GIRLS' OVERACHIEVEMENT IN THE MATHEMATICS NATIONAL EXAMINATION: CAN TYPE OF ITEMS BE THE CAUSE?

Hanizah Hamzah
Siti Rahayah Ariffin
Ruhizan Mohd Yassin
Universiti Kebangsaan Malaysia, Bangi, Malaysia

This study explored the differential performances of mathematics test items used to test secondary school girls and boys in the national examination. The main purpose was to find out whether type of items is the reason for girls' overachievement in the Malaysian mathematics national examination. To investigate seven types of items, Differential Item Functioning (DIF) analysis was used. DIF is a statistical method to identify items that function differentially for different groups of students of the same ability. Forty-six multiple choice items, categorised into seven types (Text, Graphics, Knowledge, Skills, Numbers, Shape and Space, Relationship) were obtained and certified by a panel of experts. The sample respondents were 1213 students aged around 16 to 17 years old. There were 529 boys and 684 girls who participated in this study. Data gathered were analysed using Winsteps version 3.48, a Rasch-based item analysis program. This study showed that gender-related DIF (GDIF) exists in the test used. GDIF in favour of girls was exhibited in 25 items, whereas 21 items were in favour of boys. Girls excelled in items categorised as Text and Numbers, whereas items categorised as Graphics, and Shape and Space were shown to be in favour of boys. Algebra seemed to be easier for the girls, while Statistics seemed to be easier for the boys. Results showed 14 items with significant GDIF, with sizes ranging from 0.29 to 0.65 (logits). Six items were significantly easier for the girls and eight for the boys. Although more items

in the test were easier for girls, the study did not show evidence that girls' overachievement is due to imbalance usage of item type in the test. However, findings indicate that test developers should be sensitive of the occurrence of DIF and observe the proportion of item types showing DIF in all subjects tested in the national examination.

Introduction

Fairness and equality has been a major educational theme for many years. Much emphasis has been put upon acknowledging diversity in students' backgrounds and characteristics to ensure effective education. The stress on fairness has been extended to assessment tools, in countries like Australia, the United States, the United Kingdom, Germany and Sweden. Test fairness is the motivation that encouraged researchers to undertake this type of study. Of all that are used to differentiate individuals, gender is one of the most studied variables in educational researches. Previous studies have shown boys' advantage in multiple choice tests, items involving numerical, spatial or higher reasoning (Breland, 1991; Walstead & Robson, 1997). Other studies have found girls' better performance in multiple choice tests (Bellar & Gafni, 1995; Wester & Henriksson, 2000) and in tests that require writing ability (Kleinfeld, 1998). Boys show better performance in items concerning science, sports, and mechanics (Lawrence & Curley, 1989; Wild & Mc Peek, 1986), whereas girls perform better in items related to social science, humanities, philosophy and human relationships (Wild & Mc Peek, 1986; O'Neil Wild, & Mc Peek, 1989). In mathematics, boys do better in geometry items, whereas girls do better in algebra items (O' Neil & Mc Peek, 1993; Ryan & Fan, 1996; Halpern, 1997). Fennema (2000) found that boys are more superior in tasks requiring complex mathematics calculations and high cognitive functions. A study conducted in Malaysia found that boys scored better in items containing diagrams, whereas girls scored better in items without diagrams (Othman, 2003).

Information on how Malaysian boys and girls perform in different types of items in national examinations is still not available. For more than ten years, the results of the Malaysian national examinations have shown a pattern which consistently illustrates the dominance of female students over the male students in majority of subjects, including those which have been considered the 'male' subjects. Due to this, Malaysians in general have raised these questions: Has the achievement of boys really deteriorated, or is there a test-related factor that actually caused this? The continuous overachievement of one gender over the other has driven the researchers to look into the assessment aspect to clarify this uncertainty. The questions are: (1) Do the items function so differently towards girls and boys, resulting in the lopsided performance gap? (2) Does type of items used in a test contribute to the success of girls? Information that can be used to answer these questions is still not available as far as the Malaysian national examination is concerned. Therefore, empirical evidence on how items behave towards candidates of different genders need to be investigated.

