GENDER RELATED DIFFERENTIAL ITEM FUNCTIONING IN MATHEMATICS TESTS:

A META-ANALYSIS

By

MO ZHANG

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

MASTER OF ARTS IN EDUCATION

WASHINGTON STATE UNIVERSITY
College of Education

AUGUST 2009

To the Faculty of Washington State University:

The members of the Committee appointed to examine the thesis of MO ZHANG find it satisfactory and recommended that it be accepted.

Brian F. French, Ph. D., Chair

Michael S. Trevisan, Ph. D.

Forrest W. Parkay, Ph. D.

ACKNOWLEDGEMENT

I would like the take this opportunity to thank my thesis committee: Dr. Brian French, Dr. Michael Trevisan, and Dr. Forrest Parkay. I am indebted to Dr. French who generously provided me his expertise, advice, ideas, and accurate judgments along the way. I deeply appreciate his time and great mentorship. I am grateful for Dr. Michael Trevisan too who has always been a great teacher ever since I came to WSU 18 months ago. I also owe my gratitude to Dr. Parkay, the professor of my very first lesson at WSU, who is always kind and patient with his students, and who gave me countless help along the way of my study. I feel very fortunate to know you all and be your student. Thank you very much.

I would also like to thank my Father Zhonggang Zhang and my Mother Qiulan Bian. This thesis would not have been possible without your support and encouragement. As your only child, I hope I could make you proud now and in the future.

GENDER RELATED DIFFERENTIAL ITEM FUNCTIONING IN MATHEMATICS TESTS: A META-ANALYSIS

Abstract

by Mo Zhang, M.A.
Washington State University
August 2009

Chair: Brian F. French

This study examines the gender differential item functioning in elementary through college level mathematics tests used in the United States and international countries. The purpose of this study is to explore (a) an average gender-DIF across mathematics assessment, (b) the patterns of gender DIF in mathematics assessment, and (c) the patterns of gender by item characteristic interactions in mathematics assessment. This study employed a meta-analysis approach to explore the research questions. A pseudo-effect size of gender-DIF was computed and interpreted by using the Educational Testing Service DIF classification scheme. The results of this study are consistent with previous research. However, new evidence also is generated in the area of gender DIF in mathematics tests. For instance, arithmetic, measurement, and computations as testing subjects could be plausible sources of gender DIF in mathematics assessments. One important implication, for instance, is that more studies dedicated to gender-DIF are needed to identify reasons behind the mathematics achievement gap between males and females. Recommendations for future studies also are discussed.

TABLE OF CONTENTS

Page

APPENDIX

LIST OF TABLES

LIST OF EQUATIONS

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Gender differences at the score and item level are one of the core issues in testing regarding fairness and test score validity. Gender differential item functioning (DIF) is a constant concern on achievement tests in mathematics (e.g., Bielinski & Davison, 2001; DeMars, 1998; Gardener & Engelhard, 1999). In the past 20 years, there have been hundreds of studies that inspect differences in male and female performance in math assessments with various results. For instance, males tend to perform better on difficult items compared to females (Engelhard, 1990); males tend to perform better on items related to problem solving (Mendes-Barnett & Ercikan, 2006), females tend to perform better on free response items compared to males (Burton, 1996), to name a few. Employing the keywords of gender and math in the Education Resources Information Center (ERIC) database for the years 1990-2009 resulted in 652 publications, which demonstrates that in the past 20 years there is much concern, interest, and resources devoted to this topic. However, there is no such study that provides (a) an average estimate of DIF across mathematics tests, (b) a comprehensive review of the patterns of DIF existing in mathematics tests, and (c) a broad and overall analysis of gender and item feature interactions in mathematics tests. The main purpose of this study is to address these three topics.

The following chapter presents a statement of research problems, research deficiencies in literature, research questions, and the audience of this study.

Statement of Research Problems

The purpose of this study is to gain a thorough understanding of gender differential item functioning on mathematics tests and items: an average of gender DIF magnitude across mathematics tests, patterns of gender DIF, and patterns of item features by gender interactions in mathematics tests. The perspective on gender differences in mathematics tests would be enhanced with conducting a meta-analysis of the overall averaged gender DIF across mathematical items and statistical analysis of item characteristics by gender interactions, with the intention of providing insights into the educational field in mathematics, as well as the test development field in terms of why DIF may exist in certain mathematics items across males and females.

Research Deficiencies in the Literature

Gender differences in mathematics achievement test scores have been described, debated, and documented in the field of education for almost 5 decades (e.g., Bielinski & Davison, 1998; Fennema & Sherman, 1977; Flanagan et al, 1963; Harris & Carlton, 1993; Hilton & Berglun, 1974; Jarvis, 1964; Johnson, 1984; Kaplan & Flake, 1982; Leinhardt, Seewald, & Engel, 1979; Mendes-Barnett & Ercikan, 2006; Pederson, Shinedling, & Johnson, 1968; Randhawa & Hunt, 1987; Rock & Pollack, 1991; Ryan & Chiu, 2001; Zohar & Gershikov, 2008). Researchers have made extensive efforts to explore and identify the sources of different gender performance in mathematics assessments. The literature on math education is growing at an astounding rate. Though, one significant component in these studies is missing; that is, an overall magnitude of gender DIF across mathematics tests. This magnitude estimate is best achieved by a meta-analysis given the salient merits of research synthesis methodology, which will be addressed in Chapter Two. Furthermore, since there is no study that examines the average gender DIF across mathematics tests, a meta-analysis also would generate new evidence and an overall trend in this area that a study of an individual assessment is unable to offer. In general, this meta-analysis study would provide evidence of (a) the extent of the existence of gender DIF in mathematics test, (b) the magnitude of average gender DIF across studies, and (c) the patterns of item characteristics (i.e., content and format) by gender interactions for reported DIF items identified in mathematics tests.

This study also was intended to cover the studies both in the United States and international countries. There have been a certain amount of international studies of gender differences in mathematics achievement (Penner, 2003), with most of the findings showing males having the advantage over females across countries (Feingold, 1994; Innabi & Dodeen, 2006). The value of mathematics are recognized worldwide, so that including the gender DIF in mathematics assessment results from international countries into this study, and comparing the patterns of gender DIF in mathematics assessment of the US and international countries would be a valuable addition by broadening the scope of this study.

Audience for This Study

Gender DIF results can be used to establish guidelines for mathematics test developers to use in assembling tests. Those characteristics of items that have been frequently associated with differential performance by males and females should be considered carefully regarding inclusion vs. exclusion of certain items, and balancing the number or the amount of those items in a test. Use of such guidelines in conjunction with DIF information can be beneficial to prevent the use of construct-irrelevant characteristics in the construction of test questions (Hambleton & Jones, 1994).

As a result, the outcomes and findings of this study may have direct implications for assessment, curriculum, and instructional changes. It also may have a critical impact on mathematical test developing field, and to those who make important decisions based on test scores. With a good understanding of gender DIF in mathematics testing, teachers also may be better equipped to write test items, and further provide instruction that targets weakness of both males and females.

CHAPTER TWO

REVIEW OF THE LITERATURE

The following chapter will review the literature regarding high stake testing, testing fairness and validation, as well as the importance and usefulness of differential item functioning (DIF) in testing fields as it relates to mathematics assessments. The statistical analysis approaches of identifying DIF, computation methods and formulas of DIF effect size, consistencies among different DIF approaches, and judgmental decisions after detection of DIF also are reviewed and discussed. Additionally, a comprehensive review of gender differences in mathematics test is presented. Terms that are defined in this chapter include: differential item functioning (*p.5*), high stakes tests (*p.7*), item bias (*p.7*), fairness (*p.8*), validity (*p.8*), and item analysis (*p.9*).

Differential Item Functioning

Differential item functioning (DIF) is an analysis of performance across groups on specific test items. DIF occurs when examinees from different groups show different probabilities of success (i.e., a correct response) on a dichotomously scored item or a difference in probabilities of selecting a certain level of response on a polytomous item after matching on the underlying ability that the item is intended to measure (Zumbo, 1999).

DIF is of great interest to researchers and educators given that DIF poses a potential threat to test fairness (Zheng, Gierl, & Cui, n.d.). As an item analysis methodology different from comparing mean scores at test level, DIF plays an important role in detecting the items that function differently in a test, also the potential unfairness of a test. For instance, in Garner and Engelhard’s (1999) study, statistically significant differences in mean scores on the constructed response items in the math test were in favor of males, whereas the only significant DIF on the constructed response items was found in favor of females. This demonstrates the importance of examining DIF rather than simply computing mean score differences as a critical and indispensible step in evaluating fairness, validity, as well as equity for a test.

In DIF analyses, two groups (i.e., focal group and reference group) are compared on item performance after adjusting for overall performance on the measured traits (Holland & Wainer, 1993). The focal group is the focus of the analysis, and the reference group serves as a basis for comparison for the focal group. When the groups are matched statistically with respect to ability, the researcher may then determine whether items function differently for the two groups due to some property not related to ability or due to true ability differences between the two groups of examinees (Dorans & Holland, 1993).

DIF emerged during the civil rights era of the 1960s. It was an era marked by numerous concerns of equal opportunity, and was an important era for DIF development. DIF was one of the responses to that time period of history, was influenced by it, and took its role as a standard part of testing (Cole, 1993). Therefore it was during the mid 1960s that the public and measurement professional began to pay increasing attention to score differences between groups and issues of fairness of testing (Cole & Zieky, 2001). Measurement specialists since that time also became concerned with the fairness of their instruments and possibility that some of them might be biased (Cole & Moss, 1989).

High Stakes Tests, Testing Fairness, and Item Bias

Testing and resulting scores are routinely used to make important decisions that have important personal, social, and political ramifications, most commonly admissions to colleges and graduate schools, obtaining employment, advancement, licensure, psychological diagnosis, among others. When the result of a test plays such a highly critical role in situations mentioned above, they are called high-stakes tests. The more of high stakes tests exist in the society, the more they shape the teachers, the parents, the students, as well as the education as a whole, and eventually, our human society.

With the growing importance of testing, one of the major topics in the field of measurement that is increasingly being probed and researched is test fairness and bias. Bias is defined as differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers (Cole & Moss, 1989). Item bias occurs when examinees of one group are less likely to answer an item correctly compared to examinees of another group because of some characteristics of the test item or testing situation that is not relevant to the test purpose (Clauser & Mazor, 1998; Zumbo, 1999). There are generally two types of measurement bias: external relationship bias and internal relationship bias, both of which should be examined and serve as a function of providing evidence when validating tests. External evidence of measurement bias is suggested when the relationship between the scores and criterion is different for various groups (Zumbo, 1999). Researchers studying external test relationships think about interpretations between test scores and other variables external to the test (Cole & Moss, 1989), such as SAT and GPA in a sense that SAT predicts college GPA. DIF examines the internal relationship bias by providing evidence of relationships between individual items and the test as a whole. Internal bias can be detected through DIF analysis when a test item or subset of items has a different interrelationship to the total test for two groups of interest (Cole & Moss, 1989). However, it is important to know that DIF identified items are not necessarily treated as biased items, which will be discussed and elaborated on in section "Judgments After Detecting DIF" later in this chapter.

