Chapter 4 Results and Discussion

4.1 Results

Hypothesis 1: The performance of the mathematic test in basic competence test for junior high school students has difference between male and female in the three different categories of mathematics content.

The raw scores summary statistics for the male and female groups are given in Table 1. The effect sizes in whole test, algorithm, algebra, and geometry are .024, .072, .022, and .0045 respectively. It indicates that the differences are not very much between genders. Females scored better than males on the test as a whole, algorithm, algebra, and geometry.

As shown in Table 1, the reliability coefficients of the different strands and the test as a whole were different but the difference is small for males and females, ranging from .55 for algorithm for female to .89 for the test as a whole for male.

Table 1

Summary Statistics for Raw Scores Based on Multiple-Choice Items

	Total	Algorithm (5)	Algebra (9)	Geometry (18)
Total				
M	18.62	3.51	5.61	9.5
SD	6.74	1.37	2.16	3.97
Alpha	.87	.58	.68	.78
Male				
M	18.54	3.46	5.59	9.49
SD	7.03	1.41	2.23	4.11
Alpha	.89	.61	.71	.80
Female				
M	18.71	3.56	5.64	9.51
SD	6.40	1.32	2.08	3.82
Alpha	.86	.55	.66	.77
ES	.024	.072	.022	.0045

Note. Alpha = Cronbach's coefficient alpha; ES = effect size, used here as the difference in the means in pooled (male-female) standard deviation units. The sample size for male is 2599, and for female is 2401.

The summary of MANOVA was shown on the Table 1-1. The value of Wilks is 1 ($p<.05$). It means the difference of raw mean scores between males and females is significant for $\alpha=.05$. By the ANOVA of the three dependent variables - algorithm, algebra, and geometry, the results show the difference between male and female is not significant in the categories of algebra and geometry. But the difference is significant

in algorithm category and female is advantage over male. The effective sizes are small for the whole test, algorithm, algebra, and geometry. The whole test and algorithm are significant because the sample size is big.

Therefore, I accept my research hypothesis 1 that the whole test and algorithm category mathematic performance of basic competence test for junior high school students between genders are significant but the effect size is small. As for the algebra and geometry categories, there are not significant differences between genders. But females did better than males in whole test and the three categories in raw mean scores.

Table 1-1 Summary of MANOVA

Adjusted Hypothesis Sum-of-Squares and Cross-Products				Wilks		F	
	ALGE	ALGO	GEO		ALGE	ALGO	GEO
ALGE	2.73						
ALGO	5.73	12.01		1.00*	.58	6.44*	.02
GEO	1.02	2.15	.39				
WITHIN+RESIDUAL Sum-of-Squares and Cross-Products							
	ALGE	ALGO	GEO				
ALGE	23323.61						
ALGO	9277.28	9321.53					
GEO	31051.55	17264.40	78945.61				

*p<.05

The scores distribution of males and females in the mathematic subject of student's basic competence test for junior high school students are shown in table 2. The mode of males is 26 and 21 for females. The

median is 19 for both groups. The mean is 18.54 for males and 18.7 for females. The values of median and mean are very close for the two groups.

Table 2

The scores distribution of male and female

	Male		Female	
Scores	Frequency	Percent(%)	Frequency	Percent(%)
0	8	0.3	5	0.2
1	1	0	0	0
2	1	0	1	0
3	2	0.1	1	0
4	11	0.4	4	0.2
5	17	0.7	9	0.4
6	27	1	22	0.9
7	45	1.7	39	1.6
8	77	3	42	1.7
9	86	3.3	77	3.2
10	127	4.9	88	3.7
11	125	4.8	92	3.8
12	124	4.8	103	4.3
13	107	4.1	96	4
14	93	3.6	113	4.7
15	108	4.2	117	4.9
16	107	4.1	104	4.3
17	108	4.2	119	5

18	108	4.2	107	4.5
19	108	4.2	114	4.7
20	105	4	135	5.6
21	125	4.8	145	6
22	102	3.9	116	4.8
23	117	4.5	113	4.7
24	111	4.3	115	4.8
25	124	4.8	125	5.2
26	129	5	94	3.9
27	105	4	85	3.5
28	73	2.8	83	3.5
29	83	3.2	64	2.7
30	68	2.6	46	1.9
31	44	1.7	19	0.8
32	23	0.9	8	0.3

Hypothesis 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.

Hypothesis 3: The results are not consistent for the different ways of detecting DIF.

Unidimensionality

For male, the factor analysis of the mathematic subject of student's competence test for junior high school students yielded 4 eigenvalues larger than 1, with the largest eigenvalue of 7.506 accounting for 23.5% of the total variance. For female, the results of factor analysis yielded 5

eigenvalues larger than 1, with the largest eigenvalue of 6.5313 accounting for 20.4% of the total variance. The percent of total variance associated with the largest eigenvalue for both males and females of the mathematic subject of student's basic competence test for junior high school students did meet Reckase's minimum criterion of 20 percent for unidimensionality (Reckase, 1979).