Objectives

This study attempted to investigate any signs of unfairness in items used to assess student achievement in the Malaysian mathematics national examination. This was done by conducting Differential Item Functioning (DIF) procedure on the studied items. The principle underlying a DIF study states that groups of people of the same ability in a test should perform equally well in each item in the test, regardless of gender, ethnic and other factors (O' Neill & Mc Peek, 1993). An item will show DIF when the responses of students of equal or approximate ability differ systematically based on their membership with a particular subgroup. Statistical analysis of a test can detect whether any items function differently for identified subgroups of students. For further bias analysis, judgment of

curriculum specialists and psychometricians is required to determine whether items exhibiting DIF is due to unfairness or bias. In this study, Mathematics items were used due to the subject's educational importance.

The objective of the study was to explore the existence and intensity of gender-related DIF (GDIF) in mathematics items used in a Malaysian national examination. This study was also extended to identify any item type that can be associated with the existence of GDIF in items, hence determining whether overachievement of girls can be attributed to the type of items used in the examination paper. In this study however, items showing significant GDIF were not subjected to expert review for further scrutiny.

Methodology

This was an exploratory study to detect any mathematics items showing DIF when boys and girls of equal ability were compared. The first phase of the study involved constructing studied items. Table 1 displays the types and quantity of items used as sample in this study. Table 2 displays the specifications of each studied item in the mathematics test. A test consisting of 46 multiple choice items were constructed by trained national examination item writers and certified by assessment experts from the Malaysian Examinations Syndicate, Ministry of Education. Items were categorised into seven types, according to physical presentation (Text, Graphics), construct tested (Knowledge, Skills), and area in mathematics (Numbers, Shape and Space, Relationship). These are the seven types of items used in the assessment of mathematics at the national level; four of which are common types found in the assessment of other subjects. The test covers 13 topics: Standard Form, Basic Numbers, Polygon, Circle, Transformation, Trigonometry, Algebra, Linear Equation, Straight Line, Set, Statistics, Area under Graph and Ratio. All items were written in the Malaysian national language, the Malay language.

Table 1
Types and Quantity of Studied Items

Item Type	Quantity	Elements/Topics	Quantity
Text (text only)	22		
Graphics (text with graphics)	24		
Knowledge	19	Knowledge on numbers	3
		Knowledge on shapes	2
		Knowledge on transformation	3
		Knowledge on 2 dimensional space	2
		Knowledge on algebra	2
		Knowledge on coordinate geometry	4
		Knowledge on data handling	3
Skills	27	Estimation skills	2
		Operational handling skills	10
		Counting skills	4
		Problem Solving skills	11
Number	11	Standard Form	7
		Basic number	4
Shape and Space	10	Polygon	2
		Circle	2
		Transformation	3
		Trigonometry	3
Relationship	25	Algebra	6
		Linear Equation	2
		Straight lines	4
		Set	5
		Statistics	6
		Area Under Graph	1
		Ratio	1

Table 2
Item Specifications

Item No	Presentation	Construct Area	Topic	
1	Text	Knowledge	Number	Standard Form
2	Text	Skills	Number	Standard Form
3	Text	Skills	Number	Standard Form
4	Text	Skills	Number	Standard Form
5	Text	Skills	Number	Standard Form
6	Text	Knowledge	Number	Basic number
7	Text	Knowledge	Number	Basic number
8	Text	Skills	Number	Basic number
9	Text	Skills	Number	Basic number
10	Graphics	Skills	Shape and Space	Polygon
11	Graphics	Skills	Shape and Space	Polygon
12	Graphics	Knowledge	Shape and Space	Cirle
13	Graphics	Knowledge	Shape and Space	Cirle
14	Graphics	Knowledge	Shape and Space	Transformation
15	Graphics	Knowledge	Shape and Space	Transformation
16	Graphics	Knowledge	Shape and Space	Transformation
17	Text	Knowledge	Shape and Space	Trigonometry
18	Graphics	Skills	Shape and Space	Trigonometry