Testing Validation

There is a link between fairness and validity (Cole & Zieky, 2001). Fairness is an aspect of validity. Bias has been characterized as “a source of invalidity that keeps some examinees with the traits or knowledge being measured from demonstrating the ability” (Shepard, Camilli, & Williams, 1985). In short, item bias is a potential threat to validity. Regarding validity specifically, the concept, method, and process of validation are central to evaluating measures, for without validation, any inferences made from a measure are meaningless (Zumbo, 1999). The current view of validity suggests that validity is not a property of the measurement tool but rather of the inferences made from results scores (Cizek, Rosenberg, & Koons, 2008; Zumbo, 1999). If score-based inferences are not equally valid for all relevant sub-populations, decisions derived from score inferences will not be fair (Langenfeld, 1997). The focus of the validity of a test, likewise, is no longer on the test per se, but on the application being made and on the associated interpretation of the result (Cronbach, 1988; Messick, 1988; Zumbo, 1999). The consequences of testing are very important (Cizek et al., 2008), as Messick (1988) stated, it is not the obvious misuse of measures that is the issue but rather that we need to think about the unanticipated (negative and positive) consequences of the legitimate use and/or interpretation of decision making from measures that can be traced back to test invalidity.

The Importance of Detecting DIF

In the field of education, the absence of DIF is regarded as an important aspect of test fairness by most educational researchers (Rudas & Zwick, 1997). In that the presence of an item that functions differently on a test indicates that the item is measuring some nuisance dimension or dimensions unrelated to the remainder of the test (Ackerman, 1992), detecting such items that are identified as having DIF is critical to maintain the tests’ fairness and validity. Yet, discovering the reasons for DIF is not an easy task. As the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) stated, although DIF procedures may hold some promise for improving test quality, there has been little progress in identifying the causes or substantive themes that characterize item exhibiting DIF. In other words, once an item on a test has been statistically identified as functioning differently from one examined group to another, it has been difficult to specify the reasons for the differential performance or to identify a common deficiency among the identified items. For DIF is a consequence of a complex interaction between the item and examinees, it is unlikely that any single identifiable cause of DIF can be detected (Scheuneman, 1987). To conclude, it is hard to offer explanations of why DIF occurs once it is identified.

Item Analysis and Techniques to Detect DIF

Item analysis is a set of statistical techniques to examine the performance of individual items. This is important when developing a test or when adopting an established and known measure (Zumbo, 1999).

In terms of DIF techniques, it is significant that all DIF detection methods available are designed to match the groups, either directly or indirectly, on the proficiency measured by the items (Angoff, 1993), and all DIF measurements investigate how different groups perform on individual test items to determine whether the test items are creating problems for a particular group (Zumbo, 1999). They are all based on such principles that if different groups of test-takers (e.g., males vs. females, Caucasian vs. African American, English as first language learner vs. English as second language learner) have approximately the same level of ability, they should perform similarly on individual test items regardless of group membership (Zumbo, 1999).

An review on the various methods detecting DIF was conducted by Mapuranga, Dorans, and Middleton (2008). Statistical techniques that are most commonly used by researchers to quantify the magnitude of DIF include Item Response Theory (IRT) methods (Thissen, Steinberg, & Wainer, 1993); Mantel-Haenszel (M-H) statistics (Holland & Thayer, 1988); Simultaneous Item Bias Test (SIBTEST) procedure (Shealy & Stout, 1993); Standardization method (STD) (Dorans & Kullick, 1986; Dorans & Schmitt, 1993); and the Logistic Regression (LR) procedure (Swaminathan & Rogers, 1990).

For IRT methods, there is no single IRT model that can be employed, yet there are numerous approaches that are based on a range of IRT models (e.g., Rasch index, DFIT, TestGraf), and they share the use of a matching variable that is an estimate of latent ability rather than the observed score (Clauser & Mazor, 1998). Latent ability, also called latent variable in this case, is unobservable, and is the variable (e.g., intelligence, verbal ability, spatial ability, etc.) researchers are trying to get at with their indicators (e.g., items) (Zumbo, 1993). Mantel-Haenszel statistics are the most widely used approaches for identifying DIF (Holland & Thayer, 1988). They differ from IRT approaches in that examinees are matched on an observed variable, such as a total score, then the counts of examinees in the focal and reference group getting the studies items correct and incorrect are compared. M-H procedures estimate the constant odds ratio that yields a measure of effect size for evaluating the magnitude of DIF. SIBTEST procedure is an alternative statistical method for detecting DIF. The matching criteria for SIBTEST is a latent score as IRT methods do. SIBTEST tests the significance based on the ratio of the weighted difference in proportion correct for focal and reference group to its standard error (Clauser & Mazor, 1998). The Standardization method computes the standardized difference in the proportion correct (Clauser & Mazor, 1998), and it is quite similar to M-H in the analytic procedures, as well as other respects (Angoff, 1993). Logistic regression (LR) procedure also is a model-based approach. It may be conceptualized as a link between the contingency table methods (M-H, STD, and SIBTEST) and IRT methods (Clauser & Mazor, 1998). The equations of effect size computation for each method are presented in the next section. The reasons of providing formulas of computing effect size rather than the test in general is that combining effect sizes of different studies is the focus of this study.

Each of these DIF identifying techniques has their own advantages and disadvantages: IRT methods, (e. g., Rasch method) have some advantages over other methods in that it provides more information regarding psychometric properties of individual assessment items; M-H procedures are efficient in statistical power, yet may not be able to detect non-uniform DIF; STD method is quite simple to perform, but lacks of an associated test of significance(Clauser & Mazor, 1998), to name a few.

Effect Size and Consistencies among DIF Methods

Effect size value, as it is different from the *p* value in null hypothesis significance testing, can be used to inform judgments regarding the "practical significance" of the study result (Kirk, 1996). The 5th edition of Publication Manual of the American Psychological Association views the failure to report effect size value as a defect in the design and the reporting of research results (APA, 2001).

One of the formulas of computing effect size of DIF from IRT methods is adopted from Raju (1988) as below. *Area* refers to the difference between the item characteristics curves, and *a*, *b*, and *c* are the parameters of the items, and D is a scaling constant (Raju, 1988). Item parameter *a*, *b*, and *c* are defined in the next paragraph. The difference is substantial if the *Area* is larger than .107 (Raju, 1988).

$$Area=(1-c)\left|\frac{2(a_2-a_1)}{Da_1a_2}ln(1+exp(\frac{Da_1a_2(b_2-b_1}{a_2-a_1}))-(b_2-b_1)\right| \quad (1)$$

The formula above is for item response theory three parameter logistic (3PL) model across groups for computing the difference between groups on item performance for a single item. The "3PL" refers to 3 item parameters in item characteristic curves (ICCs) (Figure 1). Figure 1 was modified based on Zumbo (1999) in order to better illustrate the issue. In the Figure, parameter *a* is the slope proportion to the curve; it represents the item discrimination. For example, a flat ICC does not differentiate among test takers (Zumbo, 1999). Parameter *b*, positioned along X-axis, represents item difficulty, which also indicates the amount of a latent variable needed to respond correctly to an item. Parameter *c*, positioned along Y-axis, represents the effect of guessing or pseudo-guessing on the probability of a correct response in multiple choice items.

Figure 1

Three Parameter Logistic Item Response Theory Model Display Item Characteristic Curve

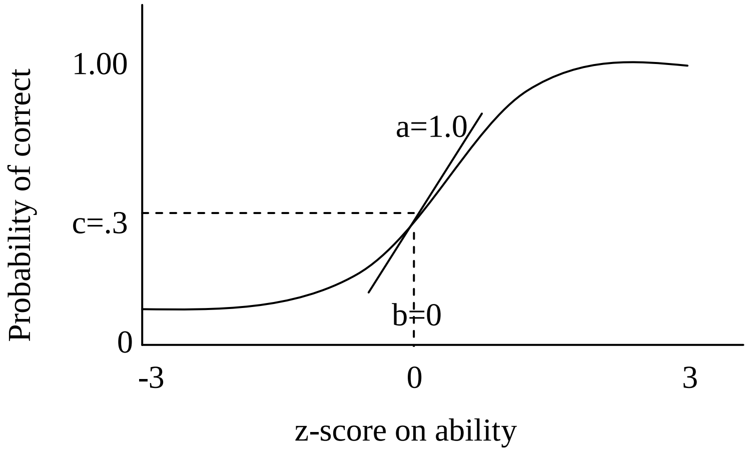

Figure 1. X-axis: the amount of latent variable needed to respond correctly to an item. Y-axis: the probability of getting a correct response in multiple choice items. The slope of the curve represents item discrimination.

The ICCs of items that do not display DIF and items that display DIF are briefly presented in Figure 2 and Figure 3 (Zumbo, 1999). In Figure 2, the area between the two ICCs are very small, and the parameters of each curve are nearly equivalent. In Figure 3, the area between the two ICCs are quite large, and group 2 is favored in this case.

Figure 2

An Example of An Item That Does Not Display DIF

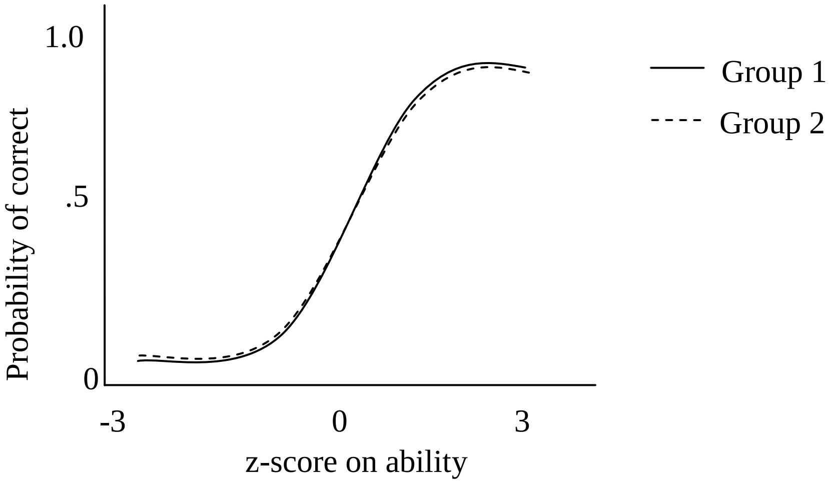

Figure 2. X-axis: the amount of latent variable needed to respond correctly to an item. Y-axis: the probability of getting a correct response in multiple choice items.

Figure 3

An Example of An Item That Displays Uniform DIF

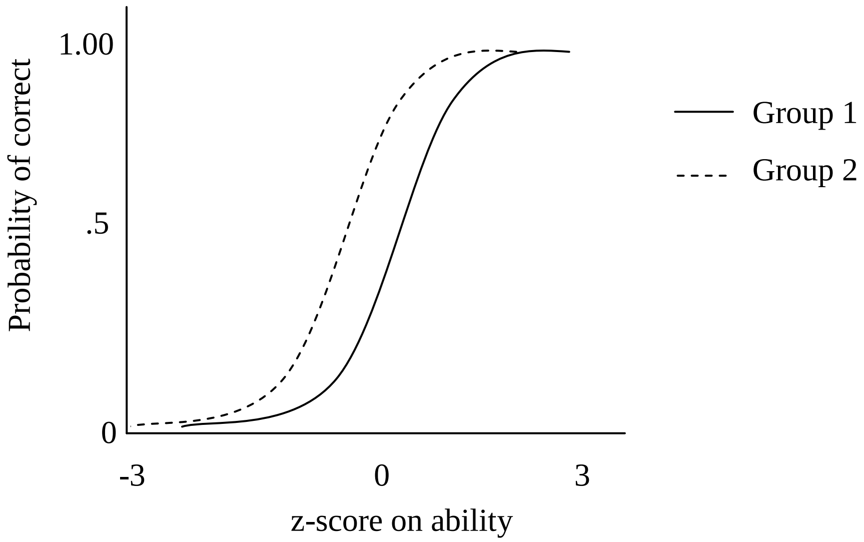

Figure 3. This is called Uniform DIF because the two curves are not crossed. X-axis: the amount of latent variable needed to get the item right. Y-axis: the probability of getting a correct response in multiple choice items.

