This is to certify that the
dissertation entitled

A ATUDY ON DIFFERENTIAL ITEM FUNCTIONING (DIF) OF THE BASIC MATHEMATICAL COMPETENCE TEST FOR JUNIOR HIGH SCHOOLS IN TAIWAN

presented by

Chien-Ming Cheng

has been accepted towards fulfillment
of the requirements for the

Ph.D	degree in	Counseling Educational Psychology, and Special Education

Mark W. Roehrig
Major Professor's Signature

April 29, 2005
Date

DATE DUE	DATE DUE	DATE DUE
MAR 2 8 2012		

A STUDY ON DIFFERENTIAL ITEM FUNCTIONING (DIF) OF THE BASIC MATHEMATICAL COMPETENCE TEST FOR JUNIOR HIGH SCHOOLS IN TAIWAN

By

Chien-Ming Cheng

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

2005

ABSTRACT

A STUDY ON DIFFERENTIAL ITEM FUNCTIONING (DIF) OF THE BASIC MATHEMATICAL COMPETENCE TEST FOR JUNIOR HIGH SCHOOLS IN TAIWAN

By

CHIEN-MING CHENG

This study investigates the relationship between gender group membership and performance on test items using four differential item functioning procedures – Area Measure, Likelihood Ratio test, Mantel-Haenszel, and SIBTEST. The analysis of DIF is important because of concern that the basic competence test for junior high schools be fair and impartial for every student. In this study, the presence of DIF for gender groups is investigated for this new system of testing. The results of this study are the identification of items that show evidence of DIF, a determination of which methods are the most accurate for detecting DIF, and an investigation of the possible causes of DIF.

Both real and simulation data are analyzed to compare the four DIF methods. From the results, synthesis and discussion of effect size, frequency, consistency, and Type I error rate, of the four methods, SIBTEST was deemed the most appropriate to detect DIF items for the basic mathematical competence test for junior high schools in Taiwan.

ACKNOWLEGEMENTS

My special appreciations go to my dear wife Suh-Ling and mother in law for their willingness to support my studies and my work, especially on this dissertation. They have been the people who have sacrificed the most as I have worked on this dissertation.

My sincere acknowledgement goes to Dr Mark Reckase, my dissertation advisor. His insight and friendship have been invaluable for me to finish this dissertation.

My sincere gratitude is also extended to Dr. Glenda Lappan, Dr. Edward Wolfe, Dr. Frederick Oswald, and Dr. Betsy Becker for spending time reading this dissertation and giving me valuable suggestions and comments on my work.

Moreover, I want to express my sincere appreciation to my dear friends and classmates, De-Ping, Yan-Ling, and Shu-Chuan for their help during the journey of my dissertation writing. I would like to express my deepest gratitude to the Committee of the Basic Competence Test for Junior High Schools for their providing data.

Finally, I would like to deliver my deepest thanks to my dear parents, families, and colleagues for their consistent support, love, and concerns during my graduate study. With these people's assistance, the accomplishment of this dissertation has become a sweet memory of my academic learning and personal growth.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Research Background

For the past 40 years, high schools in Taiwan have administered an entrance examination to select students best suited to enter particular schools. In 2001 the entrance exam was replaced by the Basic Competence Test for Junior High Schools. This was the first time the Basic Competence Test for Junior High Schools was put into practice. The test is critical to students' future development, because students will use the test results to apply to senior high schools. There are concerns about the quality of the test item, including their validity as achievement indicators and their fairness for the different demographic groups of the student population. An important question is whether performance on the test items is related to the demographic characteristics of the student population. The issue is significant for the junior high schools, students, and their parents. The demographic characteristics of interest include race, sex, and age. In Taiwan, the proportion of minority population is about 5%. The students who participate in the basic competence test are almost the same age. In addition, the difference in mathematics performance between genders is always a controversy issue in education (Noddings, 1992). Society is also concerned about the issue of sex equality. Therefore, the purpose of this study is to check the relationship between demographic group membership and performance on test items. Gender is the specific demographic feature that is considered, and performance differences are investigated in mathematics for junior high school students in Taiwan.

Tests have existed in China since ancient times. Their purpose was to choose the

elite from among the people. Even today, similar tests are used throughout the world. Why would such a system of testing endure so long? Because the test is believed to be an impartial, fair, and open system. Many problems, however, have arisen in education development over the past number of years because teachers' instruction is influenced by material included on the test; teachers' teaching and behavior often follow the contents of a test. The contents of tests were institutionalized and dominated teaching activities (Chen, 2003; Liu, 2004; Shen, 2003; Ye, 2003). That is, teacher's instruction is influenced by the content of test.

In the past a student who wanted to enter senior high school or vocational high school had to pass the Senior High School Entrance Examination. The textbook was the bible of learning. Students spent a great deal of time memorizing the contents of a textbook. However, these materials were far from real-life experience, and if the learning experience cannot link with real-life experience, learning activities lack significance. Knowledge learning is only information accumulation and does not cultivate problem-solving abilities. Students who learn material by rote do not know how to apply their knowledge to real-life experiences, affecting their level of creativity (Wu & Xie, 2001).

Due to the above shortcomings of the traditional test system, the Ministry of Education has begun to concentrate on education reform. The entrance examination is a very important part of this reform. The senior high school entrance examination has been administered throughout schools in Taiwan for about fifty years. Since 2001 it has been replaced by the Basic Competence Test for Junior High Schools. Parents, teachers, and other educators are very concerned about this reform. The results of tests are related to

many significant matters such as individual prospects, fame for the family, the ranking of schools, and the honor of teachers; the fairness of the test is a universal concern.

The Basic Competence Test for Junior High Schools is a very important breakthrough in education because it adopts item response theory (IRT). Many topics in IRT have been investigated, such as building item banks, calibration of items with an examinee's ability, setting standards, equating, and examination of differential item functioning (DIF). Differential item functioning (DIF) may be defined as the performance difference on a particular test question between individuals of comparable ability or performance who belong to different groups (Dorans & Holland, 1993). The DIF topic is important because of concern that the Basic Competence Test for Junior High Schools be fair and impartial for every student. This is a very important issue. In this study the presence of DIF for gender groups is investigated for this new system of testing. The results of this study are the identification of items that show evidence of DIF and that are judged to be due to bias, a determination of which methods are the most accurate for detecting DIF, and an investigation of the possible causes of DIF. The results can be important reference resources for parents, teachers, other educators, and the institution of Basic Competence Test for Junior High Schools.

1.2 Purpose of the Study

Most large testing programs have a formal review in order to ensure that tests are fair for all examinees. Formal review is part of the test development process where items are inspected by content experts for text that might be inappropriate or unfair to relevant test subgroups, including female examinees, minority group examinees, and disabled

examinees. But reviews are conducted before the tests are administered. Statistical measures of DIF can also help test designers identify items that may be biased against examinees. Typically, DIF analyses are conducted after the tests are administered using large samples of examinee data.

Many psychometric experts have tried to give a clear and concrete definition for item bias. Cleary and Hilton (1968) gave the definition of item bias from the analysis of variance viewpoint. They reasoned that item bias was an interaction between an item and group membership. Angoff and Ford (1973) defined item bias as the difference in difficulty parameters between two groups. Today researchers distinguish DIF from "item bias." Holland and Thayer (1988) used DIF (differential item functioning) or DIP (differential item performance) to describe the performance difference between two comparable ability groups.

Group differences in test performance should not be interpreted automatically as evidence of bias because score differences might be valid reflections of group differences in knowledge and experience, so the concept of relative difficulty was devised. Camilli and Shepard (1994) refer to the raw or uninterpreted relative difficulty as differential item functioning or DIF. DIF statistics would be used to identify all items that function differently for different groups; then, after logical analysis to determine why the items seem to be relatively more difficult, a subset of DIF items might be identified as "biased" and presumably then eliminated from the test. The item is called biased if it is determined through logical analysis to be the result of factors unrelated to the construct that is the target for the test. That is, bias is operationalized as relative item difficulty that exaggerated or distorted group differences. Bias means that some dimension other than

the target of the test affects performance, and the groups differ on that dimension. DIF is just a statistical measurement and its presence does not necessarily mean that an item should be deleted (Angoff, 1993). DIF may indicate that there is some curriculum or instruction difference that results in differences in performance rather than some biasing factor (Harris & Carlton, 1993; Lane, Wang, & Magone, 1996). In this study the researcher will adopt the perspective of Camilli and Shepard (1994) to distinguish between DIF and bias.

The detection of DIF has been included in item analysis procedures by test practitioners around the world. The objective is to identify items that show DIF and eliminate those that likely represent item bias in order to improve fairness for examinees of different backgrounds. Data and experience from analyzing DIF can contribute to future reference materials to improve the quality of items. If there are no DIF items in the test, test impartiality for examinees will increase as will the validity evidence for the test as an instrument to measure latent abilities of examinees.

Research on item bias can be traced back to 1905. A. Binet and T. Simon administered the original version of the intelligence test. They found there was significant difference between children of working class backgrounds compared to those from middle class families (Tai, 1994). From that time, cultural bias has become a topic of research.

In the late 1960s, American society experienced the rise of women's liberation and the civil rights movement. Since that time, most of school admissions, diploma granting, and employment and personnel selection have depended on test results to achieve equality and fairness for groups (Tai, 1994). American educators are especially interested

in differences in test results by gender and race. For instance, Jensen's (1968) research found the difference between whites and blacks' intelligence is about one standard deviation. Williams (1971) also believed traditional education and professional tests were advantageous to middle class whites. Freedle and Kostin (1988) researched whether the items in a test exhibited differential item functioning for different groups. Their findings established that some items were advantageous to whites. The problems arising from item bias are still a concern worldwide. Walstad and Robson (1997) used the DIF method to detect male-female differences on multiple-choice tests in economics and found DIF exhibited on particular items. Maller (2001) found that in the Wechsler Intelligence Scale for Children — Third Edition, of the 151items studied, 52 were found to function differently for boys and girls. Ryan and Chiu (2001) found that the gender DIF for the word problem category was an issue. Gibson and Harvey (2003) investigated the Armed Services Vocational Aptitude Battery and found DIF was commonplace at the individual item level and could be found to favor each sub-groups in some cases. Lane, Wang, and Magone (1996) researched gendered-related DIF on a middle-school mathematics performance assessment and found four tasks favored female students and two tasks favored male students with respect to uniform DIF. In the past decades in Taiwan, Wu, Houng, Shu, Chen, and Chen (1994), Tai (1994), Chien, Liu, Sheu, Kuo, and Yin (1995), Chen (1996), Huang (1999), and Huang and Li (1999) used data to study DIF. Chien, Liu, Sheu, Kuo, and Yin (1995) suggested that test designers increase the analysis of DIF when developing a test. Thus, it is imperative to study whether the ability test items are fair with regard to differences of gender, race, geographic location, and socio-economic status in Taiwan.

A mathematics test is administered to measure an examinee's mathematics ability. It would be disadvantageous to examinees who have low reading ability and socio-economic status if the influential factors on mathematics test scores were to include reading ability and cultural differences as well as mathematics ability of examinees. Similarly, it would not be appropriate if the test scores that measure examinees' mathematics ability were also an index of attitude toward mathematics and discriminated against examinees with negative attitudes.

1.3 Research Motivation and Goals

Many different methods for detecting DIF are available. The earliest method of detecting DIF is the "transformed item difficulty method" provided by Angoff (1972). The method compares the correct answer probability for two groups. Because the transformed item difficulty method only considers item difficulty, it cannot detect the relationship between group membership and discrimination (Merz & Grossen, 1979).

The χ^2 test procedure was developed by Scheuneman (1979). The χ^2 test procedure usually separates examinees into many subgroups by the total test scores and assumes that the ability of examinees in the subgroups is approximately equal. Although research supports the χ^2 test procedure as superior to the transformed item difficulty method, it still has many disadvantages. For instance, the arbitrariness involved in the designation of intervals for the total score indicates that results may vary with the characteristics of the data, that the index of DIF could be easily confounded by the sample size, and that valuable information may be lost by treating a continuous variable as a categorical variable (Ironson, 1982).

The item characteristic curve method is based on item response theory. Three of the ways to use IRT to detect DIF are (Camilli & Shepard, 1994; Ironson, 1983):

1. Comparison of the measured difference between two item characteristic curves for different groups — the method calculates the area between the two item characteristic curves for different groups as the index of DIF. There are three kinds of such DIF measure: first is the area measure; second is the squared differences measure; and the third is a weighted area and squared differences measure.
2. Comparison of the vectors of item parameters: the essential point of this method is to test the equality of item parameters.
3. Comparison of the goodness of fit between the item response model and data: the method for detecting DIF is to compare the item fit statistic across the groups of interest, differences in fit may indicate DIF.

Many researchers have found the transformed item difficulty method inferior to the item characteristic curve method and χ^2 test procedure (Ruder, Getson, & Knight, 1980; Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1985; Subkoviak, Mack, Ironson, & Craig, 1984). And many researchers also have found that detecting DIF based on item response theory is superior to the transformed item difficulty method and the χ^2 test procedure (Ironson, 1977; Ironson & Subkoviak, 1979; Merz & Grossen, 1979; Runder & Convey, 1978; Runder, Getson, & Knight, 1980; Shepard, Camilli & Averill, 1981; Shepard, Camilli & Williams, 1985; Subkoviak, Mack, Ironson, & Craig, 1984). Therefore, this study will include the signed area measure between two item characteristic curves and the comparison of the goodness of fit between the item response model and data to detect DIF.

Mantel-Haenszel (M-H) is another method used to detect DIF. The earliest person to present M-H is Cochran (1954), and later Mantel and Haenszel (1959) and Mantel (1963) expanded the method. M-H possesses a logical and concise concept that is the extension and application of the χ^2 test procedure. Because the distribution of χ^2 of M-H is known in advance and the probability of the value of a test statistic or one more extreme can be computed, the M-H method is used as a significance test. Many researchers prefer to use M-H to detect DIF. The study will also adopt the M-H method because it is an economical, convenient, and easy to calculate measure.

SIBTEST (Simultaneous Item Bias Test) developed by Shealy and Stout (1993a, 1993b) is the latest and a nonparametric method to detect DIF. It has comparative performance to the M-H method. It has good Type I error rates across varying levels of item discrimination and sample sizes when group mean abilities differ (Chang, Mazzeo, & Roussos). Therefore, SIBTEST will be adopted in this study.

DIF means the performance difference on a particular test question between individuals of comparable ability who belong to different groups. The appearance of DIF indicates that an item may be affected by content unrelated to the construct that is the focus of the assessment. Such effects may have an unfavorable influence on the validity of the item. The entrance examination for high school was administered for more than 40 years in Taiwan. It was the only available way to choose appropriate students from junior high school for senior high school. The reliability and validity of the entrance examination had never been evaluated. However, the test was still accepted and trusted by the public. But from a psychometric viewpoint, there was concern about the fairness and rationality of the entrance examination when a single test result was the only measure.

No pilot test or measure of the test's validity and appropriateness for the examinees was available because in order to keep all test items confidential they were designed in an imperial examination hall. A concern for test quality has led the Ministry of Education to follow the current trends in measurement procedures. Beginning in 2001 the Ministry of Education, in an attempt to reform the entrance examination program, introduced the Basic Competence Test for Junior High Schools to replace the traditional entrance test. This was a milestone for the examination system of Taiwan because in the past the entrance examination was administered once a year and adopted the theory of classical test theory.

Today the Basic Competence Test for Junior high Schools is administered twice a year and adopts item response theory. There is as yet no research to determine the fairness of the mathematics items on the Basic Competence Test for Junior High Schools. Therefore, this study's research will use the M-H, area measure, likelihood ratio test, and SIBTEST to investigate DIF by gender for the first Basic Competence Test for Junior High Schools in 2001. The goal will be to determine the consistency of all the methods employed, power and accuracy, and characteristics of any detected item bias. In addition, the Type I error rate of all methods employed in this study will be investigated.

Based on the above research motivation, the research questions are as follows:

1. Is there significant DIF in test items from the basic mathematics competence test based on gender groups?
2. How consistent are results for the different ways of detecting DIF?
3. Which is the best procedure for detecting DIF?
4. What is the Type I error rate of detecting DIF for the different methods?

5. Which of the items detected as showing significant DIF are considered to be biased after logical analysis?
6. What should be done to determine if the DIF is due to instructional differences or some biasing feature of the items with the results after identifying DIF items?

1.4 Definitions

1.4.1 Item response theory

Item response theory is also called latent trait theory. It is a kind of mathematical model. The mathematical model is a mathematical function used to describe the conditional probability of a response given the level of the latent ability (Thissen & Steinberg, 1986). There are many item response models that are developed from this theory, but the one-parameter logistic model, two-parameter logistic model, three-parameter logistic model, and four-parameter logistic model are the basic models. The format used in the basic mathematical achievement test is multiple-choice. Multiple-choice items have some non-zero probability of responding correctly even when the examinee has very little knowledge. The instrument was a 32-item mathematics test with four choices for each item. Examinees were asked to select an option that is the best or exact answer. For multiple-choice items, the three-parameter logistic model is the most suitable (Yu, 1993). Thus the study will adopt the three-parameter logistic model. The three item parameters are a - discrimination, b - difficulty, and c – pseudo guessing.

1.4.2 Item characteristic curve method

An item characteristic curve (ICC) is the graph of an item characteristic function, a mathematical model used to describe the relationship between examinee item performance and trait. The shape of ICC is an S-shaped curve that describes the relationship between the probability of correct response to an item and the ability scale. An ICC can be used to predict the probability of answering an item correctly from the examinee's ability level. The ICC can show the item's difficulty, discrimination, and guessing parameters. The item characteristic curve method is a kind of DIF detection method. That is, the item characteristic curve method is used to compare the item characteristic curves with different groups of interests. The difference between the ICCs for different groups is the index of DIF.

1.4.3 Mantel-Haenszel

Mantel-Haenszel is a method used to detect the item performance difference for different groups. In the beginning, the estimation of a common odds ratio α or log α was the M-H statistic used as the index of DIF. But this method is limited to the 2x2 contingency table. It is not practical because M-H is only used to detect the difference of the two groups. Thus Landis, Heyman, and Koch (1978) improved upon the Mantel-Haenszel method with the Cochran-Mantel-Haenszel (CMH) in order to use it with multi-level data. In this study, the researcher will adopt SAS software to run the CMH statistical analysis to compute the general association statistic. If the statistic is significant, it means the item exhibits DIF. Then the average of the DIF item for each group will be determined to see whether the item is advantageous to any group.

1.4.4 Likelihood ratio test: Model comparison measures (Neyman & Pearson, 1928)

The likelihood ratio test (LR) (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993) is used to compare two different IRT models in order to test whether the IRFs of the two groups are the same. Thissen et al. (1988) noted that this approach is preferable for theoretical reasons. One of the models is called the compact model, and the other is called the augmented model. The augmented model includes all the parameters of the compact model and additional parameters. The LR tests whether or not the additional parameter in an augmented model is significantly different from 0.

1.4.5 SIBTEST

SIBTEST (Shealy & Stout, 1993a, 1993b) is a non-parametric DIF detection method. SIBTEST is similar to the standardization method in concept. But SIBTEST has some unique characteristics. SIBTEST has a statistical significance test. The matching variable in SIBTEST is a latent score rather than an observed score. Although SIBTEST was developed from an IRT framework, it does not require item calibration. However, the method assumes that the abilities of examinees who have the same score are equal. The DIF estimate from SIBTEST uses the number correct score as the matching variable for detecting subgroup differences. It is noteworthy that the scores of matching subsets do not include the studied item score. That is the examinee is assigned to subgroups based on a total scores that does not include the studied item. This is obviously different from the M-H method.

1.4.6 Basic Competence Test for Junior High Schools in Taiwan

The Basic Competence Test for Junior High Schools is a result of education reform policy created by the Ministry of Education in Taiwan. The purpose of the test is to measure the basic abilities of students in junior high schools and how much they have learned by the time they complete junior high school. The content of the test covers the basic, important, and core knowledge and ability of students. "Basic competence" means the comprehensive, basic, and important ability of the learner who was systematically instructed for the duration of the three-year junior high school program. The score they achieve is used to help students decide which school to attend, senior high school or vocational senior high school. The test is designed by the Institution of Basic Competence Test for Junior High Schools.

CHAPTER 2

LITERATURE REVIEW

Historian DuBois (1970) indicated that ancient China had the idea of ability measurement since 2,200 B.C. But China didn't conduct scientific research on the measurements. Psychometrics, which was developed in the West, was not practiced in China. Only later did researchers in China begin to pay attention to psychometric research.

Psychometrics is a division of science that studies psychological testing and assessment. The research area includes quantitative psychology, individual difference, and mental test theories (Cohen, Montague, Nathanson, & Swerdlik, 1988). In the late 19th century, scientific psychology was born, and psychologists were interested in quantifying psychological traits. As a result Binet-Simon developed the first intelligence test in 1905.

Test theory can be divided into classical and modern test theory, depending on how scores are analyzed and interpreted. These theories use different mathematical models.

Classical test theory (CTT) is still regarded as practical test theory. Many tests are still built using relationships in data based on classical test theory. CTT hypothesizes that an examinee has a true score and observed scores. The true score is the expected value of the observed scores obtained over an infinite number of independent repeated testing using the same test (Croker & Algina, 1986). The observed scores are computed from examinees' responses to items and true scores are hypothetical values that can be estimated. CTT attempts to evaluate the association between observed scores and true scores. CTT is built on the true score model.

2.1 Item Response Theory

IRT was developed to overcome some of the shortcomings of classical test theory. The relationship between an examinee's item performance and ability can be described by a monotonically increasing function, called an item characteristic function. Different item response formats correspond to different item response models. An item response model is composed of a mathematical formulation and basic assumptions. Due to the robustness of IRT models to violations of assumptions, IRT seems to be preferred to CTT and it is respected by current psychometric researchers. Basic concepts, basic assumptions, item response models, advantages and disadvantages, and applications of IRT are described below:

2.1.1 Basic concepts of IRT

The basic concepts of IRT are as follows (Hambleton, 1989; Hambleton & Swaminathan, 1985; Yu, 1997):

1. The performance of an examinee on a test item can be predicted or explained by a single factor. Because the factor cannot be observed, it is called a latent trait or ability which is the desired measurement objective.
2. The relationship between performance and ability can be expressed by a monotonically increasing mathematical function called an Item Characteristic Function (ICF). The ICF provides the right response probability at each examinee ability level. A graph of the ICF is called Item Characteristic Curve (ICC).
3. Different types of response data have different item response measurement models and ICCs because of different requirements and assumptions.

4. Every ICC includes one or more parameters to describe item characteristics and an examinee's ability. Therefore, the shapes of ICC are different if the number of parameters is different. The most often seen shape is a non-linear regression line.

2.1.2 Basic assumptions

Because the relationship between an examinee's item performance and ability can be expressed by a mathematical function, an item response model is also called a mathematical model (Hambleton, 1989). Item response models have common basic assumptions. Support for the assumptions should be established before an IRT model is applied to test data. The basic assumptions of IRT are as follows:

1. Unidimensionality: All of the items in the test measure the same ability or latent trait. The items on a test should measure the same ability. The meaning of unidimensionality is simple, but it is not easy to find data that meet the unidimensionality assumption. Hambleton and Swaminathan (1985) thought data with a dominant first factor would meet the requirements for unidimensionality. In fact, many testing situations require multiple abilities. That is, they require an assumption of multidimensionality (Bock & Aitkin, 1981; Hambleton, 1989; Reckase, 1985). Multidimensional models are still in development. Unidimensionality is still the principal basic assumption.
2. Local independence: The examinee's response to each item is independent in a prabilistic sense from responses to other items. That is, the response to one item does not influence responses to any other items regardless whether an examinee's response is correct or incorrect. Usually, local independence will be true if the

assumption of unidimensionality can be supported (Lord, 1980).

3. Non-speed test: An examinee's bad test performance is due to a lack of ability rather than to a time constraints.

If the assumptions of the IRT models are met, they have the property of invariance, which includes the invariance of item and examinee parameters. The invariance of an examinee's ability estimate means that it doesn't change when measuring with different test items, except due to measurement error, once the scale for the parameters has been set to a common metric. The invariance of item parameter estimates is that they do not change with subgroups (e.g., sex, race, or area) except for measurement error. The invariance property provides the basic theory for test linking and equating.

