Chapter 1 Introduction

The entrance exam had been conducted for the past 40's years in Taiwan in order to select appropriate students to enter high school in Taiwan. The entrance exam 2001 year was replaced by the basic competence test for junior high school students. This was the first time the basic competence test for junior high school students was put into practice. However, how is the performance of students and how is the quality of the items? The issue is substantive for the junior schools, students and their parents. The purpose of the study is to investigate the performance and differential item functioning by gender in the subject of mathematics.

The test was derived from the ancient times in China. Its purpose was to choose the elitist from among the people. Even today, it is still used throughout the entire world. Why would such a system of testing be able to last so long? It is because the test is an impartial, fair, and a public system. The development of the system has resulted in an improved economic and political system in Taiwan, which is spoken favorably about by the entire world. Why does Taiwan have these unrivalled results produced by the test system? The reason is because of prevalent education. Prevalent education enhances the knowledge of citizens, motivation of achievement, and research development. Therefore, it induces the advantageous condition of politics, economics, and social development.

Many problems, however, have arisen in the education development over the past years; because teachers' instruction of students is determined by the material related the test. Students just recite the

material of books and do not know how to apply the knowledge to real life experiences, also causing the creativeness of students to decrease day by day. Due to the above shortcomings of the test system the Minister of Education has begun to concentrate on the education reform in recent years. The entrance examination is a very important part in the education reform. The high school entrance examination has been ministered throughout schools for around fifty years. It was replaced by the basic competence test for junior high school students 2001 year. Therefore, the parents of students, teachers, and related educators are very concern about the education reform. The basic competence test for junior high school students is a very important breakthrough in education because it adopts the concept of item response theory (IRT). There are many topics of IRT that have been investigated, such as building of item banks, calibration of items and examinee's ability, setting standards, equating, and differential item functioning (DIF). But is the student's basic competence test for junior high school fair and impartial for every student? This is a very important issue. In the study, I want to investigate the gender differences in performance on mathematics achievement items and DIF of gender relating to this system of testing. The results of this study can be an important reference for the institution of basic competence test for junior high school students, parents, teachers, and educators.

1.1 The properties of question

The mathematics achievement difference between genders is the subject investigated by many mathematics education experts. I want to divide the mathematic test into three categories by the content. In the study, I will investigate the difference of mathematics achievement performance between genders in the three categories and total scores and compare to the past studies.

The research of item bias can be traced back to 1905. A. Binet & T. Simon administered the original version of the intelligent test. They found there is significant difference between children of working class backgrounds compare to those from middle class families. From that time on, cultural bias has always become a factor in research.

Since the late of 1960's, American society experienced the rise of women's liberation and the civil rights movement. Since that time, all of the entrance examination, diploma grant, employment, and personnel setting selection depended on the results of tests in order to achieve equality and fairness for groups. Therefore, America education practitioners are especially interested in the difference of test results of 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 test were advantageous to the middle class whites. Freedle & Kostin (1988) researched whether the items of test exited item differential functioning for different groups. Their findings established that the items were advantageous to whites. The problems arising from item bias are still concerned by the America society.

Differential item functioning (DIF) means the performance

difference between two groups of comparable ability or performance. (Dorans & Holland,1993) The appearance of DIF indicates item probably tests to determine unrelated construct factor which we want to measure. It will have a disadvantageous influence on the validity of item. This is just a statistical observation. DIF is different from the item bias, as distinguished by experts, and is unfair for examinee groups (ex. sex or race).

In the past decades, Taiwan just Wu, Houng, Hsu, Chien, & Chien (1994), Tai (1994), Chien, Liu, Hseu, Kuo, & Yin (1995), Chen (1996), Huang (1999), & Huang & Li (1999) used practical data to study DIF. Thus, it is an imperative issue to study, i.e. whether the ability test items are fair with regards to differences of gender, race, geographic location, and socio economic status in Taiwan. Chien, Liu, Hseu, Kuo, & Yin (1995) pointed out test designers have to beef up the analysis of DIF in designing the test.

Mathematics test is administered to measure the examinees' mathematics ability. It will be disadvantageous to the examinees who have a poor reading ability and socio economic status if the influential factors of mathematics test scores include mathematic ability, reading ability, and cultural difference of examinees. Similarly, if the test scores are used as the index of mathematics attitude to discriminate against examinees.

The detecting of DIF has been included in the item analysis in the test practitioners of America. Why does the detecting of DIF become a part of item analysis? The purpose of the study is to objectively select out the item bias in order to reach fairness for examinees of different background. Also, we can accumulate the data and experiences of analyzing DIF for future reference.

1.2 Research motivation and goals

We find the female performance is better than the male in the elementary school period by the studies relating the performance of mathematics achievement. But the results are just opposite when the students are in the high school or higher levels. I am curious how about the performance of mathematics achievement between genders in the junior high school period in Taiwan. So I want to compare the performance of mathematics achievement between genders by the outcomes of the test of basic competence test for junior high school students.

In the item bias study, the methods of detecting DIF are variable. The earliest method of detecting DIF is "transformed item difficulty method" provided by Angoff (1972). The method is to compare the correct answer probability of two groups. Because transformed item difficulty method only manipulate the item difficulty, it cannot observe the relationship between item bias and discrimination (Merz & Grossen, 1979). Many researchers have found the transformed item difficulty method was 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).