Also, Maller (2001) proposed a root mean squared probability difference (RMSD) as the magnitude of the DIF effect. The formula is presented below.

$$\text{RMSD}=\sqrt{\frac{\sum_{j=1}^{n}\left[P_2(\theta_j)-P_1(\theta_j)\right]^2}{n}} \tag{2}$$

In the formula above, *Θ* is the estimated ability; *j* is the estimated theta value for a given examinee; *n* is the total number of estimated theta values or examinees; P refers to the probability of getting a correct answerer in multiple choices situation.

The formula of computing effect size of DIF from Mantel-Haenszel statistics is as below (Zheng, Gierl, & Cui, n.d.).

$$\Delta_{MH}= -(2.35)ln(\alpha_{MH}) = -(2.35)\, ln\left[\left(\Sigma_m R_{rm} W_{fm}/N_{tm}\right)/\left(\Sigma_m R_{fm} W_{rm}/N_{tm}\right)\right] \tag{3}$$

In this formula, α_{MH} refers to the ratio of odds that a reference group examinee will get the item correct compared to the odds for a matched focal group examinee. Also in this formula, R_{rm} and R_{fm} refer to the odds of getting right for the reference group and focal group in multiple choices items , R_{fm} and R_{rm} refer to the odds of getting wrong for the reference group and focal group in multiple choice items, and N_{tm} refers to the total number of people at ability level *m*.

The formula of computing effect size of DIF from logistic regression (LR) procedure is expressed as $R^2\Delta$ and is presented below (Jodoin & Gierl, 2001).

$$R^2\Delta=R_2^2-R_1^2 \tag{4}$$

In the formula above, $R^2\Delta$ represents the effect size of DIF. It is calculated as the difference between R_2^2 and R_1^2. R_2^2 refers to the sums of the products of the standardized regression coefficient for each explanatory variable, and R_1^2 *refers to* the correlation between the response and each explanatory variable of the two statistical models of LR analysis: $f(\varphi ,g)=\tau_0+\tau_1\varphi$ and $f(\varphi ,g)=\tau_0+\tau_1\varphi+ \tau_2 g$, where τ refers to the intercept and weight for the ability, $f(\varphi ,g)$ refers to the linear combination of the predictor variables, in which *g* represent the group membership and *φ* represent the observed ability. Notice that strictly speaking R^2in this case should be concerted to $\mathrm{R}^2/(1-\mathrm{R}^2)$ for the interpretation of effect size, yet it won't have fundamental difference between the two in terms of the result (Jodoin & Gierl, 2001).

The formula of computing effect size of DIF from SIBTEST is as below (Zheng, Gierl, & Cui, n.d.).

$$\hat{\beta}=\frac{\sum_k \hat{P}_k(\bar{Y}^*{}_{P_k}-\bar{Y}^*{}_{F_k})}{\left[\sum_k \widehat{P_k}^2\left[\frac{1}{J_{R_k}}\hat{\sigma}^2(Y|k,R)+\frac{1}{J_{F_k}}\hat{\sigma}^2(Y|k,F)\right]\right]^{1/2}} \tag{5}$$

Here $\hat{P}_k$ refers to the proportion of examinees in the focal group, $\bar{Y}^*{}_{P_k}$ and $\bar{Y}^*{}_{F_k}$ refer to the adjusted means of examinees in subgroup k, $\hat{\sigma}^2(Y|k,R)$ and $\hat{\sigma}^2(Y|k,F)$ refer to the sample variance for examinees in item from group g for reference group and focal group with a total score k on the valid subtest, and $J_{R_k}$and J_{F_k}refer to the sample size of reference and focal group with a total score of k on the valid subtest (Zheng, Gierl, & Cui, n.d.).

Finally, the formula of computing the effect size of DIF from Standardization method is as below (Dorans & Holland, 1993), where $P_f{}^*$ refers to the performance of the focal group predicted from the reference group's item test regression curve, and P_f refers to the proportion correct observed in the focal group.

$$|SMD|=\text{-}2.35ln\{[P_f{}^*/(1-P_f{}^*)]/[P_f/(1-P_f)]\} \tag{6}$$

There is some evidence showing the consistencies among different DIF measurements in regard to the magnitude/effect size of DIF. Researchers have shown that the correlation between M-H index and STD index, when expressed on the same scale, is .99 (Wright, 1987). Zheng, Gierl and Cui (n.d.) in their study of investigating the consistencies of DIF detection and effect size measurements among M-H, SIBTEST, and LR procedures using gender DIF data from 2000 examinees across grade 3, 6, and 9 found out that in terms of DIF magnitude, the matching percentage between M-H and LR, M-H and SIBTEST, and SIBTEST and LR are 83.30%, 75.86%, and 90.00%, respectively, and in terms of effect size from each three measurements, the correlation between each other range from .79 to .92. There also is evidence showing that the DIF indexes generated by M-H procedures and those generated by Rasch measurement, an IRT method, are equivalent (Engelhard et al., 1990; Schulz et al., 1996). However, there is no common method for placing effect sizes from different DIF approaches on a same scale; consequently only combined effect sizes under every single approach were calculated in this study. All in all, high positive correlations found among different DIF measures indicate different procedures provide consistent estimates on the magnitude and direction on DIF, even though the magnitude, at this time, cannot be expressed on the same scale or combined to do so. This evidence is very important to conducting this meta-analysis study, in that different studies use different approaches to identify DIF; in order to compare an average DIF across these studies, equivalent quality of methods and results is very essential.

Judgments after Detecting DIF

It is important to realize that DIF is a necessary, but not a sufficient condition, for item bias (Clauser & Mazor, 1998; Zumbo, 1999). DIF, a statistical finding, may not necessarily warrant removal of items that are identified as DIF (Angoff, 1993); rather, one would have to apply follow-up item bias analyses (e.g., content analysis, empirical evaluation) to determine the presence of item bias (Zumbo, 1999).

Since DIF is not a synonym for bias (Zieky, 1993), appropriate judgmental procedures may be indispensible to determine whether or not the difference in difficulty detected by DIF is unfairly related to a specific group (Zieky, 1993). However, the analysis of DIF provides a convenient starting point for the study of item bias (Wang & Lane, 1996). One of the reasons is that in some cases, examinees from different groups may be, in fact, expected to perform differently due to ability differences, which is called item impact (Clauser & Mazor, 1998). It also indicates that DIF results must be interpreted within individual context. Decisions of the items that are identified as DIF must be made regarding the final test content, or scoring, and interpretations of the results must be placed with the context of the overall test development process (Clauser & Mazor, 1998), especially DIF statistics are used to make decisions in the controversial and emotionally charged contexts of item and test bias (Zieky, 1993).

It is a matter of policy as to what should be done when an item is identified as displaying DIF. The issue is whether identified items are considered biased until proven valid or valid until proven biased (Clauser & Mazor, 1998). As Clauser and Mazor argued, the formal approach requires rules to be established restricting use of DIF items until they are revised or additional evidence can be collected to support the validity of the items; by contrast, items identified as displaying DIF could be targeted for review by content experts. Items could be deleted or maintained based on their judgments. Each institute/testing company has their own judgmental process. For instance, decisions made at Educational Testing Service (ETS) is that the appropriate use of DIF requires procedures that incorporate the judgments of trained test developer and subject-matter specialists (Zieky, 1993).

To further illustrate the issue of judgmental decisions after detecting of DIF, standard DIF detection procedures focus on only one categorical variable at an aggregated group level, such as gender or ethnicity/race (Clauser & Mazor, 1998). Yet, deleting items due to DIF can have unintended consequences on the focal group in this traditional one-way DIF analysis (Dorans & Holland, 1993). Zhang, Dorans and Matthews-Lopez (2005) proposed a two-way DIF classification scheme in which each item was examined for DIF effect at the subgroup level, i.e., gender DIF within ethnicity/race and ethnicity/race within gender. They found that this two-way analysis is a more informative approach to DIF analysis by not only confirming finding from one-way DIF approach but also enhancing our understanding of the behavior of DIF items, which can offer valuable assistance to the decision making process (Zhang et al., 2005).

Gender Differences in Mathematics Tests

Within the fields of education and psychology, gender differences in mathematics performance have been studied intensively. There has been some consensus on patterns of the differences (Hyde et al., 1990). An extensive body of research has found males perform better in mathematics achievement compared to females (Mendes-Barnett & Ercikan, 2006). The latest meta-analysis of gender differences in mathematics performance at the score level was conducted by Hyde et al. That large-scale review which included 259 independent effect size values of male and female performance difference in mathematics concludes with the statement that .20 as the mean magnitude of the gender difference in mathematics performance indicated better performance of males on average (Hyde et al.). Differences in mathematics performance between males and females were found as early as the fifth grade (student aged about 9), and increased in the higher grades (Seegers & Boekaerts, 1996). Despite changes in the curriculum over the past 30 year, gender differences have remained on college admission' tests of mathematical aptitude (Langenfeld, 1997) at the mean level. If you examine the mean scores of males and females in mathematics SAT in the past 36 years reported by College Board (2008), males’ scores are consistently higher than females' by more than 30 points.

Yet, there is no easy explanation for this gender gap in mathematics achievement. The causes of gender differences in math achievement are essentially unknown. Attempts at understanding the underlying causes of DIF using substantive analyses of statistically identifies DIF items also have, with few exceptions, met with overwhelming failure (Roussos & Stout, 1996a).

Even if it is mainly the case that the causes of gender differences in mathematics achievement are essentially unknown, a number of explanations have been given to account for these differences. Several plausible explanations in literature are described below. Some researchers contributed gender differences in mathematics to cognitive aspects and differences between males and females. Dai (2006) suggested that cognition is inseparably bound in any meaningful measure of aptitude in mathematics. One of them is self-perception. Male and female students at as early as first and second have already had different perceptions of self-competence in mathematics: male second graders rated higher evaluation on math skills compared to female second graders(French & Mantzicopoulos, 2007). That males and females differed in performance in Geometry could be partly due to males and females spatial visualization ability, which is an important factor in Geometry achievement (Battista, 1990). Males are more variable in their visuospatial performance compared to females (Delgado & Prito, 2003). Studies also have reported that females suffer higher levels of anxiety compared to males while taking the SAT (Katzman, Loewen, & Rosser, 1988). Sex-role socialization of females may prevent a high level of quantitative and analytic accomplishment which makes them avoid quantitative activities (Fennema & Sherman, 1977). Males and females have different self-academic concept, since males generally receive less encouragement at home and at school to pursue math, because technological fields are perceived as being primarily for males (Gross, 1988; Seegers & Boekaerts, 1996). The patterns of course taking for males and females are not the same (Davenport, et al., 1998; Feingold, 1992): females take fewer quantitative courses in high school (Alexander & Pallas, 1982; Ayalon, 2003), and take fewer years of courses in mathematics, physics, and computer science in high school (Navarro, 1989).

