US-China Education Review B 5 (2011) 734-744
Earlier title: US-China Education Review, ISSN 1548-6613

Detecting a Gender-Related DIF Using Logistic Regression and Transformed Item Difficulty

Nabeel Abedlaziz, Wail Ismail, Zaharah Hussin
University of Malaya, Kuala Lumpur, Malaysia

Test items are designed to provide information about the examinees. Difficult items are designed to be more demanding and easy items are less so. However, sometimes, test items carry with their 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 DIF (differential item functioning). The purpose of this study was to examine gender differences in performance on multiple-choice mathematical ability test, designed to match six grade curriculums. The LR (logistic regression) method and transformed item difficulty were used to detect a gender related DIF. A random sample of 800 tenth grade students was selected. DIF analysis indicated that: (1) Females showed a statistically significant and consistent advantage over males on numerical ability, whereas men showed a consistent advantage over females on spatial ability and deductive ability; (2) The percentage of agreement between the two approaches in detecting DIF is relatively low; and (3) Gender differences in mathematics may well be linked to content.

Keywords: DIF (differential item functioning), transformed item difficulty, LR (logistic regression), mathematical ability

Introduction

Standardized tests and measurements are used primarily to distinguish between ability levels of examinees. As a part of the determination of validity for these tests, differential item analysis is employed to evaluate the degree to which measurements distinguish true abilities among examinees in an unbiased manner. Psychometricians and test developers use DIF (differential item functioning) analysis to determine if there is a possible bias in a given test item. DIF is determined in a two-step process. The first step is the comparison of two groups' outcome on an item and determining the presence of DIF. The second step includes a decision of whether there is a large enough difference between the groups to eliminate or change the item of interest.

DIF is said to be present when examinees from different groups have differing probabilities of success on an item after controlling for overall ability (Clauser & Mazor, 1998). If an item is free of bias, responses to that item will be related only to the level of the underlying trait that the item is trying to measure. If item bias is present, responses to the item will be related to some other factors as well as the level of the underlying trait (Camilli & Shepard, 1999). The tight relationship between the probability of correct responses and ability or

Nabeel Abedlaziz, Ph.D., senior lecturer, Faculty of Education, Department of Educational Psychology and Counseling, University of Malaya.
Wail Ismail, Ph.D., senior lecturer, Department of Educational Foundation, Faculty of Education, University of Malaya.
Zaharah Hussin, Ph.D., senior lecturer, Department of Educational Foundation, Faculty of Education, University of Malaya.

trait levels is an explicit assumption of IRT (item response theory) (Hambelton, Swaminathan, & Rogers, 1991) and an implicit assumption of classical test theory (McDonalds, 1999). The presence of large numbers of items with DIF is a severe threat to the construct validity of tests and the conclusions based on test scores derived from items with and items without DIF.

Test items with content bias may: (1) contain content that is differentially familiar to matched groups of examinees; (2) contain sources of difficulty that are irrelevant to the construct adversely affecting test performance; (3) contain material that may be offensive, demeaning, or emotionally charged which can lower examinees' motivation and attention for the remainder of the test, thereby, decreasing performance on other questions apart from the offending items; and (4) ask for information that students have not had equal opportunity to learn. Test items with gender bias may contain: (1) tasks which perpetuate undesirable role stereotypes, race stereotypes or gender stereotypes; (2) materials or references that may be offensive to members of one gender; and (3) references to objects and ideas that are likely to be more familiar to men or to women (Pedrajita, 2009).

Several techniques have been promulgated for the statistical assessment of DIF. Several excellent reviews are available (Clauser & Mazor, 1998; Camilli & Shepard, 1999; Millsap, 1993). Most techniques for DIF assessment has been developed in educational settings in which items are generally dichotomously scored as correct or incorrect.

Methods for detecting DIF have proliferated and have been reviewed in recent years. 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 ICCs (item-characteristic curves) for different groups, differences in the distribution of incorrect responses for various groups and differences in multivariate factor structures among groups.

