploring further, to find out how to make both the methods aligned in their abilities to assess the magnitude of identified DIF items.

Third, ***cognitive demand was found to be unrelated to the DIF patterns.*** An equal amount of gender DIF items favoring both males and females were identified at each of the low, middle, and high levels of cognitive demand. Neither males nor females had an overall advantage on test items at any level of cognitive demand. In other words, the gender DIF patterns were independent of cognitive demand levels; or, there was no interaction between the gender DIF patterns and cognitive demand levels. This suggests that cognitive demand might not to be a factor contributing to gender DIF in favor of either males or females. Therefore, when developing mathematics items, cognitive demand does not seem to be a potential source of gender DIF that test item developers need to deal with.

Fourth, ***the gender DIF pattern was found replicable across countries. However, the gender DIF items that shaped the gender DIF pattern for each country were different***. In this study, the gender DIF pattern of items differing in levels of cognitive demand in the U.S. sample was found similar to that in the Taiwan sample. However, the gender DIF items identified to shape that gender DIF pattern were different for each country. Put another way, items that showed differential correct response rates between genders were country specific. In the practice of test development, when considering removing gender DIF items for a translated test used in different countries, country-specific gender DIF studies are required because gender DIF items seem to be country dependent. Even interesting, when gender DIF items are identified and removed, each country may end up with a different set of items, which may have implications for equating and comparing scores across countries.

References

Armstrong, J. M. (1981). Achievement and participation of women in mathematics: Results of two national surveys. *Journal for Research in Mathematics Education*, *12*(5), 356-372.

Acar, T., & Kelecioğlu, H. (2010). Comparison of differential item functioning determination technique: HGLM, LR and IRT-LR. *Educational Science: Theory & Practice, 10* (2), 639-649.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin*, *105*(2), 290-301.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons.

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*(2), 231-251.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association, Inc.

Angoff, W. H. (1982a). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. R. Rubin (Eds.), *Testing equating* (pp. 55-79). New York: Academic Press.

Angoff, W. H. (1982b). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk. (Ed.) *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 3-33). Hillsdale, NJ: Lawrence Erlbaum Associates.

Becker, B. J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. *American Educational Research Journal*, *27*(1), 65-87.

Beller, M., & Gafni, N. (2000). Can item format (multiple-choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles, 42*(1/2), 1-21.

Bennett, R. E. (1993). On the meaning of constructed-response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ.

Ben-Shakhar. G., & Sinai, Y. (2005). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23-35.

Beretvas, S., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200-223.

Bielinski, J. S. (1999). *Sex difference by item difficulty: An interaction in multiple-choice mathematics achievement test items administered to national probability samples.* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Bielinski, J. S., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal*, *35*(3), 455-476.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain.* New York: David McKay Co.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastics achievement. *Journal of Educational Measurement, 27,* 165-174.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*, 307-327.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*, 67-95.

Boughton, K. A., Gierl, M. J., & Khaliq, S. N. (2000). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance*. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Edmonton, Alberta, Canada.

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Burkes, L. L. (2009). Identifying differential item functioning related to student socioeconomic status and investigating sources related to classroom opportunities to learn (Unpublished doctoral dissertation). University of Delaware, Newark, DE.

Burton, N. W. (1996). Have changes in the SAT affected women's mathematics performance? *Educational Measurement: Issues and Practice*, *15*(4), 5-9.

Calvert, T. N. (2002). *Exploring differential item functioning (DIF) with the Rasch model: A comparison of gender differences on eight grade science items in the United States and Spain.* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 221-256). Westport, CT: American Council on Education and Praeger Publishers.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34*(2), 123-39.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Carlton, S. T., & Harris, A. M. (1989). *Female/male performance differences on the SAT: Causes and correlates.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Chaimongkol, S. (2005). *Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective.* (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333–353.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An $_{R}$ package for detecting differential item functioning using iterative hybrid ordinal logistic regression/Item response theory and Monte Carlo simulation. *Journal of Statistical Software, 39* (8), 1-30.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*(4), 269-279.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202-214.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55-77.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*, 131-154.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates.

Doolittle, A. E. (1989). Gender differences in performance on mathematics achievement items. *Applied Measurement in Education, 2*, 161-177.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement tests. *Journal of Educational Measurement, 24*, 157-166.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland and H. Wainer (Eds.) Differential item functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-Bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4),465-484.

Du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Engelhard, G. (1990). Gender differences in performance on mathematics items: Evidence from the United States and Thailand. *Contemporary Educational Psychology, 15*, 13-26.

Feng, Y. (2008). *Difference in gender differential item functioning patterns across item format and subject area on diploma examinations after change in administration procedure.* (Unpublished master's thesis). University of Alberta, Alberta, Canada.

Fennema, E., & Tartre, L. A. (1985). The use of spatial visualization in mathematics by girls and boys. *Journal for Research in Mathematics Education, 16*(3), 184-206.

Fidalgo, A. M., Ferreres, D., & Muniz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *The Journal of Experimental Education, 73*(1), 23-39.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*, 565-582.

Foy, P., Arora, A., & Stanco, G. M. (Eds.) (2013). *TIMSS 2011 user guide for the International Database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315-332.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection**.** *Educational and Psychological Measurement, 67*(3), 373-393.

Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematics tasks. *Review of Educational Research, 59*(2), 185-213.

Gallagher, A. M., & DeLisi, R. (1994). Gender differences in Scholastic Aptitude Test-Mathematics problem solving among high-ability students. *Journal of Educational Psychology, 86,* 204-211.

Garner, M., & Engelhard. G. J. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29-51.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and practice*, *24*(1), 3-14.

Gierl, M. J., & McEwen, N. (1998). *Consistency among statistical methods and content review for identifying differential item functioning*. Presented at the Measurement and Evaluation: Current and Future Research Directions for the New Millennium conference, Banff, AB.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis. *Journal of Educational Measurement*, *40*(4), 281-306.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20,* 26-36.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000). *Performance of Mantel-Haenszel, simultaneous items bias test, and logistic regression when the proportion of DIF items is large*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Glass, G, V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice Hall.

GraphPad Software. (2014). *Quantify agreement with kappa*. Retrieved from http://www.graphpad.com/quickcalcs/kappa1/?K=3

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine, 13*, 1665,-1677.Acar, T., & Kelecioğlu, H. (2010). Comparison of differential item functioning determination technique: HGLM, LR and IRT-LR. *Educational Science: Theory & Practice, 10* (2), 639-649.

Grouws, D. A. (Ed.). (1992). *Handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics.* New York: Macmillan Publishing.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

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

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137-151.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing, 8*(3), 237-250.

Henderson, D. (April, 2001). *Prevalence of gender DIF in Mixed format high school exit examinations.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Hess, K. K. (2006). Exploring cognitive demand in instruction and assessment. Retrieved from http://www.nciea.org/.

Hidalgo, M. D., & Lopez-Pina, J, A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903.

Holland, P. W., & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.) *Test Validity* (pp. 129-145). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: John Wiley.

Hox, J. (1998). Multilevel modeling: When and why. In R. Mathar & M. Schader (Eds.), *Classification, data analysis, and data highways*. Berlin, Germany: Springer-Verlag.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: *A meta-analysis. Psychological Bulletin, 107*(2), 139-155.

Ilich, M. O. (2013). *Differential item functioning (DIF) among Spanish-speaking English language learners (ELLs) in state science tests* (Unpublished doctoral dissertation). University of Washington, Seattle, WA.

Innabi, H., & Dodeen, H. (2006). Content analysis of gender-related differential item functioning TIMSS items in mathematics in Jordan. *School Science and Mathematics, 106*(8), 328-337.

Jeansonne, A. (2002). *Loglinear model*. Retrieved from http://userwww.sfsu.edu/efc/classes/biol710/loglinear/Log%20Linear%20Models.htm

Jian, H., & Stout, W. (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*, 291-322.

Jodoin, G. M. & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.

Kamata, A. (1998). *Some generalizations of the Rasch model: An application of the hierarchical generalized linear model*. (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Kamata, A. (1999). *Multilevel DIF Analysis via hierarchical generalized linear modeling*. Paper presented at the Annual Meeting of the Florida Educational Research Association, Deerfield Beach, FL.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79-93.

Kamata, A. (2002). *Procedure to perform item response analysis by hierarchical generalized linear model.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Kamata, A., & Binici, S. (2003). *Random-effect DIF analysis via hierarchical generalized linear models.* Paper presented at the annual meeting of the Psychometric Society, Sardinia, Italy.

Kaplan, B. J., & Plake, B. S. (1982). Sex differences in mathematics: Differences in basic logical skills? *Educational Studies*, *8*(1), 31-36.

Kelecioglu, H., & Acar, T. (2010). Comparison of differential item functioning determination techniques: HGLM, LR, and IRT-LR. *Educational Science: theory & Practice*, *10*(2), 639-649.

Kiess, H. O. (2002). *Statistical concepts for the behavioral science* (3rd ed.). Boston, MA: Allyn & Bacon.

Kim, W. (2003). *Development of a differential item functioning (DIF) procedure using the hierarchical generalized linear model: A comparison study with logistic regression procedure.* (Unpublished doctoral dissertation). Pennsylvania State University, University Park, PA.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology Education*, *16*(3), 385-402.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle school mathematics performance assessment. *Educational Measurement: Issues and Practices, 15*(4), 21–27, 31.

Langenfeld, T. E. (1997). Testing fairness: Internal and external investigation of gender bias in mathematics testing. *Educational Measurement: Issues and Practice, 16*, 20-26.

Li, H-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher, 18*(8), 17-19, 22-27.

Lowry. R. (2001). Fisher exact probability test: 2x4. Retrieved from http://vassarstats.net/fisher2x4.html

Lu, S-M., & Dunbar, S. B. (1997). *The effects of item characteristics of the Mantel-Haenszel and standardization DIF statistics*. Paper presented at the Annual Meeting of American Educational Research Association, Chicago.