IRT provides the standard error of estimate for every examinee's ability. The inverse of the square of the measurement standard error is defined as the information for the ability estimate. Information is an index for evaluating the accuracy of the ability estimate. Information is a new measurement concept. It aids in designing items, building item banks, implementing computerized adaptive testing, and test equating. Information and invariance are the main factors that make a distinction between IRT and CTT (Yu, 1997).

2.1.3 Item response model

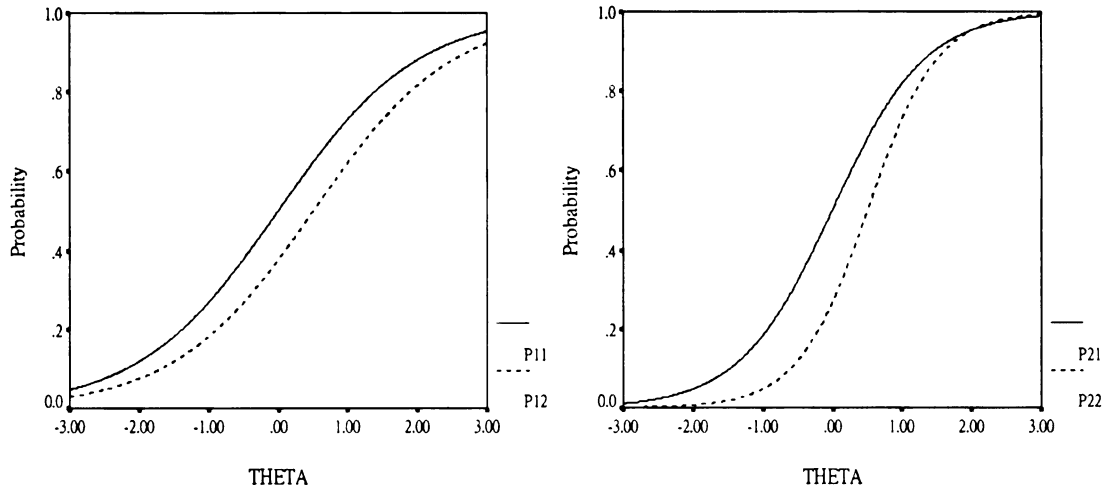
Item response models can be classified by different scoring methods. In general, there are three types of scoring: dichotomous, multicategory, and continuous. The models can also be distinguished by the mathematical form of the characteristic curve. Using the three methods of measurement scoring, Table 2.1 presents IRT models (Hambleton & Swaminathan, 1985):

Table 2.1 *Theoretical model of Item Response Theory*

Data property	Theoretical model	Reference
Dichotomous	Latent Linear	Lazarsfeld and Henry (1968)
	Perfect Scale	Guttman(1944)
	Latent Distance	Lazarsfeld and Henry (1968)
	1-, 2-, and 3-Parameter Normal Ogive Model	Lord (1952)
	1-, 2-, and 3-Parameter Logistic Model	Birnbaum(1957, 1958a, 1958b, 1968); Lord and Novick (1968); Lord (1980); Rasch (1960); Wright and Stone (1979)

The shape of the item characteristic curve will be different if the item response model is different. The following Figure 2.1 is three variants of ICCs for the IRT models.

(a) 1-parameter (p11:b=0, p12:b=.5) , (b) 2-parameter (p21:b=0, a=1.5; p22:b=.5, a=2)



(c) 3-parameter (p31:b=0, a=1.5, c=.2; p32:b=.5, a=2, c=.3)

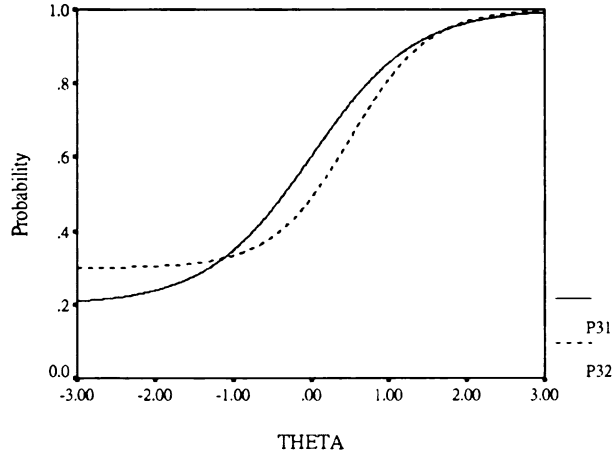


Figure 2.1 *Three ICCs of IRT*

The S shape ICCs of Graph (a), (b), and (c) are respectively for the one-, two-, and three-parameter logistic model.

Birnbaum presented the three-parameter logistic model in 1968. Its item characteristic function is as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

$P_i(\theta)$ means the probability of the ability θ to answer correctly the item i. D is a constant and equal to 1.7. b_i is the ith item difficulty. The larger the b_i value, the more

difficult will be the item. a_i is the ith item discrimination. The greater the slope is, the greater the discrimination. c_i is the ith lower asymptote parameter. The three-parameter logistic model considers simultaneously difficulty, discrimination, and lower asymptote parameters. Therefore, it is appropriate for the multiple-choice format test. Graph e is the item characteristic curve of the three-parameter logistic model. The guessing parameter can improve the fit to multiple-choice test data.

2.1.4 The application of IRT on Detecting DIF

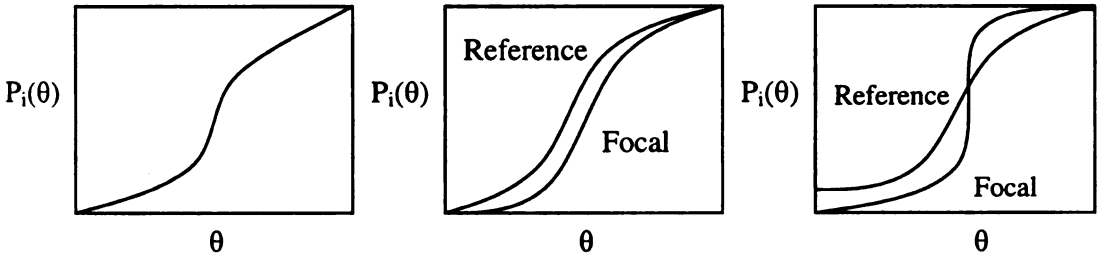
IRT can offer effective application for detecting DIF. Its applications are described as follows:

DIF is of concern to test users and developers. DIF indicates whether an item functions differently for different subgroups of a population. This is a minimal requirement for determining if a test item is potentially unfair to a certain group. The methods for detecting DIF based on IRT are the Lord χ^2 test, the ICC area measure and the likelihood ratio test. Among these methods, the ICC area measure and likelihood ratio test are better supported by empirical research. This study uses the ICC area measure and the likelihood ratio test to analyze the data from the basic mathematical competence test given in 2001 to detect DIF in the test.

2.2 Types of DIF

From the IRT viewpoint, when the IRFs of two groups are identical, that is, the ICCs of two groups are the same, it indicates the item does not exhibit DIF, as in graph 1 of Figure 2.2. If the ICCs of two groups are different, it shows that there is differential

functioning. That is, the item exhibits DIF. Mellenberg(1982) differentiated DIF into uniform and non-uniform types. Uniform indicates that the answer performance of the focal group (or the reference group) consistently maintains a relative advantage. In other words, the ICCs of two groups are different and do not intersect at any ability level, as in graph 2 of Figure 2.2. The item is advantageous to the reference group in graph 2 of Figure 2.2. Non-uniform DIF indicates that the ICCs of two groups are different and intersect as in graph 3 of Figure 2.2. The graph indicates that the item handicaps the focal group at low and intermediate ability levels, and the reference group at high ability levels.



Graph 1: No DIF Graph 2: Uniform DIF Graph 3: Non-uniform DIF

Figure 2.2 *Types of DIF*

2.3 DIF and Impact

To distinguish between DIF and impact is important. DIF means differences in item functioning after groups have been matched with respect to the ability that the item purportedly measures. Item impact can be described as any group discrepancy in item performance that reflects actual knowledge and experience differences on the construct of interest (Clauser & Mazor, 1998). DIF is different from a performance difference caused by differences in the ability level of the two groups. A difference in performance between two intact groups is called impact (Lu, 1999). Dorans and Holland (1993) used Simpson's Paradox to make a distinction between DIF and impact. Table 2.2 shows the performance

of group A and Group B students. Suppose 1440 students answered correctly in group A of 2400 students. The answer right proportion is 60 %. 1200 students answered correctly in group B of 2400 students. The answer-right proportion is 50 %. The difference in proportion of correct responses between the two groups is 10 %. The answer right proportion of group B is 0.1 higher than group A on the low, middle, and high ability level. In fact, the item is advantageous to group B. The example shows the importance of the ability grouping for detecting DIF.

Table 2.2 *The performance of two groups on an item*

	Group A			Group B		
Group	Number	Right #	Proportion	Number	Right #	Proportion
Low	400	40	0.1	1000	200	0.2
Middle	1000	500	0.5	1000	600	0.6
High	1000	900	0.9	400	400	1.0
	2400	1440	0.6	2400	1200	0.5

2.4 DIF Detecting Methods

2.4.1 IRT methods

IRT DIF procedures primarily detect if there are differences between item parameters for the reference group and the focal group. That is, are the two item response functions the same? The most commonly used IRT DIF procedures are Lord's χ^2 test (Lord, 1980), a measure of the area between ICCs (Camilli & Shepard,1994; Lord, 1980; Millsap & Everson, 1993), and the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988, 1993).

1. Lord χ^2 test method

Lord (1980) provided a statistical procedure to test whether the item parameters of two groups are different. In applying Lord's χ^2 to examine DIF, the first step is to employ IRT software, like the BILOG program, to calibrate the item response data from the focal and reference groups. The item parameters estimated always have their individual scale origin and unit because of IRT scale indeterminacy. Therefore, a direct comparison of the item parameter estimates cannot be made. Before making a comparison, item parameter estimates must be transformed to the same scale through a linking procedure. Then DIF analyses can proceed.

Lord suggested fixing the *c* parameter from the three-parameter model or using the two-parameter model. The null hypothesis for Lord's χ^2 test is: $a_F=a_R$, $b_F=b_R$. The difference between the two item parameters estimate can be expressed as the following vector,

$$V'=[\, a_F-a_R,\ b_F-b_R] \,.$$

The formula for Lord's χ^2 test is as follows,

$$\chi^2= V'S^{-1}V \,.$$

where S is the variance-covariance matrix of the item parameter estimate differences. By the large sample theorem, Lord's χ^2 estimates follow the χ^2 distribution with 2 degrees of freedom under the null hypothesis. If the χ^2 estimate reaches a significant level, the null hypothesis is rejected indicating that the item has DIF. If the variance and covariance matrices cannot be estimated accurately, errors may result in subsequent identification of DIF items. Under the above conditions, use of Lord's χ^2 would not be justified (Lane, Stone, Ankenmann, & Liu, 1995).

2. IRT measure of DIF (area measure)

Another method for detecting DIF is to calculate the area between the IRFs or ICCs for the two groups. The larger this area, the more serious is the DIF. When applying the area measure to detect DIF, the item parameters of the two groups have to be calibrated, then linked to the same scale. If $P_R(\theta)$ and $P_F(\theta)$ respectively represent the IRFs of the reference and the focal group, the area between two ICCs can be defined as:

$$A=\int_S (P_R(\theta) - P_F(\theta)) \text{ , where S indicates the range of ability } \theta.$$

If the area measure retains a positive or negative sign after calculating, it is called the signed area measure. If the value is positive or 0, it is called unsigned area measure. That is, unsigned area measure is the absolute value of the area between two ICCs.
The formula for the signed area measure is,

$$SA=\int_S [P_R(\theta) - P_F(\theta)]d\theta.$$

(Rudner, 1977)

SA>0 indicates the item is advantageous to the reference group; SA<0 indicates the item is advantageous to the focal group. This measure is easily interpreted in real application. Its disadvantage is that the IRF differences at different points on the θ-scale that may be calculated would offset each other when the ICCs of two groups intersect. Then the real value of DIF would be underestimated.
The formula of unsigned area measure is:

$$UA=\int_S |P_R(\theta) - P_F(\theta)| d\theta.$$

(Raju, 1988, 1990)

$$\text{or UA}=\sqrt{\int_{S}[P_R(\theta)-P_F(\theta)]^2 d\theta}\,.$$

(Camilli and Shepard, 1994)

The unsigned area measure is usually bigger than the signed area measure when an item presents non-uniform DIF.

3. Likelihood ratio test: Model comparison measures (Neyman & Pearson, 1928)

The statistic for the LR test can be expressed as follows,

G^2_j=-2 log [Likelihood (Compact model)/ Likelihood (Augmented model)]

The above Likelihood(·) expresses the maximum likelihood estimates of the parameter estimates in the compact or the augmented model. j refers to the parameter number difference between the augmented and the compact models. The distribution of G^2 is χ^2 distribution with j degrees of freedom under the null hypothesis. In applying the LR test to detect DIF, all the item parameters are assumed equal when estimating the parameters of the compact model. In the augmented model, all the item parameters except the studied item are assumed equal. The DIF test is employed to compare the maximum likelihood function of the two models in order to check whether there is a significant difference.

Following the terminology of Judd and McClelland (1989) and its application to IRT by Thissen et al. (1993), the model comparison approach is implemented to compare the relative fit of the two models. The first is called the compact model (C) and the second, the augmented model (A). Model (A) is an elaboration of model (C). The model (A) has all the parameters of model (C) plus a set of additional parameters. In this study, there are 3 parameters because there are one more additional item in model (A). The goal

of the comparison is to determine if the additional parameters in model (A) are necessary.

Suppose the null hypothesis is $H_0 : \Gamma = Set_C$ (where Set_C contains N parameters), and the alternative hypothesis is $H_a : \Gamma = Set_A$ (where Set_A contains N+M parameters), where Γ stands for the true set of parameters. Model C, the compact model, has M fewer parameters than Model A. The likelihood ratio (LR) of interest for the two models is

$$LR = \frac{L^*(Model - C)}{L^*(Model - A)}$$

$$\chi^2(M) \approx -2\ln(LR) = [-2\ln L^*(Model - C)] - [-2\ln L^*(Model - A)]$$

$G(C) = -2\ln L^*(Model - C)$ and $G(A) = -2\ln L^*(Model - A)$ are defined.

Then $\chi^2 \approx -2\ln(LR) = G(C) - G(A)$

Utilizing the Camilli and Shepard (1994) steps for estimating DIF, the Model Comparison Approach is as follows:

1. With a 3PL IRT model, estimate item parameters and obtain χ^2 goodness-of-fit statistic G(1) for a 32-item test.
2. Choose Item 1 to study.
3. Create two items for item 1:

   Code Item 1R as answered by the Reference (male) group and not reached by the Focal (female) group. Code Item 1F is answered by the Focal group and not reached by the Reference group.

   Original coding for Step 1

Item	1	2	3	...	32
Response variable	u_1	u_2	u_3	...	u_{32}

   Recoding for estimation run for item 1 in step 4

	2	3	4	...	33	34
Reference	u_2	u_3	u_4	...	u_{1R}	–
Focal	u_2	u_3	u_4	...	–	u_{1F}

"–" means not reached

4. Re-estimate parameters and obtain χ^2 transformation of the likelihood ratio G(2) for the 34-item test.
5. Compute G(1) – G(2). This is approximately χ^2 with 3 degrees of freedom.
6. If G(1) – G(2) exceeds the critical value, flag Item 1 as showing statistically significant DIF.
7. Repeat Steps 2-6 with all the other items.

In the above three types of IRT DIF procedures, Lord's χ^2 test and LR test are significance tests. The results show only the information whether or not there is a statistical difference between the two IRFs of the two groups, but it cannot specify the magnitude of difference. Raju (1990) provided two sampling distributions for the mean and standard deviations of the infinite interval area measure. He also provided the statistic for the signed area measure – Z(EST) and the statistic for unsigned area measure – Z(H). Both of them have a *z* distribution. In addition, the item parameters of the Lord χ^2 test and the ICC area measure in practical application need to be linked before comparison. However, for the LR test the item parameters are simultaneously estimated and there is no requirement for linking of parameter scales. Therefore, the LR test is the best of the three methods. The LR test is gradually becoming more accepted for detecting DIF. Kim and Cohen (1995) compared the performance of three methods and found the consistency of results very high for those DIF test procedures. A limitation of

IRT based procedures is that the data have to correspond to the unidimensionality assumption of the models. The above methods also require large samples for accurate parameter estimation if the two- or three-parameter model is used (Clauser & Mazor, 1998).

2.4.2 Non-IRT methods

Non-IRT procedures include the Mantel-Haenszel method (Mantel & Haenszel, 1959; Holland & Thayer, 1988), the standardization method (Dorans & Kulick, 1986), the logistic regression analysis method (Swaminathan & Rogers, 1990), and the SIBTEST procedure (Shealy & Stout, 1993a).

1. Mantel-Haenszel method

M-H is a variant of the contingency tables analysis method. The M-H method usually uses the total scores of the test as a matching variable for the reference and the focal groups. For each of the k score levels, a 2x2 contingency table can be produced (as shown in Table 2.3). Every item has a 2×2×k contingency table, at least in theory. Sometimes score groups have to be combined because of low frequencies.

Table 2.3 *2×2 contingency table for total score k*

	Item Score		
Group	1	0	Total
Reference	A_k	B_k	N_{Rk}
Focal	C_k	D_k	N_{Fk}
Total	M_{1k}	M_{0k}	T_k

T_k : frequency of score k. N_{Rk}: the number of persons in the reference group. N_{Fk} : the number of persons in the focal group. M_{1k} : the number answering the item correctly. M_{0k} : the number answering the item incorrectly.

The null hypothesis of M-H method is: The value of common odds-ratio parameter

α_{MH} of the reference and the focal groups is equal to 1. The estimate of α_{MH} is as follows:

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k D_k / T_k}{\sum_k B_k C_k / T_k}$$

If the value of α_{MH} is larger than 1, the reference group more easily answers correctly. If the value of α_{MH} is smaller than 1, the focal group more easily answers correctly.

Mantel and Haenszel (1959) provided a χ^2 statistic to test the hypothesis of α_{MH}=1.0. The formula is:

$$MH\chi^2 = \frac{\left(\left|\sum_k A_k - \sum_k E(A_k)\right| - \frac{1}{2}\right)^2}{\sum_k Var(A_k)}$$

In the formula, $E(A_k)$ and $Var(A_k)$ are respectively defined as:

$$E(A_k) = \frac{N_{Rk} M_{1k}}{T_k}, \quad Var(A_k) = \frac{N_{Rk} N_{Fk} M_{1k} M_{0k}}{T_k^2 (T_k - 1)}$$

Under the null hypothesis, χ^2_{MH} is χ^2 distribution with 1 degree of freedom. Rejecting the null hypothesis means the item response data support the existence of DIF. In real applications, α_{MH} is converted to another DIF type of measure called MH D-DIF. The conversion formula is:

$$\text{MH D-DIF} = -2.35 \ln(\alpha_{MH}) ,$$

After α_{MH} is transformed to MH D-DIF, the difficulty scale (MH D-DIF) used by ETS is used to interpret the difference between the two groups. Positive values of MH D-DIF mean items advantage the reference group. Negative values of MH D-DIF mean

the items advantage the focal group. Because the result of a significance test is easily influenced by the sample size, ETS designed a DIF critical classification system that simultaneously considers two criteria — the results of the significance test and the value of MH D-DIF. If the value of MH D-DIF is not significantly different from 0 and the absolute value of MH D-DIF is smaller than 1.0, it is classified as A type DIF. If the absolute value of MH D-DIF is bigger than 1.5 and the value of MH D-DIF is significantly larger than 1.0, it is classified as C type DIF. Items that do not fall into the A and C classifications are classified as B type DIF. Type A indicates insignificant or slight DIF. Type B means medium DIF. Type C means serious DIF (Lu, 1999). The M-H method has been shown to be effective with reasonably small samples (e.g., 200 examinees per group). The major limitation of the M-H procedure is that it is unable to detect non-uniform DIF (Narayanan & Swaminathan, 1996).

2. Standardization method

The standardization method is also used to identify DIF. The standardization method uses the proportion of correct responses (p- value) to identify differences in item difficulty between two groups. The null hypothesis for the standardization method is that the proportion of correct responses is the same for the reference and the focal groups for all values of the number-correct score. The DIF index for the standardization method is called STD P-DIF. Its value is equal to the difference in the proportion of correct responses for the two groups times the relative frequency of the focal group at each score level number. The formula is:

$$\text{STD P-DIF} = \sum_{k} W_k (P_{Fk} - P_{Rk})$$

In the formula, $W_k = n_{Fk}/n_F$, n_{Fk} is the number of correct answers at score level k; n_F is

the number in the focal group; P_{Fk} is the proportion of correct answers at score level k in the focal group. P_{Rk} is the proportion of correct answers at score level k in the reference group.

The proportion correct scale is used to describe the magnitude of DIF. The range of values for STD P-DIF is between –1 and +1. Negative values mean the item advantages the reference group. Positive values mean the item advantages the focal group. ETS also has a criterion to classify the level of DIF. If the absolute value of STD P-DIF is smaller than .05, it indicates the magnitude of DIF can be ignored. If the absolute value of STD P-DIF is between .05 and .10, it means the item needs to be checked. If the absolute value of STD P-DIF is bigger than .10, it means the item shows evidence of serious DIF and needs to be investigated carefully. Dorans and Holland (1993) provided the standard error of STD P-DIF to quantify the stability of STD P-DIF estimate. Although the ratio of STD P-DIF to its standard error can be computed, the standardization method still has no formal statistical test procedure.

3. Logistic regression analysis method

Swaminathan and Rogers (1990) applied the logistic regression analysis method to detect DIF. They considered logistic regression analysis method to be a link between contingency table and IRT methods. The difference between the logistic regression analysis method and contingency table methods (e.g., M-H) is that logistic regression analysis method considers the total scores as a continuous variable. The contingency table methods consider the total scores as an categorical variable. That is, the total scores are limited in number in contingency table methods. But the total scores represent the observed ability level. Logistic regression analysis method, on the other hand, uses total

test scores and groups to predict the probability of correct response. The basic model of logistic regression analysis method is:

$$P(u=1) = \frac{e^z}{1+e^z}$$

$$Z = \tau_0+\tau_1\theta+\tau_2G+\tau_3(\theta G).$$

In the equation, θ is the observed ability level that is represented by the test total score. G is the group that is coded as 0 and 1. τ_1 is the combined log odds ratio. Regression coefficient τ_2 corresponds to the item performance difference for the groups. τ_3 corresponds to the interaction of group and ability. The full model 1 is then

$$Z = \tau_0+\tau_1\theta+\tau_2G+\tau_3(\theta G).$$

A situation with uniform DIF would not need the interaction term and could be represented by the simpler model 2

$$Z = \tau_0+\tau_1\theta+\tau_2G.$$

Finally, there is another situation with no interaction and no DIF, only the ability term would be necessary. This would result in the simplest model 3

$$Z = \tau_0+\tau_1\theta.$$

Logistic regression analysis method uses maximum likelihood to estimate the regression coefficient of the model. The procedure for applying logistic regression analysis method to detect DIF is similar to the likelihood ratio test for IRT. Model 1 and model 2 are compared to check whether or not the τ_3 is significantly different from 0. The value is tested by using the χ^2 distribution with 1 degree of freedom. If the hypothesis that $\tau_3=0$ is rejected, it means the interaction between group and ability is significant. The item therefore has non-uniform DIF. If the hypothesis of $\tau_3=0$ is not rejected, model 2 and

model 3 are compared to check whether or not the τ_2 is significantly different from 0. It is also tested by using the χ^2 distribution with 1 degree of freedom. If the hypothesis of $\tau_2=0$ is rejected, it means the item has uniform DIF. Simulated and real data studies have shown that the logistic regression analysis method procedure produces results similar to M-H when testing for uniform DIF. It is superior to the M-H statistic for identifying nonuniform DIF (Clauser, Nungester, Mazor, & Ripkey, 1996; Rogers & Swaminathan, 1993).