19	Graphics	Knowledge	Shape and Space	Trigonometry
20	Text	Knowledge	Relationship	Algebra
21	Text	Knowledge	Relationship	Algebra
22	Text	Skills	Relationship	Algebra
23	Text	Skills	Relationship	Algebra
24	Text	Skills	Relationship	Algebra
25	Text	Skills	Relationship	Algebra
26	Text	Skills	Relationship	Linear Equation
27	Text	Skills	Relationship	Linear Equation
28	Text	Knowledge	Relationship	Straight Line
29	Graphics	Knowledge	Relationship	Straight Line
30	Text	Knowledge	Relationship	Straight Line
31	Graphics	Knowledge	Relationship	Straight Line
32	Graphics	Knowledge	Relationship	Set
33	Graphics	Knowledge	Relationship	Set
34	Text	Skills	Relationship	Set
35	Text	Skills	Relationship	Set
36	Graphics	Skills	Relationship	Set
37	Graphics	Knowledge	Relationship	Statistics
38	Graphics	Skills	Relationship	Statistics
39	Graphics	Skills	Relationship	Statistics
40	Graphics	Skills	Relationship	Statistics
41	Graphics	Skills	Number	Standard Form
42	Graphics	Skills	Number	Standard Form
43	Graphics	Skills	Relationship	Statistics
44	Graphics	Skills	Relationship	Statistics
45	Graphics	Skills	Relationship	Area Under Graph
46	Graphics	Skills	Relationship	Ratio

During the second phase, the test was administered to a sample of 1213 students, between 16 to 17 years old from five secondary schools. The two subgroups compared were 684 female students (reference group) and 529 male students (focal group). The sample consisted of three major races in Malaysia: Malay, Chinese and Indian. These students have undergone two years of upper secondary Malaysian mathematics curriculum and all of them were scheduled to sit for the Sijil Pelajaran Malaysia (SPM), an examination equivalent to the O-Levels. Of the five schools, three schools were located in urban areas and two schools were in rural areas. In terms of gender composition, two schools were categorised as single-sex, and three schools were categorised as co-ed (mixed). The test was administered under standardised and controlled conditions in the respective schools. Students were given 1 hour and 25 minutes to answer all 46 items. Data gathered were analysed for GDIF using Winsteps (version 3.48), a Rasch-based item analysis program. For further analysis, students were also identified according to their mathematics ability levels (high, moderate, low), based on student ability measures provided by the program. Analysis at different levels of ability illustrates whether GDIF in studied items was uniform or non-uniform. In this study, GDIF analysis was extended to compare students of different school types and locations. These analyses were conducted in order to look at the consistency of finding that would strengthen the conclusions made from earlier analyses.

Results

This study was designed to provide answers to three questions: (1) Does GDIF exist in mathematics items used in Malaysian national examination, and what is its intensity? (2) What are the types of items that can be associated with the existence of GDIF? In other words, what are the types of item that consistently provide advantage to the boys or the girls? (3) Based on the findings, can we conclude that the overachievement of girls in the mathematics

national examination is caused by the type of items used?

The first stage of analysis was conducted to test the instrument's unidimensionality and reliability. Item polarity test showed that all 46 items used were working coherently to test the same dimension. Fit statistics showed that all items were within the Infit Mean Square range of 0.7 logits to 1.3 logits, as recommended by Wright and Linacre, in Bond and Fox (2001). These results indicated that the test used was unidimensional, and therefore Rasch analysis assumption of unidimensionality was fulfilled. Test reliability index was 0.91. The girls showed a mean ability measure of 0.54 logits, while the boys showed a mean ability measure of 0.49 logits. Referring to the average ability measures, the girls performed slightly better in the test. However, t-test showed that the two groups of interest were statistically comparable in terms of ability in the test.