Magis, D., & Facon, B. (2012). Item purification does not always improve DIF detection: A counterexample with Angoff's delta plot. *Educational and Psychological Measurement, 73*(2), 293-311.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from respective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social science*. Monterey, CA: Brooks/Cole.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 Technical Report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2013). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martinez, M. E., (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*, 284-291.

Mazor, K. M., Hambleton, R. K., and Clauser, B. E. (1998). Multidimensional DIF analysis: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*(4), 357-367.

McKenzie, S. (2009). *Differential bundle functioning utilizing theory-based matching subtests for investigating cognitive differences in mathematics.* (Unpublished doctoral dissertation). University of Arkansas, Fayetteville, AR.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: John Wiley & Sons.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining Sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289-304.

Metcalf, L. A. (2002*). Curriculum-sensitive assessment: A psychometric study of tracking as a distributor of opportunity to learn high school mathematics* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign, IL.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*(4), 381-388.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *52*, 92-109.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2013). *TIMSS 2011 international results in mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*(1), 41-68.

O'Connell, A. A. (2006). Logistic regression models for ordinal response variables. Thousand Oak, CA: Sage.

O'Neil, H. F., Jr., & Brown, R. S. (1988). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education*, *11*(4), 331-351.

O'Neil, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland and H. Wainer (Eds.), *differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

OECD. (1999). *Classifying educational programmes: Manual for ISCED-97 implementation in OECD countries* (1999 ed.). Retrieved from http://www.oecd.org/dataoecd/7/2/1962350.pdf

Paek, I. (2012). A note on three statistical tests in the logistic regression DIF procedure. *Journal of Educational Measurement, 49*(2), 121-126.

Paek, I., & Guo, H. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel-Haenszel DIF detection procedure. *Applied Psychological Measurement, 35*, 518-535.

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement, 34*(6), 453-456.

Pomplun, M., & Capps, L. (1999). Gender differences for constructed-response mathematics items. *Educational and Psychological Measurement, 59*(4), 597-614.

PQStat Software. (2009). *The McNemar test, the Bowker test of internal symmetry*. Retrieved from http://pqstat.com/?mod_f=bowker_mcnemar

Rampey, B. D., Dion, G.S., & Donahue, P. L. (2009). NAEP 2008 Trends in Academic Progress (NCES 1009-479). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D. C.

Raudenbush, S.W., Bryk, A.G., Cheong, Y.F., Congdon, R., & du Toit, M. (2004). *HLM-6: Hierarchical linear and nonlinear modeling* [Computer software]. Chicago: Scientific Software International.

Rogers, H. J. (1989). Item bias investigation with logistic regression. (Unpublished doctoral dissertation). University of Massachusetts, Amherst, MA.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*(3), 293-322.

Roussos, L. A., & Stout, W. F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371.

Ryan, K. E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practices, 15*(4), 15-20.

Ryan, K.E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*(1), 73-90.

Shealy, R. T., & Stout, W. F. (1993a). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 197-239). Hillsdale, NJ.

Shealy, R. T., & Stout, W. F. (1993b). A model-based standardization approach that separates true-bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 54*, 159-194.

Smith, W. F. (2009). *Language-related DIF in the WASL mathematics test* (Unpublished doctoral dissertation). University of Washington, Seattle, WA.

Soper, D. (2006). Fisher's exact calculator for a 2x3 contingency table. Retrieved from http://www.danielsoper.com/statcalc3/calc.aspx?id=58

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4-28.

Spray, J. A., & Miller, T. R. (1992). *Performance of the Mantel-Haenszel statistic and the Standardized Difference in Proportion Correct when population ability distributions are incongruent* (ACT Research Report No. 92-1). Iowa City, IA: American College Testing.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293-325.

Stout, W., & Roussos, L. (1996). *SIBTEST manual*. Champaign: University of Illinois, Department of Statistics.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedure. *Journal of Educational Measurement*, *27*, 361-370.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats, *Applied Measurement in Education, 25*(3), 246-280.

Thompson, N. A. (2009). *Ability estimation with Item Response Theory.* St. Paul, MN: Assessment System Corporation.

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach*. (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. Applied Measurement in Education, *9*(2), 175-199.

Wester, A., & Henriksson, W. (2000). The interaction between item format and gender differences in mathematics performance based on TIMSS data. *Studies in Educational Evaluation, 26*, 79-90.

Wikipedia. (2014). *Fisher's exact test*. Retrieved from http://en.wikipedia.org/wiki/Fisher%27s_exact_test

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous item. *Applied Psychological Measurement, 30*(1), 22-42.

Wilson, A. W., (1993). *Logistic regression and its use in detecting nonuniform differential item functioning in polytomous items* (Unpublished doctoral dissertation). University of Arizona, Tucson, AZ.

Wilson, A. W., Spray, J. A., & Miller, T. R. (1993). *Logistic regression and its use in detecting nonuniform differential item functioning in polytomous items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement 35*(2) 145-164.

Yan, S. (2005). *Gender-related differential item functioning in mathematics assessment on the Third International Mathematics and Science Study-Repeat (TIMSSS-R).* (Unpublished doctoral dissertation). University of Toledo, Toledo, OH.

Yates, F (1934). Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society 1(2)*: 217-235 Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67-91.

Zenisky, A. L., Hambleton, R. K., and Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*(1), 49-62.

Zenisky, A., L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1-2), 61-78.

Zhang, M., & French, B. F. (2010, May). *Gender related differential item functioning in mathematics tests: A meta-analysis.* Paper presented at the annual meeting of the National Council of Measurement in Education, Denver, CO.

Zhang, Y. (2001). *Differential item functioning in a large scale standardized mathematics assessment: The interaction of gender and ethnicity.* (Unpublished doctoral dissertation). Ohio University, Athens, OH.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zilberberg, A., Phan, H., Socha, A., Kong, J., & Keng, L. (2011). *The effects of matching type and sample size on the Mantel-Haenszel technique for detecting items with DIF*. Unpublished manuscript.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1997) *A measure of effect size for a model-based approach for studying DIF*. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.

Zwick, R., Thayer, D.T., & Lewis, C. (1999) An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, 36, 1, 1-28.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121-140.

Appendix A: Student Records with Missing Data in the U.S. and Taiwan Samples

For the U.S. Sample

Final Booklet	Item Block	U.S. Sample		
		Male	Female	Total
Booklet 1	M1, M2	1	2	3
Booklet 3	M3, M4	1	1	2
Booklet 5	M5, M6	0	1	1
Booklet 7	M7, M8	0	1	1
Booklet 9	M9, M10	2	1	3
Booklet 11	M11, M12	4	1	5
Booklet 13	M13, M14	3	1	4
Odd Booklet	**M1 - 14**	**11**	**8**	**19**

For the Taiwan Sample

Final Booklet	Item Block	Taiwan Sample		
		Male	Female	Total
Booklet 1	M1, M2	0	1	1
Booklet 3	M3, M4	0	0	0
Booklet 5	M5, M6	0	0	0
Booklet 7	M7, M8	0	0	0
Booklet 9	M9, M10	0	0	0
Booklet 11	M11, M12	0	0	0
Booklet 13	M13, M14	1	0	1
Odd Booklet	**M1 - 14**	**1**	**1**	**2**

Appendix B: A List of the Studied 217 Mathematics Items

Column one lists the TIMSS item ID. The item blocks and their item sequence the test items assigned to are presented in column two and three, respectively. Column four records the final booklet where the items were used. The content domain, cognitive demand, and item type the test items belong to are recorded in columns five to seven, respectively. The last two columns eight and nine list the item labels used in logistic regression DIF and SIBTest DIF analyses.

TIMSS 2011 Item ID	Item Block	Item Sequence	Final Booklet	Content Domain	Cognitive Demand	Item Type	Item ID in LR	Item ID in SIBTest
M032166	M1	01	14 & 1	Number	Knowing	MC	M01.01	bk1.1
M032721	M1	02	14 & 1	Data & Chance	Reasoning	MC	M01.02	bk1.2
M032757	M1	03	14 & 1	Algebra	Reasoning	CR	M01.03	bk1.3
M032760A	M1	04A	14 & 1	Algebra	Reasoning	CR	M01.04A	bk1.4
M032760B	M1	04B	14 & 1	Algebra	Reasoning	CR	M01.04B	bk1.5
M032760C	M1	04C	14 & 1	Algebra	Reasoning	CR	M01.04C	bk1.6
M032761	M1	05	14 & 1	Algebra	Reasoning	CR	M01.05	bk1.7
M032692	M1	06	14 & 1	Geometry	Reasoning	CR	M01.06	bk1.8
M032626	M1	07	14 & 1	Number	Knowing	MC	M01.07	bk1.9
M032595	M1	08	14 & 1	Number	Applying	MC	M01.08	bk1.10
M032673	M1	09	14 & 1	Algebra	Knowing	MC	M01.09	bk1.11
M052216	M2	01	1 & 2	Number	Knowing	MC	M02.01	bk1.12
M052231	M2	02	1 & 2	Number	Knowing	CR	M02.02	bk1.13
M052061	M2	03	1 & 2	Number	Applying	CR	M02.03	bk1.14
M052228	M2	04	1 & 2	Number	Applying	MC	M02.04	bk1.15
M052214	M2	05	1 & 2	Number	Knowing	MC	M02.05	bk1.16
M052173	M2	06	1 & 2	Algebra	Applying	MC	M02.06	bk1.17
M052302	M2	07	1 & 2	Algebra	Knowing	MC	M02.07	bk1.18
M052002	M2	08	1 & 2	Algebra	Applying	CR	M02.08	bk1.19
M052362	M2	09	1 & 2	Geometry	Reasoning	CR	M02.09	bk1.20
M052408	M2	10	1 & 2	Geometry	Reasoning	CR	M02.10	bk1.21
M052084	M2	11	1 & 2	Geometry	Applying	MC	M02.11	bk1.22
M052206	M2	12	1 & 2	Geometry	Reasoning	CR	M02.12	bk1.23
M052429	M2	13	1 & 2	Data & Chance	Reasoning	MC	M02.13	bk1.24
M052503A	M2	14A	1 & 2	Data & Chance	Reasoning	CR	M02.14A	bk1.25
M052503B	M2	14B	1 & 2	Data & Chance	Reasoning	CR	M02.14B	bk1.26
M042032	M3	01	2 & 3	Number	Knowing	MC	M03.01	bk3.1
M042031	M3	02	2 & 3	Number	Applying	MC	M03.02	bk3.2
M042186	M3	03	2 & 3	Number	Reasoning	CR	M03.03	bk3.3
M042059	M3	04	2 & 3	Number	Knowing	CR	M03.04	bk3.4
M042236	M3	05	2 & 3	Algebra	Knowing	MC	M03.05	bk3.5
M042226	M3	06	2 & 3	Algebra	Knowing	CR	M03.06	bk3.6
M042103	M3	07	2 & 3	Algebra	Knowing	CR	M03.07	bk3.7
M042086	M3	08	2 & 3	Algebra	Applying	CR	M03.08	bk3.8
M042228	M3	09	2 & 3	Algebra	Reasoning	CR	M03.09	bk3.9