Researchers also proposed that gender-related differences in performance are results of different approaches to learning mathematics (Bohlin, 1994; Garner & Engelhard, 1999; Linn & Hyde, 1989). For example, Gallagher (1992) reported that items favoring male requiring insightful strategies, whereas all the items that favor females required standard algorithmic strategies. It also indicates that females may rely on routines learned in class more than males, while males use more alternative approaches to solve problems compared to females. That also explains that although males appear to have an advantage on mathematics achievement tests over females, females usually have higher average classroom grades compared to males (Garner & Engelhard, 1999; Linn & Kessel, 1995). Among evidence of gender differences in mathematics tests also includes cultural factors (Ibarra, 2001), self-regulation (Hong, O'Neil, & Feldon, 2005) and item bias (Harris & Carton, 1993), to name a few.

Gender Related DIF in Mathematics Tests

A number of studies have revealed differential item functioning (DIF) between male and female groups on particular mathematics items (e.g., Li, Cohen & Ibarra, 2004; Ryan & Chiu, 2001; Scheuneman & Gerritz, 1990; Skaggs & Lissitz, 1992; Wang & Lane, 1996). Previous work investigating gender-related DIF on mathematics tests has identified item difficulty that may be related to differential performance by males and females on kindergarten through 12th grade (K-12) level of education. For example, a consistent pattern of item difficulty differences between males and females has been reported, such that the item difficulty difference may serve as a function of gender differences in mathematics achievement (Bielinski & Davison, 2001; Engelhard, 1990; Ryan & Fan, 1996). Specifically, females outperformed males on the easy items and males tend to outperform females on the difficult items by the evidence that item performance gap increasingly favored males as the difficulty of the item increased (Engelhard, 1990). Likewise, females also performed better compared to males on the data sufficiency items-the easiest items on the test which only need the memorization of information (Becker, 1990). For instance, Harris and Carlton (1993) classified mathematical test items from the SAT into six levels of cognitive complexity: 0 was assigned to items measuring recall of factual knowledge and 5 was assigned to items requiring higher mental activities. After controlling for overall test scores, females outperformed males on the three lowest levels, whereas males outperformed females on the two highest levels.

In examining gender-related DIF on mathematics assessments, attempts to identify item features such as item format, mathematical content, item context, and their interaction with differential performance by males and females (Harris & Carlton, 1993; Mendes-Barnett & Ercikan, 2006; O’Neil & Mcpeek, 1993; Ryan & Fan, 1994) also have been made to identify sources of DIF. For example, compared to their matched female students, high school male students tend to perform relatively better on items involving ratios, proportions, and percents (Garner & Engelhard, 1999; Jackson & Braswell, 1992), on items embedded in a real-world context (Harris & Carton, 1993; O’Neil & Mcpeek, 1993), on items that involved the solution strategies that are generally not taught in schools (Gallagher & Lisi, 1992; Harris & Carton, 1993; O’Neil & Mcpeek, 1993), on items that required higher cognitive level like problem solving (Frost, Hyde, & Fennema, 1994; Mendes-Barnett & Ercikan, 2006). Females performed better on algebra and on more abstract mathematics items than males (Doolittle & Cleary, 1987; Garner & Engelhard, 1999; O'Neil & McPeek, 1993), and worse on geometry, measurement, and data analyses. Multiple choice items tend to favor males and the constructed items tend to favor females (Burton, 1996; Garner & Engelhard, 1999).

However, in terms of attributing gender differences to item characteristics, existing studies also have produced inconsistent results over the past three decades (Becker, 1990; DeMars, 1998; Engelhard, 1990; Frost, Hyde, & Fennema, 1994; Harris & Carlton, 1993; Lane, Wang, & Magone, 1996; Ryan & Fan, 1996). For example, compared to male students, high school female students, tend to perform poorer on geometry in terms of mean scores (Doolittle & Cleary, 1987). Likewise, Harris and Carlton (1993) reported males performed better on geometry compared to a matched group of females on the SAT. However, a study conducted in 2006 did not find geometry as a source of gender DIF (Mendes-Barnett & Ercikan, 2006).

Previous researchers have made great effort in both detecting gender DIF items in mathematical assessments regarding item difficulty and find the reasons for performance disparities between males and females in mathematics by probing the interactions of item features and gender. They have achieved some degree of agreement in terms of item difficulty by gender interaction, what subject matter favors males and females, which type of items favors males and females, yet inconsistent conclusions still exist in terms of items characteristics by gender interaction. The focus of this study is on psychometric properties of mathematics test items across gender. First of all, the magnitude of the average gender DIF that is usually detected in mathematics tests is unknown. To gain a sense of DIF magnitude across mathematics tests would be critical due to the fact that the existence and effect size of DIF poses potential threat to testing fairness, though not a sufficient condition to announce bias. This study would provide insights to this question. Moreover, this study is intended to provide insights into the pattern of gender DIF in mathematics tests, as well as item features (content and format) by gender interactions in mathematics tests through a comprehensive synthesis of previous research.

Research Questions

Due to the deficiencies in literature in terms of gender DIF in mathematics assessments, three research questions were posed for this study:

1. Among those studies that reported DIF, what is the average DIF magnitude across items across studies in mathematics tests?

2. What are the patterns of gender DIF existing in mathematics tests?

3. What are the patterns of interactions between item features (content and format) and gender in mathematics tests?

CHAPTER THREE

METHODS

There is general agreement on the value of meta-analysis as a way of synthesizing study results in quantitative educational research (Harwell & Maeda, 2008). It involves the attempt to discover the consistencies and account for the variability in similar-appearing studies (Cooper & Hedges, 1994a). One of the unique contributions of "research synthesis" or "research integration" is that questions could be asked about the trends in research over time, trends in both results and how research is conducted (Cooper & Hedges, 1994b). Research synthesis aids in accumulating evidence and in generating new evidence (Hall et al., 1994) as well. These advantages with meta-analysis meet the unique and special needs of this study that other methodologies are not able to offer for answering the research questions posed in this study: serving as a complement to the primary studies, generating new evidence in the average magnitude of and generality of effects of gender DIF in mathematics testing, and identifying the trends and patterns of research in the area of gender DIF in mathematics tests.

This study takes advantage of the strengths of research-synthesis, embraces meta-analysis as a form of "exploratory research" (Dooley, 1990) to probe the theories, evidences, trends of the area of interests, which will not only benefit the mathematics educational field, but inform educational research in the particular area. Research synthesis also can be used to detect the association between independent variables (Hall et al., 1994), which is not a focus of this study, yet could be a follow-up step by examining the combined effects of various item features and difficulty on gender performance differences.

The following chapter presents the literature searching strategies that were used in this study, the study selection and inclusion criteria in this meta-analysis, coding methods, as well as delimitations and limitations within this study.

Literature Search Strategies

To locate the studies to be included in this investigation, a database search strategy was employed. First, a list of keyword terms or combinations of terms including “differential item functioning or DIF or item bias or invariance”, “mathematics or math or math aptitude”, and “gender or sex” were used to identify the research base of gender differences in mathematics tests. Literature on achievement testing also was searched, in order to identify studies that conducted gender DIF analyses on mathematics even though the primary focus of the study is not mathematics education. It is appropriate to aggregate these studies because studies may differ in form, but as long as they measure the same ability (i.e., math) they could be included. Furthermore, when a relationship remains constant though tested under variety of circumstances, it is clearly robust (Hall et al., 1994).

Both online and printed resources were used to determine which articles to be included in this study. Online electronic databases included Education Resources Information Center (ERIC), PsychINFO, and Educational Full Text (EFT). In addition, I manually searched eight high-quality peer-reviewed journals that are most prominent in the educational and testing field in order to include references that might have been overlooked during online searching. Target studies for this study are most likely published in these journals including: *Journal of Educational Measurement, Applied Psychological Measurement, Journal of Education and Behavioral Measurement, International Journal of Testing, Applied Measurement in Education,* *Educational Measurement: Issues and Practice*, *American Educational Research Journal*, and *Educational and Psychological Measurement*.

Inclusion Criteria

In this study, in order to make justifiable conclusions and references, strict and adequate study-inclusion criteria were used. Abstracts were read, and those that did not promise to provide relevant data based on the inclusion criteria were excluded.

First, three major criteria were employed for selecting studies within the review: (a) studies had to be published in APA, NCME or AERA sponsored journals, presented at APA, AERA, or NCME conferences and the eight journals listed above. The latter were manually searched between the year 1990 and 2009; This time period is sufficient in which to capture reporting practices of gender DIF studies and current trend in gender DIF in mathematics examinations; (b) studies within each journal article or conference paper had to provide data of gender differences and DIF in mathematics items; and (c) test items' characteristics should be provided in the article.

Second, the studies were carefully evaluated to determine if they should be included in this study, particularly in regard to the methodology used, because frequently deficiencies of low-quality quantitative research are included in the methodology of meta-analysis (Harwell & Maeda, 2008). So to make sound conclusions, strictly checking the methods that each study used is quite necessary. Wandt et al. (1965) made several recommendations that focused on the methodology, which were followed in this study in order to be assured that the subject population, situation, and procedures in each study are appropriate for this study. The criteria include: (a) the population of the study should be adequately described; (b) if a sample was used, the sampling method should be specified; (c) the methods and instruments that are used to gather data are adequately described; in this study, it applies to mathematics assessment instruments; (d) the methods that are used to analyze data are described; in this study, it applies to DIF approaches; and (e) the result of the analyses are clearly presented; in this study, it applies to the category of DIF and preference of the item(s).

Coding of the Studies

Coding is a critical part of research synthesis (Orwin, 1994). To ensure the consistently and systematically extraction of information from each study, methods and index were developed and used. The coding strategy for the features of each study and item codes in this study were adapted from Wortman (1994) and Stock (1994), and the scaling of coding is consistent with the goals of this study. The coding sheet is presented in Table 1.

Table 1

Code Signing Sheet

Variable and options	Code
Item Content	
Algebra	1
Geometry	2
Arithmetic	3
Reasoning, problem solving, & application	4
Data analysis	5
Measurement	6
Quantitative comparison	7
Computation	8

Variable and options	Code
Memorization	9
Item format	
Multiple Choice	1
Free response	2
Not specified	3
DIF approach	
M-H	1
SIBTEST	2
LR	3
IRT	4
STD	5
Item preference (Favor)	
Male	1
Female	2
Sample source	
U.S.	1
International	2

Note. M-H stands for Mantel-Haenszel statistics. SMD stands for standardized mean difference in Standardization statistics. LR stands for logistic regression procedure.

In terms of coding the magnitude (effect size) of DIF, I categorized the magnitudes of DIF into three levels. For cases that report the effect size of the DIF using the M-H method and the Standardization method, I used the ETS 3-level classification scheme (Dorans & Holland, 1993), which was proposed by Zwick and Ercikan (1989). Researchers had shown that the three level classification system used by ETS is about to be appropriate for categorizing DIF magnitude (Linn, 1993; Zieky, 1993), even though there are arguments among researchers, test developers and institutions about the cutting scales between each two categories. In addition, since DIF statistics are usually difficult to interpret, the 3-level category allows test developers to use DIF more efficiently (Zieky, 1993).