4. SIBTEST procedure

The theoretical framework of SIBTEST (Huang & Li, 1999)

Shealy and Stout (1993a, 1993b) used mathematical statements to describe DIF — that is, $T_{iF}(\theta) \neq T_{iR}(\theta)$. θ indicates the ability being measured. $T_{iF}(\theta)$ and $T_{iR}(\theta)$ indicate, respectively, the marginal item response function of the focal group (F) and the reference group (R) for the i^{th} item. The function is defined as, $T_{ig}(\theta)= \int P_i(\theta,\eta) f_g(\theta,\eta) d\eta$, g: the focal group and the reference group. $P_i(\theta,\eta)$ is the probability of correct response for an examinee with ability (θ,η) for certain item i, and the probability is controlled by the two ability parameters θ and η. $f_g(\theta,\eta)$ indicates the conditional density function of the distribution of ability η when θ is known.

In applying SIBTEST to detect DIF, the first step is to separate the total set of items into two subsets. One is a valid subset that is composed of non-DIF items. Another subset is designated the suspect subtest that will be the target of the DIF test. When the SIBTEST DIF statistic is calculated, the first k items out of the total of N items are the valid subset (non-DIF item). U_i represents the item score. It is a 0 or 1 score. Then the

examinee total score is X = $\sum_{i=0}^{k} U_i$ for the valid subtest. The other items, from k+1 to N, are the suspect subtest (suspected DIF item). The examinee total score is Y = $\sum_{i=k+1}^{N} U_i$ for the suspect subtest.

The formula for the SIBTEST DIF statistic is:

$$\hat{\beta}_U = \sum_{k=0}^{K} \hat{P}_k (\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*) .$$

In the formula, P_k is the proportion of scores X=k for the focal group on the valid subtest; $(\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*)$ is the difference of the adjusted average score of both groups on the suspect subset. The scores can be based on one item or a bundle of items. If it is based on one item, it provides a DIF test. If it is based on a bundle of items, it provides a DTF test. If there is no DIF or DTF, β_U will be 0. The adjusted scores are the scores after regression correction. The primary aim of regression correction is to adjust the studied subtest scores for the two groups so that they are now estimates of the same latent ability in the case of no test bias, even if group target ability discrepancy exists. The theory and method of regression correction is in Shealy and Stout (1993a).

The statistic for testing the null hypothesis using SIBTEST to detect DIF or DTF is :

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)}$$

where $\sigma(\beta_U)$ is the estimated standard error of β_U. The SIBTEST *B* statistic is an approximation to the standard normal distribution under the null hypothesis ($H_0 : B = 0$). If the observed *B* value is larger than the 100(1-α) percentage point of the *z* distribution, the null hypothesis is rejected. The *B* statistic is designed for detecting uniform DIF.

Recently SIBTEST was extended for detecting non-uniform DIF (Li & Stout, 1996). It (SIBTEST) produces Type I errors at approximately the nominal level because of regression correction, has reasonable statistical power, and performs well with relatively small examinee samples (Narayanan & Swaminathan, 1994; Roussos & Stout, 1996).

The M-H, logistic regression analysis method, and SIBTEST have generally been found to perform satisfactorily in simulation studies. The differences among the three methods are small (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993; Shealy & Stout, 1993a). However, the power of the M-H method to detect non-uniform DIF is very weak (Narayanan & Swaminathan, 1996). This is because M-H is not designed for detecting non-uniform DIF. The standardization method is usually used to help describe the extent of DIF.

DIF procedures based on observed scores usually use the total scores as the matching variables. However, the total score is not a valid matching variable when the ability distribution is different between reference and focal groups. The fact that examinees have the same total scores does not indicate they have the same ability. Some research indicated that a Type I error rate is usually higher than average when the discrepancy of examinees' ability distribution of two groups was apparent and the discrimination of items was higher or lower (Allen & Donoghue, 1996; Lu, 1996; Lu & Dunbar, 1997; Roussos & Stout, 1996); total scores cannot effectively match the examinees' ability of the two groups. The ability discrepancy is confounded with DIF.

The major advantage of the M-H and standardization methods is that they are easily calculated. In addition, these methods have complete guidelines for interpreting DIF. The major advantage of the logistic regression method is that it conveniently detects

non-uniform DIF. The major advantage of SIBTEST is the regression correction to adjust for systematic error due to the difference in ability distributions of the two groups and to reduce the confounding of ability difference and DIF in order to effectively control Type I error (Roussos & Stout, 1996; Shealy & Stout, 1993a). In addition, the SIBTEST can evaluate the DTF of a bundle of items and investigate the phenomena of amplification or cancellation of DIF in bundles of items.

The purpose of this research is to compare the two IRT methods and the two non-IRT methods; the researcher selected the "area measure," "likelihood ratio test" IRT methods, and the Mantel-Haenszel, and SIBTEST, non-IRT methods in order to investigate the research questions. The disadvantage of AM method is that the IRF differences at different points on the θ-scale that may be calculated would offset each other when the ICCs of two groups intersect (Lu, 1999). Then the real value of DIF would be underestimated. So AM method is not appropriate to detect non-uniform DIF. The major limitation of the MH procedure is that it is not appropriate to detect non-uniform DIF (Narayanan & Swaminathan, 1996). The *B* statistics in SIBTEST designed by Li & Stout (1996) is used for detecting non-uniform DIF. Therefore, the four methods used in this study, two methods – AM and M-H are fit to detect uniform DIF and the other two methods - LR and SIBTEST are appropriate to detect non-uniform DIF.

CHAPTER 3

RESEARCH METHODOLOGY

The intent of the study is to determine whether DIF is present in the basic mathematical competence test for junior high schools using gender to define groups. The second intent is to investigate how well the four detection procedures work. In order to be representative, samples were taken from the full population of all examinees attending the test. After getting the sample data, the researcher used four DIF methods to analyze the data. The related tools, processing procedure, design, and analyses were described as follows.

3.1 The Objective of the Research

The mathematics portion of the student's Basic Competence Test for Junior High Schools was administered in April 2001. The number of examinees was 299,368. For the purposes of this study, a number of random samples were selected from the full sample. Gender was the basic demographic information used in the study.

3.2 Tools of Research

The tools of research include

3.2.1 Instrument: The Basic Mathematical Competence Test for Junior High Schools in Taiwan

The major differences between the traditional entrance examination and the basic competence test are the test contents and development of that test. The basic competence test is designed to evaluate the knowledge and cultivated ability of students after they

have completed their compulsory education. The basic competence test uses a continuously developing item bank for construction of the examinations. Item bank development includes:

1. Item writing and revision: Traditional principles of item development were followed that require that a large number of items be written both by teachers who teach in middle schools and by mathematics experts. Mathematics experts who were not the writers of items and measurement experts then examine the items to determine if they meet the requirements for content validity. If they do not, the items are deleted or revised.

2. Pilot test: The items put in the item bank have to have the same scale; appropriate item and examinee samples are important. The institution uses a matrix sampling design with overlapping items to collect data and concurrent calibration in order to get the same scale of item parameters to build the item bank. Concurrent calibration can obtain smaller equating error than other methods (Li, and Yang, 1999). The ability distribution of examinees participating in the pilot test is normal. Each item in the pilot test was administered to between 240 and 320 9^{th} grade students who lived throughout Taiwan. Tests were administered in junior high school classes whose students' abilities distributions were normal.

1. Items calibration, equating, and the test of goodness of fit

The test was composed of the items that were chosen by computer from the item bank and followed the aim of the test, which had been publicly announced. The chosen program was designed by the committee of Basic Competence Test for Junior High Schools using "Delphi" language to write the program. When the test is composed, writers usually add some limitations to the program and use an iterative method to choose

an item from the item bank. After running the program, writers recheck the items. A researcher on the test development committee indicated that the selection procedure usually is run two or three times in order to develop the test. The test specifications include item content from the junior high school mathematics curriculum. Difficulty values are targeted between .5 and .75. The mathematics test specifications include,

i. Items emphasizing comprehension, application, logical reasoning, and proof of mathematics knowledge.

ii. Items avoiding memorization and emphasizing the comprehension of basic concepts.

iii. Items are chosen by the curriculum criteria emphasizing the curricular content.

iv. The stem of the item completely describes the problem. The item order has to be logical (e.g., the items should be ordered from easy to difficult). The distracters should be developed to include the errors typically made by examinees.

There were 32 items on the test. All of the items were multiple-choice. Examinees had to choose one correct answer from four options on each item. The content of the test included the mathematics covered in junior high school. In general, the curriculum covered three general topics: algorithms, algebra, and geometry. The items were classified into these three categories. The total number of items on algorithms was 5. The total number of items on algebra was 9, and the total number of items on geometry was 18.

3.2.2 The computer software used in the study are:

i. BILOG -MG

ii. DIMTEST

iii. SAS

iv. SPSS

v. MATLAB

vi. S-Plus

vii. Dimensionality-Based DIF/DBF Package

3.3 Processing Procedure

The research questions are as follows:

1. Is there significant DIF in test items from the basic mathematics competence test based on gender groups?
2. How consistent are results for the different ways of detecting DIF?
3. Which is the best procedure for detecting DIF?
4. What is the Type I error rate of detecting DIF for the different methods?
5. Which of the items detected as showing significant DIF are considered to represent item bias based on logical analysis?
6. What should be done if the DIF is due to instructional differences or some biasing features of the items after one identifies DIF items?

In order to answer the above research questions, the following procedures will be followed.

3.3.1 Obtaining data

The test for junior high schools is the responsibility of the Institute of the Basic Competence Test Center for Junior High Schools. The data for this research were obtained from the Institute. The data consisted of the item responses of the student population. The total number of examinees taking the test was 299,368. For the purposes of this study, a sample of 29,876 was selected from the full samples. The sample method was the stratified random sampling by gender and location.

First, samples of 2,000 males and 2,000 females from the 29,876 real response data were selected because in empirical study 4,000 cases are a sufficient and appropriate sample to run IRT (Baker, 1990; Cohen & Kim, 1993; Kim & Cohen, 1991; Lim & Drasgow, 1990; Raju, van der Linden, & Fleer, 1995). BILOG-MG software was used to estimate the examinees' abilities and item parameters. The distributions of examinees' ability were computed for males and females.

Based on the estimated item parameters and the ability distribution for males and females, the three-parameter logistic item response theory (IRT) model was used to simulate the students' response data for 100 sets of data. Each data set had 1000 responses of males and females, respectively. The total number of items is 32, resulting in 100 response data tables (that is, a 2000[examinees]×32 matrix for each table). The four DIF detection methods were used to analyze each response data table resulting in 100 values for each method. Because there are no DIF items for the simulation data, distribution is based on the null hypothesis with no DIF items. For every item, the 100 DIF statistics yield a distribution. There are 32 distributions for each method. There will be 128 distributions, and the mean and standard deviation for each distribution will be

computed. Guided by previous simulation research (Rogers & Swaminathan, 1993; Narayanan & Swaminathan, 1994; Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998), 100 replications are deemed enough to give some information about the sampling distribution of the statistics under the null hypothesis of no DIF. In addition, the fixed-sample-size procedure (Law & Kelton, 2000) to calculate a confidence level of 95% of an absolute error after obtaining 10 replications in order to judge whether or not the 100 replications are enough to yield stable estimates of the sampling distributions.

Second, 100 sets of response data were randomly sampled with replacement from the 29,876 real responses. Each data set comprised of 1000 responses from males and females, respectively. The total number of items is 32, resulting in 100 real response data tables (that is, a 2000[examinees]×32 matrix for each table). The four DIF detection methods were used to analyze the 100 sets of data. Each method will result in 100 values for each item. The distribution will be based on real data for the alternative hypothesis. There will be 128 distributions. Both the mean and standard deviation for each distribution will be computed.

The z-statistic is computed based on the mean and standard of the sampling distribution of the statistic under the null hypothesis of no DIF of the same item for same method, resulting in a 32(item)×4(method) table. A comparison can be made of the z-statistic of each item between four methods in order to determine which one has the biggest z-score. In addition, a one-way ANOVA was used to determine whether or not there are differences among four methods for the 32 items. Figure 3.1 is a flow chart of the processing procedure.

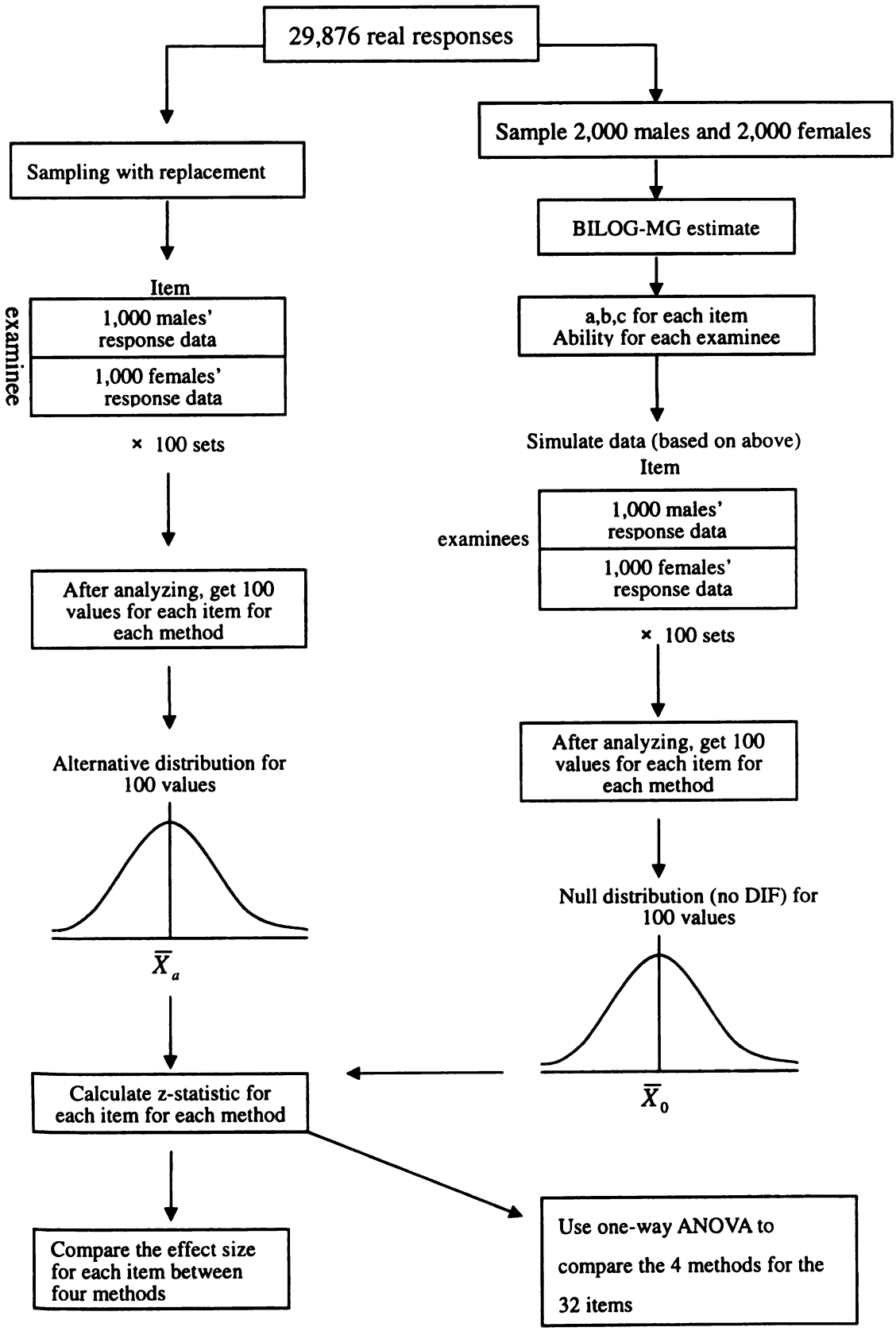


Figure 3.1 *Flow chart of the processing procedure.*

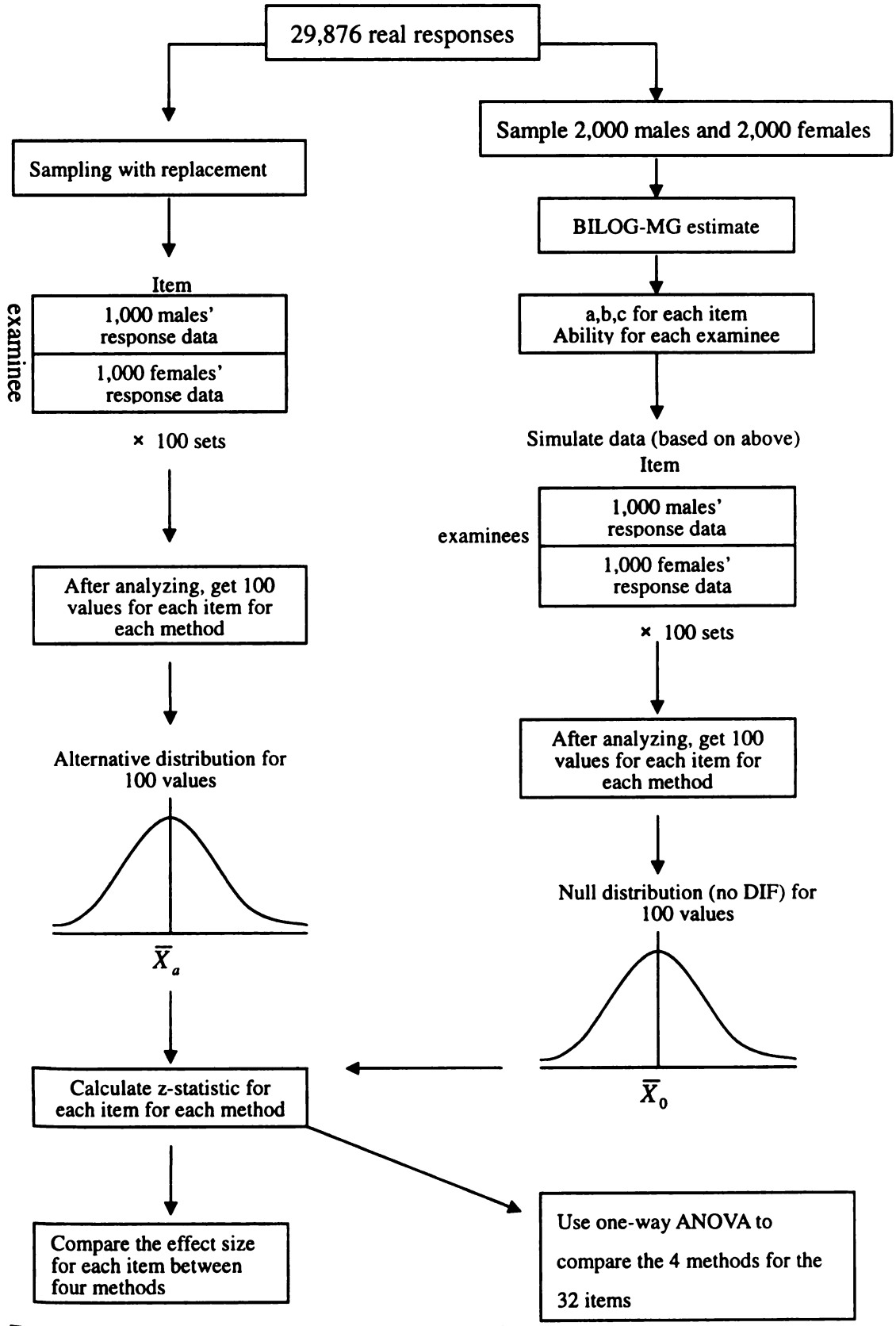


Figure 3.1 *Flow chart of the processing procedure.*

3.3.2 Type I error rate

A critical value can be obtained for testing significance once the α value is set. For instance, the statistic for testing the null hypothesis using SIBTEST to detect DIF or DTF is,

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)}.$$

In the formula, $\sigma(\beta_U)$ is the estimate standard error of β_U. The distribution of the SIBTEST *B* statistic can be approximated by the standard normal distribution under the null hypothesis. If the observed *B* value is larger than the $100(1-\alpha)$ percentage point of the *z* distribution, the null hypothesis is rejected. The *B* statistic is designed for detecting uniform DIF. If α equal to .05 is set, the critical value is equal to 1.96 for a two-tailed test. The number of values that exceed the critical value was calculated from the simulation data. For example, if 100 values are calculated and there are 7 values bigger than 1.96, the Type one error rate will be .07. Then a 32(item)×4(method) table can be got and the Type I error rate for each item among the four methods can be compared.

3.3.3 Design and analysis

The following four methods are used in the analysis: area measure (IRT-based procedure-"Sign-Area"), likelihood ratio test (model comparisons measures for identifying DIF), Mantel-Haenszel method, and SIBTEST. The groups for comparison are male versus female. The total number of items identified with significant DIF were determined for each of the four DIF methods, and then the number of common items identified among methods was determined. In addition, simulation data were used to

identify critical values that yield common α errors for the different methods.

Comparison of DIF: Area Measure (IRT-based Procedures- "Sign-Area")

Prior to the DIF analyses, several preliminary analyses will be run.

Unidimensionality. The data from the two target groups were separately factor analyzed using principle components analysis to determine the degree to which the item response data can be fit by a unidimensional model. DIMTEST (Stout, 1987) software was used to assess unidimensionality. If the unidimensionality assumption is not supported, then the IRT-based method may be not appropriately used to detect DIF items. In addition, there will be many DIF items in the test because there is not a dominant factor. Or applying multidimensionality DIF method to detect is an alternative choice.

Estimation of item parameters. BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used to estimate the item parameters and students' ability. The BILOG-MG program can estimate 1PL, 2PL, and 3PL IRT models from test data with dichotomous item response formats. There are three main estimation options for scoring examinees: (a) maximum likelihood (ML), (b) expected a posteriori (EAP), and (c) maximum a posteriori (MAP). In this study, EAP will be used to estimate the item parameters and student abilities because then a trait level estimate will be computable for all examinees, even for perfect all-endorsed and not-endorsed response patterns. BILOG-MG was also used for one of the DIF detection methods. The fit of the model to the data was evaluated using the likelihood ratio goodness-of-fit statistic, G^2, distributed as χ^2, a test of the model against a general multinomial alternative model, as discussed by Thissen, Steinberg, and Gerrard (1986). Because the area between the ICCs for males and females is infinite when the lower asymptotes are not equal for the 3PL model, only the special

case in which $c_1 = c_2 = 0.2$ will be considered here (Raju, 1990). There are four options for each item in the test. The vocabulary test in Raju's (1990) study sample also has four responses per item. Therefore, 0.2 is the appropriate choice.

Linking of item parameter estimate. The item parameter estimates were transformed for the comparisons — estimates from the female sample were transformed to the scale underlying the male sample. IRT-based DIF analyses require that the estimated item parameters from the two subpopulations be put on a common scale prior to any DIF analysis. The transformation procedure that was called "Characteristic Curve Method" described by Stocking and Lord (1983) was used because it takes into account all available information (Hambleton & Swaminathan, 1985). The item parameter estimates, *a*'s and *b*'s only, from the female group were linearly transformed so that the transformed item estimates were on the same scale as the item parameter estimates for the male group. This transformation was necessary because the item parameters were separately calibrated for the male and female groups. The transformed *a*- and *b*-parameters for the female group were used in the subsequent computations of the signed area.

Statistical tests for DIF indexes. Using the Characteristic Curve Method item parameter estimates obtained in the previous step, SA was computed by using MATLAB software. Raju's (1990) *z* statistics for SA were computed to identify items with significant DIF. Because the sample size is large, the *z* statistics associated with SA was assumed to be normally distributed and a two-tailed test of $z>2.81$ or $z<-2.81$ ($\alpha = .005$) was used to identify items with significant DIF. But in this study, the same criteria for α ($\alpha = .05$ and $\alpha = .01$) with other methods was adopted.

Computation of DIF: Likelihood Ratio Test-Model comparison measures

LR compares a compact model and an augmented model. The statistic for the LR test can be expressed as follows:

G^2_j = -2 log [Likelihood (Compact model)/ Likelihood (Augmented model)].

G^2_j is distributed as a χ^2 under the null hypothesis with degrees of freedom (df) equal to the difference in the number of parameters estimated in the compact and augmented models. For this study, G^2_j is distributed as a χ^2 with 3 df. If α=.05, then the critical value is $\chi^2_3 = 7.82$. That is, if the value of G^2_j is greater than 7.82, the item will be considered to exhibit DIF. Similarly, if α=.01, then the critical value is $\chi^2_3 = 11.34$. In the compact model, the item parameters are assumed to be the same for both the reference and focal groups. BILOG-MG permits equality constraints to be placed on items for estimation of the compact model. In the study, the parameter estimates for all 32 items for the compact model are set to be equal in both the reference and focal groups. In the augmented model, item parameters for all items except the studied item are constrained to be equal in both the reference and focal groups. These constrained items are referred to as the common or anchor set. In for the LR method, only the item parameters for the studied item are different in the reference and focal groups. For instance, in this study, for the augmented model in which Item 1 is the studied item, item parameter estimates for Item 1 will be unconstrained in the reference and focal groups. Items 2-32 form the anchor set for the augmented model and each is constrained to have the same parameter estimates in both groups. The metric used in LR is based on the set of items constrained in the anchor set. In this study, the augmented models are constrained to study a single item at a time. All items are studied sequentially for DIF.