TIMSS 2011 Item ID	Item Block	Item Sequence	Final Booklet	Content Domain	Cognitive Demand	Item Type	Item ID in LR	Item ID in SIBTest
M042245	M3	10	2 & 3	Algebra	Applying	MC	M03.10	bk3.10
M042270	M3	11	2 & 3	Geometry	Applying	CR	M03.11	bk3.11
M042201	M3	12	2 & 3	Geometry	Applying	CR	M03.12	bk3.12
M042152	M3	13	2 & 3	Geometry	Knowing	MC	M03.13	bk3.13
M042269	M3	14	2 & 3	Data & Chance	Reasoning	MC	M03.14	bk3.14
M042179	M3	15	2 & 3	Data & Chance	Applying	MC	M03.15	bk3.15
M042177	M3	16	2 & 3	Data & Chance	Applying	MC	M03.16	bk3.16
M042207	M3	17	2 & 3	Data & Chance	Applying	CR	M03.17	bk3.17
M052209	M4	01	3 & 4	Number	Knowing	MC	M04.01	bk3.18
M052142	M4	02	3 & 4	Number	Applying	MC	M04.02	bk3.19
M052006	M4	03	3 & 4	Number	Reasoning	MC	M04.03	bk3.20
M052035	M4	04	3 & 4	Number	Knowing	CR	M04.04	bk3.21
M052016	M4	05	3 & 4	Number	Applying	CR	M04.05	bk3.22
M052064	M4	06	3 & 4	Algebra	Knowing	MC	M04.06	bk3.23
M052126	M4	07	3 & 4	Algebra	Applying	CR	M04.07	bk3.24
M052103	M4	08	3 & 4	Algebra	Knowing	MC	M04.08	bk3.25
M052066	M4	09	3 & 4	Algebra	Applying	MC	M04.09	bk3.26
M052041	M4	10	3 & 4	Geometry	Reasoning	CR	M04.10	bk3.27
M052057	M4	11	3 & 4	Geometry	Reasoning	MC	M04.11	bk3.28
M052417	M4	12	3 & 4	Geometry	Applying	CR	M04.12	bk3.29
M052501	M4	13	3 & 4	Data & Chance	Reasoning	CR	M04.13	bk3.30
M052410	M4	14	3 & 4	Data & Chance	Applying	MC	M04.14	bk3.31
M052170	M4	15	3 & 4	Data & Chance	Applying	MC	M04.15	bk3.32
M032094	M5	01	4 & 5	Number	Knowing	MC	M05.01	bk5.1
M032662	M5	02	4 & 5	Number	Reasoning	MC	M05.02	bk5.2
M032064	M5	03	4 & 5	Number	Applying	CR	M05.03	bk5.3
M032419	M5	04	4 & 5	Algebra	Knowing	MC	M05.04	bk5.4
M032477	M5	05	4 & 5	Algebra	Knowing	MC	M05.05	bk5.5
M032538	M5	06	4 & 5	Algebra	Knowing	CR	M05.06	bk5.6
M032324	M5	07	4 & 5	Geometry	Reasoning	MC	M05.07	bk5.7
M032116	M5	08	4 & 5	Geometry	Applying	MC	M05.08	bk5.8
M032100	M5	09	4 & 5	Geometry	Applying	MC	M05.09	bk5.9
M032402	M5	10	4 & 5	Geometry	Reasoning	MC	M05.10	bk5.10
M032734	M5	11	4 & 5	Geometry	Knowing	CR	M05.11	bk5.11
M032397	M5	12	4 & 5	Geometry	Knowing	MC	M05.12	bk5.12
M032695	M5	13	4 & 5	Data & Chance	Applying	CR	M05.13	bk5.13
M032132	M5	14	4 & 5	Data & Chance	Knowing	MC	M05.14	bk5.14
M042041	M6	01	5 & 6	Number	Applying	MC	M06.01	bk5.15
M042024	M6	02	5 & 6	Number	Knowing	MC	M06.02	bk5.16
M042016	M6	03	5 & 6	Number	Applying	MC	M06.03	bk5.17
M042002	M6	04	5 & 6	Number	Reasoning	CR	M06.04	bk5.18
M042198A	M6	05A	5 & 6	Algebra	Knowing	CR	M06.05A	bk5.19
M042198B	M6	05B	5 & 6	Algebra	Reasoning	CR	M06.05B	bk5.20
M042198C	M6	05C	5 & 6	Algebra	Reasoning	CR	M06.05C	bk5.21
M042077	M6	06	5 & 6	Algebra	Knowing	MC	M06.06	bk5.22
M042235	M6	07	5 & 6	Algebra	Knowing	MC	M06.07	bk5.23
M042067	M6	08	5 & 6	Algebra	Applying	MC	M06.08	bk5.24
M042150	M6	09	5 & 6	Geometry	Knowing	MC	M06.09	bk5.25
M042300A	M6	10A	5 & 6	Geometry	Applying	CR	M06.10A	bk5.26
M042300B	M6	10B	5 & 6	Geometry	Applying	CR	M06.10B	bk5.27
M042300Z								
M042260	M6	11	5 & 6	Data & Chance	Knowing	MC	M06.11	bk5.29

TIMSS 2011 Item ID	Item Block	Item Sequence	Final Booklet	Content Domain	Cognitive Demand	Item Type	Item ID in LR	Item ID in SIBTest
M042169A	M6	12A	5 & 6	Data & Chance	Knowing	CR	M06.12A	bk5.30
M042169B	M6	12B	5 & 6	Data & Chance	Knowing	CR	M06.12B	bk5.31
M042169C	M6	12C	5 & 6	Data & Chance	Applying	CR	M06.12C	bk5.32
M032352	M7	01	6 & 7	Algebra	Applying	MC	M07.01	bk7.1
M032725	M7	02	6 & 7	Number	Knowing	CR	M07.02	bk7.2
M032683	M7	03	6 & 7	Algebra	Knowing	CR	M07.03	bk7.3
M032738	M7	04	6 & 7	Algebra	Knowing	MC	M07.04	bk7.4
M032295	M7	05	6 & 7	Algebra	Knowing	MC	M07.05	bk7.5
M032331	M7	06	6 & 7	Geometry	Applying	MC	M07.06	bk7.6
M032623	M7	07	6 & 7	Geometry	Applying	MC	M07.07	bk7.7
M032679	M7	08	6 & 7	Geometry	Knowing	MC	M07.08	bk7.8
M032047	M7	09	6 & 7	Algebra	Applying	MC	M07.09	bk7.9
M032398	M7	10	6 & 7	Geometry	Reasoning	MC	M07.10	bk7.10
M032507	M7	11	6 & 7	Data & Chance	Applying	MC	M07.11	bk7.11
M032424	M7	12	6 & 7	Algebra	Reasoning	MC	M07.12	bk7.12
M032681A	M7	13A	6 & 7	Data & Chance	Knowing	CR	M07.13A	bk7.13
M032681B	M7	13B	6 & 7	Data & Chance	Applying	CR	M07.13B	bk7.14
M032681C	M7	13C	6 & 7	Data & Chance	Applying	CR	M07.13C	bk7.15
M052413	M8	01	7 & 8	Number	Knowing	MC	M08.01	bk7.16
M052134	M8	02	7 & 8	Number	Knowing	MC	M08.02	bk7.17
M052078	M8	03	7 & 8	Number	Applying	MC	M08.03	bk7.18
M052034	M8	04	7 & 8	Number	Knowing	MC	M08.04	bk7.19
M052174A	M8	05A	7 & 8	Number	Applying	CR	M08.05A	bk7.20
M052174B	M8	05B	7 & 8	Number	Applying	CR	M08.05B	bk7.21
M052130	M8	06	7 & 8	Algebra	Knowing	MC	M08.06	bk7.22
M052073	M8	07	7 & 8	Algebra	Knowing	MC	M08.07	bk7.23
M052110	M8	08	7 & 8	Algebra	Knowing	CR	M08.08	bk7.24
M052105	M8	09	7 & 8	Algebra	Applying	CR	M08.09	bk7.25
M052407	M8	10	7 & 8	Geometry	Applying	MC	M08.10	bk7.26
M052036	M8	11	7 & 8	Geometry	Applying	CR	M08.11	bk7.27
M052502	M8	12	7 & 8	Data & Chance	Applying	CR	M08.12	bk7.28
M052117	M8	13	7 & 8	Data & Chance	Applying	CR	M08.13	bk7.29
M052426	M8	14	7 & 8	Data & Chance	Knowing	MC	M08.14	bk7.30
M042183	M9	01	8 & 9	Number	Knowing	MC	M09.01	bk9.1
M042060	M9	02	8 & 9	Number	Knowing	MC	M09.02	bk9.2
M042019	M9	03	8 & 9	Number	Knowing	CR	M09.03	bk9.3
M042023	M9	04	8 & 9	Number	Applying	CR	M09.04	bk9.4
M042197	M9	05	8 & 9	Number	Reasoning	CR	M09.05	bk9.5
M042234	M9	06	8 & 9	Algebra	Knowing	MC	M09.06	bk9.6
M042066	M9	07	8 & 9	Algebra	Reasoning	CR	M09.07	bk9.7
M042243	M9	08	8 & 9	Algebra	Knowing	MC	M09.08	bk9.8
M042248	M9	09	8 & 9	Algebra	Knowing	CR	M09.09	bk9.9
M042229A	M9	10A	8 & 9	Algebra	Applying	CR	M09.10A	bk9.10
M042229B	M9	10B	8 & 9	Algebra	Knowing	CR	M09.10B	bk9.11
M042229Z								
M042080A	M9	11A	8 & 9	Algebra	Knowing	CR	M09.11A	bk9.13
M042080B	M9	11B	8 & 9	Algebra	Knowing	CR	M09.11B	bk9.14
M042120	M9	12	8 & 9	Geometry	Knowing	MC	M09.12	bk9.15
M042203	M9	13	8 & 9	Geometry	Applying	MC	M09.13	bk9.16
M042264	M9	14	8 & 9	Geometry	Reasoning	CR	M09.14	bk9.17
M042255	M9	15	8 & 9	Data & Chance	Applying	MC	M09.15	bk9.18
M042224	M9	16	8 & 9	Data & Chance	Knowing	CR	M09.16	bk9.19

