Chapter2 Literature Review

The literature related to the basic competence test for junior high school students can be summarized into 6 categories.

2.1 The traits of the basic competence test for junior high school students

The basic competence test for junior high school students is different from the entrance exam for high school. The basic competence test for junior high school students has four traits.

1. It is a standardized test. Measurement experts and experts from every subject of the curriculum design the test. The test design of the procedure follows the standardized process of designing a test, for example building the two-way table, item design and selecting, administering process, scaling, reliability and validity. The aim is to produce a good result from the test.
2. It has a specific blueprint. The blueprint designed follows by the objectives of instruction, the criteria of the curriculum, and the ability of students. The blueprint is the index of the test. The two-way table is designed by ability level and content of students' learning.
3. The test is an item bank test. The item bank of the test has been built by item response theory (IRT) before the test in order to get good results from test.
4. The scores from different tests can be compared because the tests adopt IRT.

2.2 Classical test theory

Psychometrics is a division of science studying measurement and

evaluation which includes quantitative psychology, individual difference, and psychometric theory (Cohen, Montague, Nathanson, & Swerdlik, 1988). In the late of 19 century, scientific psychology was born and psychologists were interested in quantifying the psychological trait. It resulted the first intelligence test of Binet-Simon born in 1905. This is the first objective psychology test in human history.

Test theory can be divided into classical test theory and modern test theory, depending on how we interpret the measurement scores, according to different mathematical models and theories.

So far, classical test theory is still regarded as practical test theory. Many tests still build a real relationship between data by classical test theory. Classical test theory (CTT) evaluates the examinee's true scores primarily according to the examinee's observation scores, which is the examinee's response for certain item. That is, CTT attempts to evaluate the association between observation scores and true scores. Therefore, CTT is build on the true score model which is based on mathematical model.

The assumptions of CTT are to define the relationship between observation score, true score, and error score. There are main three assumptions: (Croker & Algina, 1986)

1. Observation scores are equal to true scores plus error scores. $X=T+E$
   The true score is the average of the observed scores obtained over an infinite number of repeated testing with the same test.
2. There is no relationship between error scores and true scores. $\rho_{TE}=0$
3. There is no relationship between two errors scores from two tests. $\rho_{EE}=0$

The classical test model and procedures for constructing tests and

interpreting test scores have served measurement specialists and other test users well for a long time. However, there are many shortcomings of the ways in which educational and psychological tests are usually constructed, evaluated, and used (Hambleton & Swaminathan, 1985).

1. The values of commonly used item statistics in test development such as item difficulty and item discrimination depend on the particular examinee samples where they are obtained.
2. Comparison of examinees on an ability measured by a set of test items comprising a test are limited to situations where examinees are administered the same test items. However, the problem is that because many achievement and aptitude tests are most suitable for middle-ability students, the tests do not provide very precise estimates of ability for either high or low ability examinees.
3. Test reliability is one of the fundamental concepts defined in terms of parallel forms. The concept of parallel measures is difficult to achieve in practice.
4. Classical test theory provides no basis for determining how an examinee might perform when confronted with a test item.
5. It presumes that the conditional standard error of measurement is the same for all examinees.
6. Classical test theory and associated procedures have failed to provide satisfactory solutions to many testing problems, for example, the design of tests, and the identification of biased items.

2.3 Item response theory

The birth of IRT is in order to improve the above shortcomings of classical test theory. The basic concept of IRT as follows: (Hambleton,

1989; Hambleton & Swaminathan, 1985; Yu, 1997)

1. The performance of examinee on a test item can be predicted or explained by a single factor. Because the factor cannot be observed, it is called latent trait or ability which is the desired measurement objective.
2. The relationship between performance and ability can be expressed by monotonically increasing mathematical function which is called Item Characteristic Function (ICF). ICF provides the right response probability of various kinds of examinee ability level. If we use graph to express the relationship, it is called Item Characteristic Curve (ICC).
3. Every response data has its corresponding item response measurement model and ICC because of its different type and assumption.