The second stage of analysis investigated the existence and intensity of GDIF in the test used. To analyse DIF, Winsteps performs two-tailed t-test to test significance of the differences between two difficulty indices. The confidence level was at 95% and critical t value was set at 2.0 for all GDIF analyses. Analysis revealed that all 46 items showed differential difficulties to the boys and girls, with GDIF indices ranging from 0.01 logits to 0.65 logits. Table 3 shows results of GDIF analysis on 46 studied items. Out of the 46 items, 25 items were easier for the girls and 21 items were easier for the boys. Analysis flagged 14 items (30%) with significant GDIF with indices ranging from 0.29 to 0.65 logits (* items are bold in Table 3). Out of the 14 items, six items were significantly easier for the girls (GDIF sizes from 0.33 to 0.50 logits), while eight items were significantly easier for the boys (GDIF sizes from 0.29 to 0.65 logits). Of the 14 flagged items, eight items (17%) displayed GDIF sizes of at least 0.40 logits. A DIF size of at least 0.40 logits is regarded as important and has substantive meaning (Rasch Measurement Transactions, 2004).

Table 3
GDIF Analysis of 46 Studied Items

Group	DIF Measure (Difficulty measure)	DIF S.E.	Group	DIF Measure (Difficulty measure)	DIF S.E.	DIF Contrast (DIF size)	Joint S.E	t	df	Item Label
G	-1.66	0.11	B	-1.51	0.12	-0.15	0.16	-0.92	INF	1TK StdFm
G	-0.13	0.09	B	0.12	0.10	-0.26	0.14	-1.84	INF	2TS StdFm
G	-0.73	0.10	B	-0.84	0.11	0.11	0.14	0.78	INF	3TS StdFm
G	-0.73	0.10	B	-0.62	0.11	-0.11	0.14	-0.74	INF	4TS StdFm
G	1.13	0.10	B	0.58	0.11	0.55	0.14	3.89*	INF	5TS* StdFm
G	-1.08	0.10	B	-0.75	0.11	-0.33	0.15	-2.25*	INF	6TK* BNum
G	-0.37	0.09	B	-0.27	0.10	-0.10	0.14	-0.73	INF	7TK BNum
G	-0.46	0.09	B	-0.04	0.10	-0.41	0.14	-2.97	INF	8TS* BNum
G	-0.77	0.10	B	-0.70	0.11	-0.07	0.14	-0.50	INF	9TS BNum
G	-0.34	0.09	B	-0.50	0.11	0.16	0.14	1.17	INF	10GS Polg
G	-0.13	0.09	B	-0.23	0.10	0.09	0.14	0.66	INF	11GS Polg
G	0.45	0.09	B	-0.03	0.10	0.48	0.14	3.50*	INF	12GK* Circ
G	0.97	0.09	B	0.88	0.11	0.09	0.14	0.61	INF	13GK Circ
G	-2.02	0.12	B	-2.17	0.14	0.15	0.19	0.79	INF	14GK Trans
G	-0.54	0.09	B	-0.40	0.10	-0.14	0.14	-1.01	INF	15GK Trans
G	0.60	0.09	B	0.26	0.10	0.33	0.14	2.41*	INF	16GK* Trans
G	-0.18	0.09	B	-0.16	0.10	-0.02	0.14	-0.11	INF	17TK Trig
G	1.24	0.10	B	1.53	0.12	-0.29	0.15	-1.94	INF	18GS Trig
G	-0.13	0.09	B	0.10	0.10	-0.23	0.14	-1.69	INF	19GK Trig
G	-0.39	0.09	B	-0.32	0.10	-0.06	0.14	-0.46	INF	20TK Alg
G	1.01	0.09	B	0.80	0.11	0.21	0.14	1.49	INF	21TK Alg
G	-1.19	0.10	B	-0.79	0.11	-0.39	0.15	-2.66*	INF	22TS* Alg
G	0.89	0.09	B	0.94	0.11	-0.05	0.14	-0.35	INF	23TS Alg
G	1.13	0.09	B	1.62	0.12	-0.50	0.15	-3.29*	INF	24TS* Alg
G	0.63	0.09	B	1.13	0.11	-0.50	0.14	-3.47*	INF	25TS* Alg
G	-0.41	0.09	B	-0.47	0.11	0.05	0.14	0.38	INF	26TS LEqua
G	-0.51	0.09	B	-0.57	0.11	0.06	0.14	0.42	INF	27TS LEqua
G	-0.08	0.09	B	0.09	0.10	-0.16	0.14	-1.19	INF	28TK SLine
G	0.08	0.09	B	0.14	0.10	-0.06	0.14	-0.44	INF	29GK SLine
G	0.14	0.09	B	0.07	0.10	0.07	0.14	0.52	INF	30TK SLine
G	0.85	0.09	B	0.56	0.11	0.29	0.14	2.08*	INF	31GK* SLine
G	-0.37	0.09	B	0.03	0.10	-0.40	0.14	-2.91*	INF	32GK* Set
G	0.59	0.09	B	0.82	0.11	-0.23	0.14	-1.64	INF	33GK Set