DIF Indexes for the Test

The item parameters estimate from BILOG 3 run is presented in Table 4 for both groups. Also given in this table is the transformed parameter of female.

Table 4

Item Parameter Estimate for the Male Group, Original and Transformed Parameter Estimates for the Female Group and Area Measures

				Female			Female		
	Male			Original			Transformed		
Item #	a	b	c	a	b	c	a	b	c
1	0.713	-0.971	0.2	0.681	-1.187	0.2	0.827	-0.980	0.2
2	1.518	-1.075	0.2	1.415	-1.282	0.2	1.718	-1.058	0.2
3	1.089	-0.835	0.2	0.960	-0.990	0.2	1.165	-0.818	0.2
4	0.992	-1.439	0.2	0.717	-1.564	0.2	0.870	-1.291	0.2
5	0.788	-1.117	0.2	0.587	-1.324	0.2	0.713	-1.093	0.2
6	0.941	-1.140	0.2	0.778	-1.385	0.2	0.945	-1.143	0.2
7	1.244	-0.318	0.2	1.300	-0.465	0.2	1.578	-0.385	0.2
8	1.747	-0.761	0.2	2.044	-0.838	0.2	2.481	-0.693	0.2

9	1.088	-0.801	0.2	0.845	-0.900	0.2	1.026	-0.744	0.2
10	1.028	0.180	0.2	0.863	0.104	0.2	1.048	0.083	0.2
11	1.180	0.016	0.2	0.951	0.011	0.2	1.155	0.007	0.2
12	1.890	-0.010	0.2	1.835	-0.011	0.2	2.228	-0.011	0.2
13	1.469	-0.188	0.2	1.170	-0.461	0.2	1.420	-0.382	0.2
14	0.924	0.416	0.2	0.655	0.535	0.2	0.795	0.438	0.2
15	0.895	-0.714	0.2	0.705	-0.665	0.2	0.856	-0.550	0.2
16	0.476	0.749	0.2	0.431	1.017	0.2	0.523	0.835	0.2
17	1.452	-0.023	0.2	1.225	-0.063	0.2	1.487	-0.054	0.2
18	1.725	0.412	0.2	1.680	0.366	0.2	2.040	0.299	0.2
19	0.479	0.220	0.2	0.532	-0.089	0.2	0.646	-0.076	0.2
20	1.165	0.589	0.2	1.212	0.795	0.2	1.471	0.653	0.2
21	1.079	0.624	0.2	0.878	0.646	0.2	1.066	0.530	0.2
22	1.327	0.575	0.2	1.065	0.454	0.2	1.293	0.372	0.2
23	1.099	0.307	0.2	0.974	0.400	0.2	1.182	0.327	0.2
24	0.674	1.074	0.2	0.357	1.967	0.2	0.433	1.618	0.2
25	1.675	0.037	0.2	1.594	0.090	0.2	1.935	0.072	0.2
26	0.669	0.591	0.2	1.028	0.995	0.2	1.248	0.817	0.2
27	1.123	0.665	0.2	0.906	0.765	0.2	1.100	0.628	0.2
28	0.636	0.954	0.2	0.634	0.995	0.2	0.770	0.817	0.2
29	1.083	1.150	0.2	0.984	1.250	0.2	1.195	1.027	0.2
30	1.015	1.136	0.2	0.939	1.366	0.2	1.140	1.123	0.2
31	1.284	1.660	0.2	1.315	1.862	0.2	1.596	1.531	0.2
32	0.727	1.370	0.2	0.488	1.738	0.2	0.592	1.429	0.2
Mean	1.1	0.104	0.2	0.992	0.129	0.2	1.200	0.1	0.2
S.D.	0.367	0.827	0	0.407	1.004	0	0.49	0.83	0

The χ^2 goodness-of–fit statistic difference for model comparison measure, signed and unsigned areas, and sign-z values are shown in Table 5. The signed areas vary from -.22 to .3. The unsigned areas vary between 0.003 and 0.49. For Lord' (1980) test and in Raju, Drasgon, & Slinde (1993) research, alpha level of 0.001 was used to identify items with significant DIF. Items with $\chi^2_{(3)}$ difference greater than 16.268 and Z scores greater than 3.27 or less than –3.27 are identified with two asterisks in Table 5. These items seem to indicate significantly DIF between male and female examinees. Of the three items – 13, 24, and 26 identified as DIF with the model comparison measure, two items – 13 and 26 were also identified as DIF with the signed area measure. Two of these items (Items 13 and 26) were common to both measures. 9% and 6% of items had significant DIF for the model comparison and signed area measures, respectively. All the 3 items were belonged to geometry. The 3 items were shown in the appendix. The results support my second and third hypothesis question that 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 but the proportion is low.