Every ICC usually includes one or many parameters to describe item trait and examinee's ability. Therefore, the shapes of ICC are different if the number of parameter is different. The most often seen is a non-straight regression line. Because the relationship between examinee's item performance and ability can be expressed by mathematical function, item response model is also called mathematical model. (Hambleton, 1989) Those item response models have common basic assumptions. Only could the premise of those assumptions can be established. IRT will not be misusage if we apply appropriate item response model to analyze test data. The basic assumptions of IRT will be introduced as follows:

1. Unidimensionality: All of the items in the test measure the same ability or latent trait. In here, test can be expressed as item composition of measuring the same ability. The meaning of

unidimensionality is simple, but it is not easy to correspond with unidimensioality for general measurement data. Hambleton & Swaminathan (1985) thought it will match the definition of unidimensionality if the measurement data has a key factor or component. In fact, IRT also needs various degrees of ability to interpret the assumption of the examinee's test performance. That is the assumption of multidimentionality (Bock & Aitkin, 1981; Hambleton, 1989; Reckase,1985). But its model is still in development. Unidimensionality is still in chief on current basic assumption.

2. Local independence: The examinee's response in each item is independent in statistics. That is, it will not influence his or her response to any items no matter whether examinee's response is correct or incorrect a particular item. Usually, local independence will be true if the assumption of undimensionality can be proved successfully. The two concepts are interlinked on the definition of local independence (Lord, 1980).
3. Non-speed test: The bad test performance of examinee is because of lack of ability rather than of time confinement.
4. Know-right assumption: If examinee knows the right answers on a certain item, he or she will certainly answer correct the item.

The theory of the basic competence test for junior high school students is developed from the item response theory due to the more theoretically justifiable measurement principles and the greater potential to solve practical measurement problems. It is a trend of tests to administer and analyze a test by IRT.

IRT was established in order to improve the above shortcomings of

classical test theory. Thus, IRT possess the advantages that CTT doesn't have (Hambleton, 1989; Hambleton & Swaminathan, 1985; Lord, 1980). These advantages and disadvantages are as follows.

Advantages:

1. Item parameters adopted by IRT are different from the item parameters adopted by CTT. Item parameters adopted by IRT are free from influence by the samples.
2. IRT can provide the exact index of standard error for the ability estimation of each student rather than use the identical standard error to represent all the examinees.
3. Examinee ability estimates are independent of the particular choice of test items used from the calibrated population of items. It can be significant when comparing the scores coming from different tests or examinees.
4. Item information or test information of IRT can be the index of accuracy estimation of the item or of the test.
5. IRT considers simultaneously the traits of examinee response type and item parameters when it calibrates the individual ability of examinee.
6. Statistic of goodness of fit can detect whether the model is appropriate to the data, whether the examinees' response is unusual, and whether the test or item is biased.

Taking a comprehensive view of above, IRT seems better than CTT. The effect, however, doesn't reach the expectation in real promotion and application. When inquiring the reason for this, it may be due to the following limitations.

1. The assumptions of IRT build on the strict mathematical probability

model that is difficult to be understood by the public. Thus, it is limited in promotion and application.

2. IRT is limited by strict basic assumptions and usually it is therefore necessary to have a big sample size. Therefore, the application area of IRT data is limited. It cannot be supported by general test users in promotion and application.
3. IRT has to combine with computer technology in order to estimate it. It would be more widely accessible if the price of software is cheap and time consumption is more economical. In the past two decades, the development of IRT has increased dramatically due to the computerization of tests.

2.4 Differential item performance

In the test of students' basic competence for junior high school, the items can be classified into three categories: algorithms, algebra, and geometry. In particular, algorithmic problems are considered to relate to the solution of problems that emphasize computations and other well-defined operations; algebra problems are considered to represent thinking and understanding; and geometry problems are considered to draw upon graphing and reasoning skills. Therefore, the study investigates the DIP of basic competence test for junior high school students between males and females in the three categories. One methodology to compare the performance of gender is to compare the mean score of a random sample of males against the mean score of a random sample of females. There are many researches related to this topic have been summarized in several meta-analyses (Friedman, 1994; Frost, Hyde, & Fennema, 1994; Hyde, Fennema, & Lamon, 1990). These