G	0.56	0.09	B	0.58	0.11	-0.02	0.14	-0.13	INF	34TS Set
G	0.19	0.09	B	0.18	0.10	0.01	0.14	0.10	INF	35TS Set
G	0.29	0.09	B	0.39	0.10	-0.11	0.14	-0.76	INF	36GS Set
G	0.85	0.09	B	0.83	0.11	0.02	0.14	0.14	INF	37GK Stat
G	0.66	0.09	B	0.00	0.10	0.65	0.14	4.71*	INF	38GS* Stat
G	-0.75	0.10	B	-0.76	0.11	0.01	0.14	0.09	INF	39GS Stat
G	0.24	0.09	B	0.19	0.10	0.05	0.14	0.38	INF	40GS Stat
G	-0.23	0.09	B	-0.15	0.10	-0.08	0.14	-0.55	INF	41GS StdFm
G	0.90	0.09	B	0.93	0.11	-0.03	0.14	-0.21	INF	42GS StdFm
G	0.28	0.09	B	-0.24	0.10	0.52	0.14	3.73*	INF	43GS* Stat
G	-0.05	0.09	B	-0.03	0.10	-0.02	0.14	-0.15	INF	44GS Stat
G	-0.31	0.09	B	-0.61	0.11	0.30	0.14	2.14*	INF	45GS* AUG
G	-0.18	0.09	B	-0.52	0.11	0.34	0.15	2.42*	INF	46GS* Ratio

Winsteps computes an estimate of average DIF, which is equivalent to uniform DIF. At this level of analysis, interaction between group and ability levels is not visible. To find out how GDIF occurred at different levels of ability, girls and boys of equal ability were analysed separately. The high ability group consisted of 90 boys and 123 girls, the moderate ability group consisted of 343 boys and 457 girls, and the low ability group was made up of 96 boys and 104 girls. Analyses at three ability levels revealed that 11 items (24%) showed uniform GDIF; four items (items 1, 8, 24 and 25) were easier for girls at all levels of ability and seven items (items 5, 12, 14, 21, 27, 31 and 46) were easier for boys. Results showed that GDIF exists in the items used in the mathematics national examination. Less than 20% of the studied test item needs to be considered for further analysis by content and assessment experts for judgment of bias or unfairness.

The third stage of analysis was conducted to identify any item type that can be associated with the existence of GDIF. The analysis was to find out what type of item gives more advantage to the girls or boys? This study showed that girls excelled on text items (without graphics). Five out of six text items flagged for significant GDIF (83%) were easier for girls. Boys however, received more advantage from items containing graphics such as pictures, graphs, charts,

diagrams and tables. Seven out of eight graphics items flagged for significant GDIF (87%) were easier for boys. When items were grouped according to constructs, girls showed a slightly better ability in items testing both Knowledge and Skills. Girls did better in 11 out of 19 Knowledge items and 14 out of 27 items categorized as Skills item. In terms of area in mathematics, girls did better in items categorised as Number and Relationship. Boys did better in items categorised as Shape and Space.