Table 5

Difference for model comparison measure, signed and unsigned areas, and sign-z values

	Content	Augmented-model	Difference	Sign	Unsign	Sign_z
Item1	Algorithmic	173347	6.39	-0.01	0.12	-0.18
Item2	Algebra	173350.8	2.65	0.01	0.05	0.27
Item3	Algorithmic	173352.3	1.09	0.01	0.04	0.23

Item4	Algebra	173340.8	12.60	0.12	0.12	1.37
Item5	Geometry	173345.9	7.53	0.03	0.08	0.28
Item6	Algebra	173354.5	-1.04	-0.00	0.00	-0.03
Item7	Geometry	173341.7	11.78	-0.05	0.12	-1.32
Item8	Geometry	173348.3	5.16	0.05	0.12	1.41
Item9	Algorithmic	173349.4	4.00	0.05	0.05	0.79
Item10	Algorithmic	173351.5	1.89	-0.08	0.08	-1.65
Item11	Algorithmic	173353	0.46	-0.01	0.01	-0.16
Item12	Geometry	173354.5	-1.11	-0.00	0.05	-0.03
Item13	Geometry	173325.1	28.35**	-0.15	0.15	-3.84**
Item14	Geometry	173343.9	9.52	0.02	0.11	0.27
Item15	Algebra	173342.9	10.49	0.13	0.13	2.12
Item16	Geometry	173350.8	2.63	0.08	0.11	0.74
Item17	Geometry	173352.9	0.57	-0.02	0.02	-0.65
Item18	Geometry	173351.9	1.57	-0.09	0.10	-2.84
Item19	Geometry	173344.4	9.05	-0.22	0.35	-2.94
Item20	Algebra	173342.1	11.36	0.05	0.12	1.18
Item21	Algebra	173350	3.40	-0.07	0.07	-1.54
Item22	Algebra	173341.3	12.09	-0.13	0.13	-3.20
Item23	Algebra	173352.2	1.19	0.02	0.04	0.36
Item24	Geometry	173333.1	20.37**	0.30	0.49	1.72
Item25	Geometry	173348.5	4.91	0.03	0.06	0.82
Item26	Geometry	173290.2	63.19**	0.20	0.46	3.32**
Item27	Algebra	173354	-0.60	-0.03	0.03	-0.6
Item28	Geometry	173352.3	1.11	-0.09	0.18	-1.19
Item29	Geometry	173352.5	0.88	-0.10	0.10	-1.58

Item30	Geometry	173344.8	8.64	0.01	0.07	-0.13
Item31	Geometry	173344.5	8.91	-0.10	0.13	-1.22
Item32	Geometry	173348.1	5.34	0.02	0.17	0.13

$X^2_{(3),999}=16.268$ Z(.999)=3.27

Compact- Model=173353.4

4.2 Discussion

Performance on the three categories of mathematics

For the algorithm category of mathematics, the result of this study is consistent with the previous research reported by Doolittle (1987). In regard to gender differences in the whole test and the three subcategories, females did better than males. Those results are consistently with the previous research (Friedman, 1994; Frost, Hyde, & Fennema, 1994; Hyde, Fennema, & Lamon, 1990). All the effect sizes are just slightly different, even though the effect size of the significant category – algorithmic-- is only .072. Because the examinees are junior high school students in this study and the content is limited, the results are consistent with many previous findings. The period of junior high school is a transitional stage. Before this period, the mathematic performance of female is significantly better than male. After this stage, the situation maybe will oppositely change.

Consistency in detecting DIF

There are three items – 13, 24, and 26 were detected with DIF in the model comparison measure. There is only two DIF items – 13 and 26 detected are in the signed area measure, which are the same with model comparison measure. All of them are belonged to geometry category.

With respect to Item 26 in the DIF index, most part of male had an advantage over female. Although the item 26 seems similar to graph problem, in fact, it is difficult. In addition the shape of quadrilateral is similar for the four answers, you have to know the exact proportion of every side is equal. Reasoning needs to be done to solve the item problem. We can see the b-value (.591) of male is much smaller than the transformed b-value (.817) of female. It is consistent with the previous research (Frost et al., 1994) that found males often did better than females in the content of geometry and problem solving. The Item 28 in Table 4 has the same difficulty with the Item 26. But the Item 28 is the usually form which is often seen in books or reference materials. In addition, it is a little related to calculation. So the item 28 does not belong to DIF and it favors to females.