Comparison of DIF: Mantel-Haenszel Method

The CMH procedure in SAS will be used to implement the Mantel-Haenszel Method. Thirty-three groups will be used for the analysis; score groups from 0 score to 32. But the group will not be used if the frequency in the reference or focal group is 0 for a score category. In the ideal case, there will be 33 2×2 contingency tables for total score k from 0 score to 32. For this study, χ^2_{MH} is χ^2 distribution with 1 degree of freedom. If α=.05, then the critical value is $\chi^2_{MH} = 3.84$. That is, if the value of χ^2_{MH} is greater than 3.84, the item will be considered to exhibit DIF. Similarly, if α=.01, then the critical value is $\chi^2_{MH} = 6.64$. If the average score for an item is computed for each group, the group that the item advantages can be determined. For instance, if the proportion of correctly answering item 1 is .6 and .5 for males and females respectively, then item 1 will be advantageous to the male group.

Comparison of DIF: SIBTEST

In this study, the male group is the reference group, and the female group is the focal group. The study investigates whether or not the 32 items in the Basic Mathematical Competence Test for Junior High Schools in Taiwan exhibit DIF between genders. Assuming the total number in the valid subset is k, the other items from k+1 to 32 are the suspect subtest. The process used was consistent with that suggested by Stout and Roussos (1995) to purify the valid subtest before using SIBTEST to detect DIF. The purifying procedures are as follows:

1. Do an automatic DIF analysis (ADA) for the valid subset. Choose one of the valid subtest items as the suspect subtest and to use the other items as the valid subset. To repeat the DIF detecting procedure in order to find the items that exhibit

DIF.

2. Repeat the procedure of ADA after excluding suspect DIF items.

3. Repeat the second procedure until all the items are not DIF. The valid subset items will become the purified and can be used as a basis to detect DIF.

The SIBTEST *B* statistic is an approximation to the standard normal distribution under null hypothesis ($H_0 : B = 0$). If the observed *B* value is larger than the 100(1-α) percentage points of the *z* distribution, the null hypothesis is rejected. That is, if α=.05, then the cut point is 1.96. If the value of *B* is greater than 1.96, the item will be considered to exhibit DIF. Similarly, if α=.01, then the critical value is 2.58.

Detecting item DIF

The researcher can estimate the magnitude z-score from observed data. The equation is:

$$z = \frac{\overline{X} - \mu_0}{s_X}$$

After obtaining the z-score for each item for each method, 1.65 criterion for a z-statistic is used as a reference to judge the magnitude of z-score.

If the z-score of the four methods for each item are all bigger than 1.65, then the item will be judged as exhibiting DIF. Whether or not the item is biased requires a logical analysis in addition to a significant DIF index: one has to identify the intended construct, infer the presence of a secondary construct in this particular item, and judge the latter to be irrelevant to the former. Sometimes DIF indices are unreliable. Different mathematics experts will perform logical analyses to see what characteristics of the item's content,

format, or other data may have been investigated in prior studies of the test to explain the difference in item difficulty for the different groups. That is, if the item exhibits DIF and belongs to geometry content, a mathematics expert in the field of geometry would examine whether or not the item is biased between genders. In addition, if the item is judged as biased against either gender, prior studies would be compared to try and determine the cause. Because the various causes of bias include cultural differences, different curricula, different instruction, or other factors that exist in the item content stem. It will be helpful for the institution to avoid using biased items on the test in order to enhance the validity and equality of the basic mathematical competence test. That is, DIF statistical analyses and subsequent efforts by mathematics experts to reevaluate the relationship of items to measurement of the intended construct can lead to a far greater insight about the basic mathematical competence test measures and how the items function.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter has two parts. The first part gives the results of the DIF analysis for the Basic Mathematical Competence Test for Junior High Schools in Taiwan and presents the Type I error rate for the four methods used in this study. The second part is a synthesis and discussion.

4.1 Research results for the DIF analysis of the Basic Mathematical Competence Test for Junior High Schools in Taiwan

Table 4.1 shows descriptive statistics for the total samples of 29,876 and 4,000 for both boys and girls. The means and standard deviations for 29,876 and 4,000 samples were similar.

Table 4.1 *Descriptive statistics of 29,876 and 4,000 real samples*

	Boys					Girls				
samples	n	Min	Max	M	SD	n	Min	Max	M	SD
29,876	15,411	0	32	18.44	7.55	14,465	2	32	18.41	6.93
4,000	2,051	3	32	18.4	7.64	1,949	3	32	18.5	6.86

Appendix 1 contains the three parameters estimated by BILOG-MG from the 4,000 and the 29,876 real data samples. Table 4.2 depicts the min, max, mean, and standard deviation for *a*, *b*, and *c*.

Table 4.2 *Descriptive statistics for item parameter estimates based on the 4,000 sample*

	a	*b*	*c*
min	0.65	-1.17	0.14
max	2.10	1.46	0.47
mean	1.42	0.23	0.25
S.D.	0.38	0.70	0.08

4.1

tec

an

un

nc

ap

s.

ti

c

4.1.1 Unidimensionality

Hambleton and Rovinelli (1986) used nonlinear factor analyses to be a promising technique for assessing dimensionality. Nandakumar (1994) compared Stout's procedure and nonlinear factor analyses — DIMTEST was more powerful and effective in detecting unidimensionality. In this study DIMTEST software was used to determine whether or not an assumption of unidimensionality for the data is tenable. The distribution of T is approximately the standard normal z distribution. The range of T values for the 100 real samples was between -2.3 and 1.58. It was a one-tailed test. All the values were smaller than the critical value of 1.65 for $\alpha = .05$. Therefore, all the 100 real samples' data were consistent with a hypothesis that the data can be modeled with a single person achievement parameter. The range of T values for the 100 simulation samples was between -2.58 and 1.59. It was also a one-tailed test. All the values were smaller than 1.65. Therefore, all the 100 simulation samples' data also approximately corresponded with unidimensionality. Whether real or simulated, the data sets were consistent with the assumption of unidimensionality. The expected distribution of T values for DIMTEST is a normal distribution with the values of the mean and standard deviation of 0 and 1, respectively. A one-tailed test was used. The critical value corresponding to the .05 level is 1.65. In contrast, in Table 4.3 both distributions of real and simulation samples have smaller standard deviations than the recommended distribution. Also, the distribution of the simulation group shifted farther to the left than the other two distributions. The critical values corresponding to the .05 level for the simulation, real and recommended, were 1.02, 1.36, and 1.65 respectively. That is, in general, the T values from the simulation are smaller than those from the real data. If the critical value is set by the

simulated data, there will be seven values greater than 1.02 for real data samples. This is approximately equal to the 5 that would be expected by chance for 100 samples. However, the results suggest that the distribution of the T statistic is not does not have a mean of 0.0 and standard deviation of 1.0.

Table 4.3 *Descriptive of T values for real and simulation samples*

	Real	Simulation	Recommended
min	-2.30	-2.58	-
max	1.58	1.59	-
mean	-0.03	-0.42	0
SD	0.84	0.88	1
95%C.I.	1.36	1.02	1.65

4.1.2 The results of Area Measure (AM) method

The statistical summary for the Area Measure to detect the DIF items for 100 replications from real data are shown for Table 4.4. The z values of the 32 items were between -7.25 and 7.19. The z means of 32 items for real samples were between -2.30 and 4.11. All of the 100 values of z were positive for Item 26. This demonstrates that Item 26 was advantageous for males.

Statistical summary results of the Area Measure to for detecting DIF for the 100 simulation samples are also shown for Table 4.4. The simulation results are centered around zero, and many of the standard deviations are near 1.0. Occasionally the standard deviations are much lower than 1.0. Most of the real data results are not centered around zero suggesting that there is some DIF, although much of it is not significant. The z values of 32 items were between -3.49 and 3.61. The z means of 32 items for simulation samples were between -.33 and .16.

The z-statistic for the 32 items for the Area Measure is also shown in Table 4.4. The

values are between .01 and 3.78. Item 26 has the largest z-score. The second largest z-score is for Item 21. The next largest are Items 32, and 22 with z-scores of 1.89, and 1.79, respectively. There are only five items with a z-score larger than 1.65.

After checking the histograms of the 32 items DIF statistics for AM method for real data, except Item 10, 13, and 31, all the others' absolute skew values are smaller than 0.5. All the histograms seem symmetric. For the simulation data, except Item 3 and 15, all the others' absolute skew values are smaller than 0.5. All the histograms of the 32 items DIF statistics also seem symmetric.

Table 4.4 *The z-statistic of real data to simulation data for AM method*

	AM (n=100)								
	Real				Simulation				
Item	Min	Max	M	SD	Min	Max	M	SD	z-statistic
1	-2.76	0.67	-0.74	0.67	-0.81	1.10	0.04	0.38	1.43
2	-2.24	1.88	-0.19	0.87	-1.81	1.57	0.02	0.69	0.27
3	-3.28	1.14	-0.82	0.77	-1.26	2.95	-0.04	0.64	1.11
4	-0.99	1.16	-0.01	0.51	-1.10	1.36	0.16	0.47	0.35
5	-2.23	1.10	-0.65	0.55	-1.29	1.57	0.09	0.49	1.42
6	-1.66	0.89	-0.47	0.61	-1.39	1.83	-0.02	0.61	0.74
7	-3.68	3.20	-0.16	1.48	-2.50	1.95	-0.17	1.04	0.01
8	-2.15	2.54	0.39	0.95	-2.53	2.72	-0.24	0.91	0.67
9	-0.74	2.28	0.48	0.60	-1.83	2.63	-0.02	0.72	0.75
10	-2.98	4.49	0.14	1.36	-2.24	1.90	-0.12	0.82	0.23
11	-1.55	4.16	1.07	1.04	-2.87	2.11	-0.06	0.87	1.19
12	-1.62	4.91	1.05	1.18	-2.77	2.31	-0.14	0.92	1.13
13	-3.55	3.47	-0.51	1.08	-3.00	2.48	-0.01	1.09	0.46
14	-7.25	1.10	-1.92	1.42	-3.10	2.50	-0.04	1.04	1.51
15	-1.48	1.25	-0.13	0.56	-1.49	1.50	-0.14	0.60	0.02
16	-3.30	1.68	-1.00	1.05	-2.09	1.37	-0.25	0.74	0.83
17	-1.84	4.10	1.02	1.11	-2.24	1.93	-0.13	0.94	1.12
18	-3.68	4.46	0.13	1.49	-2.19	2.13	-0.15	0.91	0.23
19	-2.05	0.83	-0.76	0.52	-1.43	1.03	-0.09	0.47	1.35
20	-1.49	3.92	1.49	1.25	-2.42	2.24	-0.22	0.86	1.60
21	-6.06	0.36	-2.30	1.26	-3.49	3.01	-0.06	0.98	1.98
22	-5.17	1.37	-1.95	1.33	-2.87	2.41	0.12	0.95	1.79
23	-3.45	3.97	0.89	1.50	-3.33	3.61	-0.17	1.18	0.79
24	-3.59	3.13	-0.45	1.06	-1.85	2.01	-0.33	0.81	0.13
25	-3.34	3.61	-0.13	1.43	-2.38	2.75	-0.16	0.90	0.03
26	0.34	7.19	4.11	1.26	-2.54	2.62	-0.19	1.01	3.78
27	-2.58	2.88	0.24	1.01	-2.24	1.76	-0.11	0.83	0.38
28	-3.85	0.99	-1.15	1.05	-2.95	2.04	-0.13	1.04	0.98
29	-3.05	2.74	0.03	1.19	-2.22	1.75	-0.19	0.83	0.21
30	-3.99	1.69	-1.39	1.21	-2.75	2.29	0.07	0.94	1.35
31	-2.27	2.94	1.02	1.00	-2.62	1.49	-0.03	0.79	1.17
32	-5.34	1.98	-1.99	1.28	-1.77	1.73	0.05	0.83	1.89

4.1.3 The results of the Likelihood Ratio (LR) method

The statistical summary results for the Likelihood Ratio method for DIF detection for 100 samples from the real data are shown for Table 4.5. The χ_3^2 values for 32 items for the real sample were between 0 and 44.24. The χ_3^2 means of 32 items for the real samples were between 1.02 and 19.9.

The statistical summary results of the Likelihood Ratio method to detect DIF items for 100 simulation samples are shown for Table 4.5. The χ_3^2 values for the 32 items were between 0 and 20.62. The χ_3^2 means of 32 items for the simulation samples were between .52 and 3.38.

The magnitudes of Likelihood Ratio z-statistic for the 32 items are also shown in Table 4.5. The values of z-statistic are between .17 and 2.89. Item 26 has the largest z-score with a value of 2.89. Item 7 has the second largest z-score with 1.72. The next largest is Item 32 whose z-score is 1.55. There are two items whose z-score is larger than 1.65.

Most of the mean chi-square values for simulated data items tend to be less than the degrees of freedom. In addition, all the SDs for simulated data tend to be less than 3. That is, the chi-square statistic for simulated data may underestimate the criteria of previous studies.

The histograms of the 32 items' DIF statistics for LR method for real data are not symmetric. Except for items 1, 7, 11, 21, 23, 24, 26, 30, 31, and 32, all the others' skew values are larger than 1. For the simulation data, except for item 5, 6, and 26, all the others' skew values are larger than 1 and larger than the skew values of real data. All the histograms of the 32 items DIF statistics also seem not symmetric. The reason is the

distribution is approximately to chi-square with d.f.=3.

Table 4.5 *The z-statistic of real data to simulation data for LR method*

	LR (n=100)								
	Real				Simulation				
Item	Min	Max	Mean	SD	Min	Max	Mean	SD	z-statistic
1	0	23.15	6.81	5.18	0	20.62	3.38	2.76	0.83
2	0	18.25	3.36	3.26	0	9.11	1.76	2.06	0.58
3	0	19.78	4.37	3.58	0	7.01	0.88	1.42	1.28
4	0	15.19	3.39	3.83	0	9.68	2.17	2.17	0.39
5	0.25	26.46	6.48	4.58	0	9.09	3.13	2.15	0.94
6	0	9.39	1.02	1.81	0	6.86	1.89	1.86	0.47
7	0.27	21.05	8.17	5.09	0	9.95	1.57	1.84	1.72
8	0	9.26	1.05	1.95	0	7.68	0.66	1.51	0.22
9	0	11.66	1.45	2.38	0	9.28	2.25	2.11	0.36
10	0	13.17	3.64	3.29	0	8.22	1.67	1.73	0.75
11	0	11.05	3.2	2.69	0	5.64	1.07	1.44	0.99
12	0	16.74	1.82	2.94	0	8.01	0.8	1.51	0.44
13	0	15.36	2.45	3.16	0	11.53	1.77	1.85	0.26
14	0	22.75	5.11	3.97	0	7.13	1.66	1.6	1.14
15	0	12.59	2.89	3.29	0	7.4	1.38	1.83	0.57
16	0	14.69	3.52	2.59	0	9.28	2.17	1.96	0.59
17	0	13.44	2.58	2.58	0	7.42	1.16	1.64	0.65
18	0	14.45	2.81	3.27	0	5.23	0.52	1.08	0.94
19	0	13.71	2.2	3.2	0	6.29	0.93	1.36	0.52
20	0	24.05	6.87	5.03	0	10.73	0.92	2	1.55
21	0	16.46	4.74	3.91	0	12.44	1.76	2.17	0.94
22	0	21.07	5	4.48	0	6.77	0.8	1.49	1.26
23	0	11.56	2.98	2.69	0	9.12	1.77	1.94	0.52
24	0	27.29	8.56	6.19	0	13.69	2.04	2.35	1.39
25	0	19.45	4.97	4.1	0	8.81	0.88	1.56	1.32
26	5.85	44.24	19.9	8.37	0	6.91	2.36	1.86	2.89
27	0	11.43	1.15	2.23	0	6.96	0.69	1.39	0.24
28	0	13.7	4.25	3.21	0	10.79	2.59	2.25	0.6
29	0	7.99	1.38	1.9	0	6.89	1.07	1.67	0.17
30	0	10.09	2.48	2.3	0	9.62	1.71	1.96	0.36
31	0	15.35	4.45	3.26	0	16.28	1.6	2.65	0.96
32	0	24.89	8.45	5.48	0	8.57	2.1	1.94	1.55

4.1.4 The results of the Mantel-Haenszel (M-H) method

The statistical summary results for the Mantel-Haenszel method for detecting DIF for 100 samples from the real data are shown for Table 4.6. The χ^2 values were between 0 and 42.84. The χ^2 mean values of 32 items for real samples were between .9 and 18.2.

The statistical summary results for the Mantel-Haenszel method for detecting DIF items for 100 simulation samples are shown for Table 4.6. The χ^2 values are between 0 and 7.14. The χ^2 mean values of 32 items for 100 simulation samples are between .23 and .97.

Table 4.6 shows the magnitude of M-H z-statistic of real data to simulation data for χ^2 with d.f.=1. The values of M-H z-statistic are between .02 and 3.12. The largest z-score is 3.12 for Item 26. The second largest z-score is 1.66 for Item 20. These are followed by Items 31, 22, 19, and 1 whose z-scores are 1.38, 1.30, 1.29, and 1.28, respectively. There are 2 items whose z-scores are larger than 1.65.

The results in Table 4.6 are similar to those in Table 4.5 in that the mean chi-square and SD values of simulated data tend to be less than the degrees of freedom. That is, the chi-square statistics for the simulated data suggest that suggested critical values may underestimate the amount of DIF.

The histograms of the 32 items DIF statistics for M-H method for real data are not symmetric. Except for items 26 and 31, all the others' skew values are larger than 1. For the simulation data, except for items 3, 11, 24, 26, 29, 30, and 31, all the others' skew values are larger than 2 and larger than the skew values of real data. All the histograms of the 32 items DIF statistics also seem not symmetric. The reason is the distribution is approximately to chi-square with d.f.=1.

Table 4.6 *The z-statistic of real data to simulation data for M-H method*

	M-H (n=100)								
	Real				Simulation				
Item	Min	Max	M	SD	Min	Max	M	SD	z-statistic
1	0.01	16.73	3.98	3.79	0	4.58	0.49	0.66	1.28
2	0	18.3	1.97	2.84	0	7.14	0.65	1.03	0.62
3	0	15.76	2.77	2.87	0	3.22	0.58	0.72	1.05
4	0	9.82	1.77	1.91	0	4.11	0.52	0.74	0.86
5	0	7.55	0.9	1.3	0	4.69	0.39	0.69	0.49
6	0	5.44	1.22	1.4	0	5.29	0.66	0.94	0.47
7	0	11.45	2.38	2.56	0	4.06	0.51	0.71	1
8	0	6.09	1.01	1.29	0	6.28	0.97	1.29	0.02
9	0	12.83	1.84	2.07	0	5.67	0.54	0.86	0.82
10	0	9.4	1.09	1.57	0	2.84	0.36	0.46	0.63
11	0	8.76	1.92	2.24	0	2.02	0.36	0.44	0.97
12	0	14.56	2.02	2.71	0	3.17	0.42	0.56	0.82
13	0	11.43	1.42	2.03	0	3.48	0.51	0.73	0.59
14	0	18	2.32	3	0	2.78	0.39	0.5	0.9
15	0.01	10.69	2.11	2.15	0	3.12	0.42	0.56	1.07
16	0	5.35	1.04	1.13	0	2.6	0.23	0.35	0.97
17	0	12.04	2.08	2.3	0	4.69	0.51	0.68	0.92
18	0	9.26	1.29	1.56	0	2.86	0.49	0.61	0.68
19	0	13.35	3.2	3.17	0	3.1	0.28	0.5	1.29
20	0	21.42	5.52	4.25	0	4.84	0.47	0.68	1.66
21	0	17.35	3.65	3.53	0	2.98	0.44	0.61	1.27
22	0.04	22.52	4.43	4.09	0	4.93	0.57	0.95	1.3
23	0	7.51	1.46	1.76	0	4.08	0.54	0.73	0.68
24	0	8.08	1.29	1.67	0	1.56	0.29	0.34	0.83
25	0	8.81	2.02	2.34	0	3.89	0.42	0.64	0.93
26	3.88	42.84	18.2	8.1	0	2	0.33	0.47	3.12
27	0	7.84	1.06	1.53	0	3.09	0.46	0.57	0.52
28	0	8.93	1.03	1.39	0	2.58	0.37	0.45	0.64
29	0	5.47	0.9	1.08	0	1.47	0.3	0.32	0.75
30	0	5.94	1.48	1.56	0	2.54	0.4	0.5	0.93
31	0	8.05	2.5	2.25	0	1.76	0.28	0.36	1.38
32	0	12.98	2.38	2.56	0	2.67	0.24	0.39	1.17

4.1.5 The results of the SIBTEST method

The statistical summary results for the SIBTEST method for detecting DIF items for 100 samples from the real data are shown in Table 4.7.

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

The B values were between -4.71 and 6.14, and the B mean values for 32 real samples items are between -2.0 and 3.83.

The statistical summary results for the SIBTEST method for detecting DIF items for 100 simulation samples are shown for Table 4.7. The B values are between -2.67 and 2.17. The B means values of 32 simulation sample items are between -.50 and .25.

Table 4.7 shows the magnitude of the SIBTEST z-statistic for the real data compared to the simulation data for B. The values of the z-statistics are between .04 and 4.67. Item 26 has the largest z-score. The second largest z-score is 2.68 for Item 20. These are followed by Items 22, 21, 19, 4, 1, and 15 with z-scores 2.46, 1.87, 1.74, 1.72, 1.69, and 1.67, respectively. All the eight z-scores are larger than 1.65.

The results in Table 4.7 are similar to those in Table 4.5 in that the mean and standard deviation of B values for the simulated data tend to be less than 0 and 1. That is, the B statistics for the simulated data suggest that the test may underestimate the amount of DIF.

Checking the histograms of the 32 items' DIF statistics for SIBTEST method for real data, most of the histograms seem symmetric. Except for items 15, 26, 31, and 32, all the others' absolute skew values are smaller than 1. But for the simulation data, except Item 5, all the others' absolute skew values are smaller than 0.5. The absolute skew values are smaller than 1for most of the histograms of the 32 items DIF statistics.