For interpreting DIF effect size for logistic regression procedure, Jodion and Gierl (2001) also developed an evaluation guideline based on ETS 3-level classification, which was followed for this study, although it's accuracy has been questioned (French & Maller, 2007). Likewise, the guidelines proposed by Roussos and Stout (1996b) which also was based on 3-level ETS classification were used to evaluate the effect size of DIF from SIBTEST procedures (Table 2). This classification, given its high matching percentage with the LR guidelines, also provided a reliable classification of DIF items (Zheng, Gierl, & Cui, n.d.). Regarding the magnitude of DIF from IRT methods, Raju (1989) proposed that the computation of the area between the two ICCs could be viewed as the effect size of 3PL Item Response Theory Model's effect size. Maller (2001) proposed the root mean squared probability difference (RMSD) to be used as effect size of DIF, and RMSD was examined to be able to matched the ETS classification by tracing back to M-H statistics.

In terms of the three category, ETS Category A refers to items with negligible or non-significant DIF; Category B refers to items with slight to moderate magnitude of statistically significant DIF; and Category C refers to items with moderate to large magnitude of statistically significant DIF. In terms of keeping or discarding the items that are identified to have DIF, items with smaller absolute DIF values should be selected in preference to items with larger values. Items that are identified as A are considered appropriate use in test construction; items that are classified as B are used only if there is no A level items available to fill the content requirement of the test; and items classified as C need to go through the judgmental process, and be examined by content experts to determine if they meet the test specifications (Clauser & Mazor, 1998), in order to determine if they should be used.

Table 2

The Categorization of Differential Item Functioning Magnitude

Category	M-H	STD	LR	SIBTEST
Category A	$\lvert\Delta\rvert < 1$	$\lvert \mathrm{SMD}\rvert < .17$	$\mathit{\Delta R}^2\text{-}U < .035$	$\lvert\hat{\beta}\rvert < .059$
Category B	$1 \le \lvert\Delta\rvert < 1.5$	$.17 \le \lvert \mathrm{SMD}\rvert < .25$	$.035 \le \mathit{\Delta R}^2\text{-}U < .07$	$.059 \le \lvert\hat{\beta}\rvert < .088$
Category C	$\lvert\Delta\rvert \ge 1.5$	$\lvert \mathrm{SMD}\rvert \ge .25$	$\mathit{\Delta R}^2\text{ -U} \ge .07$	$\lvert\hat{\beta}\rvert \ge .088$

Note. M-H stands for Mantel-Haenszel statistics. SMD stands for standardized mean difference in Standardization statistics LR stands for logistic regression procedure.

Statistical Analysis Plan

The average DIF magnitude across items was calculated for each single DIF approach, due to the fact that there is no guidelines for combining effect sizes of different DIF methods. Further, the combined effects sizes within each single DIF approaches are presented as an average score of the magnitude.

Number of items within Category A, B, and C across all items are calculated and presented. This directly provides messages of the extent of existence and severity of gender DIF in mathematics assessments.

Item format by gender interactions and item content by gender interactions are presented in tables with results of the number of items having different format and content within each DIF category, as well as among all the items.

CHAPTER FOUR

RESULTS

The results of database searching process and statistical analysis of the data are presented in the following chapter. Not only are the results of this study consistent with previous research, but new evidence was generated in the area of gender DIF in mathematics test that previous primary studies have not identified.

Database Searching Results

The database searching results are presented in Table 3, Table 4 and Table 5. Applying the inclusion criteria, the searching process resulted in 14 journal articles and AERA conferences papers to be chosen by this meta-analysis study. First, three online database, Education Resources Information Center, PsychInfo, and Educational Full Text (EFT), were searched in sequence using keywords combinations during year 1990 to 2009. The keywords used include, differential item functioning, DIF, invariance, item bias, gender, sex, math, and mathematics. 41 references were generated after excluding three duplicated references in the search results. Subsequently the same process was conducted for online database PsychInfo and EFT. In a result, there were 16 and 17 referenced generated on PyschInfo and EFT, respectively. After searching online database, eight journals were manually searched. There was no new references identified (Table 3). As a result, there were seventy-four references that were further screened in the next step of selection of studies to be included in this study.

Table 3

Brief Searching Result of Online Database

		Database		
Keywords	Number of	ERIC	PsychInfo	EFT
(DIF or differential item functioning, or item bias or invariance) and (math or mathematics) and (gender or sex)	Articles resulted	41	16	17
	Inclusion studies	12	2	0

Note. ERIC stands for Education Resources Information Center online database. EFT stands for Education Full Text online database.

After applying inclusion criteria that discussed in the Chapter Three to all 74 references, there were 14 references left and were finally included in this study: 12 from ERIC and 2 from PsychInfo. Among the 14 references, four of them were American Educational Research Association, two of them are published on National Council on Measurement in Education (NCME) journals, and eight of them are published on the eight journals that do not include NCME journals (Table 4). There are 28 references related to gender differences in mathematics assessment. However, they were not included in this study because they do not provide sufficient DIF information that this study required. There are two duplicated references with different titles yet similar content , consequently only one of them was included in this study. Moreover, there are seven studies that provided sufficient information but were excluded from this study due to the reason that they were not published by the required journals or AERA conference. Twenty-four references were not relevant to the purpose of this study, hence they were excluded. All in all, there were 18.9% of the references were included in this study, while the rest of 81.1% of the references were not included after applying inclusion criteria.

Table 4

Data Source at Study Level

Source	#
APA	0
AERA	4
NCME	2
8 Journals(not including NCME)	8
Total	14

In terms of items that are analyzed in this study, there are totally 766 items, with 732 items from the US data source, and 34 items from international countries across the 14 studies (Table 5).

Table 5

Data Source at Item Level

Country	#	%
US	732	96%
International	34	4%
Total	766	100%

Furthermore, the information of each study as well as the items that were identified as DIF were recorded, coded, and summarized in three tables. The first column of each table is coded number for each study. The number in parentheses after the code number in Table 7 represents the number of the items that the following statistics describe. Table 6 is a description of references that were included in this study. It includes the type of article, the publication year, and the author(s) (Table 6).

Table 6

Sources of References Included In The Analysis

Reference #	Year	Author(s)	Publication Type
1	1996	Wang & Lane	Journal Article
2	2006	Mendes-Barnett & Ercikan	Journal Article
3	2004	Li, Cohen, & Ibarra	Journal Article
4	1993	Harris & Carlton	Journal Article
5	2001	Ryan & Chui	Journal Article
6	2001	Henderson	Conference Paper
7	1990	Kim, Plake, Wise, & Novak	Journal Article
8	1999	Garner & Engelhard	Journal Article
9	2002	Zhang	Conference Paper
10	2003	Gierl, Bisanz, Bisanz, & Boughton	Journal Article
11	1996	Ryan & Fan	Journal Article
12	1996	Ryan & Chui	Conference Paper
13	1992	Kubiak, O'Neill, & Rayton	Conference Paper
14	1997	Williams	Journal Article

Table 7 includes the total number of participants and the average age of participants in each study (Table 7). Table 8 (Appendix A) includes (a) test statistics on gender differences in mathematical performances, such as means, standard deviation, *p*-value, effect size, item parameter differences or item parameter estimates if available; (b) the content of the mathematical item(s); (c) the format of the item(s); (d) approaches used to detect DIF; (e) the magnitude of DIF (if the number of the items are more than 1, the magnitude is an average magnitude that is provided); (f) the magnitude category of DIF under ETS classification scheme, (g) preference of the item(s), that is, which gender the item(s) is in favor of, and (h) sample source(i.e., US or international) (Table 8, Appendix A).

Table 7

Participants in Studies

Study #	Number of Male/Female	Average Age
1	886/896	12-13
2	5,069/4,335	18
3-1	500/500	>18
3-2	1,000/1,000	>18
4	181,228/198,668	17-18
5	546/520	18
6	1,382/946	18
7-1	726/721	10-12
7-2	726/721	10-12
8	1,862/2,090	18

Study #	Number of Male/Female	Average Age
9-1	4,495/4,091	9
9-2	4,522/4,291	14
10-1	6,000/6,000	15
10-2	6,000/6,000	15
11-1	Appr. 758/811	14
11-2	Appr. 758/811	14
11-3	Appr. 758/811	14
11-4	Appr. 758/811	14
12-1	303/262	>18
12-2	297/254	>18
13	4,281/6,419	>18
14	495/444	9

Note. The number of participants in study 11-1 to 11-4 is presented as approximate number because they were not clarified in the original article. Under Study # column, the first number is the reference number in an order the same as Table 6; and the number after the dash represents the sub-studies within the reference.

Results of Statistical Analysis

The patterns of gender DIF in mathematics assessments.

In terms of the severity of the gender DIF in mathematics tests, there are approximately 46% of the items fell into category A, 23% of the items fell into category B, and 31% of the items fell into category C, as seen in Table 9. Within Category A, 48% of the items favored males, and 51.4% of the items favored females. Within Category B, 88.8% of the items favored males, whereas only 11.2% of items favored females. Within Category C, 47.5% of the items favored males, and 52.5% of the items favored females.

Table 9

Number and Percentage of DIF Items in Each Category

Category	#	%	Favor (M/F) (Within Category)
A	282	≈46%	137(48.6%)/145(51.4%)
B	135(+8)	≈23%	127(88.8%)/16(11.2%)
C	190(+8)	≈31%	94(47.5%)/104(52.5%)
Total	615	100%	

Note. 151 items among 766 items do not have classification information. "+8" refers to the 8 items that are categorized as between B and C. Category A, B, and C refers to the DIF magnitude classification scheme developed by Educational Testing Service. "#" refers to number of items in each category. "%" refers to proportion of items in each category within the total 615 items. M/F refers to the Males/Females.

The average magnitude of gender DIF in mathematics assessment.

The DIF identification approaches used by each study and the combined effect size of gender DIF for each DIF approach are presented in Table 10. The combined effect size is the average magnitude of DIF within each category. In a result, there are six studies that used M-H statistics, and the combined effect size is .01, which falls into category A. The positive value indicated that averagely the DIF items identified by M-H statistics favor females. There are four studies that used SIBTEST, and combined effect size is -.02, which also falls into Category A. The negative value indicated that on average the DIF items identified by SIBTEST favored females. Likewise, there are 2 studies that used logistic regression as their DIF approach, and their combined effect size is -.54 falling into category C. And the negative value indicated that on average the DIF items identified by logistic regression favored males. There is no study among the 14 used Standardization approach to identify DIF. Two studies used IRT approaches, yet neither of them provided information of statistic parameters. In a result, it is not possible to compute the combined effect size of this category using neither computing Area between the two ICCs nor root mean squared probability difference (RMSD).

Table 10

DIF Approach at Study Level and Combined Effect Size at Item Level

Approach	#	%	Combd. *d*	Classification	Favor(M/F)
Mantel-Haenszel	6	42.90%	0.01	A	F
Simultaneous Item Bias Test	4	28.60%	-0.02	A	F
Logistic Regression	2	14.25%	-0.54	C	M

Approach	#	%	Combd. *d*	Classification	Favor(M/F)
Item Response Theory methods	2	14.25%	-	-	-
Total	14	100%			

Note. The combined effect size of IRT methods are not processed due the reason that neither of the two studies provided item parameter information. No study used Standardization procedure, so that it is not listed in the table. "#" refers to the number of studies that used each DIF approach. "%" refers to the proportion of studies that used each DIF approach among the total 14 studies. *d* refers to effect size. M/F refers to Males/Females.

The item characteristics by gender interactions in mathematics assessments.

The characteristics of items within each category and their preference to gender (favor male or favor female) are presented in Tables 11 and 12. Note that the items in these two tables do not include the items that do not have classification information and either item format or item content information. The item characteristics of the total 766 items combined and their preference to gender are presented in Tables 13 and 14.

