Chapter 3 Research Methodology

3.1 The object of research

The mathematics portion of the student's basic competence test for junior high school students was administered in April of 2001. The number of examinees was 299,368. For the purposes of this study, 5,000 students where randomly selected from the full sample.

The basic information of data is as follows:

Sex: male 2599 female 2401

Location of examinees: 1: north 1921, 2: central 1640, 3: south 1317, 4: east 122

3.2 Tools of research

The tools of research include:

1. Instrument: The mathematics test of basic competence test for junior high school students in Taiwan

The most differences between the traditional entrance examination and the basic ability test are the test content and the development of test tool. The basic ability test is designed to evaluate the basic knowledge and cultivated ability after students finish compulsory education. The basic ability test uses the item bank of continuous development to join the examinations. The test is composed of the items which are chosen from the item bank by computer. How is the test to be composed of? The principles are corresponding with the object of the basic ability test, item contents equally distributed to all the mathematic materials in senior high school, and the item difficulty between .5 and .75. The traits of mathematics are as follows:

a. The item design emphasizes on the comprehension, application, logical reasoning, and proof of mathematics knowledge.

b. The item design avoids the memory item and pays attention to the comprehension of basic concept.

c. The items are chosen by the curriculum criteria and emphasize the curricular content.

d. The stem of item has to be completely described. The question of item has to be logically arranged. The distracter of item has to be designed by the error type that students are easily to make.

There were 32 items in the test. All of the items are multiple-choice. Examinees have to choose one correct answer from the four options in every item. The content of test included the mathematics covered in junior high school. In general, the material can be divided to three parts: algorithms, algebra, and geometry. The items can also be classified into the same three categories. The total number of items on algorithm is 5. The total number of items on algebra is 9, and the total number of items on geometry is 18.

2. The computer software:

1. BILOG 3
2. SAS
3. SPSS
4. Mathlab

3.3 Processing Procedure

1. Obtaining data

The basic competence test for junior high school students is charged of the institute of the basic ability test center. The data of this research are obtained from the institute. The number of examinees was 299,368. For the purposes of this study, 5,000 students where randomly selected from the full sample by computer.

2 Design and Analysis:

A split-plot factorial design was used to investigate the effects of item category on gender differences in performance. The observed score for each examinee was the proportion correct of the items in each specific item category. Performance for a group was measured by mean proportion correct. In this design, gender will be considered between-group treatments and item category is a within-group treatment. I adopt MANOVA to analyze the data in order to investigate the research question of DIP.

Because Kim & Cohen (1995) have compared the methods of Lord χ^2 test, Z (Signed-Area), Z (Unsigned-Area), and the model comparison measures. They found the results were consistent to check DIF. The four methods belong to IRT-based procedure. The advantages of the model comparison measures are this method does not need the accurate covariance of item parameter (Thissen, Steinerg, & Wainer, 1988, 1993) and the item parameter does not need to be linked because item parameter was calibrated simultaneously (Camilli & Shepard, 1994). Therefore, I choose the IRT-based procedures Z(Sign-Area) and the model comparison measures for identifying DIF: Comparison male versus female. The total number of items identified with significant DIF was determined for each of the two DIF methods, and then the number of common items identified between methods was determined.

Unidimensionality: The tests of two groups were separately factor analyzed, using principle components analysis to determine the degree to which the items formed unidimensional sets. Reckase's (1979) eigenvalue criterion for assessing unidimensionality was used.

Estimation of item parameters. We adopted BILOG 3 (Mislevy & Bock, 1990) to estimate the items calibration and students' ability and detect the DIF. The anchor items and estimation of examinees are calibrated using BILOG 3, which uses the ML(Maximum Likelihood), EAP(Expected a Posteriori) estimation procedure. The fit of the model to the data was evaluated using the likelihood ratio goodness-of-fit statistic, G^2, which is distributed as χ^2 and is a test of the model against a general multinomial alternative model, as discussed by Thissen, Steinberg, and Gerrard (1986). Because the area between the two ICCs of male and female 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)

Linking of item parameter estimate. The item parameter estimates were transformed as follows for the test across the comparison – estimates from the female sample were transformed to the scale underling the male sample. Because IRT-based DIF analysis requires that the estimated item parameters from the two subpopulations be put on a common scale prior to any DIF analysis. The transformed procedure mean and sigma method described in Hambleton and Swaminathan (1985), the item parameter estimates – as and bs 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. $b^*_g = \alpha b_g + \beta$, $a^*_g = a_b / \alpha$, $\alpha = \sigma_{bb} / \sigma_{bg}$, $\beta = \mu_{bb} - \alpha \mu_{bg}$ The additive and multiplicative constants associated with the linear transformation were 0.0433(β)and 0.8237(α), respectively.

Statistical tests for signed area and indexes. Using the transformed item parameter estimates obtained in the previous step, SA was computed by using Matlab software. Assuming the observed SAs are normally distributed. Raju's (1990) z statistics for SA were computed to identify items with significant DIF. To judge whether the observed SAs is significantly different from 0 or not, compute $Z = (SA - 0) / \sigma(SA)$, $\sigma^2(SA) = (1-C)^2 * Var(b_2-b_1)$

Because IRT analyses typically involve large samples (500 or more) and the SD of SA is sample-size dependent. In this study, there are 2599 and 2401, respectively. It is desirable to the z statistics associated with SA, a two-tailed test of $z>3.27$ or $z<-3.27$ ($\alpha = .001$) was used to identify items with significant DIF.

A χ^2 of 16.268 ($\alpha = .001$ with 3 degrees of freedom) was used to identify items with significant DIF in model comparison measures.