authors found that females scored slightly better than males during the elementary (average d=-.06) and middle school (average d=-.07) years, but males scored substantially better than females in the high school (average d=.29) and college years (average d=.41). Male often performed better than female in geometry and problem solving, whereas female did as well as male or better in algebra, arithmetic, and computation. Geometry items and items such as word problems that emphasize reasoning skills were predicted to favor male examinees. On the other hand, algorithmic, calculation-oriented items were predicted to relatively favor females (Doolitte, 1987). Wei(1996) investigated the 482 5^{th} grade students of elementary school in the central of Taiwan. She found there was no significant relationship between the mathematic achievement and gender. The findings were close to the research of Randhawa, Beamer, & Lundberg (1993) who indicated the mathematic achievement was not significant in elementary schools but the mathematic achievement would be significant when students enter to middle schools. Wu(1997) investigated the 464 4^{th} grade students of elementary school in the south of Taiwan. He also found there was no significant between the mathematic achievement of gender in multiple-choice, completion, calculation, and application categories. The findings correspond with the research results of Lee(1983), Wei(1988), Echols(1992), Tan(1992), and Samules(1991). But the finding is different from the results of Lee(1983) and Tsai & Walberg(1983) whose research indicated that the mathematic achievement of high school students are significant between males and females. Males and females were equally capable in mathematics during elementary school. This was true in the cultures of America, Japan, and Taipei. At eleventh grade, however, scores of the boys in all three

locations were significantly higher than those of the girls (Stevenson, Chen, & Lee, 1993). In general, the researches claimed the mathematic achievement is significant between males and females when the students are in high school or more high level.

2.5 Differential item functioning

The presence of bias is a cause for concern because tests are used as a gatekeeper for educational opportunities. It is a very important issue whether or not the test items are fair for every examinee. Differential item functioning (DIF) is said to occur when a test item does not have the same relationship to a latent variable across males and females examinee groups. Under an IRT framework, a test item is showing DIF if the IRC is not the same for the boys and girls groups who are equal in level on the latent trait do not have the same probability of endorsing a test item. (Embretson & Reise, 2000) The purpose of the present investigation was to examine whether difference exists between test performance for groups of males and females of equal ability levels. This study examined the possible presence of differential item functioning (DIF) in the assessment of these groups. A DIF effect is noted when examinees from different subgroups that are equal in terms of the ability being measured by a test have an unequal probability of correctly answering a given item (Camilli & Shepard, 1994). An important distinction, however, must be maintained between DIF and item bias. Identifying an item as functioning differently for two groups is not synonymous with stating that item is biased. As Camilli and Shepard (1994) noted, "only if an item is relatively more difficult for one group than the other and the source of that difficulty is irrelevant to the test constructs is an item

considered biased" (p16). Thus, Holland and Thayer (1988) referred to the raw uninterrupted relative difficulty as DIF and the use of the DIF index in conjunction with logical analysis as an item bias procedure. The focus of the present investigation was limited to DIF. In addition, it will extend into consideration of item bias.

DIF procedures can be categorized into three broad classes: traditional classical test theory methods, chi-square methods, and latent trait theory methods. The study will conduct the latent trait methods that are theoretically preferred (Shepard, Camilli, & Williams, 1984). It is very important in IRT to exactly estimate the item parameters and examinee ability. There are many kinds of software can be conducted. Comparing to BILOG, MUTILOG & PARSCALE, BILOG is the most accuracy and steady for the estimate of item parameter (Liu, Shu, & Jeng, 1998). The study will use BILOG to estimate the item parameters and examinee ability, and detect the DIF in the assessment of male and female groups.

2.6 Detecting DIF method

DIF of IRT tendency procedure primarily detects if there is any difference between the item parameters of the reference group and the focus group. That is, to examine whether the two item response functions are the same. The most universal procedures of IRT tendency are Lord χ^2 test (Lord, 1980)、IRT measure of DIF (area measure)(Camilli & Shepard,1994; Lord, 1980; Millsap & Everson, 1993), and likelihood ratio test (Thissen, Steinberg & Wainer, 1988,1993). They are introduced as follows:

1. Lord χ^2 test method

Lord (1980) provided a statistic procedure to test whether the item parameters of two groups are different. In applying Lord χ^2 to examine DIF, the first step is to employ IRT software, like the BILOG program, to calibrate the item response data of the focus and the reference groups. The item parameters estimated every time have their individual original point and unit because of IRT scale indeterminacy. Therefore, a direct comparison of the item parameters estimate cannot be made. Before comparison, we have to transfer the item parameters estimate to the same scale by linking the original point and unit. Then we can proceed with the DIF test.