From Tables 11 and 13, multiple choice items tend to favor males with 53.1% of the total items favoring males and 46.9% of the total items favoring females. Table 12 and 14 indicate that Algebra items tend to favor females and Geometry items tend to favor males. Also, items related to reasoning, problem solving, application of math concepts and data analysis tend to favor males, while memorization items tend to favor females. Arithmetic items tend to favor males with 76.4% of the items favor this group; Measurement items tend to favor males, since all items in this study favor males; and items related to Computation tend to favor females, in that

72.6% of the items are in favor of females, while none of the previous primary studies has ever identify arithmetic, measurement, and computation to be sources of gender DIF.

Table 11

Items Format within Each Category and Item Preference Exhibited by Percentage

	Category A			Category B			Category C		
Format	#	%	Favor M/F	#	%	Favor M/F	#	%	Favor M/F
Multiple Choice	61	21.6%	29 (47.5%) /32(52.5%)	27	18.9%	13(48.1%) /14(51.9%)	161	81.3%	70(43.5%) /91(56.5%)
Free response	0	0	-	2	1.4%	1(50%)/1(50%)	4	2%	0/4(100%)
Not identified	221	78.4%	-	114	79.7%	-	33	16.7%	-
Total	282	100%		143	100%		198	100%	

Note. 151 items do not have classification information. There are 8 items that are between Category B and C. The Category A, B, and C refer to the DIF magnitude classification scheme from Educational Testing Service. "#" refers to the number of items within each format. "%" refers to the proportion of items within each format under each category. M/F refers to Males/Females.

Table 12

Item Content within Each Category and Item Preference Exhibited by Percentage

	Category A			Category B			Category C		
Content	#	%	Favor M/F	#	%	Favor M/F	#	%	Favor M/F
Algebra	114	41.6%	0/114 (100%)	27	18.9%	14 (51.8%) /13(48.2%)	48	24.2%	19(39.6%) /29(60.4%)
Geometry	22	8.0%	12(54.5%)/ 10 (45.5%)	108	75.5%	108 (100%)/0	45	22.7%	45(100%)/0
Arithmetic	120	43.8%	120(100%)/0	0	-	-	34	17.2	0/34(100%)
Reasoning, problem solving, & application	4	1.5%	3(75%)/1(25%)	3	2.1%	2 (66.7%) /1(33.3%)	32	16.2	30(93.8%) /2(7.2%)
Data analysis	0	-	-	0	-	-	0	-	-
Measurement	0	-	-	0	-	-	0	-	-
Quantitative comparison	2	0.7%	0/2(100%)	5	3.5%	3(60%)/2(40%)	0	-	-
Computation	0	-	-	0	-	-	39	19.7%	0/39(100%)

	Category A			Category B			Category C		
Content	#	%	Favor M/F	#	%	Favor M/F	#	%	Favor M/F
Memorization	12	4.4%	0/12(100%)	0	-	-	0	-	-
Total	274	100%		143	100%		198	100%	

Note. There are 8 items that do not have content information. The Category A, B, and C refer to the DIF magnitude classification scheme from Educational Testing Service. "#" refers to the proportion of items within each content area. "%" refers to the number of items within each content area under each category. M/F refers to Males/Females.

Table 13

Item Format for All Items and Their Preference Exhibited by percentage

Item format	#	%	Favor M/F
Multiple Choice	356	46.5%	189 (53.1%)/167 (46.9%)
Free Response	12	1.6%	0/12 (100%)
Not Identified	398	51.9%	-
Total	766	100%	

Note. "#" refers to the number of items within each format. "%" refers to the proportion of items with specific item format among the total number of items. M/F refers to Males/Females.

Table 14

Items Content for All Items and Item Preference Exhibited by Percentage

Item content	#	%	Favor M/F
Algebra	213	27.6%	38(17.8%)/175(82.2%)
Geometry	218	28.2%	203(93.1%)/15(6.9%)
Arithmetic	161	20.8%	123(76.4%)/38(23.6%)
Reasoning, problem solving, & application	63	8.2%	46(73.0%)/17(27.0%)
Data analysis	14	1.8%	14(100%)/0
Measurement	23	3.0%	23(100%)/0
Quantitative comparison	7	0.9%	3(42.9%)/4(57.1%)
Computation	62	8%	17(27.4%)/45(72.6%)
Memorization	12	1.5%	0/12(100%)
Total	773	100%	

Note. There are 16 items that don't have item content information. There are 23 items that are measuring both Geometry and Measurement. "#" refers to the number of items within each ontent area. "%" refers to the proportion of items with specific content area among the total number of items. M/F refers to Males/Females.

CHAPTER FIVE

DISCUSSION

This meta-analysis study investigated 14 studies with 766 items that related to gender DIF in mathematics tests. The participants ranged from elementary school examinees to college level examinees in the US as well as international countries. A pseudo-effect size of gender DIF was computed and interpreted based on ETS DIF classification scheme.

One of the goals and major contributions of this study is to identify the presence and extent of gender DIF in mathematics tests. With about 46% of the items falling in into Category A indicates that (a) gender DIF does exist in mathematics tests and (b) the extent of severity of the issue is medium. This outcome also is consistent with the statement made by O'Neil and McPeek (1993): relatively few items have DIF values that are considered high, when they conducted a research investigating the relationship between DIF and item and test characteristics of the Scholastic Aptitude Test (SAT), the Graduate Management Admission Test (GMAT), the American College Testing Program Assessment Test (ASAT), the NTE Core Battery (NTE), and the Graduate Record Examination (GRE) overall. These tests/examinations are large-scale tests with much more resources and effort spent on the development of the tests compared to smaller tests, which likewise, 11 out of 14 studies included in this study used large-scale statewide or nationwide mathematics assessments. Those 11 studies covered 568 items with DIF magnitudes, in which 264 (46.5%) items fell into category A, 133 (23.4%) items fell into category B, and 171 (30.1%) fell into category C. These results further illustrate O'Neil and McPeek's point that most items in large-scale tests are A items, and relatively few items are C items.

Due to the fact that it is not possible to convert effect size values from different DIF approaches to a common scale, combined average effect sizes for different DIF approaches are separately calculated. The results showed that more than 70% of the items went to Category A, which also demonstrated the previous point.

In terms of item format by gender interaction and item content by gender interaction, the results are largely consistent with previous research. That multiple choice items tend to favor males is consistent with the findings from previous research (Burton, 1996; Garner & Engelhard, 1999), even though the percentage difference is not very large. It is also consistent with previous research that free response items tend to favor females with all 12 free response items favoring females (Burton, 1996; Garner & Engelhard, 1999). In addition, it is consistent with previous research that Algebra items tend to favor females and Geometry items tend to favor males (Doolittle & Cleary, 1987; O'Neil & McPeek, 1993, Garner & Engelhard, 1999; Burton, 1996; Garner & Engelhard, 1999). In addition, the results of this study in terms of reasoning, problems solving, application, and data analysis items favoring males and memorization items favoring females are consistent with previous research (Harris & Carton, 1993; O'Neil & Mcpeek, 1993).

More importantly, the study has also generated some evidence that could be plausible explanation of male and female differences in performance in mathematics and that previous primary research alone did not identify. This new evidence includes: Arithmetic and Measurement items tend to favor males, while Computation items tend to favor females.

Regarding the process of conducting this study, among 766 items that are included in this study, only 34 items are from international sources. Due to the limited amount of data, the international data was not analyzed separately. Item context by gender interaction was not investigated in this study because of two reasons. First, there were no studies that are included in this review that provided information of the item context, that is, whether the setting of the item reflects a male role (e.g., sports, cars, etc.), a female role (e.g., dolls, clothes, etc.), or is neutral setting (e.g., fruits, TV, etc.). Second, even though several studies provided the item questions on the test in their article/conference papers, it is impossible for the author to code item context correctly and reliably, due to the lack of a standardized protocol that clearly define the characteristics of different item contexts. Consequently item context by gender interaction was not processed in this study.

Delimitations And Limitations

Delimitation.

This study included empirical studies ranging from elementary to higher education across the United States as well as international samples between January 1990 and March 2009. However, during the searching process, the majority of the studies that are related to gender difference in mathematical tests did not report DIF, even though the results showed there were mean differences between males and females. As a result, mean difference studies were not included in this study, as the focus of this meta-analysis is on DIF.

Limitations.

By taking meta-analysis approach for this review, two technical limitations are unavoidable due to the nature of meta-analysis methodology. First, the synthesis of researches involves an aggregation of studies that are dissimilar in a lower or higher degree, some extent of mixing “apples” and “oranges” is unavoidable. Second, the potential impact of including and excluding particular studies is substantial for allowing the readers to fully judge potential effects on references (Harwell & Maeda, 2008). In order to reduce the lack of clarity, I provided clear description of inclusion and exclusion criteria to the largest degree, and leave the judgment to the readers about the inclusion criteria.

A limitation of this study is that not all studies that reported DIF had sufficient or clear details and complete information of the characteristics of the DIF items, which made the coding process fairly complex. Moreover, there are 27 references that are related to gender differences in mathematics assessment did not adequately provide the required information for this study; consequently they were excluded from this study.

In terms of the source availability, studies focusing on gender-related DIF in mathematics assessment are rare (Garner & Engelhard, 1999) to begin with. There are more studies that are focusing on comparing mean scores differences than on gender DIF analysis. Furthermore, only studies that were published in upper-tier journals (i.e., APA and NCME) and AERA annual conference were included in this review. Seven studies with sufficient data were excluded from this study due to the publication sources do not meet the requirements for this study.

Regarding the coding of DIF magnitudes, this study used ETS 3-level DIF classification scheme. This scheme is approved by researchers to be practically right (Linn, 1993; Zieky, 1993), yet the cutting points between each two categories are still under debate by researchers and different institutions. Consequently, using ETS classification scheme could be a potential threat to the judgments and conclusions made from this study.

Studies with non-significant statistical results might have never been published or accepted by AERA conferences. And it is often suspected that published studies will show results that are more often statistically significant and have larger effect sizes compared to unpublished studies (Begg, 1994). This suspicion has directly impact on this study, because if studies with non-significant DIF identification are less likely to be published, it would further strengthen the finding of this study; that is, the majority of the DIF items in mathematics assessment are with negligible magnitude.

Inter rater reliability calculation was not processed yet in this study. Even though a basic coding index is used by the author to maintain the reliability of the coding process, high inter rater agreement values still needed to provide evidence that a variable is reliably being coded.

Implications

Although gender DIF is often found in items on mathematics assessments, little information is actually known about the sources of these differences (Mendes-Barnett & Ercikan, 2006), partly because very few studies have examined gender-related DIF on mathematics (Lane, Wang, & Magone, 1996), even though there are mean score differences between males and females in most empirical studies that examine such gender differences in mathematics tests. Thus, an important implication generated from this study is that more analytical studies dedicated to the gender gap in mathematics may be needed in order to identify the sources of gender DIF. By not including items that have large potential DIF based on research, this kind of study also would be of benefit in the test developing industry and mathematics educational systems. Moreover, in the future, with sufficient data generated adequate amount of studies leading to the suggestion that DIF is merely a minor problem in gender differences in mathematics assessment, researchers should probably focus more on the true gender gaps in mathematics assessments and education.