Table 4.7 *The z-statistic of real data to simulation data for SIBTEST method*

SIBTEST (n=100)									
	Real				Simulation				
Item	Min	Max	M	SD	Min	Max	M	SD	z-statistic
1	-4.53	0.83	-1.65	1.09	-1.78	2.17	-0.07	0.75	1.69
2	-4.14	1.71	-1.01	1.1	-2.67	1.2	-0.34	0.77	0.71
3	-4.24	0.47	-1.39	0.96	-1.88	1.12	-0.33	0.73	1.24
4	-2.07	3.5	1.2	1.06	-2.06	1.13	-0.32	0.66	1.72
5	-2.5	3.28	0.13	1	-2.22	1.67	-0.31	0.62	0.53
6	-2.56	2.31	-0.34	1	-2.41	1.65	-0.26	0.82	0.09
7	-3.78	1.79	-1.58	1.09	-2.06	1.5	-0.25	0.76	1.42
8	-2.53	2	-0.31	1.03	-2.63	1.38	-0.5	0.9	0.20
9	-1.24	3.88	1.05	0.96	-2.28	1.72	-0.26	0.77	1.51
10	-2.76	2.33	-0.7	1	-2.05	1.52	-0.1	0.67	0.70
11	-1.9	2.95	0.55	1.07	-2.43	1.52	-0.08	0.63	0.72
12	-1.5	3.33	0.84	1.04	-1.44	1.42	-0.25	0.7	1.23
13	-3.26	1.89	-0.78	0.96	-2.05	1.74	-0.24	0.75	0.63
14	-4	1.23	-1.07	1.08	-1.45	1.86	0.05	0.64	1.26
15	-1.05	3.68	1.27	0.96	-1.58	1.32	-0.14	0.71	1.67
16	-2.96	2.43	-0.41	0.92	-1.7	1.17	-0.1	0.58	0.40
17	-1.7	3.45	1.07	0.96	-2.05	1.35	-0.17	0.72	1.46
18	-3.42	1.37	-0.74	1.01	-2.11	1.83	-0.04	0.75	0.79
19	-3.73	0.64	-1.5	1.01	-1.64	1.32	-0.07	0.58	1.74
20	0	4.67	2.11	0.93	-2.05	1.65	-0.07	0.68	2.68
21	-4.24	0.86	-1.6	0.97	-2.62	1.29	-0.01	0.71	1.87
22	-4.71	0.8	-2	1	-1.81	2.1	0.15	0.73	2.46
23	-2.24	2.76	0.4	1.04	-2.19	1.62	0.01	0.74	0.43
24	-2.22	3.83	0.46	1.11	-1.48	1.17	-0.01	0.63	0.52
25	-2.05	2.95	0.86	1.13	-1.79	1.58	-0.2	0.68	1.14
26	1.39	6.14	3.83	0.98	-1.5	1.32	0.05	0.59	4.67
27	-1.95	2.67	0.12	1.05	-1.86	2.14	0.08	0.72	0.04
28	-2.57	2.48	-0.44	0.88	-1.91	1.43	0.1	0.63	0.71
29	-2.86	1.95	-0.44	0.98	-1.23	1.67	0	0.67	0.52
30	-2.71	1.76	-0.71	1.04	-1.41	2.09	0.25	0.65	1.11
31	-1.6	3.1	1.01	0.94	-1.6	1.6	0.09	0.6	1.17
32	-3.14	3.27	-0.98	1.06	-1.27	1.41	0.17	0.55	1.36

4.1.6 Type I error rate of the four methods

The results of Type I error rate of the four methods are based on the assumed sampling distribution for the significance tests for each method. Table 4.8 shows the significant number for each α=.05 and α=.01 for the four methods based on the simulated data. The significant number is the frequency of items whose values are larger than the critical value recommended by previous studies. For the AM method, the z is the tabled value from the unit normal distribution. For example, for Item 11, four significant z values were obtained for α=.05 and 1 for α=.01. Because 100 replications were generated, the expected number of significant z values due to chance at a normal α of .05 and .01 for a single item was 5 and 1, respectively. For this sample size condition, 98 significant z values were obtained across all items for α=.05 and, 20 for α=.01. Therefore, the Area-Measure Type I error rate was .031 and .006 for α=.05 and α=.01, respectively. Both values are far smaller than .05 and .01.

For the Likelihood Ratio method, G(1)-G(2) is approximately χ^2_3. Thirty-three significant χ^2_3 were obtained across all items for α=.05 and 6 for α=.01. Therefore, the Likelihood Ratio Type I error rate was .01 and .0019 for α=.05 and α=.01, respectively. Both values are far smaller than .05 and .01. This is not consistent with Kim and Cohen's (1998) finding that indicates Type I error rates of the LR procedures were within expected values at each of the nominal α levels. In this study, it seemed that the observed values are much than the expected values.

For the M-H method, Table 4.8 shows the number of significant χ^2 with $d.f.$=1 for each item at α=.05 and α=.01. The χ^2 with $d.f.$=1 was the tabled value from the chi-square distribution. Twenty significant χ^2 with $d.f.$=1 were obtained across all items

for α=.05 and 0 for α=.01. The error rates for α=.05 and α=.01 are .00625 and 0, respectively. Both values are far smaller than .05 and .01. Results in this study doesn't support the claim of Narayanan and Swaminathan (1994), Rogers and Swaminathan (1993), Roussos and Stout (1996), and Shealy and Stout (1993a) that the performance of Type I error control of the M-H was generally satisfactory. But in this study, it was really much lower than expected.

For the SIBTEST method, Table 4.8 shows the number of significant *p* for each item at α=.05 and α=.01. Thirty-two significant *p* were obtained across all items for α=.05 and three for α=.01. The error rates of the SIBTEST method for α=.05 and α=.01 are .01 and .001, respectively. Both are smaller than .05 and .01.

All of the four detecting DIF methods seem to be well under the expected number of identified values. Therefore, the Type I error rate of the M-H method had the smallest value among the four methods. SIBTEST, LR, and AM methods, in that order, followed. Thus, the non-IRT for detecting DIF methods has a smaller Type I error rate than IRT methods. But the IRT methods are closer to expected number than non-IRT methods.

Table 4.8 *Type I error of four methods for simulation data*

	Area-Measure		Likehood Ratio		Mantel-Haenszel		SIBTEST	
Item	α=.05	α=.01	α=0.05	α=0.01	α=.05	α=.01	α=0.05	α=0.01
1	0	0	2	1	1	0	2	0
2	0	0	1	0	1	0	3	1
3	1	1	0	0	0	0	0	0
4	0	0	1	0	1	0	2	0
5	0	0	1	0	2	0	1	0
6	0	0	0	0	2	0	4	0
7	5	0	2	0	1	0	1	0
8	4	0	0	0	3	0	5	1
9	1	1	2	0	2	0	2	0
10	1	0	1	0	0	0	1	0
11	4	1	0	0	0	0	1	0
12	7	2	1	0	0	0	0	0
13	7	2	1	1	0	0	1	0
14	5	2	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0
16	2	0	2	0	0	0	0	0
17	4	0	0	0	1	0	1	0
18	3	0	0	0	0	0	1	0
19	0	0	0	0	0	0	0	0
20	4	0	2	0	1	0	1	0
21	4	2	2	1	0	0	1	1
22	4	1	0	0	3	0	3	0
23	12	4	2	0	1	0	0	0
24	1	0	3	1	0	0	0	0
25	5	1	1	0	1	0	0	0
26	5	1	0	0	0	0	0	0
27	1	0	0	0	0	0	1	0
28	7	1	4	0	0	0	0	0
29	3	0	0	0	0	0	0	0
30	6	1	1	0	0	0	1	0
31	2	0	3	2	0	0	0	0
32	0	0	1	0	0	0	0	0
total	98	20	33	6	20	0	32	3

4.1.7 Results of the empirical research

In this study, four null distributions were created using 100 simulated data sets with non-DIF items for the four DIF detection methods. For every item, 100 values were determined for every method. The study used the 100 values to determine the distribution under the null hypothesis of non-DIF and determined the limits for 95% and 99% confidence intervals. Then, the criteria for values were used to judge whether or not the item exhibited DIF using the real data. The values of $\pm 1.96SD$ and $\pm 2.58SD$ for the null distributions were used as the 95% and 99% confidence intervals, respectively, in the AM and SIBTEST methods. After sorting the 100 values of the 100 simulated data, the 95th and 99th values were used as critical values. The 95% and 99% confidence intervals for the χ_1^2 and χ_3^2 distributions were between 0 and the 95th and 99th values, respectively. The range of confidence intervals for the four methods is shown in Appendix 2.

The study used criteria from Appendix 2 to count the frequency of DIF items for 100 replications of real data sets for the 32 items. The results are shown in Table 4.9.

The α=.01 case was used because of the large samples: 1000 males and females. For the Area Measure method, the largest frequency of identified DIF was 93 for Item 26. The second largest was for Item 21 whose frequency was 43. These are followed by Items 20, 1, 22, and 32 whose frequencies are 39, 37, 35, and 35 respectively.

For the Likelihood Ratio method, the largest frequency of DIF detection is 99, also for Item 26. The second largest is Item 32, whose frequency is 56. These are followed by Items 24 and 7 with frequencies 42 and 41, respectively.

Using the M-H method, the largest frequency for a DIF detection is 100, also for

Item 26. The second largest is Item 20 whose frequency is 74. These are followed by Items 31 and 1 with frequencies of 59 and 57, respectively.

For the SIBTEST method, the largest DIF item frequency is 99 for Item 26. The second largest is for Item 20 with a frequency of 66. These are followed by Items 22 and 19 with frequencies of 65 and 50, respectively. The summaries of percentage of time that each item is detected by each method are shown in Table 4.10.

Table 4.9 *The frequency of DIF detection in the 100 samples for real data for four methods using simulation cut score criteria*

	Simulation cut_score							
	Area-Measure		Likelihood Ratio		Mantel-Haenszel		SIBTEST	
Item	α=.05	α=.01	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01
1	53	37	44	30	63	57	56	41
2	15	6	14	11	25	16	28	14
3	23	11	48	31	45	37	35	20
4	9	0	24	17	32	21	60	20
5	34	16	42	31	32	3	24	10
6	11	1	3	3	21	2	15	5
7	6	3	69	41	37	28	45	28
8	11	3	11	5	8	0	9	2
9	5	1	5	3	37	11	39	24
10	21	11	29	17	29	23	26	20
11	27	11	34	21	46	35	32	16
12	27	15	25	5	39	21	32	22
13	6	2	14	10	23	13	18	10
14	51	29	43	26	46	34	47	28
15	5	0	25	15	46	38	54	32
16	28	17	15	3	47	35	26	13
17	27	12	16	8	46	37	45	24
18	37	11	41	18	28	14	25	11
19	29	12	26	15	67	41	59	50
20	53	39	67	19	78	74	80	66
21	55	43	30	18	67	46	61	41
22	55	35	49	31	63	37	79	65
23	26	8	15	1	29	13	19	9
24	14	4	51	42	39	28	31	16
25	21	11	59	22	36	29	42	29
26	96	93	99	99	100	100	100	99
27	15	7	10	6	18	12	18	8
28	16	8	12	10	30	12	22	10
29	16	7	3	4	39	23	26	11
30	39	22	14	4	42	27	42	26
31	38	16	30	1	65	59	48	28
32	61	35	61	56	62	46	55	42
M	29.06	16.44	32.13	19.47	43.28	30.38	40.56	26.25
SD	20.61	18.61	22.55	19.84	19.33	21.28	20.89	20.70

Table 4.10 *The percentage of time that each item is detected by each method using simulation cut score criteria for* $\alpha = .01$

Item	AM	Item	LH	Item	M-H	Item	SIBTEST
26	93	26	99	26	100	26	99
21	43	32	56	20	74	20	66
20	39	24	42	31	59	22	65
1	37	7	41	1	57	19	50
22	35	3	31	21	46	32	42
32	35	5	31	32	46	1	41
14	29	22	31	19	41	21	41
30	22	1	30	15	38	15	32
16	17	14	26	3	37	25	29
5	16	25	22	17	37	7	28
31	16	11	21	22	37	14	28
12	15	20	19	11	35	31	28
17	12	18	18	16	35	30	26
19	12	21	18	14	34	9	24
3	11	4	17	25	29	17	24
10	11	10	17	7	28	12	22
11	11	15	15	24	28	3	20
18	11	19	15	30	27	4	20
25	11	2	11	10	23	10	20
23	8	13	10	29	23	11	16
28	8	28	10	4	21	24	16
27	7	17	8	12	21	2	14
29	7	27	6	2	16	16	13
2	6	8	5	18	14	18	11
24	4	12	5	13	13	29	11
7	3	29	4	23	13	5	10
8	3	30	4	27	12	13	10
13	2	6	3	28	12	28	10
6	1	9	3	9	11	23	9
9	1	16	3	5	3	27	8
4	0	23	1	6	2	6	5
15	0	31	1	8	0	8	2

The correlation between the four methods is shown in Table 4.11. The M-H and SIBTEST methods have the highest correlation: .87 for the $\alpha = .01$ critical value. The LR and M-H methods have the lowest correlation: .61 for $\alpha = .01$ critical value. The correlation between non-IRT methods is higher than between IRT methods. The

correlation between AM and LR was .72. Kim and Cohen's (1995) finding that the consistency performance of AM and LR procedures to detect DIF was high is not consistent with this study.

Table 4.11 *The correlation among the four methods for the frequency of DIF Items in the 100 real data samples using simulation cut score criteria*

	1	2	3	4
1. AM	-	0.72**	0.81**	0.84**
2. LR	-	-	0.61**	0.68**
3. M-H	-	-	-	0.87**
4. SIBTEST	-	-	-	-

**Correlation is significant at the .01 level (2-tailed).

The results of an ANOVA comparing the results of the four methods for detecting DIF items in the 100 real data samples by simulation cut score criteria for α=.01 and α=.05 is shown in Table 4.12 and Table 4.13, respectively.

There were significant differences among the four methods for detecting the frequency of DIF with F=3.17, $p<.05$ for $\alpha = .01$. The results are shown in Table 4.12. After a posteriori comparison, there was significant difference between the AM and M-H methods. In addition, the frequency of detecting DIF using M-H method was larger than when using the AM method.

Table 4.12 *Summary of* ANOVA *and A Posteriori comparison for the frequency of DIF items in the 100 real data samples among the four methods using simulation cut score criteria,* $\alpha = .01$

Source of Variation	*SS*	*df*	*MS*	*F*
Between Groups	3853.40	3	1284.46	3.17*
Within Groups	50265.34	124	405.37	
Total	54118.74	127		

*$p<.05$

Method I	Method J	Mean Difference (I-J)
1	2	-3.03
	3	-13.94*
	4	-9.81
2	1	3.03
	3	-10.91
	4	-6.78
3	1	13.94*
	2	10.91
	4	4.13
4	1	9.81
	2	6.78
	3	-4.13

Note. Means of the four different methods differ significantly at $p < .05$ using the Tukey honestly significant difference comparison. 1:AM, 2:LR, 3:M-H, 4:SIBTEST
*$p<.05$

There was also a significant difference among the four methods for detecting the frequency of DIF items with F=3.35, $p<.05$ for $\alpha = .05$. The results are shown in Table 4.13. After a posteriori comparison, there was significant difference between AM and M-H methods. The frequency of DIF items on M-H method was larger than in the AM method. The results were basically the same for $\alpha=.05$ and $\alpha=.01$.

Although there was no significant difference between M-H and SIBTEST, SIBTEST tended to identify fewer DIF than M-H. This is not consistent with Gierl, Khaliq, and Boughton's (1999) finding.

Table 4.13 *Summary of* ANOVA *and A Posteriori comparison of the frequency of DIF items in the 100 real data samples among the four methods using simulation cut score criteria,* $\alpha = .05$

Source of Variation	*SS*	*df*	*MS*	*F*
Between Groups	4374.77	3	1458.26	3.35*
Within Groups	54043.72	124	435.84	
Total	58418.49	127		

*$p<.05$

Method I	Method I	Mean Difference (I-J)
1	2	-3.06
	3	-14.22*
	4	-11.50
2	1	3.06
	3	-11.16
	4	-8.44
3	1	14.22*
	2	11.16
	4	2.72
4	1	11.50
	2	8.44
	3	-2.72

Note. Means of the four different methods differ significantly at $p < .05$ using the Tukey honestly significant difference comparison. 1:AM, 2:LR, 3:M-H, 4:SIBTEST
*$p<.05$

The correlation between the z-statistic in this study and the frequency of DIF detection for four methods using simulation cut score criteria is shown in Table 4.14. The results show that the SIBTEST method for $\alpha = .01$ and LR method for $\alpha = .05$ have the highest correlation, .95, between the z-statistic and the frequency of DIF detection for four methods using simulation cut score criteria. LR method for $\alpha = .01$ and AM method for $\alpha = .05$ have the lowest correlation, .89, between the z-statistic and the frequency of DIF detection for four methods using simulation cut score criteria. The researcher has hypothesized that the higher the z-score, the higher the frequency, and that the correlation must be high. The results show the correlations are high for the M-H and

SIBTEST methods. The correlations of non-IRT methods are a little higher than the IRT methods; all correlations are high, almost all are higher than .90. Using the magnitude of z-score to replace the frequency of DIF items will yield similar results.

Table 4.14 *The correlation between the z-statistic and the frequency of DIF detection for four methods using simulation cut score criteria*

Method	α	Pearson
AM	0.05	0.89**
	0.01	0.90**
LR	0.05	0.95**
	0.01	0.89**
M-H	0.05	0.93**
	0.01	0.94**
SIBTEST	0.05	0.94**
	0.01	0.95**

**Correlation is significant at the 0.01 level (2-tailed).

Using the critical values of previous research of the four methods

If the frequency of the DIF item is high, it means that the item has a greater possibility to be a DIF item in a real situation. The critical values of previous research seem greater than the critical values of simulated data for the four methods. Therefore, the frequency of a DIF item using the critical values of previous research should be smaller than the frequency of a DIF item using the critical values of simulated data. In addition, this study will investigate the consistency between the two criteria for each method.

Using previous research on the AM method, if α=.05 and α=.01, the criteria of z values is 1.96 and 2.58, respectively. The frequency of DIF items in the 100 duplication samples from real sample is shown in Table 4.15. The results for α=.01 were used because of the large samples in every group: 1000 males and 1000 females. The highest

frequency of DIF detection is 93 for Item 26. The second largest is for Item 21, with a frequency of 43. These are followed by Items 14, 32, and 20 whose frequencies were 34, 27, and 24, respectively.

Using previous research on the Likelihood Ratio method, if α=.05 and α=.01, then the criteria of χ_3^2 values are 7.82 and 11.34, respectively. The largest frequency for a DIF item is 86 for Item 26. The second largest is Item 24, with a frequency of 29. These are followed by Items 7 and 32 with frequencies of 27, and 24, respectively.

Using previous research on the M-H method, if α=.05 and α=.01, the criteria of χ_1^2 values are 3.84 and 6.64, respectively. The largest frequency for a DIF item is 97 for Item 26. The second largest is Item 20, whose frequency is 35. These are followed by Items 22 and 1, whose frequencies are 24 and 21, respectively.

Using previous research on the SIBTEST method, if α=.05 and α=.01, the criteria of *B* values are 1.96 and 2.58, respectively. The largest frequency for a DIF item is 84 for Item 26. The second largest is for Item 20 with a frequency of 31. These are followed by Items 22 and 1, with frequencies of 22 each.

After comparing the frequencies of DIF items in Table 4.9 with those in Table 4.15, the study has found that the frequencies of DIF items for the criteria of simulation data are larger than the frequencies of items for the previous studies. The researcher believe this is because the criteria of simulation data are more restrictive than in the previous studies. Similarly more restrictive criteria for Tables 4.5 and 4.6 result in chi-square values for the simulated data that tend to be less than the degrees of freedom.

Table 4.15 *The frequency of DIF detection in the 100 real data samples among four methods using the criteria from previous research*

	Area-Measure		Likelihood Ratio		Mantel-Haenszel		SIBTEST	
Item	α=.05	α=.01	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01
1	4	2	34	18	43	21	38	22
2	2	0	8	3	16	5	21	7
3	6	2	14	4	27	10	26	9
4	0	0	15	4	14	3	23	13
5	1	0	32	14	3	2	5	3
6	0	0	0	0	8	0	7	2
7	19	8	47	27	20	9	37	18
8	5	0	1	0	7	0	5	0
9	1	0	3	1	11	4	17	3
10	12	6	11	4	7	1	14	1
11	20	4	6	0	16	7	11	1
12	19	11	4	2	16	9	15	7
13	11	2	7	4	9	3	12	1
14	53	34	18	6	19	7	18	9
15	0	0	10	1	18	6	18	8
16	17	5	4	2	3	0	3	1
17	19	7	4	2	17	4	15	4
18	19	6	10	2	7	1	11	2
19	1	0	8	3	33	13	34	14
20	40	24	29	14	57	35	54	31
21	56	43	20	8	35	17	33	15
22	45	30	22	9	43	24	54	22
23	33	17	7	1	10	3	9	2
24	7	3	47	29	8	2	9	2
25	14	7	21	7	25	5	20	7
26	96	93	93	86	100	97	91	84
27	7	2	3	1	10	1	6	1
28	19	11	12	5	6	1	4	0
29	8	4	1	0	2	0	6	1
30	36	14	3	0	11	0	13	1
31	18	3	17	2	29	7	13	5
32	47	27	46	24	20	10	18	7
M	19.84	11.41	17.41	8.84	20.31	9.59	20.63	9.47
SD	21.48	18.62	19.40	16.18	19.55	17.81	18.49	15.59

ANOVA results comparing the results for the four methods for detecting the DIF items, α=.05 and α=.01 are shown in Table 4.16. There is no significant difference among the four methods.

Table 4.16 ANOVA *for the frequency of DIF items in the 100 real data samples among four methods using the criteria from previous research*

	$\alpha = .05$			
Source of Variation	*SS*	*df*	*MS*	*F*
Between Groups	206.02	3	68.67	0.18
Within Groups	48385.59	124	390.21	
Total	48591.62	127		

	$\alpha = .01$			
Source of Variation	SS	df	MS	F
Between Groups	116.59	3	38.86	0.13
Within Groups	36227.63	124	292.16	
Total	36344.22	127		

The correlations among the four methods using the criteria of previous research are shown in Table 4.17. The M-H and SIBTEST methods have the highest correlation, .98 for $\alpha = .01$. The AM and LR have the lowest correlation, .76 for $\alpha = .01$. That is, the correlations between the non-IRT methods were very high and larger than the IRT methods. The result was similar to the empirical study.

Table 4.17 *The correlation among the four methods for the frequency of DIF Items in the 100 real data samples by previous research*

	1	2	3	4
1. AM	-	0.76**	0.85**	0.82**
2. LR	-	-	0.86**	0.86**
3. M-H	-	-	-	0.98**
4. SIBTEST	-	-	-	-

**Correlation is significant at the 0.01 level (2-tailed).

4.2 Synthesis Discussion

Item response models assuming a single latent ability that adequately accounts for examinee test performance are referred to as unidimensional. Of course, this assumption cannot be strictly met because there are always other cognitive, personality, and test-taking factors those impacts on test performance, at least to some extent. DIF is an indicator of violation of unidimensionality. In this study, DIMTEST software is used to assess unidimensionality. All sets whether real or simulation data sets, both approximately corresponded with the assumption of unidimensionality. But Item 26 is still detected as exhibiting DIF. DIMTEST software is used to analyze whether or not data have a dominant first factor. The dominant factor may only explain 25% of the data. There is still more than 75% that cannot be explained by the dominant factor. Therefore, significant DIF between two groups can be expected whenever group distributions on a secondary trait do not mirror those on the primary trait measured by the test.

The study found that Item 26 exhibited significant DIF because the frequencies were 93, 99, 100, and 99 with simulation cut score criteria and 93, 86, 97, and 84 with previous research criteria for AM, LR, M-H, and SIBTEST, α=.01. The results in Table 4.4 and 4.7 show that all the SA and *B* values for Item 26 are positive. That is, Item 26 favors males.

From Table 4.4, the simulated distributions do not seem to be consistent with the expected null hypothesis distributions, which approximate the standard normal distribution. The mean of the 32 simulation items is -0.08, and the standard deviation is .82. There are only 6 items with SDs greater than 1. That is, most of the simulated distributions were concentrated at zero. And the 95% and 99% critical values are smaller than the expected critical values of the null hypothesis distribution. Therefore, there is

frequency discrepancy for DIF items between Tables 4.9 and 4.15. The discrepancy is more apparent if the SD of the simulated distribution is very small. For example, the frequencies of DIF in the two tables for Item 1 using the AM method are 2 and 37 respectively for $\alpha = .01$. This situation occurs also in the other three methods. The expected distributions of LR and M-H methods are approximately χ_3^2 and χ_1^2. But almost all means and SDs of the items of are smaller than expected for the null hypothesis except the mean of Items 1 and 5 of the LR method. Therefore, the discrepancy occurs in Tables 4.9 and 4.14 for Items 3, 11, 14, and 22, $\alpha = .01$, if the SD of the simulated distribution is small and the discrepancy of the mean is large. The similar situation occurs with Items 3, 11, 14, 15, 16, 17, 19, 31, and 32 for the M-H method as well as Item 15, 30, and 32 for the SIBTEST method. In general, the more discrepancy in the SD, the more different the frequency of DIF detection. The extreme example is that of the M-H method. Its mean SD for the real distribution of the M-H method is 2.44 and the mean SD of the simulation distribution for the M-H method is only .63. Similarly, the mean frequency of DIF items decreased greatly from 30.38 to 9.59 for $\alpha = .01$.

Because the simulated distributions do not seem to be consistent with the expected null hypothesis distributions, the simulated distributions appear more concentrated and have a smaller SD. For this reason, all of the Type I error rates of the four DIF methods for $\alpha = .05$ and .01 are smaller than the expected values.

The reason may be that the assumption of the statistics of previous studies is different from the results of this study. For example, the inferential statistics usually assume the distribution of data is standard normal distribution but if the distribution of empirical data is not normal there will be a different result. In Appendix 3, it is shown

that the distributions for boys and girls are not normal distributions. That is, if it is desirable to detect DIF item on the Basic Mathematical Competence Test for Junior High Schools in Taiwan by the four methods, more serious criteria should be adopted than the indicated in previous studies. The correlation between the empirical results and the previous research for the frequency of DIF items in the four methods is shown in Table 4.18. The results show that the LR method for $\alpha = .01$ has the highest correlation, .92, between the empirical results and the previous research; and the M-H method for $\alpha = .01$ has the lowest correlation, .80, between the empirical results and the previous research. In our analysis, the correlation is high between the empirical result and the previous research for the AM, LR, and SIBTEST methods for $\alpha = .01$. That is, the higher correlation indicates that the method has higher consistency between the two criteria for detecting the frequency of DIF items.

Table 4.18 *The correlation between the frequency of DIF detection for the empirical critical values and those based on previous research for the four methods*

Method	α	Pearson
AM	0.05	0.84**
	0.01	0.91**
LR	0.05	0.88**
	0.01	0.92**
M-H	0.05	0.85**
	0.01	0.80**
SIBTEST	0.05	0.88**
	0.01	0.88**

**Correlation is significant at the 0.01 level (2-tailed).

Table 4.19 shows the z-statistic for the four methods. Since a common variance cannot be assumed in this case, z is more difficult to estimate. In this case, the root mean square as an average, within-population standard deviation can be used to standardize the

difference between means. The larger the z-score, the more discrepancy between the observed distribution based on the real data and the distribution based on the simulated data. If the item has a large z-score, it means the item exhibits DIF. By comparing the magnitude of z-statistics for each item for the methods in Table 4.19 the results show that SIBTEST has 12 items whose z-scores are the largest among the four methods, the AM method has 10 items, LR has six items, and M-H has four items. The SIBTEST method also had the largest mean z-score for the thirty-two items, 1.2. The mean z-scores of thirty-two items of the other three methods are 0.96, 0.96, and 0.86 for M-H, AM, and LR methods, respectively. From Table 4.19, AM seems to identify more DIF items then M-H. Both of them have the same mean z-score but M-H has a smaller SD than AM. Therefore, the AM method has more extreme values than M-H method. This explains the fact that AM method has 10 items – 5, 6, 8, 11, 14, 21, 23, 28, 30, 32 whose z-score is the largest among the four methods but M-H only has four items – 16, 27, 29, 31. In fact, from the frequency of DIF items in Table 4.9, M-H has more frequency of DIF detection than the AM method.

If one adopts the criterion that an item is considered to exhibit DIF if all the four z-scores are larger than 1.65, the item is considered as suspect for DIF if all the four z-scores are larger than the mean z-score of the method, and the item is considered as non-DIF if all the four z-scores are smaller than the mean z-score of the method. For example, the four z-scores of Item 2 are .27, .58, .62, and .71. All of the values are smaller than their mean z-score - .96, .86, .96, and 1.2, respectively. Regarding the agreement of the methods, Items 2, 6, 8, 10, 13, 23, 27, and 29 generally do not exhibit significant DIF, Items 3, 20, 21, 22, 31, and 32 have suspect DIF, and Item 26 is

identified as exhibiting significant DIF. Therefore, the total agreement items for no DIF, suspect DIF, and significant DIF are 15. The agreement percentage was only 46.88% (15/32). If the methods are divided into IRT and non-IRT, the agreement of non-IRT is 78.13% (25/32), and the agreement of the IRT method is 68.75% (22/32). Therefore, the agreement between non-IRT methods is larger than IRT methods.

If items detected as DIF adopt the Cohen's rule-of-thumb criteria, d=.80 is a large z-score. Then, Item 1, 3, 20, 21, 22, 26, 31, and 32 will be detected as DIF rather than only Item 26. The criteria z=1.65 was adopted in this study because only 5% probability the z-score will larger than 1.65 in order to avoid many DIF items. It will induce debate for society if there are too many DIF items in the test because the test has been administered already and the results are reported to about 30,000 examinees. In addition, Scheuneman (1987) found that DIF effects were much more complex than originally anticipated, resulting in complex interactions rather than simple main effects between different groups.

Table 4.19 *The z-statistic values for the four methods*

Item	A-M	LR	M-H	SIBTEST
1	1.43	0.83	1.28	1.69
2	0.27	0.58	0.62	0.71
3	1.11	1.28	1.05	1.24
4	0.35	0.39	0.86	1.72
5	1.42	0.94	0.49	0.53
6	0.74	0.47	0.47	0.09
7	0.01	1.72	1	1.42
8	0.67	0.22	0.02	0.2
9	0.75	0.36	0.82	1.51
10	0.23	0.75	0.63	0.7
11	1.19	0.99	0.97	0.72
12	1.13	0.44	0.82	1.23
13	0.46	0.26	0.59	0.63
14	1.51	1.14	0.9	1.26
15	0.02	0.57	1.07	1.67
16	0.83	0.59	0.97	0.4
17	1.12	0.65	0.92	1.46
18	0.23	0.94	0.68	0.79
19	1.35	0.52	1.29	1.74
20	1.6	1.55	1.66	2.68
21	1.98	0.94	1.27	1.87
22	1.79	1.26	1.3	2.46
23	0.79	0.52	0.68	0.43
24	0.13	1.39	0.83	0.52
25	0.03	1.32	0.93	1.14
26	3.78	2.89	3.12	4.67
27	0.38	0.24	0.52	0.04
28	0.98	0.6	0.64	0.71
29	0.21	0.17	0.75	0.52
30	1.35	0.36	0.93	1.11
31	1.17	0.96	1.38	1.17
32	1.89	1.55	1.17	1.36
Min	0.01	0.17	0.02	0.04
Max	3.78	2.89	3.12	4.67
M	0.96	0.86	0.96	1.2
SD	0.78	0.57	0.51	0.9

The consistency of the magnitude of z-statistic between IRT methods and non-IRT methods results in a correlation of .57 between IRT methods and .91 between non-IRT methods. Therefore, non-IRT methods are much more consistent than IRT methods.

The results of ANOVA comparing the four methods for z-scoores are shown in Table 4.20. The results show there is no significant difference at $p < .05$ between the four methods.

Table 4.20 *Summary of* ANOVA *of AM, LR, M-H and SIBTEST methods for z-statistic*

Source of Variation	*SS*	*df*	*MS*	*F*
Between Groups	2.03	3	0.68	1.35
Within Groups	62.11	124	0.50	
Total	64.15	127		

The distribution based on simulated data has smaller variance than the expected sampling distribution under the null hypothesis of no DIF. That is, almost the entire mean and SD of simulation distribution for each item on the four methods are smaller than the mean and SD of the expected sampling distribution. That causes the results of the magnitude of z-statistic in the study to be larger than the expected z-scores between the observed distribution based on real data and the expected sampling distribution based on previous studies. The results are shown in Table 4.21.

Table 4.21 *The z-statistic for the four methods by the expected distribution*

Item	A-M	Likelihood	M-H	SIBTEST
1	0.87	0.68	0.98	1.58
2	0.20	0.07	0.39	0.96
3	0.92	0.28	0.71	1.42
4	0.01	0.08	0.40	1.16
5	0.80	0.65	0.06	0.13
6	0.56	0.45	0.13	0.34
7	0.13	0.93	0.60	1.51
8	0.40	0.44	0.00	0.31
9	0.59	0.34	0.41	1.07
10	0.12	0.13	0.05	0.70
11	1.05	0.04	0.43	0.53
12	0.96	0.25	0.43	0.82
13	0.49	0.11	0.21	0.80
14	1.56	0.42	0.52	1.03
15	0.16	0.02	0.54	1.30
16	0.98	0.11	0.03	0.43
17	0.97	0.09	0.50	1.09
18	0.10	0.04	0.16	0.74
19	0.95	0.17	0.83	1.49
20	1.32	0.70	1.36	2.19
21	2.02	0.34	0.92	1.62
22	1.65	0.38	1.07	2.00
23	0.70	0.00	0.24	0.39
24	0.44	0.91	0.16	0.44
25	0.10	0.38	0.47	0.81
26	3.62	2.32	2.92	3.87
27	0.24	0.41	0.04	0.12
28	1.13	0.26	0.02	0.47
29	0.03	0.36	0.06	0.44
30	1.25	0.11	0.27	0.70
31	1.02	0.30	0.71	1.04
32	1.73	0.95	0.60	0.95
Min	0.01	0.00	0.00	0.12
Max	3.62	2.32	2.92	3.87
M	0.85	0.40	0.51	1.01
SD	0.74	0.44	0.56	0.74

Using Table 4.22, correlations are computed between the frequency of DIF using the criteria in previous studies for 100 real samples and the z-statistic for AM, LR, M-H, and

SIBTEST methods for $\alpha = .05$ and .01. The order of the magnitude of correlations corresponds with the results in Table 4.14. The correlations of non-IRT methods were higher than IRT methods.

Table 4.22 *Correlation between the frequency of DIF in previous research and z-statistic for each method*

Method	α	Pearson
AM	0.05	0.79**
	0.01	0.81**
LR	0.05	0.90**
	0.01	0.84**
M-H	0.05	0.92**
	0.01	0.91**
SIBTEST	0.05	0.94**
	0.01	0.91**

**Correlation is significant at the 0.01 level (2-tailed).

To judge whether an item is biased depends on the support of qualitative and quantitative evidence. In general, DIF detection is the result of statistic analysis and quantitative evidence; DIF is the necessary condition rather than the sufficient condition for identifying an item as biased. There are many circumstances that can cause DIF problems, including instruction, material in textbooks, policy, and the item itself. An item may function differently if it contains content or language that is differentially familiar to subgroups of examinees, or if the item structure or format is differentially familiar to subgroups of examinees.

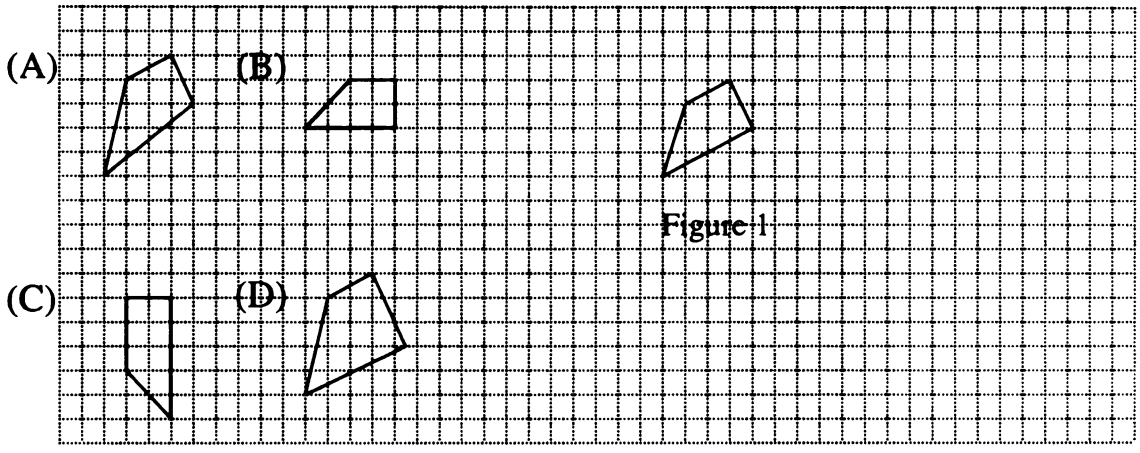
There is a continuing controversy about gender performance differences in mathematics test scores (Ryan & Chiu, 2001; Lane, Wang, & Magone, 1996; Noddings, 1992). The research literature documents many performance differences between males and females. These differences appear in many mathematical fields. They have been

identified by investigating item characteristics that include content, type, and cognitive background of items. For instance, females were found to perform less well on items that measure geometry, computation, and ratio and proportion (Doolittle & Cleary, 1987; Jackson & Braswell, 1992). Application problems, multiple-choice items, and some specific terminology were also probably disadvantageous to females (Harris & Carlton, 1993; O'Neil & McPeek, 1993; Burton, 1995). Items that required solution strategies not taught in class and the real world are more disadvantageous to females (Harris & Carlton, 1993), whereas males are more at a disadvantage when it comes to algebra, calculation related to book content, pure mathematics, short-answer, comparative, and abstract property items (Doolittle & Cleary, 1987; O'Neil & McPeek, 1993; Burton, 1995; Scheuneman & Grima, 1997). For graph or table items, there are no common conclusions (Harris & Carlton, 1993; Scheuneman & Grima, 1997).

Item 26 was detected as exhibiting significant DIF by the four methods. The frequencies were 93, 99, 100, and 99 for AM, LR, M-H, and SIBTEST, respectively, for α=.01. After checking the magnitude of z-statistic for the four methods, the values of Item 26 were all found to be larger than 1.65. In conclusion, Item 26 can be classified as exhibiting DIF. In fact, the magnitude of z-statistic for Items 20, 21, and 22 were close to the criterion, 1.65, except that one or two values were smaller. That is, every item has four z-scores. Items 20, 21, and 22 have two or three values larger than 1.65 and merit further discussion.

Item 26:

Which of the following quadrilaterals is similar to the quadrilateral of Figure 1?



The stem and answer for Item 26 are very clear and impossible to misunderstand. However, it is possible to produce DIF because of the form and content of the item. The question is seldom seen in a textbook, reference book, or practice problem in Taiwan. It is consistent with the research of Doolittle & Cleary (1987) that male high school students perform relatively better than females on geometry items, which usually contains figures. Although Item 26 seems similar to graph problems, it is actually more difficult: the shapes of the quadrilaterals are similar, the directions are different for each of the four answers, the exact proportion of every side is equal, and the length of a side is an irrational number. Reasoning may be used to solve the problem. The second way to solve the problem for some students may be to fold the graph paper in order to find the answer. The third way to solve the item is to use the elimination method. After checking the results, all the SA and *B* values of item 26 were positive. Therefore, the question for Item 26 may be resolved to the advantage of all male samples rather than only by certain samples.

As a result, although the question for Item 26 exhibits DIF, it is not a biased item,

according to expert opinion. The way to avoid items like 26 exhibiting DIF is to change the mathematics instruction to females. As a start, mathematics teachers should use a different mode of instruction in order to change the thinking and learning of female students rather than to delete the item from the test in order to reduce DIF problems. Item 26 should be kept on the test; more inter-gender discrepancies in mathematics ability will be produced if this kind of DIF item is removed. Findings in this study support the claim of Ryan and Fan (1996) that changes in curriculum, instruction, and assessment may be a part of the solution to resolve gender difference in mathematics performance.

Item 20:

A gang of pirates hides three boxes of treasure on Unknown Island. First, they hide a box of treasure at the A place of Unknown Island. Then they walk x km east and 5 km south to arrive at the B place to hide the second box of treasure. Then they go back to the A place and walk 6 km west. They then walk 10 km north to arrive at the C place to hide the third box of treasure. If A, B, and C are exactly on the same line, x=?

(A)3 (B)6 (C)$\frac{25}{3}$ (D)12

Item 20, does not seem to be an example of bias within the context of the discussion with mathematics experts and teachers. The stem and the answer for the item are clear. At least two methods can be used to solve Item 20. One is the traditional way — students have to understand the concept of coordinates and orientation as well as how to match them. Then students have to express the three points — a, b, c — as coordinates and use the concept that if three points are on the same line, they have the same slope. In the second method, the first half of this approach is the same as it is with the first half of the previous method. Then, using the concept of similar triangular angles to solve for x, the

content of Item 20 is related to the direction of the concept. Males usually have a stronger concept of direction than females. Checking the mean of Item 20 for the AM and SIBTEST methods reveals that in both cases the values for males were larger than for females. This item supports the result from the previous item that instruction on similar figures and right triangle geometry is different for the two genders.

Item 21:

If a rectangle whose two edges are bigger than 1 is composed of n squares whose length of sides are equal to 1 and no square is left, then which of the following numbers cannot be n?

(A)81 (B)85 (C)87 (D)89

In Item 21, it is a very easy concept to distinguish which number is a prime number. The major task is to understand the meaning of the question. Then it is necessary to decide whether to use a multiple of 3 or 5, at which point it becomes easy to choose the right answer. A number is a multiple of 5 if the unit number is 0 or 5. A number is a multiple of 3 if the sum of all the digits is a multiple of 3. For example, the sum of 8 and 1 is equal to 9. Therefore, 81 is a multiple of 3. If the above concept is understood and the wrong answer is chosen, there may be a careless understanding of the word "cannot." Males usually are more careless than females. After checking the mean of Item 21 for AM and SIBTEST methods, both the values for females were larger than for males.

Item 22:

Figure 4.1 (below) presents the following mathematical game: Enter from the left side and follow the instruction in the frame to determine the right path. Then where is the final site?

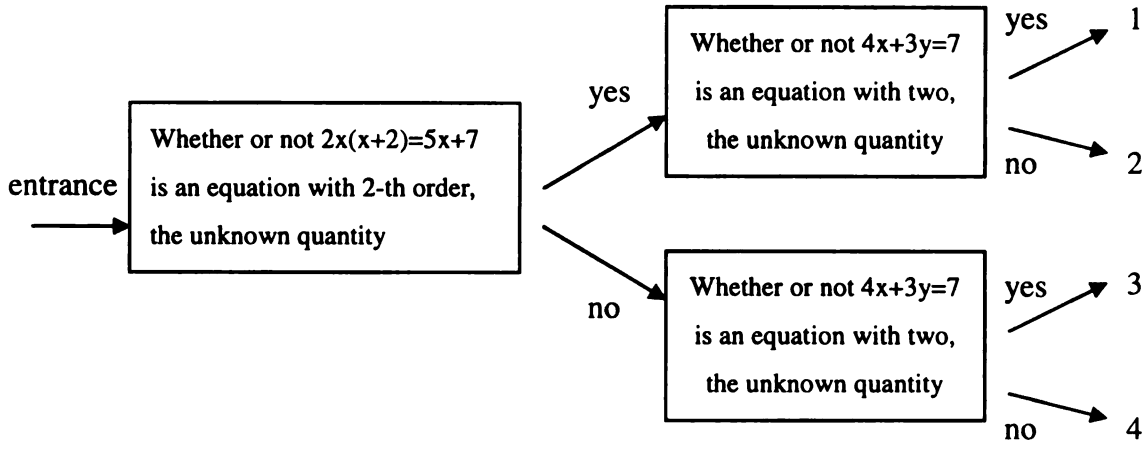


Figure 4.1 the following mathematical game

(A)1 (B)2 (C)3 (D)4

For Item 22, one must know an equation for one squared unknown quantity and an equation with two unknown quantity. It is also important to pay attention to the Chinese character between "one" and "two." "One" and "two" are very similar Chinese characters. The item is also related to "careless." After checking the mean of Item 22 for AM and SIBTEST methods, both the values for females were larger than for males.

Since Item 26 is detected as exhibiting DIF, if it is deleted from the test, the total score mean for 15,411 males will shift from 18.44 to 17.91. Similarly, the 14,465 females will shift from 18.41 to 17.96. In the beginning, the total score mean for males is larger than females. But the result is just opposite after Item 26 is deleted. The reason is because Item 26 is a non-uniform DIF and is advantageous for males.

The test was reviewed by three experts, one mathematics professor and two mathematics teachers from different junior high schools. The professor had 25 years

teaching experience. Both mathematics teachers had taught more than 18 years. Table 4.23 presents the results of items reviewed by them. There was 81.25% (26/32) agreement between the three experts' opinions. This is similarly consistent with Hambleton and Jones's (1994) finding whose agreement is 78.67% (59/75). Items 6 and 9 were suspected of DIF by some of the experts, but they showed no-DIF after statistical analysis. On the other hand, Item 26 was not considered as a biased item by the experts, but it was determined to exhibit DIF after statistical analysis.

Table 4.23 *The results of the items review*

Item	Content	expert_1	expert_2	expert_3	favor
1	Algorithmic	X	X	X	
2	Algebra	X	X	X	
3	Algorithmic	X	X	X	
4	Algebra	X	X	X	
5	Geometry	X	X	X	
6	Algorithmic	O	X	O	male
7	Geometry	X	X	X	
8	Geometry	X	X	X	
9	Algorithmic	X	O	X	male
10	Algebra	X	X	X	
11	Algorithmic	X	X	X	
12	Geometry	X	X	X	
13	Algebra	X	X	X	
14	Geometry	X	O	X	male
15	Algebra	X	X	X	
16	Algebra	X	X	X	
17	Geometry	X	X	X	
18	Algebra	X	X	X	
19	Geometry	X	X	X	
20	Algebra	X	X	O	male
21	Algorithmic	X	X	X	
22	Algebra	X	O	X	male
23	Algebra	X	X	X	
24	Geometry	X	X	X	
25	Geometry	X	X	X	
26	Geometry	X	X	X	
27	Algorithmic	X	X	X	
28	Geometry	X	X	X	
29	Geometry	X	X	X	
30	Geometry	X	X	X	
31	Geometry	X	X	X	
32	Geometry	X	X	O	female

X: no DIF, O: DIF

Item	Reason
6	The content of item is related to basketball. Its relation to sports will be advantageous for males.
9	The content of item is a creative operation, which is advantageous for males.
14	This needs higher thinking including geometry and algebra. It will be advantageous for males.
20	The direction concept is necessary to solve the item. Males usually have the stronger direction concept.
22	The mode of the item is like a video game. It will be advantageous for males.
32	It is related to folding paper. It is usually advantageous for females because females like to fold paper.

A synthesis of the above results indicates that SIBTEST has the biggest mean z-statistic, the highest mean DIF frequency, and the most consistency between the z-statistics and the frequency of DIF detection using the simulation cut score and the previous study. The SIBTEST method was found to be the most appropriate among the four methods for detecting DIF items for the Basic Mathematical Competence Test for Junior High Schools in Taiwan. The study found that the consistency was low among the four methods for detecting DIF items. Although the SIBTEST was the most appropriate in this study, if there were sufficient time and financial resources, of the four non-IRT methods, using both the M-H and SIBTEST methods would likely be recommended to detect DIF items for the Basic Mathematical Competence Test for Junior High Schools in Taiwan. This is consistent with Shealy and Stout's (1993a) recommendation that in the case of DIF item analyses, practitioners use SIBTEST and M-H simultaneously, taking special care to note when the two procedures give different answers regarding the potential presence of bias.

CHAPTER 5

CONCLUSION AND SUGGESTION

The content of this chapter includes two parts — research conclusions and suggestions for future research.

5.1 Research conclusions

The purpose of the study was to evaluate four methods — Area Measure, Likelihood Ratio, Mantel-Haenszel, and SIBTEST — for detecting Differential Item Functioning (DIF) for gender groups for the Basic Mathematical Competence Test for Junior High Schools in Taiwan.

Concrete research questions are as follows:

1. To investigate the fairness for gender groups of the basic mathematical competence test items.
2. To compare the consistency of results for the different methods of detecting DIF.
3. To identify the best method for detecting DIF.
4. To investigate Type I error rate for the different methods of detecting DIF.
5. To investigate item bias in the basic mathematical competence test items.
6. To investigate if DIF is due to instructional differences or to some biasing feature expressed in the results of identified DIF items.

The objects of the study were the 32 items of the mathematics portion of the student's Basic Competence Test for Junior High Schools. Data from 29,876 examinees' were used to evaluate the 32 items. The 29,876 examinees, randomly selected from the full 299,368 examinees who attended the mathematics portion of the student's Basic Competence Test for Junior High Schools, were administered the test in April, 2001. The

tools of research included a. Instrument: the Basic Mathematical Competence Test for Junior High Schools in Taiwan and b. computer software: DIMTEST, BILOG-MG, SAS, SPSS, S-PLUS, MATLAB, and Dimensionality-Based DIF/DBF Package. The following conclusions came from statistical analyses:

1. All the real and simulated data sets met the assumption of unidimensionality.

2. Item 26 had the largest z-scores in the four detecting DIF methods. Thus, Item 26 was considered to exhibit DIF, and Items 20, 21, and 22 were suspected of exhibiting DIF.

3. The SIBTEST method had the largest mean z-score, followed by Mantel-Haenszel, Area Measure, and Likelihood Ratio methods. There was no significant difference among the four methods. The non-IRT methods were much more consistent than the IRT methods.

4. Mantel-Haenszel method had the smallest Type I error rate, followed by SIBTEST, Likelihood Ratio, and Area Measure method. The non-IRT DIF detection methods were better than IRT methods, with a small Type I error rate.