TIMSS 2011 Item ID	Item Block	Item Sequence	Final Booklet	Content Domain	Cognitive Demand	Item Type	Item ID in LR	Item ID in SIBTest
M052017	M10	01	9 & 10	Number	Knowing	MC	M10.01	bk9.20
M052217	M10	02	9 & 10	Number	Reasoning	CR	M10.02	bk9.21
M052021	M10	03	9 & 10	Number	Reasoning	CR	M10.03	bk9.22
M052095	M10	04	9 & 10	Number	Knowing	CR	M10.04	bk9.23
M052094	M10	05	9 & 10	Number	Reasoning	CR	M10.05	bk9.24
M052131	M10	06	9 & 10	Algebra	Applying	MC	M10.06	bk9.25
M052090	M10	07	9 & 10	Algebra	Applying	MC	M10.07	bk9.26
M052121A	M10	08A	9 & 10	Algebra	Reasoning	MC	M10.08A	bk9.27
M052121B	M10	08B	9 & 10	Algebra	Reasoning	CR	M10.08B	bk9.28
M052042	M10	09	9 & 10	Geometry	Applying	CR	M10.09	bk9.29
M052047	M10	10	9 & 10	Geometry	Applying	CR	M10.10	bk9.30
M052044	M10	11	9 & 10	Geometry	Reasoning	MC	M10.11	bk9.31
M052422A	M10	12A	9 & 10	Data & Chance	Applying	MC	M10.12A	bk9.32
M052422B	M10	12B	9 & 10	Data & Chance	Applying	MC	M10.12B	bk9.33
M052505	M10	13	9 & 10	Data & Chance	Knowing	MC	M10.13	bk9.34
M042015	M11	01	10 & 11	Number	Knowing	MC	M11.01	bk11.1
M042196	M11	02	10 & 11	Data & Chance	Knowing	MC	M11.02	bk11.2
M042194	M11	03	10 & 11	Number	Knowing	CR	M11.03	bk11.3
M042114A	M11	04A	10 & 11	Number	Knowing	CR	M11.04A	bk11.4
M042114B	M11	04B	10 & 11	Number	Applying	CR	M11.04B	bk11.5
M042112	M11	05	10 & 11	Algebra	Applying	MC	M11.05	bk11.6
M042109	M11	06	10 & 11	Algebra	Applying	MC	M11.06	bk11.7
M042050	M11	07	10 & 11	Algebra	Knowing	CR	M11.07	bk11.8
M042074A	M11	08A	10 & 11	Algebra	Reasoning	CR	M11.08A	bk11.9
M042074B	M11	08B	10 & 11	Algebra	Reasoning	CR	M11.08B	bk11.10
M042074C	M11	08C	10 & 11	Algebra	Reasoning	CR	M11.08C	bk11.11
M042151	M11	09	10 & 11	Geometry	Applying	CR	M11.09	bk11.12
M042132	M11	10	10 & 11	Geometry	Reasoning	MC	M11.10	bk11.13
M042257	M11	11	10 & 11	Geometry	Reasoning	MC	M11.11	bk11.14
M042158	M11	12	10 & 11	Data & Chance	Knowing	MC	M11.12	bk11.15
M042252	M11	13	10 & 11	Data & Chance	Applying	MC	M11.13	bk11.16
M042261	M11	14	10 & 11	Data & Chance	Knowing	MC	M11.14	bk11.17
M052079	M12	01	11 & 12	Number	Applying	MC	M12.01	bk11.18
M052204	M12	02	11 & 12	Number	Knowing	MC	M12.02	bk11.19
M052364	M12	03	11 & 12	Number	Applying	CR	M12.03	bk11.20
M052215	M12	04	11 & 12	Number	Knowing	CR	M12.04	bk11.21
M052147	M12	05	11 & 12	Number	Applying	MC	M12.05	bk11.22
M052067	M12	06	11 & 12	Algebra	Knowing	MC	M12.06	bk11.23
M052068	M12	07	11 & 12	Algebra	Knowing	MC	M12.07	bk11.24
M052087	M12	08	11 & 12	Algebra	Applying	CR	M12.08	bk11.25
M052048	M12	09	11 & 12	Geometry	Applying	CR	M12.09	bk11.26
M052039	M12	10	11 & 12	Geometry	Applying	CR	M12.10	bk11.27
M052208	M12	11	11 & 12	Geometry	Reasoning	MC	M12.11	bk11.28
M052419A	M12	12A	11 & 12	Data & Chance	Knowing	MC	M12.12A	bk11.29
M052419B	M12	12B	11 & 12	Data & Chance	Knowing	MC	M12.12B	bk11.30
M052115	M12	13	11 & 12	Data & Chance	Applying	MC	M12.13	bk11.31
M052421	M12	14	11 & 12	Data & Chance	Reasoning	CR	M12.14	bk11.32
M042182	M13	01	12 & 13	Number	Applying	MC	M13.01	bk13.1
M042081	M13	02	12 & 13	Number	Knowing	CR	M13.02	bk13.2
M042049	M13	03	12 & 13	Algebra	Knowing	MC	M13.03	bk13.3
M042052	M13	04	12 & 13	Number	Knowing	MC	M13.04	bk13.4
M042076	M13	05	12 & 13	Algebra	Knowing	MC	M13.05	bk13.5

TIMSS 2011 Item ID	Item Block	Item Sequence	Final Booklet	Content Domain	Cognitive Demand	Item Type	Item ID in LR	Item ID in SIBTest
M042302A	M13	06A	12 & 13	Number	Applying	CR	M13.06A	bk13.6
M042302B	M13	06B	12 & 13	Number	Applying	CR	M13.06B	bk13.7
M042302C	M13	06C	12 & 13	Number	Reasoning	CR	M13.06C	bk13.8
M042100	M13	07	12 & 13	Algebra	Knowing	MC	M13.07	bk13.9
M042202	M13	08	12 & 13	Algebra	Applying	MC	M13.08	bk13.10
M042240	M13	09	12 & 13	Algebra	Applying	MC	M13.09	bk13.11
M042093	M13	10	12 & 13	Algebra	Applying	CR	M13.10	bk13.12
M042271	M13	11	12 & 13	Geometry	Applying	MC	M13.11	bk13.13
M042268	M13	12	12 & 13	Geometry	Reasoning	MC	M13.12	bk13.14
M042159	M13	13	12 & 13	Data & Chance	Applying	CR	M13.13	bk13.15
M042164	M13	14	12 & 13	Data & Chance	Reasoning	CR	M13.14	bk13.16
M042167	M13	15	12 & 13	Data & Chance	Reasoning	CR	M13.15	bk13.17
M052024	M14	01	13 & 14	Number	Knowing	MC	M14.01	bk13.18
M052058A	M14	02A	13 & 14	Number	Applying	CR	M14.02A	bk13.19
M052058B	M14	02B	13 & 14	Number	Applying	CR	M14.02B	bk13.20
M052125	M14	03	13 & 14	Number	Reasoning	MC	M14.03	bk13.21
M052229	M14	04	13 & 14	Number	Knowing	CR	M14.04	bk13.22
M052063	M14	05	13 & 14	Algebra	Applying	MC	M14.05	bk13.23
M052072	M14	06	13 & 14	Algebra	Knowing	MC	M14.06	bk13.24
M052146A	M14	07A	13 & 14	Algebra	Reasoning	CR	M14.07A	bk13.25
M052146B	M14	07B	13 & 14	Algebra	Reasoning	CR	M14.07B	bk13.26
M052092	M14	08	13 & 14	Algebra	Applying	MC	M14.08	bk13.27
M052046	M14	09	13 & 14	Geometry	Reasoning	MC	M14.09	bk13.28
M052083	M14	10	13 & 14	Geometry	Applying	MC	M14.10	bk13.29
M052082	M14	11	13 & 14	Geometry	Applying	MC	M14.11	bk13.30
M052161	M14	12	13 & 14	Data & Chance	Applying	MC	M14.12	bk13.31
M052418A	M14	13A	13 & 14	Data & Chance	Applying	MC	M14.13A	bk13.32
M052418B	M14	13B	13 & 14	Data & Chance	Applying	MC	M14.13B	bk13.33

Notes: Items M042300Z and M042229Z were deleted and not included in final mathematics booklets.