A number of approaches have used item difficulty as the focus of analysis. An item is considered biased in this approach if, compared to other items on the test, it is relatively more difficult for one group than for another. One of the more widely implemented techniques of this type is TID (transformed item difficulty). LR (Logistic Regression) is based on transforming data by taking their natural logarithms so as to reduce nonlinearity. In other words, LR uses the logistic curve that best approximates the distribution of the data. LR estimates parameters using maximum likelihood estimation (Pedrajita, 2009). LR has been known for some time to be useful for the assessment of effect modification in observational studies and enables analyses of continuous predictor variables without requiring stratification. Not surprisingly, simulation studies from educational testing experts have found that LR-based DIF detection techniques enables the detection of both uniform and non-uniform DIF. Uniform DIF is said to apply when differences between groups in item of responses are found at all trait levels, while in non-uniform DIF an interaction is found between trait level, group assignment and item responses (Camilli & Shepard, 1999; Jodin & Gierl, 2001).

Gender Differences in Mathematics

In the past few decades, research has repeatedly reported gender differences in mathematics performance on a number of standardised mathematics tests such as the SAT-M (Scholastic Assessment Test-Mathematics) (Gallagher, 1990, 1992; Gallagher & DeLisi, 1994; Willingham & Cole, 1997; Hyde, Royer, Tronsky, Chan, Jackson, & Marchant, 1999). The test scores on these standardized tests have been regarded as an important measure of abilities to do mathematics problems (Casey, Nuttall, Pezaris, & Benbow, 1995; Halpern, 2000;

Stumpf & Stanley, 1998). But results from these studies are not consistent: Some found that males generally outperformed females on mathematical tasks (Maccoby & Jacklin, 1974; Fennema & Carpenter, 1981; Halpern, 2000); some showed different sizes of gender differences with respect to types of mathematical tasks (D. Voyer, S. Voyer, & Bryden, 1995). Hyde, Fennema and Lamon (1990) suggested that there was very small or null gender difference in mathematical ability on these tests. T. B. Caplan and P. J. Caplan (2005) even argued that the link between gender and the mathematical ability was very weak.

Battista (1990) conducted a study among 145 high school geometry students from middle-class communities. This research examined the role that spatial visualization and verbal-logical thinking played in gender differences in geometric problem-solving in high school. The findings suggested that males and females differed in the level of discrepancy between spatial and verbal abilities.

Gallagher, De lisi, Holst, McGillicuddy-De Lisi, Morely, and Cahalan (2000) suggested that males tended to be more flexible than females in applying solution strategies. Kessel and Linn (1996) and Gallagher (1998) reported that females were more likely than males to adhere to classroom-learned procedures to solve problems, so they might be less likely to use shortcuts and estimation techniques for solving unfamiliar and complex problems quickly.

Current education reform in general and mathematics education reform in particular emphasize the importance of thinking, understanding, reasoning and mathematical ability in students' learning (e.g., NCTM (National Council of Teachers of Mathematics), 1989, 1991, 2000; National Research Council, 1989). Such reform effort in mathematics curriculum and instruction requires examination of male and female students' thinking, reasoning, problem-solving and mathematical ability rather than merely computation and symbol manipulation. This study provided an opportunity to examine issues in mathematics learning in general and issues in gender-related differential item functioning of mathematical ability in specific.

Purpose

This study aimed to detect DIF of mathematics ability test. This study can significantly contribute to educational research. Test experts and developers may: (1) gain insights on the applicability of DIF detection method(s); (2) realize the validity of DIF methods in detecting a gender biased test items; (3) use DIF methods in developing valid and equitable tests; and (4) employ DIF methods in purifying their assessment instruments.

This study sought answers to the following questions: To what extent do the two methods (i.e., transformed item difficulty and LR) agree or disagree in detecting a gender-related DIF? A second question was: What is the nature of cognitive ability of those items identified as revealing DIF? A third question was: Are gender differences linked to content areas within mathematics?