Table 4 shows GDIF analyses results when comparing boys and girls according to ability levels, type of school and school location. Analysis at three ability levels showed that the boys in the high ability group did better in all seven types of items, compared with the girls of the same ability. Analyses of the moderate and low ability groups showed boys' better performance in items categorised as Graphics and Shape and Space and girls' better in items categorised as Text and Number. Seven items were flagged for significant and uniform GDIF. Analysis showed that girls did better on items 8, 24 and 25 across the three ability levels. These items are all commonly characterized as Text and Knowledge items. On the other hand, the boys uniformly did better on items 5, 12, 31 and 46. These items are commonly characterised as Graphics, with the exception of item 5. GDIF analyses were also conducted to compare boys and girls of the same type of school and same school location. Based on four analyses, it was found that girls consistently did better in Text and Number items. Graphics and Shape and Space items consistently were easier for boys.

Table 4
GDIF Analyses According To Ability Level, School Type, and Location

Type of Item	Ability Levels			Type of School		Location of School	
	High	Mod	Low	Single	Coed	Urban	Rural
Text	B	G	G	G	G	G	G
Graphics	B	B	B	B	B	B	B
Knowledge	B	G	B	B	G	G	G
Skills	B	B	G	B	G	B	G
Number	B	G	G	G	G	G	G
Shape and Space	B	B	B	=	B	B	B
Relationship	B	G	B	B	G	B	G

Note:
B : Easier for boys
G : Easier for girls
= : Equally easy for boys and girls

Eight GDIF analyses provided evidence to show that girls consistently do better in items labeled as Text and Numbers, whereas boys consistently do better in items labeled as Graphics and Shape and Space. GDIF analysis also showed topics in which boys and girls are inclined. Generally, this study showed that the girls found topics such as Standard Form, Basic Numbers, Trigonometry, Algebra and Set to be easier, whereas the boys found Polygon, Circle, Statistics, Linear Equation, Straight Line, Transformation, Area Under Graph and Ratio items to be easier. It was observed that findings concerning Algebra, Standard Form, Basic Numbers, and Statistics are rather sound. Conclusions made about the other topics are not considered stable due to less number of items used to represent the topics.

Discussion

The constant overachievement of girls in the Malaysian mathematics national examination has driven the researchers to explore this phenomenon from the assessment aspect, in order to find out whether it could be attributed to type of item used. Looking into each studied test item, the existence and intensity of gender-related differential item functioning (GDIF) were explored. By investigating which type of item systematically gives an advantage to each gender group, the question of whether type of item used in the test contributes to the success of girls, is answered.

The findings showed that girls did better in the mathematics test on the average. This finding supports studies done by Bellar and Gafni (1995) and Wester and Henriksson (2000). GDIF existed in the test, with more items showed to be in favour of the girls. The types of items that systematically gave advantage to the girls were items categorised as Text and Number. The types of items that systematically gave advantage to the boys were items categorised as Graphics and Shape and Space. Findings are coherent with previous studies which concluded that girls performed better in items presented without diagrams and boys performed better in items presented with diagrams (Othman, 2003); that girls outperformed boys in items that require verbal skills and boys outperformed girls in items that require spatial skills (Halpern 1992; Walstead & Robson, 1997; Kleinfeld, 1998). One reason for these observations is gender differences in cognitive abilities, as explained by the Theory of Multiple Intelligences by Gardner (1983). Although inconsistent at times, girls are generally found showing superior verbal ability, whereas boys are consistently more superior in terms of logic-mathematics and spatial ability (Gipps & Murphy, 1994; Elliot, Kratochwill, Littlefield, & Travers, 1996; Gardner, 1999). Despite general conclusion made about girls' superiority in verbal tasks, it was noted that one item (item 5), categorised as Text and

Number showed significant GDIF index in favour of the boys (DIF size = 0.55 logits; t = 3.89). Figure 1 shows item 5, translated to English for the purpose of illustration.

5. A rectangular piece of land with a length of 10.2 km and a width of 6 km, is to be divided equally among 20 land buyers. Calculate the area obtained by each buyer, in m^2.