Appendix C: Codes of Ordinal Logistic Regression DIF Methods

```
* SPSS SYNTAX written by:      .
* Bruno D. Zumbo, PhD      .
* Professor of Psychology and Mathematics,      .
* University of Northern British Columbia      .
* e-mail: zumbob@unbc.ca     .

* Instructions .
* Copy this file and the file "ologit2.inc", and your SPSS data file into the same folder .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file.
* Run this entire syntax command file.

include file='ologit2.inc'.
execute.


GET
FILE='multicategory.sav'.
EXECUTE .

compute item= item2.
compute total= total.
compute grp= group.

* Regression model with the conditioning variable, total score, in alone.
ologit var = item total
/output=all.
execute.

* Regression model adding uniform DIF to model.
ologit var = item total grp
/contrast grp=indicator
/output=all.
execute.

* Regression model adding non-uniform DIF to the model.
ologit var = item total grp total*grp
/contrast grp=indicator
/output=all.
execute.
```

Appendix D: Sample SPSS Data File Developed for Logistic Regression DIF Analysis

	IDBOOK	ITSEX	m01.01	m01.02	m01.03	m01.04a	m01.04b	m01.04c	m01.05	m01.06	m01.07	m01.08	m01.09
1	1	1	1	0	2	0	0	0	0	0	0	0	1
2	1	1	1	0	2	0	0	0	0	0	1	0	0
3	1	2	1	1	2	2	1	1	2	2	1	1	1
4	1	1	1	1	2	1	0	0	2	2	1	1	1
5	1	2	1	0	2	0	0	0	0	0	0	0	1
6	1	2	1	1	2	2	0	0	0	0	0	1	1
7	1	1	0	0	1	0	0	0	0	0	0	0	0
8	1	2	1	1	2	2	1	1	0	0	1	1	1
9	1	2	1	1	2	1	1	1	1	2	1	1	0
10	1	2	0	0	2	0	0	0	0	0	1	1	1
11	1	2	0	1	2	0	0	0	0	0	1	0	0
12	1	2	1	0	2	2	1	0	0	0	1	1	0
13	1	2	1	0	0	0	0	0	0	0	1	1	0
14	1	1	1	0	0	0	0	0	0	0	0	0	0
15	1	2	0	1	0	0	0	0	0	0	0	1	0
16	1	1	1	1	2	2	0	0	0	0	1	0	0
17	1	1	1	0	2	2	0	0	0	0	1	0	0
18	1	2	1	0	0	2	0	0	0	0	0	1	0
19	1	1	0	0	0	0	0	0	0	0	0	0	1
20	1	2	1	0	0	0	0	0	0	0	0	0	0
21	1	2	1	1	2	0	0	0	0	0	1	0	0
22	1	2	0	1	0	0	0	0	0	0	1	0	0
23	1	2	1	1	2	0	0	0	1	0	1	1	1
24	1	1	1	0	0	0	0	0	0	0	1	1	1
25	1	2	1	1	2	2	1	0	1	1	1	1	0
26	1	2	1	0	1	0	0	0	0	0	0	0	1
27	1	2	0	1	2	0	0	0	0	0	0	0	1
28	1	2	1	0	0	0	0	0	0	0	0	1	1
29	1	2	1	0	0	0	0	0	0	0	0	0	0
30	1	2	1	0	2	2	1	0	0	0	1	1	0
31	1	1	1	0	0	0	0	0	0	0	0	0	0
32	1	1	1	0	2	0	0	0	0	0	1	0	0
33	1	2	0	0	0	0	0	0	1	0	1	1	1
34	1	2	1	1	2	2	1	0	1	0	1	1	1
35	1	2	1	1	0	0	1	0	0	0	0	1	1
36	1	1	1	0	2	0	0	0	0	0	1	0	0
37	1	2	1	0	2	1	1	0	0	0	1	1	1
38	1	1	1	0	0	0	0	0	0	0	1	0	0
39	1	1	1	0	2	2	1	0	0	1	1	1	1
40	1	1	1	0	2	2	0	1	1	0	1	0	1
41	1	2	1	0	2	0	0	0	0	0	0	0	1
42	1	1	1	0	2	2	0	1	2	2	1	1	1
43	1	1	1	0	0	0	0	0	0	0	0	1	0
44	1	2	1	1	2	2	1	0	2	0	1	1	1
45	1	2	1	0	2	2	1	0	1	0	1	1	1
46	1	1	0	0	0	0	0	0	0	0	1	1	0

Data View | Variable View

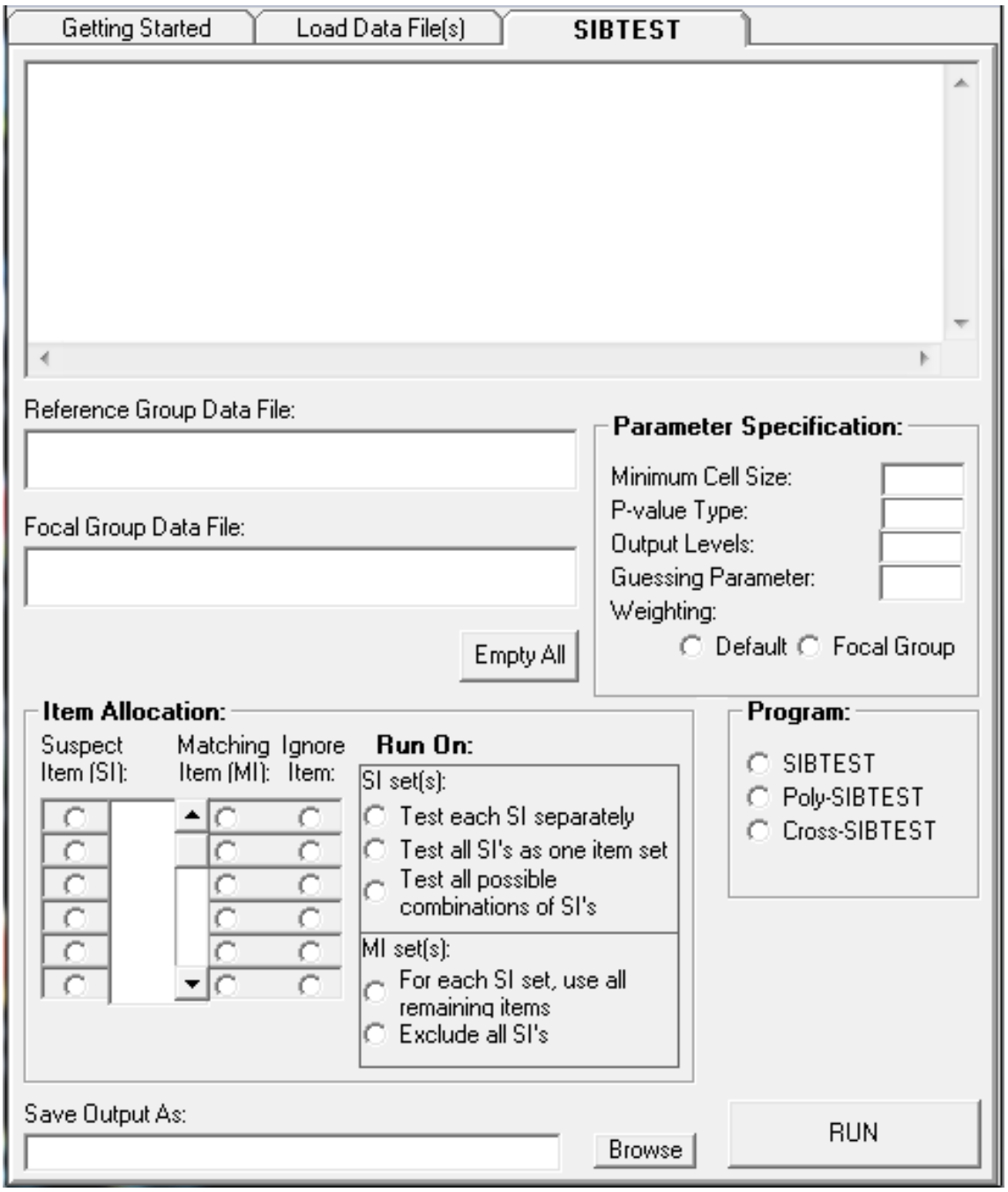
	m02.01	m02.02	m02.03	m02.04	m02.05	m02.06	m02.07	m02.08	m02.09	m02.10	m02.11	m02.12	m02.13	m02.14a	m02.14b	BSMMAT01
1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	482.72
2	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	430.33
3	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	625.44
4	1	1	1	1	1	0	1	1	0	0	1	1	1	0	0	609.33
5	1	1	0	0	1	0	1	0	0	0	0	0	1	1	1	470.69
6	0	1	1	1	0	0	1	0	0	0	1	1	1	1	0	509.11
7	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	310.45
8	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	538.32
9	1	1	1	0	0	0	1	0	1	1	1	0	1	1	0	551.50
10	1	1	0	0	1	1	1	0	0	0	0	0	1	0	0	441.21
11	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	386.39
12	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	513.86
13	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	425.24
14	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	427.13
15	1	1	1	0	0	0	1	0	0	0	1	0	0	0	0	444.06
16	1	1	0	0	0	0	1	0	1	0	1	0	1	0	0	453.42
17	1	1	1	0	1	0	1	0	1	0	0	0	0	0	0	443.05
18	1	1	1	0	1	0	1	0	0	0	0	0	1	0	0	470.75
19	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	331.18
20	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	349.93
21	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	420.46
22	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	395.95
23	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	469.07
24	1	1	0	0	1	0	0	0	0	0	0	0	1	0	1	488.69
25	1	1	1	1	0	0	1	0	1	1	1	0	1	1	0	578.73
26	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	337.96
27	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	371.59
28	1	1	1	1	0	0	1	0	0	0	1	0	1	0	0	454.68
29	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	453.46
30	1	1	1	0	1	0	1	0	0	0	1	0	1	0	1	531.65
31	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	444.68
32	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	367.06
33	1	0	1	0	0	0	1	0	1	0	1	0	1	0	1	497.90
34	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	640.65
35	1	1	1	0	0	0	0	0	0	0	1	0	1	0	0	493.00
36	1	1	1	0	0	0	1	0	0	1	0	0	1	0	0	514.94
37	1	1	1	1	0	0	1	0	1	0	1	0	1	0	0	561.15
38	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	371.66
39	0	1	1	1	0	1	1	0	0	1	1	0	1	1	0	559.32
40	1	1	1	0	0	0	1	0	1	0	1	0	0	1	0	502.56
41	0	1	1	1	0	0	1	0	1	1	1	1	1	1	1	513.24
42	1	1	1	0	1	1	1	2	1	1	1	1	1	1	1	646.75
43	1	1	0	0	0	0	1	0	0	0	1	0	1	1	1	482.99
44	1	1	1	1	0	0	1	0	0	0	1	0	1	0	0	591.48
45	1	1	0	0	1	0	1	0	1	0	1	0	1	1	0	570.17
46	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	446.95