Considerations should be given to teachers, instructors, as well as any other test developers regarding test writing by being aware that there are certain types of items in terms of item format and item content that are most likely to show gender DIF in mathematics tests. Consequently they should be cautious of including those questions when designing an assessment. Also, to further maintain the quality of the measurement, before placing items into the item pool, DIF analysis should be conducted if possible, as DIF items poses a potential threat to test fairness. To decision makers, they also should be aware of the existence of gender-DIF in mathematics tests, and understand that the un-equivalent performance of males and females in a math assessment may due to the portion of DIF items.

Recommendations

There are actions researchers can take in reporting DIF analysis results, which will benefit the field to better understand the extent of the problem of gender DIF in mathematics and beyond. First, researchers should provide item parameters' information to the readers, so that readers are able to make their own judgment if they want. Second, researchers should provide the type of DIF approach and formula they used to compute DIF, which will benefit the readers and other researchers to repeat the calculation process and make judgments if they want. Third, researcher should provide the effect size of DIF, the formula they used to calculate effect size, preferably also the Category of DIF magnitude based on ETS classification scheme. It would aid the conducting future large-scale reviews in the area of gender DIF in mathematics education to a large extent. The fourth and final recommendation, item characteristics should be identified. Providing the actual items would be a good addition. So readers and other researchers would be able to conduct the analysis, repeat the study, or make their own judgments if they want. There were 28 studies excluded from this study because they don't have above information. It was quite a loss of potential influence of DIF in mathematics tests.

More extensive insight into gender differences in mathematics achievement might be gained by extending the study with a qualitative component. For instance, interviewing students or observing students while they solve some of the same problems would help to determine different approaches or patterns of thinking, which may further explain observed gender difference (Garner & Engelhard, 1999). Follow-up focus group studies with DIF items to males and females also would be helpful to understand the reasons for different performance between males and females.

During the database searching process, there are seven articles/research reports with that do not qualify for this study, because they are not published by the required journals/AERA conference. However they are dedicated to gender DIF in mathematics tests and have adequate information. It would be valuable to include them into this study and compared the results. They cover totally 1291 items identified as showing DIF with 96.3% of items are multiple choice items. Primary data analysis results show that 82.3% of the items belong to category A, which have negligible effect size, only 2.2% belong to category B, and less than 1% belong to category C. 40% of the items favor males, and 60% of items favor females. All the samples were from the United States. These preliminary results are consistent with the results of this study. At this point, it looks like adding these data into this study would at least strengthen one of the major findings of this study: the majority of DIF items fall into category A with negligible effect size. All in all, further analyzing these studies, including into and compare them with the results of this study would be a good addition.

A statistical method that could place effect size values from different DIF approaches on the same scale, generate a common effect size or common metric of differences among different DIF approaches is needed to conduct future studies on topics similar to this study, With such method, a combined effect size generated from different DIF approaches could be calculated, which will allow researchers go beyond the ETS pseudo-effect size, therefore help researchers probe, understand, and discuss this topic in a deeper manner.

Studies about the distractors options by differential distractor functioning (DDF) analysis (Doran and Holland, 1993) that students usually choose would also be a valuable and useful research to the field of mathematical education. Since most standardized test items are dichotomously scored as either correct or incorrect, the distractors refer to the options that examinees are choosing incorrectly (Middleton & Laitusis, 2007). DDF analysis are used to determine whether different distractor, or incorrect option choices, attract various group differently (Green, Crone, & Folk, 1989). The importance of studying DDF lies in its advantage of providing additional source of information about student performance and modifying an assessment by eliminating of distractors. However, applying keywords combination of differential distractor functioning (or DDF), gender (or sex), and mathematics (or math) to search within ERIC, PsychInfo, and Educational Full Text resulted in only one reference, which was not relevant to Mathematics assessment. As a result, the need for more studies of distracters in mathematics assessment after detection of DIF has been identified (Garner & Engelhard, 1999), and such studies would be valuable additions to DIF studies in terms of making contributions to the area of mathematics assessment and education.

Summary

This study found that the majority of DIF items in mathematics assessments are negligible DIF items, which indicated that the severity level of gender DIF issue is not high in mathematics tests. Also, not only the results from this study are largely consistent with previous research, but were new evidence also generated from this study in regard to the possible sources of gender DIF in mathematics tests including Arithmetic, Measurement and Computation.

Future works should further explore the topics above in order to aid and enrich the understanding of gender DIF in mathematics assessment and education. Researchers should continue such works, in that the outcomes of such studies would help a large group of audiences, mostly the researchers, educators, test developers, decision makers in Mathematics education as well as education as a whole, to maintain the fairness and validity of a test(s), correctly use a test score for making decisions and references, and further maintain the equity of society, as it relates to test score inferences.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 2*9, 67-91.

Alexander, K., & Pallas, A. (1982). *Sex differences in quantitative SAT performance: new evidence on the differential coursework hypothesis.* Unpublished manuscript, John Hopskin University at Baltimore.

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

American Psychological Association. (2001). Publication manual (5th ed.). Washington, DC: Authors.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Erlbaum.

Ayalon, H. (2003). Women and men go to university: Mathematical background and gender differences in choice of field in higher education. *Sex Roles, 48,* 277-290.

Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal of Research in Mathematics Education, 21*(1), 47-60.

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

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399-409). New York: Russell Sage Foundation.

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

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple choice mathematics items administered to national probability samples. *Journal of Educational Measurement*, *38*(1), 51-77.

Bohlin, C. F. (1994). Learning style factors and mathematics performance: Sex-related differences. *International Journal of Educational Research, 21*, 387-397.

Burton, N. W. (1996). How have changes in the SAT affects women's math scores? *Educational Researcher*, *15*(4), 5-9.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*(3), 397-412.

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31-43.

Cole, N. S. (1993). History and development of DIF . In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25-33). Hillsdale, NJ: Erlbaum.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-219). New York: American Council in Education/Macmillan.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369-382.

Cooper, H., & Hedges, L. V. (1994a). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 3-14). New York: Russell Sage Foundation.

Cooper, H., & Hedges, L. V. (1994b). Potentials and limitations of research synthesis. In H. Cooper& L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 521-529). New York: Russell Sage Foundation.

Crobach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Bruan (Eds.), *Testing validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Dai, D. Y. (2006) There is more to aptitude than cognituve capacities. *American Psychologist*, *61*, 723-724.

Davenport, E. C., Davison, L., Kuang, H., Ding, S., & Kwak, N. (1998). High school mathematics course-taking by gender and ethnicity. *American Educational Research Journal*, *35*, 497-514.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education*, *11*, 279-299.

Delgado, A. R. & Prieto, G. (2004). Cognitive mediators and sex-related differences in mathematics. *Intelligence*, *32*, 25-32.

Dooley, D. (1990). *Social research methods* (2nd ed.). Englewood Cliff, NJ: Prentice-Hall.

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: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kullick, E. (1986). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Erlbaum.

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

Engelhard, G., Anderson, D., & Gabrielson, S. (1990). An empirical comparison of Mantel Haenszel and Rasch procedure for studying differential item functioning on teacher certification tests. *Journal of Research and Development in Education*, *23*, 172-179.

Feingold, A. (1992). Sex differences in variability in intellectual ability: A new look at an old controversy. *Review of Educational Research*, *62*, 61-84.

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles*, *30*, 81-92.

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal*, *14*, 51-71.

French, B. F., & Mantzicopoulos, P. (2007). An examination of the first/second-grade form of the pictorial scale of perceived competence and social acceptance: Factor structure and stability by grade and gender across groups of economically disadvantaged children. *Journal of School Psychology, 45*, 311-331.

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*, 373-393.

Frost, L. A., Hyde, J. S., & Fennema, E. (1994). Gender, mathematics performance, and mathematics-related attitudes and affect: A meta-analytic synthesis. *International Journal of Educational Research, 21*, 373-384.

Gallagher, A. M., & Lisi, R. D. (1992, April). *Gender differences in mathematics problem solving strategies.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

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

Green, P. W., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26*(2), 147-160.

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.

Gross, S. (July 1988). *Participation and performance of women and minority in mathematics.* Department of educational accountability, Montgomery County Public School.

Hall, J. A., Rosenthal, R., Tickle-Degnen, L., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 17-28). New York: Russell Sage Foundation.

Hambleton, R. K., & Jones, R. W. (1994). Comparison of empirical and judgmental procedure for detecting differential item functioning. *Educational Research Quarterly*, *18,* 21-36.

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

Harwell, M., & Maeda, Y. (2008). Deficiencies of reporting in meta-analysis and some remedies. *The journal of experimental education*, *76*, 403-428.

Henderson, D. L. (2001, April). *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.

Hilton, T. L., & Berglund, G. W. (1974). Sex differences in mathematics achievement: A longitudinal study. *Journal of Educational Research*, *67,* 231-237.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Testing validity* (pp. 129-145). Hillsdale, NY: Erlbaum.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hong, E., O'Neil, F. Jr., & Feldon, D. F. (2005). Gender effects on mathematics achievement: Mediating role of state and self regulation and test anxiety. In A. M. Gallagher & J. C. Kaufman, (Eds.). *Gender differences in mathematics: An integrative psychological approach* (pp. 264-293). New York: Cambridge University Press.

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

Ibarra, R. A., (2001). *Beyond affirmative action.* Madison, WI: University of Wisconsin press.

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

Jackson, C., & Braswell, J. (1992, April). *An analysis of factors causing differential item functioning on SAT Mathematics items.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Jarvis, O. T. (1964). Boy-girl ability differences in elementary school arithmetic. *School Science and Mathematics*, *64*, 657-659.

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

Johnson, E. S. (1984). Sex differences in problem solving. *Journal of Educational Psychology, 76*, 1359-1371.

Kubiak, A., O'Neil, K., & Payton, C. (1992, April). *The effects of using educational background variables in DIF analyses.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Katzman, J., Loewen, J., & Rosser, P. (1988, April). *Gender bias in SAT items.* Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA.

Kim, H., Plake, B. S., Wise, S. L., & Novak, C. D. (1990). A longitudinal study of sex-related item bias in mathematics subtests of the California Achievement Test. *Applied Measurement in Education, 3*(3), 275-284.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practice, 15*, 21-27.

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

Leinhardt, G., Seewald, A. M., & Engel, M. (1979). Learning what's taught: Sex differences in instruction. *Journal of Educational psychology*, *71*, 432-439.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing*, *4*, 115-136.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Erlbaum.

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

Linn, M. C. & Kessel, C. (1995, April). *Participation in mathematics courses and careers: climate, grades, and entrance examination scores.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Maller, S. J. (2001). Differential item functionin in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement*, *61*, 793-817.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning*. (Educational Testing Service Rep. No. RR-08-43). Princeton, NJ: Educational Testing Service.

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*, 289-304.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33 45). Hillsdale, NJ: Erlbaum.

Middleton, K., & Laitusis, C. C. (2007). *Examining test items for differential distractor functioning among students with learning disabilities*. (Educational Testing Service Rep. No. RR-07-43). Princeton, NJ: Educational Testing Service.

Navarro, C. (1989, March). *Why do women have lower average SAT-Math scores than men.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 139-162). New York: Russell Sage Foundation.

Pederson, D. M., Shinedling, M. M., & Johnson, D. L. (1968). Effects of sex of examiner and subject on children's quantitative test performance. *Journal of Personality and Social Psychology*, *10*, 251-254.

Peenner, A. M. (2003). International gender × item difficulty interaction in mathematics and science achievement tests. *Journal of Educational Psychology, 95,* 650-655.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495 502.