A. 8.50_10^2

B. 3.06_10^3

C. 8.50_10^5

D. 3.06_10^6

Figure 1. Example of item easier for boys.

Item 5 was a difficult item (item measure = 0.90 logits), testing problem-solving skills in the context of Standard Form. Looking at its characteristics, this item should be easier for the girls. This item required students to transform a unit from kilometer to meter. Based on conclusion made by Fennema (2000), the complexity in calculations to solve the problem presented in item 5, could explain the observed boys' advantage in this particular item. This finding also supports the conclusions that type of verbal test makes a difference (Hyde & Linn, 1988) and item's superficial appearance alone cannot explain differential performances between gender groups (Santacreu, 2004).

This study provides considerably stable results regarding Text, Graphics, Number and Shape and Space types of items. Due to lack of agreement in the analysis results, strong conclusions about the other types of items (Knowledge, Skills and Relationship) cannot be made. Although the results showed that there were more items in favour of girls, multiple analyses did not provide enough evidence to show that type of item alone can be the cause for girls' overachievement in the test. Four reasons for this conclusion are

presented: (1) only two out of seven types of items were shown to consistently give advantage to the girls. This is the same number as those shown to consistently give advantage to the boys, indicating that there is no imbalance usage of item type in the test; (2) Analyses at different ability levels revealed that the number of items showing uniform, significant GDIF in favour of girls is less than the number that uniformly favour the boys. Boys from high ability group performed better on all seven types of items. This shows that girls did not receive too much advantage from the items and girls' advantage in Text and Number items is not uniform across the three ability levels; (3) Inclination of girls and boys towards certain topics could add one crucial factor that can explain student differential ability to answer item correctly – topic, which probably interacts with the type of items used. Garner and Engelhard (1999) have shown that findings regarding content of items are more consistent than those regarding type of items. However, more studies investigating mathematics topics should be conducted to confirm these preliminary findings. (4) This study also showed the influences of external factors such as type and location of school. There is a possibility that external factors interact differently with boys and girls, resulting in differential ability to answer certain types of item.

Conclusion and Recommendations

Results show that GDIF exists, even in a test of high quality. Some items may function well with one group, but otherwise with the other. Some items with certain characteristics show inclination towards one gender. However, this study did not show that the type of items used in the test is giving extra advantage to the girls. Although the findings of this study are reliable, they may not be overgeneralised without further studies. Items exhibiting high and significant DIF indices should be reviewed by content specialists before a decision to either use or discard is made. Future studies are needed to understand why boys and girls perform differently

on GDIF items, especially when the explanation is not apparent from inspecting the content of an item.

Findings carry implications for both test developers and educators. Test developers must be sensitive to the occurrences of DIF and observe the types of items showing DIF in all subjects tested in the national examinations. Information on how items 'behave' towards different groups of students can help test developers to enhance test specifications, so that the test is not going to be too lop-sided in terms of design. Test developers who are aware of DIF would be able to control, to a certain extent, the proportion of item types in a particular test which will be best for the groups taking the test. This study is just the starting point of studying DIF existence in a Malaysian national examination. Test developers may not be able to understand the nature of DIF occurrences with just one study. The effects of different item presentations to test one particular construct still need to be investigated. This is what we have to explore in Malaysia and in the Asian context. With future DIF studies, we would soon discover which type of item presentation that would not produce such high DIF. With DIF analyses results and much experience, it is not impossible that a well-informed test item developer or a trained item writer would be able to anticipate how an item would perform when administered.

Test development work will need to take into account gender differences in test items if equivalent and fair tests are desired. The use of this instrument can be extended to investigate other factors such as ethnic groups, socioeconomic status or other types of schools that may contribute to DIF. DIF analysis can be applied to tests of other subjects. Researchers also recommend that DIF analysis is included in the test construction process in any institution responsible for developing tests and examinations. Educators can use information from DIF analyses to identify the strengths and weaknesses of their students so that more meaningful teaching and learning activities can be planned.