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

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

The formula of Lord χ^2 test is as follows:

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

S indicates the variance of item parameter estimate difference – covariance matrix. By large sample theorem, Lord χ^2 estimates present χ^2 distribution with degree freedom 2 under the null hypothesis. If the χ^2 estimate reaches a significant level, we reject the null hypothesis. It indicates that the item is DIF. If the variance and covariance matrices cannot be estimated accurately, errors may result in subsequent identification of DIF items. Under such conditions, use of Lord's Lord χ^2 would not be justified (Lane,Stone, Ankenmann, & Liu, 1995).

2. IRT measure of DIF (area measure)

The other method of detecting DIF is to calculate the area measure

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

$A=\int_s (P_R(\theta) - P_F(\theta))d\theta$, S indicates the scope of ability θ.

If the area measure has positive, negative, or 0 after calculating, it is called signed area measure. If the value is positive or just 0, it is called unsigned area measure.

The formula of signed area measure is: $SA=\int_s [P_R(\theta) - P_F(\theta)]d\theta$

(Ruder, 1978)

If SA>0 indicates the item is advantageous to the reference group, SA<0 indicates the item is advantageous to the focus group. Its advantage is that it is easily interpreted in real application. Its disadvantage is the IRF difference calculated would offset each other when the ICCs of two groups intercept. Then the real value of DIF will be underestimated.

The formula of unsigned area measure is: $UA=\int_s | P_R(\theta) - P_F(\theta) | d\theta$

(Raju, 1988, 1990)

or $UA=\sqrt{\int_s [p_R(\theta)-P_F(\theta)]^2 d\theta}$ (Camilli & Shepard, 1994)

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

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

The procedure of likelihood ratio test (LR) (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988 & 1993) is used to compare two different IRT model 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 compact model and additional parameters. The procedure of LR test is primarily to test whether or not the additional parameter in augmented model is significantly different from 0.

The statistic of 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. However, j refers to the parameter number difference between the augmented and the compact model. The distribution of G^2 value is χ^2 distribution with j degree freedom under null hypothesis. In applying the LR test in detecting DIF, all the item parameters are supposed to be equal in compact model when you proceed calibrating parameters. In the augmented model, all the other item parameters are also supposed to be equal except the item parameters that have been detected. The DIF test is employed to compare the maximum likelihood function of the two models in order to check whether there is significant difference or not.

The terminology of Judd and McClelland (1989) and its application to IRT by Thissen et al. (1993), is the model comparison approach that 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). The model (A) is an elaboration of the 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 we have one more additional item in model (A). The goal of the comparison is to determine if the additional parameters in model (A) are necessary.

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

1. With a 3P IRT model, estimating 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. Reestimate parameters and obtain χ^2 transformation of the likelihood ratio G(2) for the 33-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 normal 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 tendency DIF procedure, Lord χ^2 test and LR test belong to the procedure of significant statistic test. The results only show whether or not there is a statistic difference between the two IRFs of the two groups but it qualifies difference. Raju (1990) provided two kinds of sampling distribution of mean and standard deviation on infinite interval area measure. He also provided the statistic for signed area measure – Z(EST) and the statistic for unsigned area measure – Z(H). Both of them are Z distribution. But the effect of Z(EST) and Z(H) on the DIF test still needs more research to evaluate it (Lu,1999) In addition, the item parameters of the Lord χ^2 test and the ICC area measure in real application need to be linked before comparison. However, in the LR test in practice the item parameters are simultaneously estimated and there is therefore no need to proceed with the procedure. Therefore, in the actual application, the LR test is the best in the three methods. LR test is a gradually more accepted application of detecting DIF. Kim & Cohen (1995) compared the performance of detecting DIF in the three methods and found the consistency of detecting results was very high in those DIF test procedures. The limitation of IRT 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).

The entrance exam for high school had conducted the classical test theory for the past 50 years. The entrance exam was replaced by student's basic ability test for junior high schools because of adapting to the trend of world and education reform from last year in Taiwan. There were around 300,000 examinees. Many schools, teachers, parents, and students were concerned about the quality of basic competence test for junior high school students. Is it fair and impartial? The purposes of the study were to focus on the fair issue – DIF for gender in mathematics subject and to compare the achievement performance of mathematics in different categories.