5. When the cut score criteria from the simulated data were adopted, the frequencies for detecting DIF for Items 26 were 93, 99, 100, 99 for the AM, LR, M-H, and SIBTEST methods, respectively, for $\alpha = .01$. Cut score criteria frequency was the evidence for identifying Item 26 as exhibiting DIF.

6. The average of the frequency of DIF on α=.01 and α=.05 for 100 real samples by the criteria of simulation cut score was significant for the four methods. After a posteriori comparison, the frequency of detected DIF for the M-H method was higher than the AM method.

7. The order of correlation between z-score and frequency of DIF using the empirical

study, $\alpha = .01$ and .05, is the same: SIBTEST>M-H>LR>AM. The correlations of non-IRT methods were a little higher than correlations of the IRT methods.

8. When the criteria of previous studies were adopted, the frequencies of detecting DIF for Item 26 were 93, 86, 97, and 84 for the AM, LR, M-H, and SIBTEST methods, respectively — $\alpha = .01$. This, too, was evidence to support Item 26 as exhibiting DIF.

9. The average of the frequency of DIF, α=.01 and α=.05, for 100 real samples using the criteria of previous studies was not significant for the four methods.

10. The order of correlation between z-score and frequency of DIF using previous studies, $\alpha = .01$ and .05, is the same: SIBTEST>M-H>LR>AM. This order of correlation corresponds with the result in conclusion 7, above.

11. The correlation between the non-IRT methods for the frequency of DIF was higher than the IRT methods for the two criteria.

12. The logical analysis of Item 26 by mathematics experts indicated that the identified DIF was not bias.

13. The consistency of the magnitude of z-score was greater for non-IRT methods than for IRT methods.

14. The distribution based on simulated data seems to underestimate the spread of the expected sampling distribution under the null hypothesis of no DIF. The reason may be that the assumption of the statistics of previous studies is different from the results of this study. In this study, the distributions for boys and girls are not normal distribution based on the histograms for males and females. That is, if a researcher wants to detect DIF item in the Basic Mathematical Competence Test for Junior High Schools in Taiwan by the four methods, he or she should adopt more serious criteria than the previous studies. That

caused the results of the magnitude of z-scores in the study to be larger than the expected z-scores between the observed distribution based on real data and the expected sampling distribution based on the previous studies. That also caused the frequency of the DIF item by the criteria of simulated data to be larger than by the criteria of the previous studies.

15. From the results and synthesis discussion of z-scores, frequency, consistency, and Type I error rate, of the four methods, SIBTEST was determined to be the most appropriate to be applied to detect DIF items for the Basic Mathematical Competence Test for Junior High Schools in Taiwan. But if time and financial resources were adequate, the non-IRT methods — M-H and SIBTEST methods together — would be recommended among the four methods to accurately detect DIF for the Basic Mathematical Competence Test for Junior High Schools in Taiwan.

5.2 Suggestion

This section will offer some suggestions for DIF researchers based on this study's results.

5.2.1 The suggestions for test application

It is seldom that there are no items exhibiting DIF in a test. In this study, there was an item that exhibited DIF in the Basic Mathematical Competence Test for Junior High Schools in Taiwan, but there was not 100 % agreement using four methods to detect DIF items. Detection of differential item functioning in the test is essential to ensure that mathematical competence test is measured equally across gender groups. Therefore, the committee for the Basic Competence Test for Junior High Schools has to discreetly examine the process and quality of the test items in order to ensure the items are impartial.

In addition, the committee has to periodically investigate the items to identify any biased items and remove them. Study results show that the Mantel-Haenszel method has the smallest Type I error rate, followed by SIBTEST. SIBTEST has the largest mean z-score, followed by Mantel-Haenszel; and Mantel-Haenszel has the largest frequency of DIF items in the 100 samples of real data among the four methods using the criteria of simulation cut score followed by SIBTEST. Therefore, the non-IRT methods are recommended in order to detect items exhibiting DIF in the Basic Mathematical Competence Test for Junior High Schools in Taiwan. The M-H method is effective for detecting uniform DIF but its statistical power is weak for detecting non-uniform DIF. Cross-SIBTEST can be used to assist in detecting non-uniform DIF.

5.2.2 Future research direction and suggestion

1. In the study, the researcher used an S-Plus program to simulate the data. The distribution of simulated data is not same with the variation of the expected distribution. The future research should try to determine which is accurate.

2. One hundred replications were used in the study. In the future, more replications can be used to investigate the four DIF methods and again compare the simulated distribution and expected distribution.

3. The way to generate simulation data: The simulation data in the study were generated assuming no DIF. In future studies, part of the simulation data should be generated with some level of DIF to compare power for the four detecting DIF methods.

4. Data type: In the study, only dichotomous data was used. In the future, polytomous data to investigate DIF should be used in order to investigate whether the results

demonstrate any discrepancy with a prior study.

5. In the study, only the mathematics test was used. In the future, the four methods can be applied to the other subjects in the Basic Competence Test for Junior High Schools, for example, English or Science, in order to investigate for DIF items in the test.

6. In the study, only gender was considered. In the future, DIF can be investigated for different races, country and city, and social-economic status in the mathematics test.

7. The DIF factor has been neglected in related studies of test design in Taiwan. But the DIF factor is important; for the future, DIF analysis should be emphasized when designing a test.

7. The mathematics content of the test can be divided into three categories: algorithms, algebra, and geometry. To detect the DBF (Differential Bundle Functioning) by the three categories is also a research direction.

APPENDICES

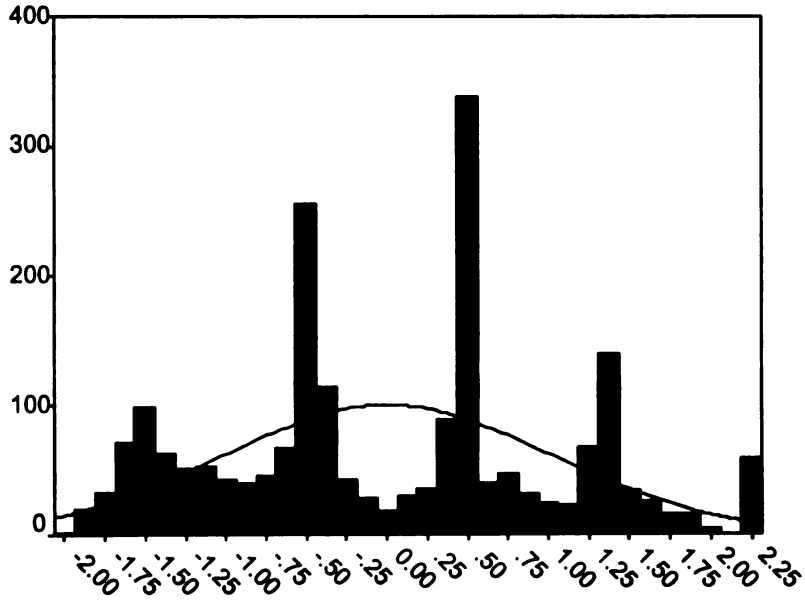
Appendix 1 *Item Parameter Estimates Based on the 4000 Sample*

Item	*a*	*b*	*C*
1	0.91	-0.27	0.46
2	1.57	-0.90	0.27
3	1.24	-0.64	0.22
4	0.98	-1.17	0.20
5	1.21	-0.28	0.47
6	1.09	-0.90	0.18
7	2.05	-0.02	0.33
8	2.06	-0.63	0.16
9	1.18	-0.60	0.24
10	1.53	0.44	0.30
11	1.50	0.20	0.26
12	1.86	0.03	0.15
13	1.76	-0.06	0.26
14	1.55	0.62	0.29
15	0.97	-0.49	0.17
16	0.95	1.12	0.35
17	1.59	0.10	0.22
18	2.10	0.35	0.14
19	0.65	0.09	0.16
20	1.39	0.60	0.19
21	1.41	0.59	0.22
22	1.55	0.48	0.19
23	1.68	0.42	0.27
24	0.95	1.12	0.27
25	1.89	0.12	0.21
26	1.42	0.86	0.33
27	1.07	0.56	0.15
28	1.55	1.07	0.34
29	1.18	0.92	0.16
30	1.27	1.04	0.20
31	1.92	1.46	0.23
32	1.38	1.28	0.27
min	0.65	-1.17	0.14
max	2.10	1.46	0.47
mean	1.42	0.23	0.25
S.D.	0.38	0.70	0.08

Appendix 2 *95% and 99% Confidence Interval for the four methods from simulation data*

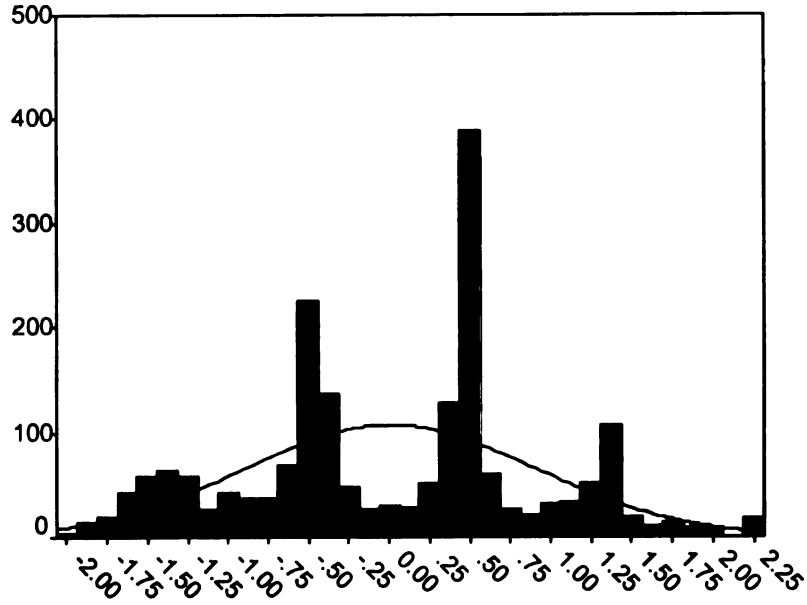
	AM		LH		MH		SIBTEST	
Item	95%CI	99%CI	95%CI	99%CI	95%CI	99%CI	95%CI	99%CI
1	(-.71,.78)	(-.95,1.02)	(0,6.90)	(0,8.78)	(0,1.56)	(0,2.18)	(-1.53,1.39)	(-2.0,1.86)
2	(-1.32,1.36)	(-1.75,1.79)	(0,6.24)	(0,7.30)	(0,2.66)	(0,3.55)	(-1.85,1.18)	(-2.33,1.66)
3	(-1.29,1.22)	(-1.69,1.62)	(0,3.49)	(0,5.39)	(0,2.00)	(0,3.02)	(-1.75,1.10)	(-2.20,1.55)
4	(-.76,1.09)	(-1.05,1.38)	(0,6.31)	(0,7.63)	(0,2.21)	(0,3.01)	(-1.61,.97)	(-2.02,1.38)
5	(-.87,1.05)	(-1.17,1.35)	(0,6.87)	(0,7.79)	(0,1.03)	(0,4.18)	(-1.53,.92)	(-1.91,1.3)
6	(-1.22,1.18)	(-1.60,1.56)	(0,5.75)	(0,6.43)	(0,2.29)	(0,5.06)	(-1.87,1.34)	(-2.38,1.85)
7	(-2.21,1.86)	(-2.85,2.50)	(0,4.94)	(0,8.97)	(0,2.05)	(0,2.94)	(-1.75,1.25)	(-2.22,1.72)
8	(-2.03,1.56)	(-2.6,2.12)	(0,4.67)	(0,5.62)	(0,3.05)	(0,6.20)	(-2.25,1.26)	(-2.81,1.82)
9	(-1.44,1.4)	(-1.88,1.85)	(0,6.45)	(0,8.68)	(0,1.83)	(0,3.92)	(-1.78,1.25)	(-2.26,1.73)
10	(-1.72,1.48)	(-2.23,1.99)	(0,4.93)	(0,7.05)	(0,1.22)	(0,1.86)	(-1.41,1.22)	(-1.83,1.63)
11	(-1.76,1.63)	(-2.30,2.17)	(0,4.17)	(0,5.55)	(0,1.18)	(0,1.87)	(-1.32,1.16)	(-1.71,1.55)
12	(-1.94,1.67)	(-2.51,2.24)	(0,3.41)	(0,6.73)	(0,1.42)	(0,2.62)	(-1.61,1.12)	(-2.04,1.55)
13	(-2.14,2.12)	(-2.82,2.79)	(0,5.06)	(0,7.17)	(0,2.09)	(0,2.92)	(-1.7,1.22)	(-2.16,1.68)
14	(-2.07,1.99)	(-2.72,2.63)	(0,5.15)	(0,6.80)	(0,1.41)	(0,2.24)	(-1.2,1.29)	(-1.59,1.69)
15	(-1.31,1.04)	(-1.68,1.41)	(0,5.05)	(0,6.79)	(0,1.54)	(0,2.13)	(-1.53,1.26)	(-1.97,1.7)
16	(-1.7,1.19)	(-2.16,1.65)	(0,5.71)	(0,8.82)	(0,0.76)	(0,1.11)	(-1.24,1.03)	(-1.6,1.39)
17	(-1.97,1.71)	(-2.55,2.29)	(0,4.27)	(0.6.05)	(0,1.67)	(0,2.12)	(-1.57,1.24)	(-2.02,1.69)
18	(-1.94,1.65)	(-2.5,2.21)	(0,2.58)	(0,5.18)	(0,1.53)	(0,2.84)	(-1.51,1.42)	(-1.97,1.89)
19	(-1.01,.82)	(-1.30,1.11)	(0,3.31)	(0,5.28)	(0,1.14)	(0,3.00)	(-1.21,1.08)	(-1.58,1.44)
20	(-1.91,1.47)	(-2.45,2.0)	(0,4.29)	(0,10.63)	(0,1.61)	(0,2.46)	(-1.40,1.26)	(-1.82,1.68)
21	(-1.98,1.86)	(-2.59,2.47)	(0,5.73)	(0,8.87)	(0,1.48)	(0,2.70)	(-1.40,1.37)	(-1.83,1.81)
22	(-1.74,1.97)	(-2.32,2.56)	(0,3.99)	(0,6.57)	(0,2.34)	(0,4.92)	(-1.28,1.57)	(-1.73,2.02)
23	(-2.49,2.15)	(-3.23,2.89)	(0,5.67)	(0,8.94)	(0,1.92)	(0,3.58)	(-1.43,1.45)	(-1.89,1.91)
24	(-1.92,1.26)	(-2.42,1.76)	(0,7.07)	(0,8.26)	(0,0.95)	(0,1.45)	(-1.24,1.23)	(-1.63,1.62)
25	(-1.93,1.61)	(-2.49,2.17)	(0,3.33)	(0,7.75)	(0,1.61)	(0,3.03)	(-1.54,1.13)	(-1.96,1.55)
26	(-2.17,1.78)	(-2.79,2.41)	(0,5.97)	(0,6.38)	(0,1.42)	(0,1.79)	(-1.11,1.22)	(-1.48,1.59)
27	(-1.75,1.52)	(-2.26,2.04)	(0,3.96)	(0,5.63)	(0,1.56)	(0,2.49)	(-1.34,1.49)	(-1.78,1.93)
28	(-2.17,1.91)	(-2.81,2.55)	(0,7.77)	(0,8.35)	(0,1.20)	(0,2.15)	(-1.12,1.33)	(-1.51,1.72)
29	(-1.81,1.44)	(-2.33,1.95)	(0,4.60)	(0,6.63)	(0,0.89)	(0,1.45)	(-1.32,1.32)	(-1.74,1.74)
30	(-1.77,1.90)	(-2.35,2.48)	(0,5.08)	(0,7.10)	(0,1.25)	(0,2.06)	(-1.03,1.53)	(-1.44,1.94)
31	(-1.58,1.52)	(-2.08,2.01)	(0,5.87)	(0,14.06)	(0,1.05)	(0,1.49)	(-1.08,1.25)	(-1.45,1.62)
32	(-1.58,1.68)	(-2.09,2.20)	(0,5.90)	(0,6.64)	(0,0.94)	(0,1.74)	(-.91,1.24)	(-1.25,1.59)

Appendix 3 *The ability distributions of males and females for 4000 real data*



BOY

S.D.=1.03 Mean=0



GIRL

S.D.=0.90 Mean=0

Appendix 4 *The first Basic Mathematical CompetenceTtest for Junior High Schools for 2001*

1. After calculating $(-\sqrt{\frac{5}{6}})\times\sqrt{\frac{24}{25}}\div(-\sqrt{\frac{3}{5}})$, which one result can you get?

(A) $-\sqrt{\frac{4}{3}}$ (B) $\sqrt{\frac{4}{3}}$ (C) $-\frac{\sqrt{4}}{3}$ (D) $\frac{\sqrt{4}}{3}$

2. Sho-Sho goes to post-office and buys two kinds of stamps that are 5 dollars and 12 dollars, respectively. The total number of stamps is 29 and stamps cost 250 dollars. If there is x pieces for 5 dollars and y pieces for 12 dollars. Which one of the following linear combination equation is correct?

(A) $\begin{cases} x+y=250 \\ 5x+12y=29 \end{cases}$ (B) $\begin{cases} x+y=29 \\ 5x+12y=250 \end{cases}$

(C) $\begin{cases} x+y=250 \\ 12x+5y=29 \end{cases}$ (D) $\begin{cases} x+y=29 \\ 12x+5y=250 \end{cases}$

3. After rounding off $\sqrt{x}$ and $\sqrt[3]{10x}$, choosing the approximate values to the first digit of decimal are 7.5 and 8.3. Then, how much is it for x?

Tables of power and extraction of a root

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N^3	$\sqrt[3]{N}$	$\sqrt[3]{10N}$	$\sqrt[3]{100N}$
55	3025	7.416198	23.45208	166375	3.802952	8.193213	17.65174
56	3136	7.483315	23.66432	175616	3.825862	8.242571	17.75808
57	3249	7.549834	23.87467	185193	3.848501	8.291344	17.86316
58	3364	7.615773	24.08319	195112	3.870877	8.339551	17.96702

(A)55 (B)56 (C)57 (D)58

4. If a store has a promotion activity, it costs 105 dollars to buy 3 packages of cookie and 2 breads. How much can Xiao-Fen get back, if she uses 500 dollars to buy 6 packages of cookie and 4 breads?

(A) 290 (B) 395 (C) 105 (D) 210

5. For Figure 1, $\overline{BC}$ is cut to equivalent 4 sections by D, E, F in $\triangle ABC$. $\overline{AG} : \overline{AC} = 1 : 3$, H is midpoint of $\overline{AB}$. Which point is the center of gravity of $\triangle ABC$?

(A)X (B)Y

(C)Z (D)W

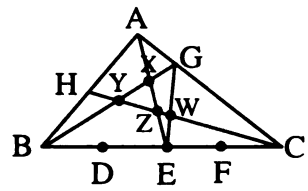


Figure 1

6. Chuang-Chuang shot 10 balls and got 7 as well as Shou-Shou shot 20 balls and got 14 in a basketball game. Which statement is wrong?

(A) The ratio between Chuang-Chuang got and shot is 7 : 10

(B) The ratio values between Shou-Shou got and shot is $\frac{14}{20}$

(C) $\because 7 : 10 = 7\times2 : 10\times2 = 14 : 20$, then both of them have the same shooting average.

(D) $\because$ Chuang-Chuang got 7 and Shou-Shou got 14, Shou-Shou has higher shooting average.

7. For Figure 2, ABCD is a rectangle. If the coordinate of A is (- 2 , 3), B is (- 2 , - 3) and D is (4 , 3), then which option is the equation of line BC?

(A) y - 3 = 0 (B) y + 3 = 0

(C) x - 1 = 0 (D) x - 4 = 0

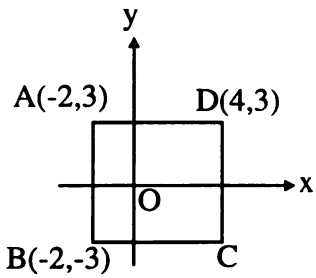
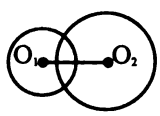


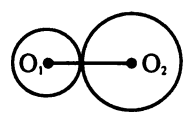
Figure 2

8. If the radius of O_1 and O_2 is 2 centimeters and 4 centimeters in a plane as well as $\overline{O_1O_2} = 7$ centimeters. Then, which graph indicates the position relationship between O_1 and O_2?

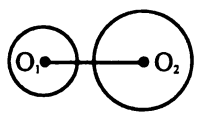
(A)



(B)



(C)



(D)

9. If $\ulcorner\oplus\lrcorner$ is a new operation symbol for use with 1 and 0, the rules are as follows:

$1\oplus1=0$

$1\oplus0=1$

$0\oplus1=1$

$0\oplus0=0$

Then, which application of the operation is correct?

(A) $(1\oplus1)\oplus0=1$

(B) $(1\oplus0)\oplus1=0$

(C) $(0\oplus1)\oplus1=1$

(D) $(1\oplus1)\oplus1=0$

10. Which is the solution of $91x^2 - 53x + 6 = 0$?

(A) $-\frac{2}{7}$ (B) $-\frac{3}{7}$ (C) $\frac{2}{13}$ (D) $\frac{3}{13}$

11. a is a natural number. Its positive factors are 1、2、4、7、14、28. Then which number is the highest common factor of a and 210?

(A) 4 (B) 7 (C) 14 (D) 28

12. For Figure 3, line L1 parallels line L2. If $\angle1 = 80°$, $\angle2 = 60°$ and $\overline{BO}$ bisect$\angle$DBC, then $\angle3$ = ?

(A)10°

(B)15°

(C)20°

(D)25°

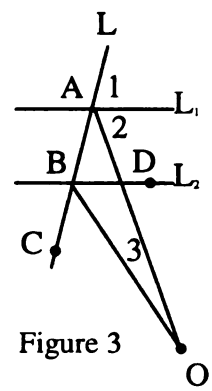


Figure 3

13. If $y = 2x^2 + 1$ and $y = 2x^2 - 1$ are drawn on the same coordinate plane, which statement is <u>wrong</u> about the relationship between the two function graphs?

(A) with the same open direction (B) both graphs are parabola

(C) with the same apex coordinate (D) with the same symmetrical axis

14. If Figure 4 is composed of 1 square with edge length a and 4 squares with edge length b (b>a), rectangle is composed of $\overline{AB}$、$\overline{BC}$、$\overline{CD}$、$\overline{AD}$. How much the area of rectangle ABCD?

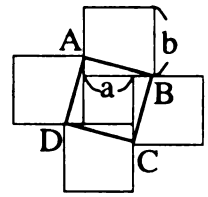


Figure 4

(A) $b^2 + (b - a)^2$

(B) $b^2 + a^2$

(C) $(b + a)^2$

(D) $a^2 + 2ab$

15. Every cake has the same weight and every candy also has the same weight. Shou-Shou takes a weighing scale to measure the weight of cake and candy. Get the results as follows :

First: For Figure 5, the scale is balance if putting two cakes in left hand and putting three candies in the right hand.

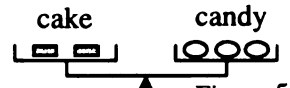


Figure 5

Second: For Figure 6, the scale is balance if putting 10 grams weight in left hand and putting a cake and a candy in the right hand.

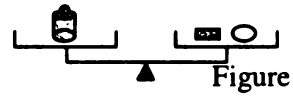


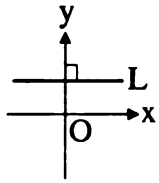
Figure 6

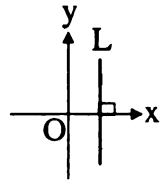
Third: Which way can make the scale balance again if a candy has been put in the left hand and a cake has been put in right hand?

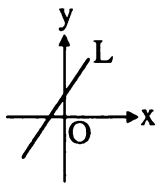
(A) adding 2 grams weight in the left hand (B) adding 2 grams weight in the right hand

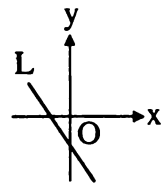
(C) adding 5 grams weight in the left hand (D) adding 5 grams weight in the right hand

16. If ab>0, which one is the graph of $x + ay = b$?

(A)


(B)


(C)


(D)


17. For Figure 7, if the line CD is the perpendicular bisector of $\overline{AB}$ and intersect $\overline{AB}$ on D, then which statement is wrong?