Method

Participants

A total of 800 (380 males and 420 females Grade 10) students in Jordan were targeted as participants in this study at the end of the first semester in the school year of 2009-2010.

Instrument

A mathematical ability scale was developed as a part of this study. The scale compressed of 30 multiple-choice items to measure three components of mathematical ability (i.e., numerical ability, deductive ability and spatial ability). Psychometric properties of the test reveal some items needing revision. Nonetheless,

reliability is reported KR (Kuder-Richardson)-20 indices to be 0.91. Spearman-Brown correction on split-half reliability for odd even comparison also show similar results $r = 0.89$. Validity of the instrument was shown using inter-correlation of the scale (0.19 to 0.855). Factor analysis reveals that the test measure one trait (unidimensionality).

Logistic Regression (LR)

Swaminathan and Rogers (1990) applied the 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 MH (Mantel-Haenszel) and SIB (Simultaneous Item Bias) test 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.

LR 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 comparison strategy. However, LR 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.

LR uses the examinee as the unit of analysis and has the following form:

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: β_1, β_2 and β_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_1, β_2 and β_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 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 one 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 non-uniform 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 ($\alpha = 0.05$).

Transformed Item Difficulty (TID)

Angoff (1972) offered the delta-plot or TID method which involves computing the difficulty or p-value

(proportion of subjects getting item right) for each item separately for each group. Using tables of the standardized normal distribution, the normal deviate z is obtained corresponding to the (1-p) the percentile of the distribution, i.e., z is the tabled value having proportion (l-p) of the normal distribution below it. Then to eliminate negative z-values, a delta value is calculated from the z-value by the equation $\Delta = 4z + 13$. A large delta value indicates a difficult item. For two groups, there will be a pair of delta values for each item. These pairs of delta values can then be plotted on a graph, each item represented by a point on the graph. Δ Line can be fitted to the plot of points and the deviation (distance) of a given point from the line is taken as measure of that item's bias, large deviations indicating much bias. In the present study, the distance that each point deviates from the major axis of the ellipse was calculated. The equation used for the major of the ellipse was $Y = AX + B$ (the best fitting line) in which: Y represents males delta values (Δ_M), X represents females delta values (Δ_F), and:

$$B = \mu_x - A\mu_y$$

Where:

A: Represents a line slope;

B: The line sector of Y-*axis*;

μ_y: The mean of delta values for females (Δ_F);

$\mu_x \rightarrow$ The mean of delta values for males (Δ_M); and

$$A = \frac{(\sigma_Y^2 - \sigma_X^2)^2 \pm \sqrt{(\sigma_Y^2 - \sigma_X^2)^2 + 4r_{XY}\sigma_Y^2\sigma_X^2}}{2r_{XY}\sigma_Y^2\sigma_X^2}$$

Where:

σ_X : The standard deviation of the deltas for males group;

σ_Y : The standard deviation of the deltas for females group;

σ_{XY}: The correlation between males and females' deltas.

The Perpendicular distance that each point deviates from the major *axis* was calculated from the formula:

$$D_i = \frac{AX_i - Y_i + B}{\sqrt{A^2 + 1}}$$

Where:

X_i : Represents males' delta value for item i;

Y_i : Represents females delta value for item i. (D_i)

Those items with (D_i) values in excess of one standard deviation reveal DIF (Osterlind, 1983). In this study, the larger (D_i) is, the more biased the item. A signed transformed difficulty measure of DIF, which preserved both the direction and magnitude of DIF, was obtained by attaching a positive sign to (D_i) if the item reveals DIF in favor of females and a negative sign if the item reveals DIF in favor of males **(**Abedalazeez, 2010).

Results and Discussion

Table 1 shows the summary results of the LR method to identify DIF on the mathematical ability scale for each of 30 items. