The International Journal of Educational and Psychological Assessment
May 2012, Vol. 10(2)

Related DIF in Cognitive Components of Solving Routine Problems

Nabeel Abedalaziz
University of Malaya, Kuala Lumpur, Malaysia

Abstract

DIF may be attributed to item bias but may also reflect performance differences that the test is designed to measure (Pedrajita, 2009). This study examined the gender related DIF of Malaysian students in their cognitive solution processes of solving routine mathematical problems using Maentel-Haenszel (M-H) and Logistic Regression approaches. A total of 300 sixth grade Chinese and 400 sixth grade Malay students participated in the study. The Chinese sample consisted of 144 female and 156 male students, and the Malay sample consisted of 165 female and 235 male students. A set of 31 routine items was developed. Results of the study showed that overall there were a small gender differences (favoring females) on routine problem solving for Malay and Chinese samples. Examination of students' component solutions (translation, integration, planning, and execution) of solving routine problems revealed that for the both samples there were a significant DIF on the execution component items, whereas , no significant DIF on the translation, integration, and planning components items. Furthermore, Results indicated that the percentage of agreement between the two approaches in detecting DIF is relatively high.

Keywords: Execution, Integration, Logistic Regression, Mantel-Haenszel, Planning, Translation.

Introduction

Test items are designed to provide information about the examinee. Difficult items are designed to be more demanding, and easy items are less so. However, sometimes test items carry with them demands other than those intended by the test developer (Scheuneman & Gerritz, 1990). When personal attributes, such as gender systematically affect examinee performance on an item, the result can be differential item functioning (DIF).

Psychometricians define DIF more precisely as a situation where individuals who have the same ability, but are members of different subgroups, do not have the same probability of a correct response to an item (Hambleton et al., 1991). Operationally, when the item characteristic curves for two or more subgroups are different, the item is showing DIF (Hambleton et al., 1991).

Gender related differential item functioning is a constant concern on large-scale standardized achievement tests in mathematics because differences between females and males are often found (e.g., Bielinski & davison, 2001; Boughton et al., 2000; DeMars, 1998; Gamer &Engelhard, 1999; Scheuneman & Grima, 1997; Willingham & Cole, 1997, Abedalaziz, 2010, 2011). Presumably, because of the complexity of gender-related issues, results reported from a variety of studies are inconsistent and often even contradictory (Willingham & Cole, 1997; Hyde, 1991; Cleary, 1992; Cleary (1992) suggested that such contradictory results may be accounted for by disentangling effects of different cohorts, construct, and selectivity of the sample.

Reviews of research led to the conclusions that there were gender differences in mathematical problem solving that favored males based on the fact that male samples outperformed female samples in their studies (for example, Benbow & Stanley, 1980, 1983; Benbow, 1988; Casey et al., 1995; Gallagher & DeLisi, 1994; Royer, et al., 1999). However, these conclusions were often limited to an atypical population, normally talented or highly motivated or college bound students, and relying on the selection of measures and the particular experimental situations (Caplan & Caplan, 2005). The opposite evidence found among these high-ability populations even sometimes challenged these conclusions. For example, Pajares (1996) found that gifted girls outperformed gifted boys in mathematical problem solving.

Hyde et al. (1990) meta-analysis of 100 studies suggested that gender differences in mathematics performance were small but gender differences in mathematical problem solving with lower performance of women existed in high school and in college. Many factors such as cognitive abilities, speed of processing information; learning styles, socialization were suggested to have contributions to gender difference in mathematical problem solving (for example, Duff, Gunther, & Walters 1997; Kimball 1989; Linn & Petersen, 1985; Maccoby & Jacklin, 1974; Royer, et al., 1999).

Problem situations can establish a need to know, and foster the motivation for the development of concepts (NCTM, 1989). Therefore, students should be placed into classroom problem-solving situations from the very earliest stages of mathematics learning. Thus, problem solving is a major method for mathematics knowledge acquisition rather than merely applying the new learned mathematics knowledge to solve problems. NCTM advocates that learning is led by the search to answer questions: first at an intuitive, empirical level, then by generalizing, and finally by justifying (proving).

Routine problem solving stresses the use of sets of known or prescribed procedures (algorithms) to solve problems. Gradually, students are asked to solve more complex problems that involve multiple steps and include irrelevant data. Commencing with the concrete level, students are asked to develop their own story problem situations and demonstrate the solution process with manipulative and/or pictures and later with symbols. Such problems are later presented to the class for solution One-step, two-step, or multiple-step routine problems can be easily assessed with paper and pencil tests typically focusing on the algorithm or algorithms being used.

Mayer (1987) developed a model for analyzing cognitive components in solving word problems. In his model four cognitive components involved in solving mathematical word problems were classified and analyzed: translation, integration, planning, and execution. In order to solve a problem, a student must be able to translate each statement of the problem into a mathematical sentence or an equation. This translation process requires that the student understand English sentences. Second, the student must be able to integrate each of the statements of the problem into a coherent problem representation. This integration process requires schematic knowledge. Third, the student needs to find an adequate algorithm in order to solve the problem. This solution planning requires the student's strategic knowledge. The last component requires the student to flawlessly execute the algorithm. This solution execution requires the student's procedural knowledge. This model was

successfully applied to assess students' routine problem-solving skills in other studies (Mayer, Tajika, & Stanley, 1991).

In the past, researchers have explored how the gender differences in mathematics were related to various levels of tasks and age groups. Researchers consistently found that male students are superior in geometry and visualization (Geary, 1996). On the other hand, female students show superiority in computation based on the data available. With respect to the gender differences in mathematical problem solving, however, there are mixed results. For example, Marshall (1984) examined general differences of sixth-grade students' mathematical performance in solving computation (involving whole numbers, fractions, and decimals) and word problems. She found that female students are more likely than male students to perform computations successfully, while male students are more likely than female students to solve word problems successfully are.

In another study, Marshall and Smith (1987) explored the gender differences of third grade and sixth grade students on various tasks, including computation problems, word problems, and nontraditional problems. According to Marshall and Smith (1987), third grade female students performed better than male students for both computation tasks and nontraditional problems, but there is no significant gender difference on word problems. Sixth grade female students again performed better than male students for computation tasks did, but there were not significant differences on word problems and nontraditional problems.

Hough (2003) examined the gender differences of U.S. and Chinese students in their solution processes of solving routine and non-routine mathematical problems. Results of the study showed that overall there were statistically significant gender differences (favoring males) on both routine and non-routine problem solving for the U.S. sample, but not for the Chinese sample. However, examinations of students component processes (translation, integration, planning, and execution) for solving routine problems revealed that significant gender differences only exist for the execution component (computation skills) for the U.S. sample.

In conclusion, a large body of literature reports that there are gender differences in mathematical problem solving favoring males. The literature has consistently reported that males outperform females on mathematics problem solving among high ability students on standardized mathematics tests. These genders related differences are generally obvious in high school and in college and can be traced back to the very early stage of elementary schooling. Furthermore, these gender differences are varying across mathematical tasks. It is found that students' strategy use is related to cognitive abilities, speed of processing information, physiological differences in brains, influences of sex hormones, learning styles, learners' attitudes, and stereotype threat in mathematics tests, differences in socialization, and the impact of socioeconomic variables (Hembree, 1992)).

The significance of this study is related to the importance of problem solving in the current mathematics education reform and the goal of achieving equal educational outcomes in students' learning of mathematics (National Council of Teachers of Mathematics (NCTM), 1989, 2000). Since mathematics is no longer just a prerequisite subject for prospective scientists and engineers but is a fundamental aspect of literacy for the twenty-first century (Mathematics

Sciences Education Board, 1993; NCTM, 1989), male and female students should have equal opportunity to learn mathematics, have equal treatment within classrooms, and achieve equal mathematics educational outcomes (Fennema & Leder, 1990). The examination of gender-related performance differences on routine allows for investigating gender differences in their thinking and reasoning as they solve these problems

Current education reform in general and mathematics education reform in particular emphasize the importance of thinking, understanding, reasoning, and problem solving in students' learning (e.g., NCTM, 1989, 1991, 2000; National Research Council, 1989). Such reform effort in mathematics curriculum and instruction requires examination of male and female students' thinking, reasoning, and problem solving rather than merely computation and symbol manipulation. The uniqueness of this study was its investigation of gender differences of relatively large samples of Malaysian students using routine problem solving test. Moreover, the present study tries to detect a gender related DIF of cognitive processes of solving routine mathematical problems.

This study provided an opportunity to examine issues in mathematics learning in general and issues in gender – related differential item functioning of routine problem solving in specific. The present study sought answers to the following questions: To what extent does the two methods (i.e. Mantel-Haenszel & Logistic Regression) agree or disagree in the identification DIF? (2) Are there gender differences in cognitive components of solving routine word problems? (3) Are gender differences linked to content areas within mathematics?

Method

Samples

A total of 300 sixth grade Chinese and 400 sixth grade Malay students participated in the study. The Chinese sample consisted of 144 (48%) female and 156 (52%) male students, and the Malay sample consisted of 165 (41%) female and 235 (59%) male students. Chinese and Malay samples are students from different public schools in Kuala Lumpour / Malaysia. Schools and students had been selected randomly during the second semester of the school year 2009- 2010.

Instrument

Mayer's model was used to examine gender differences in the processes of solving routine problems. A set of 31 multiple-choice items was used to assess component processes of solving routine problems: five items for the translation component, five items for the integration component, 5 items for the planning component, and 16 items for the execution component. The execution component involved students' computation skills (addition, subtraction, multiplication, and division) on different types of numbers (whole numbers, decimals, and fractions).

The test was tried out on a sample of 200 students-males and females, to make sure that the items of the test are clear and are understood by those being tested, and to find out the psychometric properties of the test. Accordingly, the item analysis revealed levels of difficulty from .31 to .92. Besides, it revealed

that the detractors were reversal to the item discriminate. Data about validity of the scale were collected through three methods: Internal consistency, item analysis, content validity.

The Cronbach's alpha coefficients calculated for the Execution, Translation, Integration, and Planning subscales were .86, .70, .70, and .71, respectively, and it was calculated to be 0.87 for the entire scale. The scale correlation coefficients ranged between .37 and .47 on execution component, between .39 and .58 on translation component, between .39 and .46 on integration component, and between .37 and .62 on planning component. It is generally agreed that correlations in the range of .38 to .63 are useful and statistically significant beyond the 1% level, whereas correlations less than .25 are not useful and statistically non-significant (Brown, 1983). Thus, the results show that the alpha coefficients for all subscales were significantly high, suggesting that the internal reliability index of the four constructs and the entire scale is adequate.

DIF Detection Procedure

Methods for detecting DIF have proliferated in recent years and have been reviewed. The various methods include techniques that tested differences in relative item difficulty among different groups, differences in item discrimination among different groups, differences in the item-characteristic curves (ICCs) for different groups, differences in the distribution of incorrect responses for various groups, and differences in multivariate factor structures among groups (Abedalaziz, 2010, 2011). A plausible but not exhaustive classification of DIF detection techniques is as follows: Classical Test Theory CTT-based methods, Factor Analysis FA-based methods, χ^2-based methods, and Item Response Theory IRT-based methods.

According to Cole and Moss (1989), the work on DIF has focused upon the last two approaches, namely, those based upon χ^2 and IRT. Unlike CTT-based methods, these last two approaches are conditional methods. In turn, χ^2-based methods can be divided into four different groups: (1) the χ^2- based methods in the strict sense, i.e., the χ^2 correct (Scheuneman, 1979) and the χ^2_{full} (Camilli, 1979); (2) theMantel-Haenszel (MH) procedure (Holland &Thayer, 1988), a natural outgrowth of the former χ^2 methods, which is widely used and easy to implement; (3) the Loglinear Models (LM) (Mellenbergh, 1982) to test the conditional independence of group membership and the score on the studied item given the matching variable, and (4) the Logistic Regression (LR) procedure (Rogers &Swaminathan, 1993). This paper compared the potential of some of these methods for detecting DIF (i.e. Mantel-Haenszel and Logistic Regression). Test experts and developers should use contingency table (CT) methods, particularly the LR and MH methods, in item DIF detection. These two methods are viable in the detection of DIF and are widely implemented in both test construction and research settings (Pedrajita & Talisayon, 2009).

Mantel-Haenszel (M-H) procedure

.

The Mantel-Haenszel (M-H) procedure was originally used to match subjects retrospectively on cancer risk factors in order to study current cancer rates (Mantel & Haenszel, 1959). The procedure has since been adapted to

study differential item functioning and is now the primary DIF detection device used at the Educational Testing Service (ETS; Dorans & Holland, 1992). The M-H method works by first dividing subgroups into the reference group and the focal group. The focal group is of primary interest in the analysis and is compared to the reference group after being matched on θ (Uttaro & Millsap, 1994). The total test score usually serves as the θestimate and the performance of the reference and focal groups is compared at unit intervals of θ weighted by the number of examinees at each level (Scheuneman & Gerritz, 1990). From this comparison, an odds-ratio estimator can be calculated, and a χ^2 test of significance can be carried out to assess the presence of DIF.

To assess the degree of DIF present, the odds-ratio estimator can be transformed onto the ETS "delta metric" (Δ; Dorans & Holland, 1992). The Δ statistic represents the difference in item difficulty for the reference and focal groups after the total score has been taken into account (Scheuneman & Gerritz, 1990). The advantage of using the Δ statistic to classify degree of DIF present is that the ETS has defined the values of it into a classification scheme delineated by Dorans and Holland (1992). A Δ value of 0.0 indicates no DIF, a positive value indicates DIF favoring the focal group, and a negative Δ value reflects DIF that favors the reference group.

In the present study, the M-H technique was implemented. First, subgroup members were matched based on their total scores for each scale. The matching variable was the 0.2 z unit divisions of the continuous θ scale, as was done in Harvey and Greenberg (1996). Once it was verified that there were no empty cells along the θ scale in any subgroups, a χ^2 statistic was calculated for each item to see if there were any overall differences in item endorsement rates across subgroups. Then, the χ^2's p value, as well as the M-H odds-ratio estimator (transformed to lie on the ETS delta [Δ] metric described by Dorans, & Holland, 1992), were examined to assess the degree of DIF present. Positive values of Δ favored females (i.e. they were more likely to endorse the item in the keyed direction), and negative Δ values favored males. More specifically, there are three possible degrees of DIF: (a) negligible DIF, where χ^2 is non-significant or the absolute value of Δ is less than 1.0; (b) intermediate DIF, where χ^2 is significant and Δ is between 1.0 and 1.49 in absolute value; and (c) large DIF, where χ^2 is significant and the absolute value of Δ is 1.5 or larger (Dorans & Holland, 1992).

Logistic Regression Procedure

Swaminathan and Rogers (1990) applied the Logistic Regression (LR) procedure to DIF detection. This was a response, in part, to the belief that the identification of both uniform and non-uniform DIF was important. The strengths of this procedure are well documented. It is a flexible model-based approach designed specifically to detect uniform and non-uniform DIF with the capability to accommodate continuous and multiple ability estimates. Furthermore, simulation studies have demonstrated comparable power in the detection of uniform and superior power in the detection of non-uniform DIF compared to the Mantel-Haenszel (MH) and Simultaneous Item Bias Test (SIB) procedures (Rogers & Swaminathan, 1993; Swaminathan & Rogers,

1990). These studies also identified two major weaknesses in the LR DIF procedure: 1) the Type I error or false positive rate was higher than expected, and 2) the lack of an effect size measure.

Logistic Regression (LR) is based on transforming data by taking their natural logarithms so as to reduce nonlinearity. In other words, logistic regression uses the logistic curve that best approximates the distribution of the data. Logistic regression estimates parameters using maximum likelihood estimation (Pedrajita, 2009). Logistic regression has a formal mathematical equivalence to the log linear model approach of Mellenbergh (1982): Coefficients for group, total score, and interaction terms are estimated and tested for significance with a model of comparison strategy. However, logistic regression is highly similar to standard ordinary least squares regression. It can be conceptualized as an equation that uses group, ability, and group-by-ability terms to predict whether an item response is right (1) or wrong (0). This property is desirable for didactic purposes.

Logistic regression uses the examinee as the unit of analysis, and has the following form:

$$P(u/x,g) = \frac{e^{(1-u)[-\beta_0-\beta_1 x-\beta_2 g-\beta_3(xg)]}}{1+e^{[-\beta_0-\beta_1 x-\beta_2 g-\beta_3(xg)]}}$$

Where:

g: represents group membership (0 for focal group (female) and 1 for reference group (male)).

x: the matching group (the observed total test score).

u: represents the item response value (0 for an incorrect answer and 1 for correct answer).

xg: represents the interaction between the matching variable and the group variable..

β_0: $\beta_1\beta_2\ and\beta_3$: Parameters to be estimated.

The above equation is used for predicting the probabilities of correct and incorrect responses to each dichotomously scored item, given an observed total test score and its associated group membership. Once the estimates of the four coefficient parameters, β_0: $\beta_1\beta_2\ and\ \beta_3$, for an item are obtained from a sample of test responses, the usual likelihood ratio chi-square tests of significance of the estimates of $\beta_2\ and\ \beta_3$ are conducted to examine if DIF exist. The null hypothesis is that $\beta_2 = \beta_3 = 0$. An item shows uniform DIF if $\beta_2 \neq 0$ and $\beta_3 = 0$ with 1 degree of freedom and non-uniform DIF if $\beta_3 \neq 0$ (whether or not $\beta_2 = 0$) with 1 degree of freedom (Swaminathan & Rogers, 1990).

In the present study, the item reveals uniform DIF when the significant odd ratio is for the group, whereas the item reveals nonuniform DIF when the significant odd ratio is for the interaction between the group and total score. The item reveals DIF in favor of males when the significant odd ratio is greater than one, whereas the item reveals DIF in favor of females when the significant odd ratio is less than one (α=0.05).

Results

Several items were found to contain DIF in the comparisons made. There were two comparisons: Malay males versus Malay females, and Chinese males versus Chinese females.

Tables 1 shows the group means, Δ, and χ^2 statistic obtained in the Malay males versus Malay females comparison (i.e. The summary results of the M-H method to identify Differential Item Functioning on the mathematics routine problems for each of the thirty-one items). According to the ETS criteria, six items or 19 percent of the items revealed "large DIF" (i.e. the items: 4, 21, 24, 25, 28 were in favor of females, whereas the item: 20 was in favor of males). Combining the items that exhibited either "intermediate" or "large" DIF shows that there were eight or 26 percent of the items revealed DIF (i.e. the items: 4, 5, 21, 24, 25, 26, 28 were in favor of females, whereas the item: 20 was in favor of males). The value of Δ signifies DIF in favor of males was -1.54, whereas the significant range of Δ for female students were from 1.03 to 6.75.

Tables 2 shows the group means, Δ, and χ^2 statistic obtained in the Chinese males versus Chinese females comparison (i.e. The summary results of the M-H method to identify Differential Item Functioning on the mathematics routine problems for each of the thirty-one items). According to the ETS criteria, two items or 6 percent of the items revealed "large DIF" (i.e. the items: 13, 18 were in favor of females). Combining the items that exhibited either "intermediate" or "large" DIF shows that there were seven or 23 percent of the items revealed DIF (i.e. the items: 2, 6, 8, 13, 14 18, 21 were in favor of females). The range of Δ signifies DIF in favor of females were from 1.03 to 1.96.

Appendix 1 shows the summary results of the Logistic Regression method to identify Differential Item Functioning for each of the thirty-one items for Malay sample. Four items or 13 percent of the items revealed DIF (i.e. the items: 24 and 26 were revealed uniform DIF in favor of females, whereas the item: 4 and 11 were revealed non-uniform DIF in favor of females).

Appendix 2 shows the summary results of the Logistic Regression method to identify Differential Item Functioning for each of the thirty-one items for Chinese sample. Five items or 16 percent of the items revealed DIF (i.e. the items: 2, 8 and 14 were revealed uniform DIF in favor of females, whereas the item: 6 and 13 were revealed non-uniform DIF in favor of females).

Table 1
Summary Result of the M-H Analysis: Malay Males versus Malay Females

Item	Group mean		Δ	χ^2	P. value
	Male	Female			
1.	0.14	0.30	0.29	3.26	0.02
2.	0.29	0.53	0.18	2.40	0.12
3.	0.42	0.56	0.27	5.73	0.02
4.	0.24	0.41	6.76**	28.75	0.00
5.	0.57	0.78	1.03*	51.77	0.00
6.	0.76	0.83	-0.02	0.01	0.92
7.	0.50	0.64	0.42	13.14	0.00
8.	0.37	0.48	-0.44	13.16	0.00
9.	0.35	0.51	-0.17	1.69	0.19
10.	0.24	0.32	-0.73	29.75	0.00
11.	0.91	0.93	-0.14	0.42	0.52
12.	0.45	0.65	-0.22	2.63	0.10
13.	0.29	0.48	-0.18	1.35	0.25
14.	0.73	0.82	-0.06	0.13	0.72
15.	0.25	0.21	-0.88	45.57	0.00
16.	0.82	0.88	-0.10	0.31	0.58
17.	0.04	0.06	-0.41	2.52	0.11
18.	0.30	0.43	-0.19	2.21	0.14
19.	0.16	0.22	-0.53	13.43	0.00
20.	0.50	0.41	-1.54**	141.57	0.00
21.	0.25	0.41	1.91**	45.10	0.00
22.	0.16	0.13	-0.96	31.39	0.00
23.	0.42	0.42	-0.30	5.34	0.02
24.	0.43	0.62	2.04**	51.01	0.00
25.	0.46	0.64	2.41**	341.66	0.00
26.	0.27	0.34	1.13*	49.64	0.00
27.	0.19	0.26	0.95	52.10	0.00
28.	0.29	0.42	1.67**	197.22	0.00
29.	0.78	0.76	-0.41	9.06	0.00
30.	0.57	0.65	0.85	49.11	0.00
31.	0.44	0.91	-0.43	15.61	0.00

Note. * The item revealing intermediate DIF.
** The item revealing large DIF.

Table 2
Summary Result of the M-H Analysis: Chinese Males versus Chinese Females

Item	Group mean		Δ	χ^2	P. value
	Male	Female			
1.	0.34	0.40	0.30	4.82	0.03
2.	056	0.62	1.03*	77.68	0.00
3.	0.50	0.57	0.43	33.08	0.00
4.	0.70	0.78	0.72	29.43	0.00
5.	0.61	0.55	-0.89	75.81	0.00
6.	0.26	0.81	1.12*	73.36	0.00
7.	0.70	0.72	-0.21	2.62	0.11
8.	0.26	0.81	1.12*	0.39	0.53
9.	0.41	0.40	-0.84	37.30	0.00
10.	0.48	0.50	-0.22	3.29	0.07
11.	0.26	0.78	-0.42	7.59	0.01
12.	0.62	0.71	0.84	42.63	0.00
13.	0.43	0.57	1.81**	28.90	0.00
14.	0.63	0.73	1.09*	56.03	0.00
15.	0.38	0.44	0.40	6.76	0.01
16.	0.75	0.81	0.51	10.05	0.00
17.	0.27	0.36	0.82	44.69	0.00
18.	0.12	0.18	1.96**	123.21	0.00
19.	0.52	0.60	0.53	20.88	0.00
20.	0.45	0.54	0.42	11.11	0.00
21.	0.64	0.70	1.11*	90.31	0.00
22.	0.49	0.48	-0.91	41.76	0.00
23.	0.85	0.89	0.66	15.05	0.00
24.	0.57	0.61	-0.07	0.28	0.60
25.	0.50	0.50	-0.49	17.55	0.00
26.	0.53	0.57	-0.02	0.03	0.87
27.	0.63	0.73	-0.30	5.33	0.02
28.	0.18	0.58	0.78	28.13	0.00
29.	0.43	0.53	-0.67	28.67	0.00
30.	0.67	0.82	0.21	1.47	0.22
31.	0.14	0.30	0.29	3.26	0.07

Note. * The item revealing intermediate DIF.
** The item revealing large DIF

In order to inspect the consistency between the two approaches in detecting DIF for Malay sample, the percentage of agreements between the two approaches was computed (i.e. the percentage of items revealing or not revealing DIF). Table 3 summarizes the consistency in which the two approaches flagged the items. The percentage of agreement between the Mantel-Haenszel and Logistic Regression approaches is 81% (i.e. 3+22/31 =78%).

Table 3
The Agreement between the Mantel-Haenszel Approaches in Detecting DIF

	Results From Logistic Regression		
Results From Mantel-Haenszel approach	No. of Non-flagged Items	No. of flagged Items	Marginal total
No. of non-flagged items	22	1	23
No. of flagged items	5	3	8
Marginal total	27	4	31

In order to inspect the consistency between the two approaches in detecting DIF for Chinese sample, the percentage of agreements between the two approaches was computed (i.e. the percentage of items revealing or not revealing DIF). Table 4 summarizes the consistency in which the two approaches flagged the items. The percentage of agreement between the Mantel-Haenszel and Logistic Regression approaches is 81% (i.e. 5+24/31 =93%).

Table 4
The Agreement between the Mantel-Haenszel Approaches in Detecting DIF for Chinese Sample

	Results From Logistic Regression		
Results From Mantel-Haenszel approach	No. of Non-flagged Items	No. of flagged Items	Marginal Total
No. of nonflagged items	24	0	24
No. of flagged items	2	5	7
Marginal total	26	5	31

Discussion

This study examined the gender-related DIF of Malaysian (i.e. Malay. and Chinese) students in cognitive components of solving routine problems. The results of the differential item functioning analysis showed that there were statistically gender related DIF present in cognitive process of solving routine problems. Overall, it appears that male and female examinees performed fairly. Results of the study showed that overall there were gender differences (favoring females) on routine problem solving for the Malay and Chinese samples. Examination of students' component solutions (translation, integration, planning, and execution) of solving routine problems revealed that for both samples there were a significant DIF on the execution component items, but not on the translation, integration, and planning components items.

The Logistic Regression and the Mantel-Haenszel Statistic yielded very similar results with respect to uniform differential item functioning (DIF). The two procedures result in similar number and identity of items being identified. Hence, there is high degree of correspondence between these two procedures. In summary, the percentage of agreement between the two approaches in detecting DIF for both samples are relatively high (78% for Malay sample and

93% for Chinese sample), however, this may due to: both methods related to classical theory of measurement. This finding seems to be consistent with the previous studies (e.g., Hambleton &Rogers, 1989; Skaggs & Lists, 1992; Hakim & Cohen, 1995; Abedalaziz, 2010).

The MH and LR techniques share some convenient features, i.e., ease of implementation or applicability, associated tests of statistical significance and, finally, the requirement of the number of examinees needed to get satisfactory results is not quite as large as with other conditional methods.

Further, results of the study showed that overall there were a small gender related DIF (favoring females) on Execution component items for Malay and Chinese samples. The findings in the present study seem to be consistent with the findings from previous studies. Hyde, et al., (1990) suggested that there were very small or null gender differences in mathematics performance on Scholastic Assessment Test-Mathematics (SAT-M) tests. Caplan and Caplan (2005) even argued that the link between gender and the mathematics performance was very weak.

How can we explain the finding that there were small genders related DIF for Malaysian students in cognitive components of solving routine problems? In Malaysian society, women and men tend to have equal opportunities for jobs and equal salaries. Thus, the finding that there was no gender related DIF in solving routine problems for Malaysian samples may be explained by the fact that Malaysian students are raised in relatively more uniform educational and social conditions. In addition, the mathematics curriculum of sixth grade concentrates on teaching problem solving ability.

The finding that female students outperformed male students on the execution component (measuring computation skills) in the present study seems to be consistent with the findings from previous studies. In fact, previous studies reported that in general, female students were more successful than male students were on computation tasks (Geary, 1996; Hyde et al., 1990; Marshall & Smith, 1987; Abedalaziz, 2011). Only the item 20 which measures mathematical thinking was in favor of Malay males, The finding from the present study seem to be consistent with the findings from several meta-analyses (e.g., Wilder & Powell, 1989; Hyde et al., 1990), which have revealed that male students usually score higher than female students on tasks requiring mathematical thinking and problem solving.

In conclusion, the present study provides evidence that the link between gender and mathematics routine problems was very weak. Also, the study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even content are closely tied to curriculum. Furthermore, assuming that females' better performance on computation does indicate a reliance on algorithmic learning, women might benefit even more than men from an instructional strategy that relies less on teaching algorithms and more on teaching problem solving (Abedalaziz, 2010) .

The conclusions drawn from this study have to be focused on the particular characteristics of the tests and samples of examinees considered. It would be useful to widen the comparison of DIF detection techniques to other conditions. Special attention should be paid to factors such as percentage of DIF items, test length, and the type and magnitude of DIF. These factors could have an important impact on DIF detection. Therefore, it would be interesting to check whether these results are generalizable to conditions other than those

considered here. Among the factors which may affect the DIF results, are the number of DIF items in the test, difficulty level of items, the degree of DIF in items, the number of individuals in each group, the difference between the group means.

Since the one limitation of the M-H procedure is that it may lack power to detect DIF that is not uniform across the range of θ scores (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990), further researches are needed, to detect a gender related DIF of Malaysian students in routine problems using different DIF methods. A natural extension of this study is to examine gender differences of low and high school students in their solution processes of solving routine mathematical problems.

The findings deserve further comment. *First*, the number of items exhibiting DIF with both the LR and the MH procedures seems very low. *Second*, consistent with earlier research, the MH and the LR procedures result in similar number of items (and similar items) being identified (Rogers & Swaminathan, 1993). Thus, there is a high degree of correspondence between the LR and the MH procedures when either one or two ability estimates are included in the analysis. LR has shown that under comparable conditions, when matching is based on a single test score, it produces results that are extremely similar to those produced using the MH Statistic. Methods for detecting DIF may be evaluated in terms of external evidence of validity. Some possible types of validity evidence for a bias technique would be a demonstration that: (1) the procedure is not selecting items at random; and (2) the results obtained with different methods tend to agree. The LR and MH procedures appear to have demonstrated the external validity evidence mentioned above. Hence, these two approaches are widely implemented in DIF detections (Pedrajita &Talisayon, 2009).

References

Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *The International Journal of Educational and Psychological Assessment 5*(2), 101-116.

Abedalaziz, N. (2011). Detecting DIF using item characteristic curve approaches. *The International Journal of Educational and Psychological Assessment, 8*(2), 1- 15.

Abedalaziz, N. (2010). Detecting gender related DIF using logistic regression and Mantel-Haenszel approaches. *Procedia Social and Behavioral Sciences, 7 (*c), 406-413.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences, 11*(2), 169-232.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210*(4475), 12-62.

Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science, 222*(4627), 10-29.

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38*(1), 51-77.

Boughton, K. A., Gierl, M. J., & Khaliq, S. N. (2000, Jun). Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance. Paper presented at the annual meeting of the Canadian Society in Education, Edmonton, Alberta, Canada

Brown, F. G. (1983). *Principles of educational and psychological testing*. New York: Holt, Rinehart & Winston.

Camilli, G. (1979). *A critique of the chi-square method for assessing item bias*. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Caplan, J. B. & Caplan, P. J. (2005). The preservative search for sex differences in mathematics abilities. *Sex roles, 37(*7-8), 477-494.

Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*(4), 697-705.

Cleary, T. A. (1991, May). Gender differences in aptitude and achievement test scores. Paper presented at the ETS invitational conference on sex equity in educational opportunity, achievement and testing, Princeton, NJ.

Cole, N. S., & Moss, P. A. (1989). *Bias in test use: Educational measurement*. New York: Macmillan

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(3), 279-299.

Dorans, N. J., & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenszel and standardization*. (RR-92-10). Princeton, NJ: Educational Testing Service

Duffy, J., Gunther, G., & Walters, L. (1997). Gender and mathematical problem solving. *Sex Roles, 37*(7), 477-494.

Fennema, E., & Leder, G. C. (1990). *Mathematics and gender*. PO, Wiliston: Teachers College Press.

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

Gallagher, A. M., & De Lisi, R. (1994). Gender differences in Scholastic Aptitude Test: Mathematics problem solving among high-ability students. *Journal of Educational Psychology, 86*(2), 204- 211.

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences, 19*(2), 229-246.

HaKim, S., & Cohen, A. (1995). Comparison of Lord chi-square and Rajus measures and the likelihood method on detecting of differential item functioning. *Applied Measurement In education, 14*, 291-312.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Harvey, R. J., & Greenberg, S. E. (1996). *Gender-based differential item functioning in the Myers-Briggs type indicator: Implications for employee*

selection and big-five inventories. Unpublished manuscript. Virginia Polytechnic Institute and State University.

Hembree, R. (1992). Experiments and relational studies in problem solving: A meta-analysis. *Journal for Research in Mathematics Education, 23*(3), 242-273.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, *23*, 129-145.

Hough, David (2003). Exploring Gender Differences of U.S, and Chinese Students in Their Solution Processes of Solving Routine and Nonroutine Mathematical Problems, *Research in Middle Level Education (RMLE), 26*(1), 1-25.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological bulletin, 107*(2), 139.

Hyde, J.S. (1991). Gender differences in mathematics performance: A meta analysis .*Psychological Bulletin, 10*(792), 139-155.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological bulletin, 105*(2), 198- 214.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, *56*, 1479-1498.

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences: Stanford University Press (Stanford, Calif.).

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute, 22,* 719-748.

Marshall, S. P. (1984). Sex differences in children's mathematics achievement: Solving computations and story problems. *Journal of Educational Psychology, 76*(2), 194-204.

Marshall, S. P., & Smith, J. D. (1987). Sex differences in learning mathematics: A longitudinal study with item and error analyses. *Journal of Educational Psychology, 79*, 372-383.

Mathematics Sciences Education Board (1993). A conceptual guide for mathematics assessment. Washington, DC: National Academy Press.

Mayer, R. E. (1987). *Educational psychology: A cognitive approach*. Boston: Little, Brown.

Mayer, R. E., Tajika, H., & Stanley, C. (1991). Mathematical problem solving in Japan and the United States: A controlled comparison. *Journal of Educational Psychology, 83*(1), 69-72.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational and Behavioral Statistics, 7*(2), 105-118.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Research Council. (1989). *Everybody counts.* Washington, DC: National Academy of Sciences.

Pajares, F. (1996). Self-efficacy beliefs and mathematical problem-solving of gifted students. *Contemporary Educational Psychology, 21*, 325-344.

Pedrajita, J. Q. (2009). Using logistic regression to detect biased test items. *The International Journal of Educational and Psychological Assessment, 2, 54-73*.

Pedrajita, J. Q., & Talisayon, V. M. (2011). Identifying biased test items by differential item functioning analysis using contingency table approaches: A comparative study. *Education Quarterly, 67*(1), 21- 43.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105- 116.

Royer, J. M., Tronsky, L. N., Chan, Y., Jackson, S. J., & Marchant, H. (1999). Math-fact retrieval as the cognitive mechanism underlying gender differences in math test performance. *Contemporary Educational Psychology, 24*(3), *181-266*.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*(3), 143-152.

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*(2), 109-131.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examinees. *Applied Measurement in Education, 10*(4), 299-319.

Skaggs, G., & Lissitz, R. W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement, 29*(3), 227-242.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27,* 361-370.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement. 18,* 15-25.

Wilder, G. Z., Powell, K., & Board, C. E. E. (1989). *Sex differences in test performance: A survey of literature*. New York: College Entrance Examination Board.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* London: Lawrence Erlbaum, Associate Publishers.

The International Journal of Educational and Psychological Assessment
May 2012, Vol. 10(2)

Appendix 1

Summary Result of the Logistic Regression Analysis for Malay Sample

Item	Variable	Statistical significance	Odds-Ratio	Type of DIF
1	Group	0.12	1.70	
	Interaction	0.11	0.94	
2	Group	0.17	1.74	
	Interaction	0.34	1.04	
3	Group	0.60	0.81	
	Interaction	0.20	1.05	
4	Group	0.64	1.23	
	Interaction	0.03	0.87	Nouniform
5	Group	0.00	0.14	
	Interaction	0.06	1.26	
6	Group	0.16	0.55	
	Interaction	0.09	1.08	
7	Group	0.70	0.86	
	Interaction	0.58	1.03	
8	Group	0.21	2.70	
	Interaction	0.11	0.94	
9	Group	0.00	0.25	
	Interaction	0.07	1.14	
10	Group	0.08	2.62	
	Interaction	0.08	0.88	
11	Group	0.08	0.17	
	Interaction	0.00	1.25	Nonuniform
12	Group	0.13	1.99	
	Interaction	0.48	0.97	
13	Group	0.10	2.69	
	Interaction	0.22	0.92	
14	Group	0.70	0.88	
	Interaction	0.74	0.99	
15	Group	0.25	0.57	
	Interaction	0.67	0.98	
16	Group	0.04	2.19	
	Interaction	0.32	0.99	
17	Group	0.06	3.15	
	Interaction	0.21	0.90	
18	Group	0. 50	0.77	
	Interaction	0.69	0.99	
19	Group	0.12	0.49	
	Interaction	0.71	0.01	
20	Group	0.75	0.87	
	Interaction	0.62	0.98	
21	Group	0.11	1.77	
	Interaction	0.07	0.94	
22	Group	0.32	0.13	
	Interaction	0.13	1.05	
23	Group	0.23	0.34	
	Interaction	0.12	1.04	
24	Group	0.05	0.26	Uniform
	Interaction	0.13	0.57	
25	Group	0.45	1.91	
	Interaction	0.08	0.90	
26	Group	0.04	0.66	Uniform
	Interaction	0.13	0.56	
27	Group	0.13	0.38	
	Interaction	0.09	1.07	
28	Group	0.07	3.52	Nonuniform
	Interaction	0.02	0.88	
29	Group	0.50	0.77	
	Interaction	0.69	1.00	
30	Group	0.18	2.19	
	Interaction	0.09	0.94	
31	Group	0.50	0.77	
	Interaction	0.69	1.00	

Appendix 2

Summary Result of the Logistic Regression Analysis for Chinese Sample

Item	Variable	Statistical significance	Odds-Ratio	Type of DIF
1	Group	0.07	2.45	
	Interaction	0.12	0.91	
2	Group	0.04	0.74	Uniform
	Interaction	0.31	1.09	
3	Group	0.63	0.89	
	Interaction	0.22	1.15	
4	Group	0.64	1.23	
	Interaction	0.83	1.01	
5	Group	0.00	0.14	
	Interaction	0.16	1.27	
6	Group	0.16	0.55	
	Interaction	0.03	0.88	Nonuniform
7	Group	0.70	0.86	
	Interaction	0.09	1.03	
8	Group	0.01	0.79	Uniform
	Interaction	0.11	0.94	
9	Group	0.00	0.25	
	Interaction	0.08	1.14	
10	Group	0.08	2.62	
	Interaction	0.21	0.88	
11	Group	0.07	0.19	
	Interaction	0.08	1.29	
12	Group	0.13	1.95	
	Interaction	0.48	0.97	
13	Group	0.10	2.69	
	Interaction	0.02	0.92	Nonuniform
14	Group	0.04	0.89	Uniform
	Interaction	0.72	0.95	
15	Group	0.25	0.57	
	Interaction	0.67	0.98	
16	Group	0.14	2.19	
	Interaction	0.21	0.99	
17	Group	0.06	2.16	
	Interaction	0.07	0.90	
18	Group	0. 09	0.77	
	Interaction	0.66	0.91	
19	Group	0.17	0.44	
	Interaction	0.76	0.09	
20	Group	0.78	0.87	
	Interaction	0.32	0.76	
21	Group	0.09	0.79	
	Interaction	0.52	1.02	
22	Group	0.32	0.19	
	Interaction	0.16	1.55	
23	Group	0.13	0.39	
	Interaction	0.12	1.32	
24	Group	0.09	1.07	
	Interaction	0.17	0.57	
25	Group	0.35	1.81	
	Interaction	0.23	0.85	
26	Group	0.08	1.26	
	Interaction	0.13	0.56	
27	Group	0.06	0.34	
	Interaction	0.09	1.11	
28	Group	0.07	3.52	
	Interaction	0.09	0.91	
29	Group	0.09	0.71	
	Interaction	0.69	1.05	
30	Group	0.32	1.09	
	Interaction	0.13	0.87	
31	Group	0.50	0.64	
	Interaction	0.69	1.03	

About the Author

Dr. Nabeel Abedalaziz *is a lecturer in the Department of Educational Psychology at the University of Malaya/Malaysia* where he teaches courses in measurement, evaluation, educational research, Testing theory, psychological testing, data analysis of quantitative research, quantitative approaches in evaluation, instrument design and item development, and statistics. He took his undergraduate in Yarmouk university/Jordan with the degree Bachelor of mathematics. He took his Master's degree in Education major in measurement and evaluation at Yarmouk University. He received his PhD in Educational Psychology major in Measurement and Evaluation at Amman Arab University of graduate study/Jordan. He was worked at UNRWA institutes from 1997-2009. Majority of his research uses quantitative and DIF techniques in the field of educational psychology.

e-mail: nabeelabdelazeez@yahoo.com
Mobile number:+60162501958
Office TEL:+60379675171

.