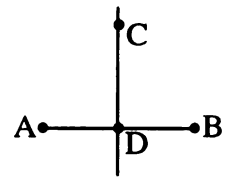


Figure 7

(A) If using C as the center of a circle and $\overline{CB}$ as the radius to draw a circle, then the circle should pass A.

(B) If using A as the center of a circle and $\overline{AB}$ as the radius to draw a circle, then the circle should pass C.

(C) If using B as the center of a circle and $\overline{AC}$ as the radius to draw a circle, then the circle should pass C.

(D) If using D as the center of a circle and $\overline{AD}$ as the radius to draw a circle, then the circle should pass B.

18. For Figure 8, if the graph of $y = x^2$ is shifted two units to right hand, then which one quadratic equation can be indicated the dotted line?

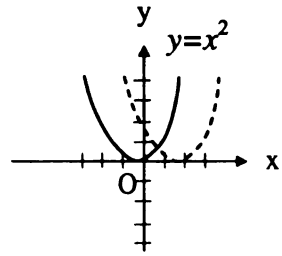


Figure8

(A) $y = x^2 + 2$

(B) $y = x^2 - 2$

(C) $y = (x + 2)^2$

(D) $y = (x - 2)^2$

19. On coordinate plane, which one point has the smallest distance with x-axis?

(A) (1 , 3) (B) (5 , - 2) (C) (- 3 , 5) (D) (0 , - 4)

20. A gang of pirates hides three blocks of treasure on Unknown Island. First, they hide a block of treasure at the A place of Unknown Island. Then they walk x km east and 5 km south to arrive at the B place to hide the second block of treasure. Then they go back to the A place and walk 6 km west. They then walk 10 km north to arrive at the C place to hide the third block of treasure. If A, B, and C are exactly on the same line, x=?

(A) 3 (B) 6 (C) $\frac{25}{3}$ (D) 12

21. If a rectangle whose two edges are bigger than 1 is composed of n squares whose length of sides are equal to 1 and no square is left, then which of the following numbers cannot be n?

(A) 81 (B) 85 (C) 87 (D) 89

22. Figure 9 (below) presents the following mathematical game: Enter from the left side and follow the instruction in the frame to determine the right path. Then where is the final site?

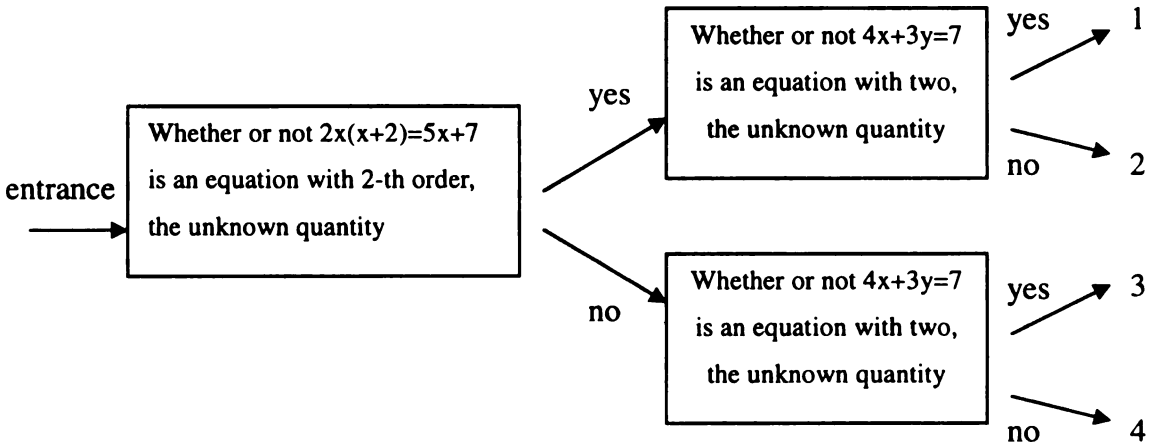


Figure 9

(A)1 (B)2 (C)3 (D)4

23. For Figure 10, there are 4 kinds of rectangle – A, B, C, and D. If every side of the 4 kinds of rectangle is positive integer and there are 2A, 1B, 2C, and 1D. If a big rectangle is composed of the 6 rectangles, how long is it for the two adjacent side of the big rectangle?

(A) 2x + 1 , x + b

(B) 2x + b , x + 1

(C) x + 2b , 2x + 1

(D) x + 1 , 2x + 2b

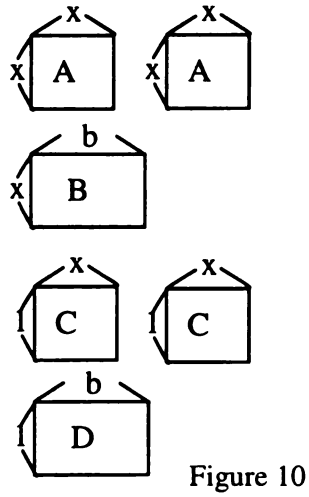


Figure 10

24. For Figure 11, in △ABC, ∠ACB = 90°and $\overline{BC}$ > $\overline{AC}$.

To do: a circle with the center O on $\overline{AB}$ and is tangent with $\overline{AC}$ and $\overline{BC}$. Which one is the correct drawing method in order to get the center O of a circle?

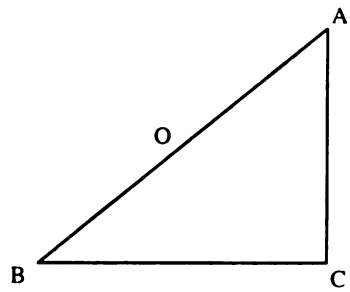


Figure 11

(A) choose the midpoint of $\overline{AB}$ as O

(B) make the perpendicular bisector of $\overline{AC}$ and intersect $\overline{AB}$ in O

(C) make the perpendicular bisector of $\overline{BC}$ and intersect $\overline{AB}$ in O

(D) make the bisector of ∠ACB and intersect $\overline{AB}$ in O

25. For Figure 12, ABCD is a square and A on the line L, $\overline{DE} \perp L$, $\overline{BF} \perp L$, perpendicular points are E、F($\overline{AE} \neq \overline{AF}$) respectively.

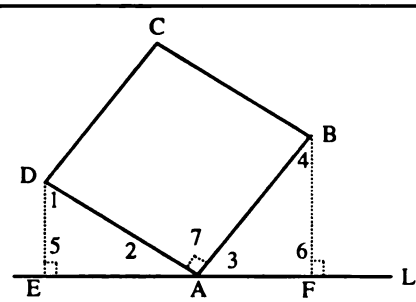


Figure 12

To prove : △ADE≅△BAF

Proof : 1.∵ABCD is a square, ∴$\overline{AB}$ = $\overline{AD}$, ∠7 = 90°

2.∵$\overline{DE} \perp L$, $\overline{BF} \perp L$, ∴∠5 = ∠6 = 90°

3._____(P)_____

4. ∴△ADE≅△BAF

Which one is the correct process and can be fill in (P).

(A) ∵$\overline{DE} \perp L$, $\overline{BF} \perp L$, ∠7 = 90° , ∴$\overline{DE}$ = $\overline{BF}$

(B) ∵$\overline{DE} \perp L$, $\overline{BF} \perp L$, ∠7 = 90° , ∴∠1 = ∠4

(C) ∵∠7 = 90° , ∠5 = ∠6 = 90°∴∠2 = ∠3

(D) ∵∠7 = ∠5 = 90° , ∴∠1 + ∠2 = ∠2 + ∠3 , ∴∠1 = ∠3

26. Which of the following quadrilaterals is similar to the quadrilateral of Figure 13?

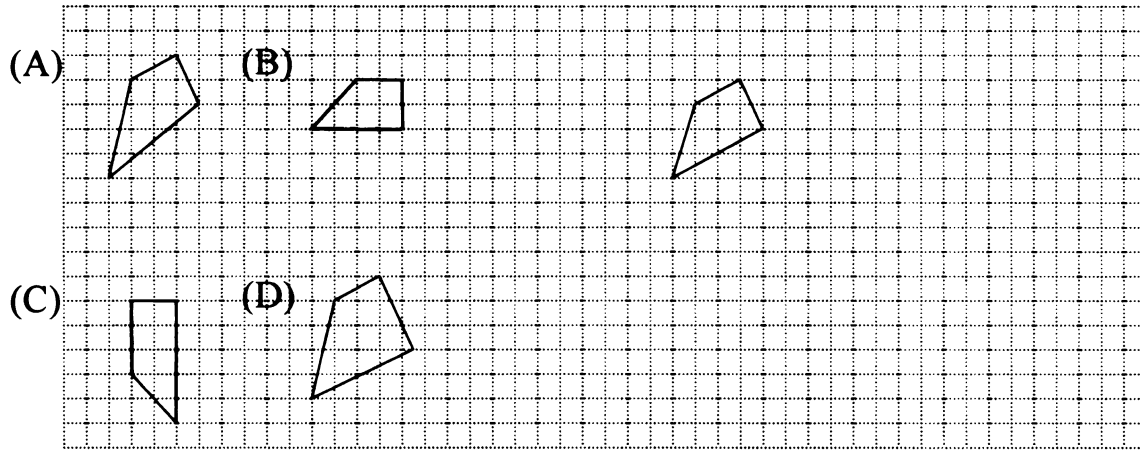


Figure 13

27. A series number a_1 , a_2 , $\cdots$, a_{100} have equal difference. If $a_{70}-a_{57}<0$, then which one is correct?

(A) $a_{43}-a_{69}>0$

(B) $a_{42}-a_{51}<0$

(C) $a_{18}+a_{51}>a_{21}+a_{48}$

(D) $a_{12}+a_{31}>a_9+a_{34}$

28. For Figure 14, Mei-Mei landscape designing company designs a rectangle garden whose longer side is 16 meters and shorter side is 12 meters. There is S area ($\triangle$ ABC is an isosceles triangle) designed as view point and T area (rectangle) designed as a pedestrian precinct inside the garden. Then the left area is 141 square meters and belongs to flowers and plants area. How long is it for the width ($\overline{EF}$) of T area?

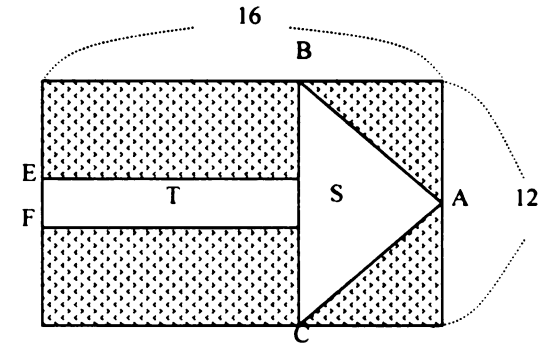


Figure 14

(A) 1 (B) $\frac{3}{2}$ (C) 2 (D) $\frac{5}{2}$

29. For Figure 15, PH is the perpendicular bisector of $\triangle PQR$, $\overline{PQ} \neq \overline{RQ}$, and M is the midpoint of $\overline{PQ}$. Which one statement is wrong?

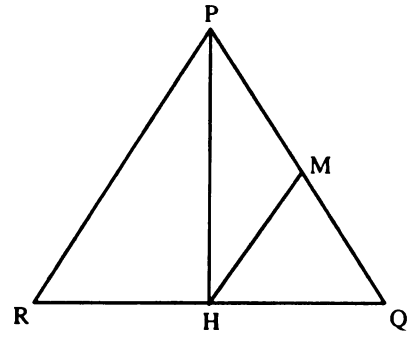


Figure 15

(A) $\overline{MH} = \overline{HQ}$

(B) $\overline{MH} // \overline{PR}$

(C) $\overline{MH} = \overline{MP}$

(D) $\triangle PQH \cong \triangle PRH$

30. For Figure 16, $\overline{AB}$、$\overline{CD}$ is the two diameters of circle O. If $\angle ACD = 2\angle AOC$ and the radius of circle O is 30 centimeter. How long is the arc BC of the opposite $\angle BOC$?

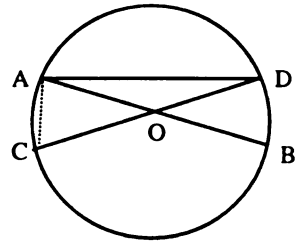


Figure 16

(A) 10π (B) 12π (C) 20π (D) 24π

31. For Figure 17, $\overline{AB}$ is the diameter of circle O, $\overline{BC}$ is the tangent line passing B, and D is on the arc of AB.

To do: choose P from $\overline{BC}$ in order to let $\overline{AP}$ divide equally the area of $\triangle ABC$.

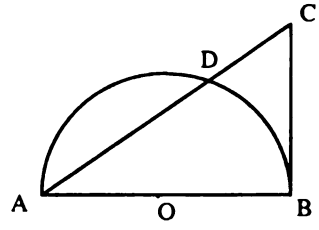


Figure 17

As following four drawing methods, which one is wrong?

(A) choose the midpoint of $\overline{BC}$, P, and connect $\overline{AP}$

(B) make the bisector of $\angle A$ and intersect $\overline{BC}$ in P

(C) make the perpendicular bisector of $\overline{BD}$, intersect $\overline{BC}$ in P, and connect $\overline{AP}$

(D) make a line pass O and parallel $\overline{AC}$, intersect $\overline{BC}$ in P, and correct $\overline{AP}$

32. For Figure 18, △ABC is an isosceles triangle. $\overline{AB} = \overline{AC} = 13$, $\overline{BC} = 10$

(1) Fold up $\overline{AB}$ to the direction of $\overline{AC}$, make $\overline{AB}$ and $\overline{AC}$ overlap, and appear folded line $\overline{AD}$, as Figure 19.

(2) Fold up $\overline{CD}$ to the direction of $\overline{AC}$, as Figure 20, match $\overline{CD}$ and $\overline{AC}$ together, and appear folded line $\overline{CE}$, as Figure 21.

Then, how much is it for the area of △AEC?

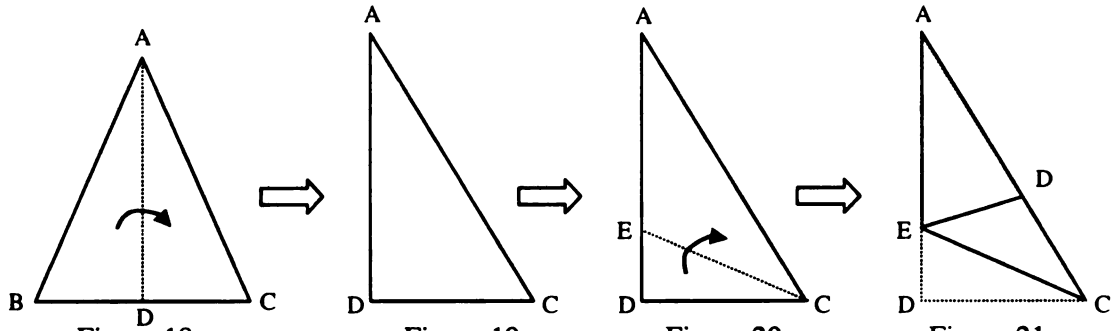


Figure 18 Figure 19 Figure 20 Figure 21

(A) 15 (B) $\frac{65}{4}$ (C) 20 (D) $\frac{65}{3}$

REFERENCE

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

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED069686)

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential Item Function* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-105.

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement, 14*, 139-150.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model.* (Research Bulletin 81-20). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems.* (Series Report No. 58-16). Randolph Air Force Base, Texas: USAF School of Aviation Medicine. (Project No. 7755-23)

Birnbaum, A. (1958a). *On the estimation of mental ability.* (Series Report No. 15). Randolph Air Force Base, Texas: USAF School of Aviation Medicine. (Project No. 7755-23)

Birnbaum, A. (1958b). *Further considerations of efficiency in tests of a mental ability.* (Technical Report No. 17). Randolph Air Force Base, Texas: USAF School of Aviation Medicine. (Project No. 7755-23)

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. P. Novick (Ed.), *Statistical theories of mental test scores* (pp. 397-422). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443-459.

Burton, N. (1995, April). *How have the changes in the SAT affected women's mathematics performances?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

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

Chen, M. C. (1996). *A study on item biases of ability tests.* Unpublished doctoral dissertation, Taiwan Normal University, Taipei, Taiwan.

Chen, Z. X. (2003). Does constructivist teaching cause students' mathematical capacities to decline? *Secondary, 54*(6), 136-149.

Chien, M. F., Liu, H. C., Sheu, T. W., Kuo, B. C, & Yin, C. W. (1995). Factors that influence the use of Mantel-Haenszel procedure to detect Differential Item Function. *Psychological Testing, 46*, 33-44.

Cleary, T. A. and Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28*, 61-75.

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

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

Cochran, W. G (1954). Some methods for strengthening the common χ^2 test. *Biometrics, 10*, 417-451.

Cohen, A. S., & Kim, S. (1993). A comparison of Lord's χ^2 and Raju's area measures on detection of DIF. *Applied Psychological Measurement, 17*, 39-52.

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik, M. E. (1988). *An introduction to tests and measurement*. Mountain View, CA: Mayfield.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

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

Dorans, N. J., & Kulick, E.(1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test, *Journal of Educational Measurement, 23*, 255-268.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Function* (pp35-66). Hillsdale, NJ: Lawrence Erlbaum.

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

DuBois, P. H. (1970). *A history of psychological testing.* Boston, MA: Allyn and Bacon.

Freedle, R., & Kostin, I. (1988). *Relationship between item characteristics and an index of differential item functioning for the four GRE verbal item types.* (Report No. 85-3P). Princeton, NJ: Educational Testing Service.

Gibson, S. G., & Harvey, R. J. (2003). Gender and Ethnicity based differential item functioning on the Armed Services Vocational Aptitude Battery. Equal *Opportunity International, 22*(4), 1-15.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999, June). *Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications.* Paper presented at Annual Meeting of the Canadian Society for the study of Education, Sherbrooke, Quebec, Canada.

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

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement* (pp. 147-200). New York: Macmillan.

Hambleton, R. K. & Jones, R. W. (1994). Comparison of empirical and judgmental procedures for detecting Differential Item Functioning. *Educational Research Quarterly, 18*(1), 21-36.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

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

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Englewood Cliffs, NJ: Lawrence Erlbaum.

Huang, T. W. (1999). *From multiple scoring to investigate the DIF of mathematic achievement test of junior high school.* Unpublished master dissertation, Changhua Normal Universality, Changhua, Taiwan.

Huang, T. W., & Li, H. H. (1999). Gender DIF/DBF analysis: application of Poly-SIBTEST. *Psychological Testing, 46*, 45-60.

Ironson, G. H. (1977). *A comparative study of several methods of assessing item bias.* Unpublished doctoral dissertation, University of Wisconsin, Madison.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. Berk (Ed.), *Handbook of Methods for Detecting Test Bias*. Baltimore, MD: The Johns Hopkins University Press.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), *Application of item response theory* (pp. 155-174). Vancouver, BC: Educational Research Institute of British Columbia.

Ironson, G. H., & Subkoviak, M. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16*, 209-225.

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

Jensen, A. R. (1968). Social class, race, and genetics: Implications for education. *American Educational Research Journal, 5*, 1-42.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach.* San Diego, CA: Harcourt Brace Jovanovich.

Kim, S., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15*, 269-278.

Kim, S., & Cohen, A. S. (1995). A comparison of Lord chi-square, Raju area measures and the likelihood ratio test in detection of differential item functioning. *Applied Measurement in Education, 8*(4), 291-12.

Kim, S., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*(4), 345-355.

Landis, R. J., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review, 46*, 237-254.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance. *Applied Measurement in Education, 8*(4), 313-340.

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

Law, A. M., & Kelton, W. D. (2000). *Simulation modeling and analysis*. Boston: McGraw Hill.

Lazarsfeld, R. F., & Henry, N. W. (1968). *Latent structure analysis*. New York: Houghton Mifflin.

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

Li, Y. H., & Yang, Y. N. (1999). Algorithms for finding equating coefficients for mixed-format tests. *Psychological Testing, 46*(1), 129-140.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters for estimating differential item functioning. *Journal of Applied Psychology, 75*, 164-174.

Liu, Y. L. (2004). The social function and effects of testing. *Secondary Education, 55*(2), 136-154.

Lord, F. M. (1952). *A theory of test scores*. Psychometric Monograph, No.7. New York: Psychometric Society.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lu, S. (1996). *The relationship between item statistics and the Mantel-Haenszel and Standardization DIF statistics when comparison groups differ in ability.* Unpublished Ph.D. Dissertation. The University of Iowa, Iowa City.

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

Lu, S. M. (1999). An overview of procedures for identifying Differential Item Functioning. *Taipei municipal teachers college academic journal, 30,* 149-166.

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Education and Psychological Measurement, 61*(5), 793-817.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690-700.

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

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Merz, W. R., and Grossen, N. F. (1979, April). *An empirical investigation of six methods for examine test item bias.* Paper presented at the annul meeting of the National Council on Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service No. ED 178566)

Millsap, R. E., & Everson, H. T. (1993). Methodology review: statistical approaches for assessing Measurement Bias. *Applied Psychological Measurement, 17,* 297-334.

Nandakumar, R. (1994). Assessing dimensionality of a set of items – Comparison of different approaches, *Journal of Educational Measurement, 31,* 17-35.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18,* 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20,* 257-274.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purpose of statistical inference: Part I and Part II. *Biometrika, 18,* 105-117.

Nodding, N. (1992). Variability: A pernicious hypothesis. *Review of Educational Research, 62,* 85-88.

O'Neil, 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 Function* (pp255-279). Hillsdale, NJ: Lawrence Erlbaum.

Raju, N. S. (1988). The area between two item characteristics curves. *Psychometrika, 54,* 495-502.

Raju, N. S. (1990). Determining the significance of estimated sign and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197-207.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353-368.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-116.

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

Ruder, L. M. (1977, April). *An approach to biased item identification using latent trait measurement theory.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Ruder, L. M., & Convey, J. J. (1978, March). *An evaluation of select approaches for biased item identification.* Paper presented at the annual Meeting of the American Educational Research Association, New York.

Ruder, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17,* 1-10.

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

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

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*, 143-152.

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. & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examinees. *Applied Measurement in Education, 10*, 299-319.

Shealy, R. & Stout, W. F. (1993a). A model-biased standardization approach that separates true bias/DIF from group ability differences and detects test/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shealy, R., & Stout, W. F. (1993b). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential Item Function* (pp197-239). Hillsdale, NJ: Lawrence Erlbaum.

Shen, T. Y. (2003). The change of senior high school admission system in Taiwan. *Secondary Education, 54*(1), 144-156.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimentionality, *Psychometric, 52*, 589-617.

Stout, W., & Roussos, L. (1995). *SIBTEST user manual.* Urbana, IL: University of Illinois.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21*, 49-58.

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

Tai, L. H. (1994). *A study on item bias of joint college entrance examination in Taiwan – The comparison between ICC approach and Manrel-Haenszel Method.* Unpublished master dissertation, Taiwan Normal Universality, Taipei.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147-169). Hillsdale NJ: Lawrence Erlbaum.

Thissen, D. and Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Function* (pp67-113). Hillsdale, NJ: Lawrence Erlbaum.

Walstad, W. B., & Robson, D. (1997). Difference item functioning and male-female differences on multiple-choice tests in Economics. *Journal of Economics Education, 28*(2), 155-171.

Williams, R. L. (1971). Abuses and misuse in testing black children. *The Counseling Psychologist, 2*, 62-77.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA.

Wu, Y. I., Houng, B. C., Shu, C. S., Chen, U. C., & Chen, L. I. (1994). *The compile report of scholastic aptitude test of senior elementary school.* Tainan: Tainan Teachers College.

Wu, B. L., & Xie, M. C. (2001). From the view of the Basic Competence Test of Junior High School to see the teaching materials and instructional methods on mathematics of the new 1-9 curriculum. *Journal of Education Research, 86*, 77-89.

Ye, C. F. (2003). The linking problems between the new and old curriculum. *The Educator Monthly, 2003*(11), 32-34.

Yu, M. L. (1993). Introduction of Item Response Theory (11). *Inservice Education Bulletin, 10*(4), 9-13.

Yu, M. L. (1997). The development trend of measurement theory. In Chinese Testing Institution (Ed), *The development and application of psychometrics – Thesis collection for 60 anniversary of Chinese testing institution* (pp23-62). Taipei: Psychology.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software.