Appendix E: Modified Codes Developed for Ordinal Logistic Regression DIF Analysis

```
** DIF analysis with LR (for ordinal items)
     timss 2011 math items, 8th, USA
     codes developed by Bruno Zumbo
*.
* SPSS SYNTAX written by:      .
* Bruno D. Zumbo, PhD      .
* Professor of Psychology and Mathematics,      .
* University of Northern British Columbia      .
* e-mail: zumbob@unbc.ca     .
*.
* Instructions .
* Copy this file and the file "ologit2.inc", and your SPSS data file into the same folder .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file.
* Run this entire syntax command file.
*.
*********************************************************.
*********************************************************.
***** gender DIF analysis for booklet 1.

get file='o:\bscUSAm5_MathItemOnly_Oddbklt,ItemRenamed,Bk01,NoMissing.sav'.

include file='ologit2.inc'.
execute.

*****************************************.

compute item = m01.01.
compute total = bsmmat01.
compute grp = itsex.

*+++++++++++++++++++++++++++.

**** Regression model with the conditioning variable, total score, in alone.
ologit var = item total
  /output = all.
execute.

**** Regression model adding uniform DIF to model.
ologit var = item total grp
  /output = all.
execute.


**** Regression model adding non-uniform DIF to the model.
```

```
ologit var = item total grp total*grp
  /output = all.
execute.

*++++++++++++++++++++++++++.
```

Appendix F: Interface of DIFPACK Software Applications



Appendix G: Sample ASCII Data File Developed for Poly-SIBTest DIF Analysis

```
For females, booklet=1, the US sample
10200000001110010100000000
10200000100110000100100000
11210022111111111010011100
00100000000100000000000100
10000000000000000011000010100
11220000100110000101010100
10220000100111010101000000
00000000001010100000010000
10000000111111001000000101
10000000000011000010000000
10200000100110000100000000
10200000100111000100100100
10000000100001010010000000
10221001111011101100110110
10220110101111100010101010
10220122111111101112111111
10000000010110000100010111
00000000110110000100000000
00200000101011010100010100
00000000000011001110000010000
10200000010110000100011111
01200000111111000000010000
11221110110110100101011100
10201000111111001010000000
00200000100000000100000000
10221100011111101010101010
00200000110100000100010000
10200000100111010100010100
11000000001111110010000000
10220100110111111010011100
00200000110010010101001000
10020000100110000100110000
11201000111111110010011010
11200000011111101010000100
11000000101010000101000011
00000000100110000000000000
11220110111111010010101010
11220000111111001000000000
11221100011111100111101100
10000000000011000010000100
10201012111111101010111100
10200000111111000100100000
11220002111110110111011110
11200000000111000100000000
10220110101111110110010100
10221120111111100011001111
01000000000010000100000001
11200000101111111010011101
10200000000100110100000100
11000000011111000100000000
10000000001011000000000000
00000000000010000000000000
10100100010111000100100001
10200000101110100100011000
11000000100110000100010000
10221000101111111010011110
11221100111111111010111010
10200000111111000101110101
10200000100110000100000000
10200000101111100010100100
10221110111111100000000000
10200000101111110010100100
10000002110111000100010110
10221112101111101012001100
10220012111111011011001011
10200000000111010111011100
10221100010111010100110010
10000000011111000010010010
10221100101111101010001011
00000000011111100010011101
10000000100110000100010000
11210100111111100010010100
```

For males, booklet = 1, the US sample
1122112211111111111111011100
1020000000111001010000011 1
1122000001101110010001111 0
1122110011111100001000100 01
1121111211011100010111011 0
0020000011111001110000010 0
0120000010001001000000000 0
1022100011011000010000110 0
1000000011011111000000010 0
0100000001011110001000100 00
1002000001011110101000001 00
1000000000000010000000010 0
1120000010011000000000010 0
0100000010011000010000000 0
1120001011101000010101000 0
1122101111011110010111011 0
1010000000100000010000000 0
0120000000101000100001010 0
1000000001111111001000101 00
1000000000011000100000000 0
1022100011011101010001010 1
0000001011110100010101010 1
1122101011111111111110100
1100100001111110000000101 00
1021100011111111001010101 00
1020000000101110010111111 1
1122102011111111001000101 00
1022101011111100101010101 10
1022110011111100101101100 00
1110000010101100010111100 0
0020000010011000010000000 1
0000000000000000000000000 0
0100000010110000000000000 0
1122011000111100010001000 0
1000000011111111101011111 00
1122012201111111012110010 0
0100100000111000011000000 0
0112000000000000000000000 0
1012000010111100010011010 0
1022111011011110010101010 1
1122100101111111001010101 10
1000000010000101010000010 0
0010000001111101010100010 0
1120000001000000010000000 0
1122001011010100010000100 10
1000000011111111001010101 00
1122001211111110011100010 10
1000000011101000010001010 0
1122110200111110001010101 00
1100000000111110001000000 1
1001000011011100010001010 0
1000000010111000000001010 0
1122012011011101010111111 0
1000000011111110001200101 00
1022100011011100010101011 0
1020000010111001010000110 0
1010000011111110101010111 00
1020101001111110101000101 00
0022100001111111001000001 00
1122010011011100110001010 0
1100000211111100010100010 0
1020000010001000000000000 0
0000000000011000000010000 0
1121011011111111001010101 00
1020000000111001010000010 0
1022110001111111001010001 00
1000000001001000010001100 0
0100000001011000010001010 0
1000000011011001010101110 0

Appendix H: Ordinal Logistic Regression Gender DIF Results, the U.S. Sample

```
** DIF analysis with LR (for ordinal items)
     timss 2011 math items, 8th, USA
     codes developed by Bruno Zumbo
*.
* SPSS SYNTAX written by:      .
* Bruno D. Zumbo, PhD      .
* Professor of Psychology and Mathematics,      .
* University of Northern British Columbia      .
* e-mail: zumbob@unbc.ca    .
*.
* Instructions .
* Copy this file and the file "ologit2.inc", and your SPSS data file into the sa
     me folder .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file
     .
* Run this entire syntax command file.
*.

************************************************************.
************************************************************.
***** gender DIF analysis for booklet 1.

get file='o:\bscUSAm5_MathItemOnly_Oddbklt,ItemRenamed,Bk01,NoMissing.sav'.

include file='ologit2.inc'.
   30    set printback off.
   31    ******************* OLOGIT 2.0 (test version) ****************************
             **
   32
   33    * By Prof. Dr. Steffen Kuehnel
   34       Institut fuer Politikwissenschaft
   35       Justus-Liebig-Universitat Giessen
   36       Karl-Glockner-Str 21, Haus E
   37       35394 Giessen
   38       Germany.
   39
   40    * contrast subcommand added by
   41       John Hendrickx <J.Hendrickx@maw.kun.nl>
   42       Department of Sociology
   43       University of Nijmegen
   44       P.O. Box 9104
   45       6500 HE Nijmegen
   46       The Netherlands.
   47
```

```
  48    * This macro is available from
  49        <http://www.socsci.kun.nl/maw/sociologie/resources/mlogist>.
  50
  51    * macro subroutine, called by class.
execute.

*****************************************.

compute item = m01.01.
compute total = bsmmat01.
compute grp = itsex.

*+++++++++++++++++++++++++++.

**** Regression model with the conditioning variable, total score, in alone.
ologit var = item total
  /output = all.

Matrix

Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE

                        (by Steffen M. KUEHNEL)

******************** Information Section **********************

Dependent variable is:
  item

Marginal distribution of dependent variable
        Value       Frequ.      Percent      %>Value
          .00       114.00        15.66        84.34
         1.00       614.00        84.34          .00

Effective sample size:
       728

Means and standard deviations of independent variables:
                    Mean      Std.Dev.
total      507.5340        74.4755

******************** Estimation Section **********************

Running Iteration No.:
```

1

Running Iteration No.:
2

Running Iteration No.:
3

Running Iteration No.:
4

Running Iteration No.:
5

..... Optimal solution found.