Randhawa, B. S., & Hunt, D. (1987). Sex and rural-urban differences in standardized achievement scores and mathematics subskills. *Canadian Journal of Education*, *12*, 137 151.

Rock, D. A., & Pollack, J. M . (1991) *Psychometric report for the NELS:88 base year test battery*. (Rep. No. NCES-91-468). Princeton, NJ: Educational Testing Service.

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

Roussos, L., & Stout, W. (1996b). Simulation studies of the effects of small sample size and studies item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, *33*, 215-230.

Rudas, T., & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics*, *22*, 31-45.

Ryan, K. E., & Chiu, S. (1996, April). *Detecting DIF on mathematics items: The case for gender and calculator sensitivity*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

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

Ryan, K. E., & Fan, M. (1994, April). *Gender differences on a test of mathematics: Multidimensionality or differential test functioning.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.

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

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement*, *24*, 97-118.

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, *27*, 109-131.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Vol. 3. Theory into practice* (pp. 65-82). Norwood, NJ: Ablex.

Seegers, G., & Boekarts, M. (1996). Gender-related differences in self-referenced cognitions in relation to mathematics. *Journal of Research in Mathematics Education, 27*(2), 215-240.

Shealy, R.T., & Stout, W. F. (1993). 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: Erlbaum.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22,* 77-105.

Skaggs, G., & Lissitz, R. W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement, 29*(3), 227-242.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 125-138). New York: Russell Sage Foundation.

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). History and development of DIF . In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-114). Hillsdale, NJ: Erlbaum.

Wandt, E., Adam, G. M., Collett, D. M., Michael, W. B., Ryan, D. G., & Shay, C. B. (1965). *An evaluation of educational research published in journals.* Retrieved April 2nd, 2009, from http://www.indiana.edu/~educy520/readings/wandt65.pdf.

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

Williams, V. S. L. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education*, *10*(3), 253-267.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 97-109). New York: Russell Sage Foundation.

Wright, D. J. (1987). Assessment of unexpected differential item difficulty for Asian-American examinees on the Scholastic Aptitude Test. In A. P. Schmitt & N. J. Dorans (Eds), *Differential item functioning on the Scholastic Aptitude Test* (Research Memorandum No. 87-1) (pp. 1-27), Princeton, NJ: Educational Testing Service.

Zhang, Y. (2002, April). *DIF in large scale mathematics assessment: The interaction of gender and ethnicity.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Zhang, Y., Dorans, N. J., & Matthews-Lopez, J. L. (2005). *Using DIF dissection method to assess effects of item deletion.* (College Board Rep. No. 2005-10). Princeton, NJ: Educational Testing Service.

Zieky, M. (1993). History and development of DIF. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zohar, A., & Gershikov, A. (2008). Gender and performance in mathematical tasks: does the context make a difference? *International Journal of Science and Mathematics Education*, *6*, 677-693.

Zumbo, B. D. (1999). *A handbook of theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of human resources, research and evaluation, department of national defense.

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

APPENDIX A

Item Characteristics and DIF Statistics

Table 8

Description of DIF Statistics of Each Study

	a									
Study #	M (M/F)	SD (M/F)	*d*	b	c	d	e	f	g	h
1(1)	1.42/1.48	-	-	Mathematical thinking and reasoning skills	2	3	-	C	2	1
1(1)	SAA			Mathematical thinking and reasoning skills	2	3	-	C	2	1
2(6)	30.2/30.1	7.47/7.12	-	Problem solving	1	2	0.098	C	1	2
2(5)	SAA			Polynomials	1	2	-0.059	B	2	2
2(3)	SAA			Quadratic systems	1	2	-0.001	A	2	2
2(8)	SAA			Logarithms & exponents	1	2	0.086	B	1	2
2(8)	SAA			Sequence & series	1	2	-0.002	A	2	2
2(4)	SAA			Geometry	1	2	0.008	A	1	2
3(5)	-	-	-	Basic algebra	1	1	-	-	1	1
3(3)	-	-	-	Arithmetic	1	1	-	-	2	1
3(2)	-	-	-	Arithmetic	1	1	-	-	1	1
3(3)	-	-	-	Intuitive geometry	1	1	-	-	1	1

	a									
Study #	M (M/F)	SD (M/F)	*d*	b	c	d	e	f	g	h
3(2)	-	-	-	Intuitive geometry	1	1	-	-	2	1
3(1)	-	-	-	Advanced algebra	1	1	-	-	2	1
3(2)	-	-	-	Advanced algebra	1	1	-	-	1	1
3(3)	-	-	-	Geometry	1	1	-	-	2	1
3(2)	-	-	-	Geometry	1	1	-	-	1	1
3(6)	-	-	-	Basic algebra	1	1	-	-	2	1
3(7)	-	-	-	Intuitive geometry	1	1	-	-	1	1
3(1)	-	-	-	Arithmetic	1	1	-	-	2	1
3(3)	-	-	-	Arithmetic	1	1	-	-	1	1
3(3)	-	-	-	Advanced algebra	1	1	-	-	1	1
3(1)	-	-	-	Basic algebra	1	1	-	-	1	1
3(3)	-	-	-	Geometry	1	1	-	-	1	1
3(2)	-	-	-	Algebra	1	1	-	-	2	1
4(108)	-	-	-	Arithmetic	3	1	-0.09	A	1	1

	a									
Study #	M (M/F)	SD (M/F)	*d*	b	c	d	e	f	g	h
4(103)	-	-	-	Algebra	3	1	0.03	A	2	1
4(98)	-	-	-	Geometry	3	1	-0.13	B	1	1
5(5)	23.58/18.74	8.14/7.53	-	Geometry	3	2	0.155	C	1	1
5(10)	SAA			geometry	3	2	-0.001	A	2	1
5(10)	SAA			Trigonometry	3	2	0.069	B	1	1
5(10)	SAA			Algebra	3	2	0.09	C	1	1
5(9)	SAA			Algebra	3	2	-0.093	C	2	1
6(3)	-	-	-	Sequence & series	3	1	-	B/C	1	1
6(1)	-	-	-	Polynomial functions	3	1	-	B/C	1	1
6(1)	-	-	-	Exponential & Logarithmic	3	1	-	B/C	1	1
6(1)	-	-	-	Permutations & combinations	3	1	-	B/C	1	1
6(1)	-	-	-	Statistics	2	1	-	B/C	2	1
6(1)	-	-	-	Polynomial functions	2	1	-	B/C	2	1
7(6)	-	-	-	Computation	3	4	-	-	2	1

	a									
Study #	M (M/F)	SD (M/F)	*d*	b	c	d	e	f	g	h
7(6)	-	-	-	Computation	3	4	-	-	1	1
7(11)	-	-	-	Concept & application	3	4	-	-	1	1
7(13)	-	-	-	Concept & application	3	4	-	-	2	1
8(11)	7.67/7.28	2.24/2.28	.11	Number & computation	1	4	0.92	-	1	1
8(14)	10.94/10.58	2.55/2.58	.14	Data analysis	1	4	0.37	-	1	1
8(23)	16.05/15/23	4.55/4.64	.18	Geometry & measurement	1	4	0.78	-	1	1
8(12)	9.00/9.41	2.58/2.30	-.17	Algebra	1	4	-3	-	2	1
8(8)	3.24/3.19	-	-	-	1	4	-0.49	-	2	1
9(2)	34.95/34.50	9.76/9.54	-	Problem solving	1	1	-0.65	A	1	1
9(1)	26.58/26.16	10.69/9.91	-	Problem solving	1	1	-0.57	B	1	1
9(1)	SAA			Problem solving	1	1	1.7	B	2	1
10(6)	34.70/33.14	10.90/11.02	-	Spatial	1	2	0.25	C	1	1
10(4)	SAA			Memorization	1	2	-0.04	A	2	1
10(8)	32.71/31.64	11.30/11.17	-	Spatial	1	2	0.29	C	1	1

	a									
Study #	M (M/F)	SD (M/F)	*d*	b	c	d	e	f	g	h
10(8)	SAA			Memorization	1	2	-0.03	A	2	1
11(6)	17.27/16.54	6.86/6.52	-	Algebra	1	2	-0.265	C	2	1
11(12)	17.27/16.54	6.86/6.52	-	Arithmetic	1	2	0.009	A	1	1
11(9)	SAA			Geometry	1	2	0.249	C	1	1
11(10)	SAA			Computation	1	2	-0.334	C	2	1
11(4)	SAA			Application	1	2	0.16	C	1	1
11(6)	15.35/15.26	6.36/6.88	-	Algebra	1	2	-0.083	B	2	1
11(12)	SAA			Arithmetic	1	2	-0.535	C	2	1
11(9)	SAA			Geometry	1	2	0.369	C	1	1
11(11)	SAA			Computation	1	2	-0.434	C	2	1
11(4)	SAA			Application	1	2	0.211	C	1	1
11(6)	16.34/16.34	6.62/7.05	-	Algebra	1	2	-0.181	C	2	1
11(11)	SAA			Arithmetic	1	2	-0.145	C	2	1
11(8)	SAA			Geometry	1	2	0.032	A	1	1

	a									
Study #	M (M/F)	SD (M/F)	*d*	b	c	d	e	f	g	h
11(9)	SAA			Computation	1	2	-0.151	C	2	1
11(8)	SAA			Application	1	2	0.217	C	1	1
11(6)	17.22/17.19	6.44/6.94	-	Algebra	1	2	-0.207	C	2	1
11(11)	SAA			Arithmetic	1	2	-0.329	C	2	1
11(8)	SAA			Geometry	1	2	0.34	C	1	1
11(9)	SAA			Computation	1	2	-0.245	C	2	1
11(8)	SAA			Application	1	2	0.384	C	1	1
12(1)	7.57/6.68	3.09/2.89	-	Algebra	3	3	-0.468	C	1	1
12(1)	SAA			Algebra	3	3	-0.538	C	1	1
12(1)	4.54/3.67	2.45/2.12	-	Algebra	3	3	-0.621	C	1	1
13(1)	39.69/32.01	11.46/10.23	-	Quantitative comparison	1	1	1	B	2	1
13(1)	SAA			Quantitative comparison	1	1	-0.76	B	1	1
13(1)	SAA			Quantitative comparison	1	1	0.69	A	2	1
13(1)	SAA			Discrete quantitative	1	1	-1.03	B	1	1

	a									
Study #	M (M/F)	SD (M/F)	*d*	b	c	d	e	f	g	h
13(1)	SAA			Data interpretation	1	1	-0.19	A	1	1
13(1)	SAA			Discrete quantitative	1	1	0.46	A	2	1
13(1)	SAA			Quantitative comparison	1	1	0.8	A	2	1
13(1)	SAA			Quantitative comparison	1	1	1.09	B	2	1
13(1)	SAA			Quantitative comparison	1	1	-1.05	B	1	1
13(1)	SAA			Quantitative comparison	1	1	-1.17	B	1	1
14(6)	-	-	-	-	1	1	0.19	A	2	1
14(2)	-	-	-	-	1	1	-0.09	A	1	1

Note. Information under column a is test statistics, b is item content, c is item format, d is DIF approach, e is DIF magnitude, f is DIF category, g is preference of the items, and h is sample source. M/F refers to Males/ Females. M refers to Mean Score, SD refers to the Standardization. *d* refers to effect size. Under the Study # column, the first number is the reference' number, it is in an order the same with Table 6. The numbers in brackets are the number of items within different analysis.