χ^2 test procedure is developed by Scheuneman (1979). It is similar to the method of using IRT as a base measure. χ^2 test procedure usually separate examinees into many subgroups by the total test scores, and assume the ability of all examinees approximately equal. Although many researches support χ^2 test procedure as superior to transformed item difficulty method, it still has many disadvantages if we use χ^2 test

procedure to detect DIF. For instance, the interval of a subgroup will change with the property of data, the total scores cannot be the index of ability, the index of DIF can be easily confounded by the sample size, and it will thus lose some valuable information to deal with continuous variable by categorical variable.

Item characteristic curve method is based on item response theory. It can be divided into three categories if we use IRT to detect DIF (Camilli & Shepard, 1994; Ironson, 1983).

1. To compare the measure difference between two item characteristic curves of different groups: The method is to calculate the area between the two item characteristic curves of different groups as the index of DIF. There are three kinds of DIF measure. First is area measure. Second is squared differences measure. The third is weighting the area and squared differences measures.
2. To compare the estimation values of item parameters: This essential point of this method is to test the equality of item parameters.
3. To compare the goodness of fit between item response model and data.

Many researchers have found the detecting of DIF based on item response theory is superior to transformed item difficulty method and 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, the study will adopt the signed area measure between two item characteristic curves to detect DIF.

The entrance examination of high school was administered more than 40 years. It was the only way to choose appropriate students from junior school to senior high school. In the past the reliability and validity

of the entrance examination had never been tested. It was, however, still accepted and trusted by the public. But, from a psychometric viewpoint, we doubt the fairness and rationality of the entrance examination by using a single test result to be the only criteria, especially since there is no pilot test for most students. Due to this the Ministry of Education wants to catch up with the globe trend of measurement. From 2001, the Ministry of Education in an attempt to reform the entrance examine introduced the basic competence test for junior high school students to replace the traditional entrance test. This was a milestone for the examination system of Taiwan. There is as yet however, no researches to detect the fairness of mathematics item of the basic competence test for junior high school students. Therefore, this researcher wants to use M-H, area measure, and likelihood ratio test to detect DIF of gender by the original mathematics answer of the first basic competence test for junior high school students on 2001. The researcher tries to determine the results consistency of all methods employed, which methods are more powerful and accurate, and what the characteristic of item bias is.

Based on the above research motivation, the goals of the study are as follows:

1. To investigate the performance difference of gender between three categories in the mathematic test.
2. To investigate the fairness in test items from the basic mathematics achievement test based on gender groups.
3. To compare the consistency of results for the different ways of detecting DIF.
4. To investigate item bias in test items from the basic mathematics achievement test.
5. To investigate the properties of DIF and item biases in the test.

1.3 Research question and hypothesis

By the above research motivation and goals, the study wants to investigate the following questions:

1. Whether the performance of mathematic test has difference between male and female in different categories of mathematic content?
2. Is there significant DIF in test items from the basic mathematics achievement test based on gender groups?
3. How consistent are results for the different ways of detecting DIF?
4. Which of the items detected as showing significant DIF are considered to be biased after logical analysis?
5. What are the properties of DIF and item biases in the test?

Hypothesis:

1. The performance of the mathematic test in basic competence test for junior high school student has difference between male and female in the three different categories of mathematics content.
2. The items of the mathematic test in basic competence test for junior high school students are with differential item functioning between male group and female group.
3. The results are not consistent for the different ways of detecting DIF.

1.4 Definition

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 latent ability (Thissen & Steinberg, 1986). There are many item response models that are developed from this theory. But one parameter logistic model, two parameter logistic model, three parameter logistic model, and four parameter logistic model are the basic model. The format used in the basic mathematical achievement test is multiple-choice. The data we got is dichotomous. Thus the study will adopt the three parameter logistic model. The three item parameters are a - discrimination, b - difficulty, and c - guessing.

1.4.2 Item characteristic curve method

Item characteristic curve (ICC) is the regression line of using examinee's ability to predict answering correct probability. ICC can show the trait of item difficulty, discrimination, and guessing. Item characteristic curve method is a kind of detecting DIF methods, which is based on item response theory. That is, item characteristic curve method is used to compare the item characteristic curve difference between different groups. And the difference measure is the index of DIF. We will compare the signed area and observe the item is favor to which group except calibrating the item parameters.

1.4.3 Item bias and DIF

Many psychometric experts have tried to give a clear and concrete

definition for item bias. Cleary & Hilton (1968) gave the definition of item bias from the variation analysis viewpoint. They reasoned that item bias was an interaction between item and groups. Angoff & Ford (1973) thought item bias was the difficulty parameters different between two groups. Today researchers distinguish DIF from "item bias", since Holland & Thayer (1988) used DIF (differential item functioning) or DIP (differential item performance) to describe the performance difference between two compatible ability groups.

Camilli & 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 as to why the items seem to be relatively more difficult, a subset of DIF items would be identified as "biased" and presumably then eliminated from the test. That is, DIF is uninterpreted relatively difficult items for two groups. The item is called item bias if it is explained by logical analysis, and interpreted as to the cause relative difficulty. DIF is just a statistic measurement result and is not necessarily used just to delete the item (Angoff, 1993). On the contrary, it means some curriculum or instruction to be changed if DIF in the test (Harris & Carlton, 1993; Lane, Wang & Magone, 1996).

1.4.4 Basic competence test for the junior high school students

The basic competence test for junior high school student is a policy of education reform created by the Minister of Education in Taiwan. The purpose of the test is to measure the basic abilities of students in junior high school 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. The "basic achievement" means the comprehensive, basic, and important ability of the learner, who was systemically instructed for the duration of the three-year junior high school period. The score they achieve is used to help students decide which school to attend, high school or vocational school. The test is designed by the institution of basic competence test for junior high school students.