********** OUTPUT SECTION **********

LR-test that all predictor weights are zero

-2 Log-Likelihood of Model with Constants only:
631.871

-2 Log-Likelihood of full Model:
499.977

LR-statistic

Chisqu.	DF	Prob.	%-Reduct
131.894	1.000	.000	.209

Estimations, standard errors, and effects

	Coeff.=B	Std.Err.	B/Std.E.	Prob.	exp(B)
total	.017690	.001802	9.817194	.000000	1.017848
Const.1	-6.790732	.834169	-8.140719	.000000	.001124

Results assuming a latent continuous variable

R-Square (%):
34.54

Standardized regression weights of the latent variable:

```
total      .5877

------ END MATRIX -----

execute.

**** Regression model adding uniform DIF to model.
ologit var = item total grp
  /output = all.

Matrix

Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE

                         (by Steffen M. KUEHNEL)

********** Information Section **********

Dependent variable is:
  item

Marginal distribution of dependent variable
        Value       Frequ.      Percent     %>Value
          .00       114.00        15.66        84.34
         1.00       614.00        84.34          .00

Effective sample size:
       728

Means and standard deviations of independent variables:
                    Mean      Std.Dev.
total      507.5340        74.4755
grp          1.5151          .5001

********** Estimation Section **********

Running Iteration No.:
     1

Running Iteration No.:
     2

Running Iteration No.:
     3
```

Running Iteration No.:
4

Running Iteration No.:
5

..... Optimal solution found.

********** OUTPUT SECTION **********

LR-test that all predictor weights are zero

-2 Log-Likelihood of Model with Constants only:
631.871

-2 Log-Likelihood of full Model:
499.923

LR-statistic

Chisqu.	DF	Prob.	%-Reduct
131.948	2.000	.000	.209

Estimations, standard errors, and effects

	Coeff.=B	Std.Err.	B/Std.E.	Prob.	exp(B)
total	.017686	.001802	9.816038	.000000	1.017843
grp	-.053318	.229398	-.232428	.816206	.948078
Const.1	-6.707615	.906255	-7.401467	.000000	.001222

Results assuming a latent continuous variable

R-Square (%):
34.53

Standardized regression weights of the latent variable:
total .5876
grp -.0119

------ END MATRIX -----

execute.

**** Regression model adding non-uniform DIF to the model.
ologit var = item total grp total*grp
 /output = all.

Matrix

Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE

(by Steffen M. KUEHNEL)

Interaction term total*grp
 int1.1 total grp

********** Information Section **********

Dependent variable is:
 item

Marginal distribution of dependent variable

Value	Frequ.	Percent	%>Value
.00	114.00	15.66	84.34
1.00	614.00	84.34	.00

Effective sample size:
 728

Means and standard deviations of independent variables:

	Mean	Std.Dev.
total	507.5340	74.4755
grp	1.5151	.5001
int1.1	769.0654	280.7082

********** Estimation Section **********

Running Iteration No.:
 1

Running Iteration No.:
 2

Running Iteration No.:
 3

Running Iteration No.:
4

Running Iteration No.:
5

..... Optimal solution found.

********** OUTPUT SECTION **********

LR-test that all predictor weights are zero

-2 Log-Likelihood of Model with Constants only:
631.871

-2 Log-Likelihood of full Model:
499.253

LR-statistic

Chisqu.	DF	Prob.	%-Reduct
132.619	3.000	.000	.210

Estimations, standard errors, and effects

	Coeff.=B	Std.Err.	B/Std.E.	Prob.	exp(B)
total	.022274	.005988	3.719551	.000200	1.022524
grp	1.308625	1.685242	.776521	.437441	3.701082
int1.1	-.002971	.003643	-.815469	.414804	.997034
Const.1	-8.810054	2.763582	-3.187911	.001433	.000149

Results assuming a latent continuous variable

R-Square (%):
34.98

Standardized regression weights of the latent variable:

total	.7375
grp	.2910
int1.1	-.3707

------ END MATRIX -----

execute.

*++++++++++++++++++++++++++.

Appendix I: Poly-SIBTest Gender DIF Results, the U.S. Sample

```
 name of input parameter file = sib.in

 number of items on test =  26
 name of file for Ref. grp. scores = o:\DIF, sibtest, t11, usa, bk1, 2=m,
reference.dat

 name of file for Focal grp. scores = o:\DIF, sibtest, t11, usa, bk1, 1=f, focus.dat

 minimum no. of examinees per matching score cell =   2
 number of runs for this data set =  26
 number of examinees in Reference Group =    375
 number of examinees in Focal group     =    353

 Examinee Test Score Summary Statistics

     Reference Group:              Mean =  12.91
                     Standard deviation =  6.08
     Focal Group:                  Mean =  12.97
                     Standard deviation =  6.21

     Standardized Score Difference = -0.01

 Item Statistics
    # = item number
    m = mean score on item
    r = point biserial (item score-test score correlation)

    #:     1     2     3     4     5     6     7     8     9    10
    m: 0.843 0.357 1.305 0.804 0.269 0.206 0.321 0.308 0.635 0.628
    r: 0.401 0.286 0.602 0.698 0.602 0.576 0.644 0.508 0.405 0.549

    #:    11    12    13    14    15    16    17    18    19    20
    m: 0.621 0.831 0.886 0.587 0.315 0.376 0.092 0.874 0.216 0.400
    r: 0.504 0.422 0.247 0.556 0.418 0.313 0.265 0.366 0.536 0.538

    #:    21    22    23    24    25    26
    m: 0.234 0.523 0.265 0.672 0.196 0.176
    r: 0.379 0.514 0.490 0.495 0.332 0.210

    p-value notation:
 R  denotes p-value for test of DIF/DBF against Ref. group
 F  denotes p-value for test of DIF/DBF against Foc. group
 E  denotes p-value for test of DIF/DBF against either the
                                        Ref. or Foc. group.

 NOTES:
         MS/SSD = Matching Subtest Standardized Score Difference.
                  Standardized difference in mean observed scores
                  between Reference group and Focal group on the
                  matching subtest.

         p-elim = proportion of Reference (R) and Focal (F) groups
                  eliminated (not used) in SIBTEST calculations.

         Positive Beta estimate indicates DIF/DBF favoring Ref. grp.
         Negative Beta estimate indicates DIF/DBF favoring Foc. grp.

         FLAG = error flag indicator.  FLAG=0 indicates a normal
                successful completion of a SIBTEST run.  All other values
                of FLAG come with short error messages.
```

SIBTEST-pooled weighting

Run no.	Suspect Subtest Item Numbers	Beta estimate	standard error	p-value		p-elim R	p-elim F	MS SSD	FLAG
1	1	-0.017	0.025	0.490	E	.02	.04	-0.01	0
2	2	0.035	0.036	0.327	E	.01	.02	-0.02	0
3	3	-0.082	0.061	0.175	E	.02	.04	0.00	0
4	4	-0.021	0.058	0.724	E	.01	.03	-0.01	0
5	5	0.046	0.029	0.107	E	.01	.02	-0.02	0
6	6	0.010	0.027	0.699	E	.02	.04	-0.01	0
7	7	-0.015	0.036	0.679	E	.01	.03	-0.01	0
8	8	-0.053	0.048	0.268	E	.01	.01	0.00	0
9	9	-0.061	0.035	0.083	E	.02	.05	0.00	0
10	10	0.126	0.031	0.000	E	.02	.04	-0.03	0
11	11	0.051	0.033	0.121	E	.01	.03	-0.02	0
12	12	-0.061	0.026	0.020	E	.02	.04	0.00	0
13	13	-0.019	0.024	0.425	E	.02	.04	-0.01	0
14	14	0.048	0.033	0.140	E	.02	.04	-0.02	0
15	15	0.062	0.033	0.061	E	.01	.03	-0.02	0
16	16	-0.029	0.036	0.415	E	.02	.04	0.00	0
17	17	0.023	0.020	0.231	E	.01	.03	-0.01	0
18	18	-0.083	0.023	0.000	E	.02	.04	0.00	0
19	19	-0.019	0.034	0.570	E	.01	.04	-0.01	0
20	20	0.045	0.034	0.186	E	.02	.04	-0.02	0
21	21	0.009	0.031	0.762	E	.01	.05	-0.01	0
22	22	0.046	0.034	0.172	E	.02	.04	-0.02	0
23	23	-0.027	0.030	0.382	E	.01	.03	-0.01	0
24	24	0.052	0.033	0.107	E	.02	.04	-0.02	0
25	25	-0.027	0.029	0.357	E	.01	.03	-0.01	0
26	26	-0.071	0.029	0.013	E	.01	.03	0.00	0

Program execution is completed.

Your output is stored on the file: o:\DIF, sibtest, t11, usa, bk1, output.dat

Appendix J: Ordinal Logistic Regression Gender DIF Results, the Taiwan Sample

```
* DIF analysis with LR (for ordinal items)
     timss 2011 math items, 8th, TWN
     codes developed by Bruno Zumbo
*.
* SPSS SYNTAX written by:      .
* Bruno D. Zumbo, PhD      .
* Professor of Psychology and Mathematics,      .
* University of Northern British Columbia      .
* e-mail: zumbob@unbc.ca     .
*.
* Instructions .
* Copy this file and the file "ologit2.inc", and your SPSS data file into the sa
     me folder .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file
     .
* Run this entire syntax command file.
*.

************************************************************.
************************************************************.
***** gender DIF analysis for booklet 1.

get file='o:\bscTWNm5_MathItemOnly_Oddbklt,ItemRenamed,Bk01,NoMissing.sav'.

include file='ologit2.inc'.
   30    set printback off.
   31    ********************* OLOGIT 2.0 (test version) ****************************
             **
   32
   33    * By Prof. Dr. Steffen Kuehnel
   34       Institut fuer Politikwissenschaft
   35       Justus-Liebig-Universitat Giessen
   36       Karl-Glockner-Str 21, Haus E
   37       35394 Giessen
   38       Germany.
   39
   40    * contrast subcommand added by
   41       John Hendrickx <J.Hendrickx@maw.kun.nl>
   42       Department of Sociology
   43       University of Nijmegen
   44       P.O. Box 9104
   45       6500 HE Nijmegen
   46       The Netherlands.
   47
```

```
   48    * This macro is available from
   49        <http://www.socsci.kun.nl/maw/sociologie/resources/mlogist>.
   50
   51    * macro subroutine, called by class.
execute.

*****************************************.

compute item = m01.01.
compute total = bsmmat01.
compute grp = itsex.

*+++++++++++++++++++++++++++.

**** Regression model with the conditioning variable, total score, in alone.
ologit var = item total
  /output = all.

Matrix

Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE

                         (by Steffen M. KUEHNEL)

******************** Information Section **********************

Dependent variable is:
  item

Marginal distribution of dependent variable
        Value       Frequ.      Percent      %>Value
          .00        91.00        25.21        74.79
         1.00       270.00        74.79          .00

Effective sample size:
       361

Means and standard deviations of independent variables:
                    Mean      Std.Dev.
total      621.6212      100.9809

******************** Estimation Section **********************

Running Iteration No.:
```

1

Running Iteration No.:
2

Running Iteration No.:
3

Running Iteration No.:
4

..... Optimal solution found.