Item 13 is also identified DIF in the two measures. The DIF index of Item 13 points to the conclusion that female had an advantageous over male. The content of this item belongs to geometry. But if we check this item, we find it seems some problems is in the stem of item. First, males are sometimes more careless than females. Males will overlook the value "+1" and "-1". Second, the answer is to choose the "wrong statement," but we cannot see the underline or another obvious mark on the word "wrong". Because of carelessness, it will be advantageous to females. We believe the item has to be revised. If we looked at the Table 4, we could find the Item 7 has the similar difficulty with Item 13 for females. If we looked the Item 7 in the appendix, we found Item 7 did not have the problem – careless, with Item13. Therefore, the evidence let us make more confidence that the reason to cause Item 13 DIF is because of careless.

Regarding to item 24, it is not significant in SA. But if we check the unsigned-area (.30) and signed-area (.49) of item 24 in Table 5, we can find the difference (.19) is large for the two areas. We believe the item may have to belong to the DIF problem if we use other measures to detect it. Therefore, item 24 is worth of being checked. The difference (.54) of b-value is larger and advantageous to the male.

In the model comparison measure, there are three items (6, 12, and 27) that have negative values between the differences of the compact model and the augmented model. The situation is because we estimate the samples rather than the population. The negative value was caused of estimate errors. In addition their value is very small and close to zero. In fact, we find the values are also very small after we check the value of signed and unsigned areas. It claims these items fit very well in the model comparison and the negative values were produced because of estimate errors.

The proportion of common DIF is just 6%(2/32) for the two measures, it is lower than the study of Budgell, Raju, & Quartetti (1995). The reason is because of the efforts of the institution of basic competence test for junior high school students. And they adopted the concept of IRT to design the item bank and calibrated the item parameters. But the mathematic subject of student's basic competence test for junior high school students will influence the fate of around 300,000 students. The proportion of the DIF item has to be reduced to zero. Otherwise, it will influence the impartiality of gender if there are still DIF items, and it is meaningful after experts analyze the DIF items in the test. We have to pay attention to the DIF problem after we administer the pilot test in order to build an item bank.

The results of DIF are not consistent for the two methods in the study. However, the results are acceptable that one is common and the other two is similar in the 32 items. The reason for difference is maybe because of the estimate error. The weight of students was used in model comparison. But it just focuses on the area difference in the SA measure. I believe the method of weight is more accurate because the distribution of examinees are similar to normal distribution and tend to concentrate to toward the b value. In the future, we hope to use more methods to efficiently detect the DIF to build a standard method. It will be helpful for the building of the impartiality of the test.

According to Table 4, the average difference in the b values for the male and female groups is .03, or almost zero. The Item 13 and 26 that were identified as DIF and the Item 24 that was close the criteria of DIF by the two methods had the highest b-value difference; item 24 and 26 were favoring the most part of male group and item 13 was favoring the female group. It appears that the items with substantial b-value differences (compared to the other items in the test) were generally identified as DIF in male-female comparisons.

In general, DIF is just the results of statistic analysis. DIF is the necessary condition rather than the sufficient condition. To judge whether or not item is bias has to be supported by the qualitative and quantitative evidence. There are many circumstances that can cause the DIF problems, which includes instruction, material of textbook, policy, and item itself. According to Item 26, it is very clear and impossible to misunderstand. It is possible to produce DIF because of the form and content of the item. The question is seldom seen in the material of a textbook, reference book or practice problems in Taiwan. It is creative. It

is consistent with the research of Doolittle & Cleary (1987) that male high school students perform relatively better than female on geometry items, which usually contain figures. Although the Item 26 seems similar to graph problems, in fact, it is difficult. In addition, the shape of the quadrilateral is similar for the four answers, and you have to know the exact proportion of every side is equal and the length of side is an irrational number. Reasoning needs to be done to solve the item problem. This is one of the solutions. The other way to solve the problem for some students' maybe is to fold the paper to find the answer. From the graph item 26, we can find females did better than males when difficulty is greater than 1.078. Therefore, this creative question is just advantageous for the most part of males rather than all.

Item 24 is similar to item 26. If we analyze the content of item 24, we can find it is similar to graph problem. It also belongs to the content of geometry and problem solving. The one way to solve this problem is students have to combine two concepts together, -- the distance is equal from circle to tangent line and from a point of equalized line of angle to the two sides of angle. The other way is to use every answer to draw the graph and check which one is the right answer. If the reason to produce DIF is because the item is too creative, I believe we have to keep it. From the graph of IRT in appendix, we find not all males did better than females. The females did better than males when the difficulty was less than .097. But the difference is small than the part of difficulty greater than .097. So in average, the creative item will favor most part of males but not all. Teachers' instruction of students is determined by the material related the test in Taiwan. Because of the feature of gender difference, females always follow the teacher's instructions and seldom

independently think. This creative item can break the traditional instruction, change the teacher's instruction and change the students' learning and thinking. If this situation occurs, this item will not be DIF in the future.