DIF analyses provide important quantitative information to the study of fairness in a test item, aimed to reduce, not to totally eliminate unfairness in a test. It is directly relevant to questions of differences in the performance of subgroups of examinees. Although it is undeniably difficult to construct a perfect test that is well-balanced and fair to every single group taking a test, DIF analysis is still a critical aspect to consider. If certain items show DIF and judged to be unfair or biased, removing them from the measurement instruments will enhance test validity. If DIF is not conducted, problematic items may not be discovered. An equal proportion of all item types may not be possible after applying DIF in test construction, but the effort would certainly produce the most well-thought and fair tests.

References

Bellar, M. & Gafni, N. (1995). International perspectives on the schooling and learning achievement of girls and boys as revealed in the 1991. *International Assessment of Educational Progress (IAEP).* Jerusalem: National Institute for Testing and Evaluation.

Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. New Jersey: Lawrence Erlbaum Associates.

Breland, H.M. (1991). A study of gender and performance on advanced placement history examinations. *College Board No. 91-4*. New York: College Entrance Examination Board.

Elliott, S.N., Kratochwill, T.R., Littlefield, J. & Travers, J.F. (1996). *Educational psychology: Effective teaching effective learning*. Dubuque: Brown and Benchmark Publishers.

Fennema, E. (2000). *Gender and mathematics: What is known and what do I wish was known.* Paper presented at the Fifth Annual Forum of the National Institute for Science Education. Detroit, Michigan, 22-23 May. [On line]. Available: http://www.wcer.wisc.edu/nise/News_Activities/Forums/Fennemapaper.htm (accessed 17 March 2003).

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. London: Heinemann.

Gardner, H. (1999). *Intelligence reframed: multiple intelligences for the 21st century*. New York: Basic Books.

Garner, M. & Engelhard, J.G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12(1)*, 29-51.

Gipps, C. & Murphy, P. (1994). *A fair test? Assessment, achievement and equity.* Buckingham: Open University Press.

Halpern, D.F. (1992). *Sex differences in cognitive abilities*. (2nd ed.). Hillsdale, N.J: Lawrence Erlbaum Associates.

Halpern, D.F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091-1102.

Hyde , J.S. & Linn, M.C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104(1)*, 53-69.

Klienfeld, J. (1998). The myth that schools shortchange girls: Social science in the serviceof deception. *ERIC Document Reproduction Service.* Washington, D.C.: Women's Freedom Network.

Lawrence, I.M. & Curley, W.E. (1989). *Differential item functioning for males and females on SAT-Verbal reading subscore items: Follow up study*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

O'Neill, K.A., Wild, C., & Mc Peek, W. M. (1989). *Gender-related differential item performance on graduate admission tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

O'Neill, K.A. & McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In Holland, P.W. & Wainer, H. (eds.) *Differential Item Functioning* (pp. 255-276). Hillsdale, N.J: Erlbaum.

Othman Lebar (2003). *Perbandingan pencapaian pelajar mengikut gaya belajar dan bentuk pentaksiran*. Paper presented at the Educational Psychology National Seminar, Langkawi. 5- 8 Mei 2003.

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

Rasch Measurement Transactions, 18:3. (2004). *When does a gap between measures matter?* [On line]. Available: http://www.209.41.153/rmt/rmt32a.htm (Accessed 17 November 2004)

Santacreu, J. (2004). Sex differences in verbal reasoning are mediated by sex differences in spatial ability. *The Psychology Record*. [On line]. Available: http://www.highbeam.com/library/doc3.asp (Accessed 9 October 2004)

Walstead, W.B. & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *The Journal of Economic Education*. 28(2):155-169. [On line]. Available: http://www.flinders.edu.au/education/iej (Accessed 1 January 2002).

Wester, A. & Henriksson, W. (2000). The interaction between item format and gender differences in mathematics performance based on TIMMS data. *Studies in Educational Evaluation, 26(1)*, 79-90.

Wild, C.L. & Mc Peek, W.M. (1986). *Performance of the Mantel-Haenszel statistic in identifying differentially functioning items*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.