********** OUTPUT SECTION **********

LR-test that all predictor weights are zero

-2 Log-Likelihood of Model with Constants only:

407.646

-2 Log-Likelihood of full Model:
328.335

LR-statistic

Chisqu.	DF	Prob.	%-Reduct
79.311	1.000	.000	.195

Estimations, standard errors, and effects

	Coeff.=B	Std.Err.	B/Std.E.	Prob.	exp(B)
total	.011716	.001507	7.774703	.000000	1.011785
Const.1	-5.931211	.892004	-6.649312	.000000	.002655

Results assuming a latent continuous variable

R-Square (%):
29.85

Standardized regression weights of the latent variable:
total .5463

------ END MATRIX -----

execute.

**** Regression model adding uniform DIF to model.
ologit var = item total grp
  /output = all.

Matrix

Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE

(by Steffen M. KUEHNEL)

********** Information Section **********

Dependent variable is:
 item

Marginal distribution of dependent variable

Value	Frequ.	Percent	%>Value
.00	91.00	25.21	74.79
1.00	270.00	74.79	.00

Effective sample size:
 361

Means and standard deviations of independent variables:

	Mean	Std.Dev.
total	621.6212	100.9809
grp	1.5235	.5001

********** Estimation Section **********

Running Iteration No.:
 1

Running Iteration No.:
 2

Running Iteration No.:
 3

Running Iteration No.:
4

..... Optimal solution found.

********** OUTPUT SECTION **********

LR-test that all predictor weights are zero

-2 Log-Likelihood of Model with Constants only:
407.646

-2 Log-Likelihood of full Model:
327.781

LR-statistic

Chisqu.	DF	Prob.	%-Reduct
79.865	2.000	.000	.196

Estimations, standard errors, and effects

	Coeff.=B	Std.Err.	B/Std.E.	Prob.	exp(B)
total	.011793	.001515	7.783581	.000000	1.011862
grp	-.206537	.278199	-.742407	.457841	.813396
Const.1	-5.659103	.959318	-5.899090	.000000	.003486

Results assuming a latent continuous variable

R-Square (%):
30.20

Standardized regression weights of the latent variable:
total .5485
grp -.0476

------ END MATRIX -----

execute.

**** Regression model adding non-uniform DIF to the model.
ologit var = item total grp total*grp
/output = all.

Matrix

Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE

(by Steffen M. KUEHNEL)

Interaction term total*grp
int1.1 total grp

********** Information Section **********

Dependent variable is:
item

Marginal distribution of dependent variable

Value	Frequ.	Percent	%>Value
.00	91.00	25.21	74.79
1.00	270.00	74.79	.00

Effective sample size:
361

Means and standard deviations of independent variables:

	Mean	Std.Dev.
total	621.6212	100.9809
grp	1.5235	.5001
int1.1	948.1291	351.7022

********** Estimation Section **********

Running Iteration No.:
1

Running Iteration No.:
2

Running Iteration No.:
3

Running Iteration No.:

4

..... Optimal solution found.

********** OUTPUT SECTION **********

LR-test that all predictor weights are zero

-2 Log-Likelihood of Model with Constants only:
407.646

-2 Log-Likelihood of full Model:
327.773

LR-statistic

Chisqu.	DF	Prob.	%-Reduct
79.872	3.000	.000	.196

Estimations, standard errors, and effects

	Coeff.=B	Std.Err.	B/Std.E.	Prob.	exp(B)
total	.012193	.004796	2.542246	.011014	1.012267
grp	-.050586	1.791061	-.028243	.977468	.950672
int1.1	-.000267	.003030	-.088135	.929769	.999733
Const.1	-5.891768	2.812812	-2.094618	.036205	.002762

Results assuming a latent continuous variable

R-Square (%):
30.23

Standardized regression weights of the latent variable:
total .5670
grp -.0117
int1.1 -.0433

------ END MATRIX -----

execute.

*++++++++++++++++++++++++++.

Appendix K: Poly-SIBTest Gender DIF Results, the Taiwan Sample

```
name of input parameter file = sib.in

number of items on test =  26
name of file for Ref. grp. scores = o:\sibtest, t11, twn, bk1, 2=m, referencee.dat

name of file for Focal grp. scores = o:\sibtest, t11, twn, bk1, 1=f, focus.dat

minimum no. of examinees per matching score cell =   2
number of runs for this data set =  26
number of examinees in Reference Group =    189
number of examinees in Focal group     =    172

Examinee Test Score Summary Statistics

    Reference Group:              Mean =  22.47
                     Standard deviation =  7.41
    Focal Group:                  Mean =  22.20
                     Standard deviation =  8.18

    Standardized Score Difference =  0.04

Item Statistics
   # = item number
   m = mean score on item
   r = point biserial (item score-test score correlation)

   #:     1     2     3     4     5     6     7     8     9    10
   m: 0.748 0.604 1.748 1.596 0.629 0.607 1.360 1.443 0.884 0.845
   r: 0.528 0.330 0.553 0.807 0.658 0.675 0.802 0.783 0.536 0.615

   #:    11    12    13    14    15    16    17    18    19    20
   m: 0.845 0.906 0.881 0.873 0.831 0.737 0.657 0.914 1.163 0.731
   r: 0.589 0.557 0.405 0.572 0.680 0.511 0.693 0.558 0.749 0.694

   #:    21    22    23    24    25    26
   m: 0.529 0.837 0.665 0.709 0.255 0.346
   r: 0.578 0.642 0.594 0.585 0.271 0.305

   p-value notation:
R  denotes p-value for test of DIF/DBF against Ref. group
F  denotes p-value for test of DIF/DBF against Foc. group
E  denotes p-value for test of DIF/DBF against either the
                                        Ref. or Foc. group.

NOTES:
        MS/SSD = Matching Subtest Standardized Score Difference.
                 Standardized difference in mean observed scores
                 between Reference group and Focal group on the
                 matching subtest.

        p-elim = proportion of Reference (R) and Focal (F) groups
                 eliminated (not used) in SIBTEST calculations.

        Positive Beta estimate indicates DIF/DBF favoring Ref. grp.
        Negative Beta estimate indicates DIF/DBF favoring Foc. grp.

        FLAG = error flag indicator.  FLAG=0 indicates a normal
               successful completion of a SIBTEST run.  All other values
               of FLAG come with short error messages.


       SIBTEST-pooled weighting                                    F
```

Run no.	Suspect Subtest Item Numbers	Beta estimate	standard error	p-value		p-elim R	p-elim F	MS SSD	L A G
1	1	-0.023	0.040	0.565	E	.13	.08	0.04	0
2	2	0.024	0.056	0.670	E	.11	.10	0.03	0
3	3	0.030	0.063	0.630	E	.10	.14	0.03	0
4	4	0.072	0.053	0.177	E	.08	.08	0.02	0
5	5	-0.018	0.041	0.668	E	.12	.08	0.03	0
6	6	0.045	0.041	0.281	E	.14	.11	0.03	0
7	7	-0.099	0.058	0.089	E	.11	.06	0.05	0
8	8	-0.065	0.064	0.309	E	.13	.10	0.04	0
9	9	-0.017	0.030	0.586	E	.11	.13	0.04	0
10	10	0.010	0.031	0.746	E	.10	.12	0.03	0
11	11	0.015	0.033	0.642	E	.06	.07	0.03	0
12	12	0.001	0.026	0.984	E	.09	.12	0.04	0
13	13	-0.003	0.031	0.917	E	.12	.10	0.03	0
14	14	0.045	0.029	0.117	E	.16	.10	0.03	0
15	15	-0.018	0.025	0.454	E	.12	.10	0.04	0
16	16	-0.037	0.042	0.379	E	.06	.12	0.04	0
17	17	-0.008	0.038	0.839	E	.08	.12	0.04	0
18	18	-0.012	0.021	0.561	E	.17	.14	0.04	0
19	19	-0.095	0.074	0.204	E	.11	.12	0.05	0
20	20	-0.065	0.033	0.049	E	.20	.09	0.04	0
21	21	-0.039	0.046	0.395	E	.10	.08	0.04	0
22	22	0.042	0.031	0.177	E	.10	.05	0.03	0
23	23	0.033	0.042	0.433	E	.11	.07	0.02	0
24	24	0.018	0.041	0.669	E	.15	.11	0.03	0
25	25	0.007	0.049	0.883	E	.13	.12	0.03	0
26	26	-0.099	0.052	0.059	E	.11	.04	0.05	0

Program execution is completed.

Your output is stored on the file: o:\bk1,ouput.dat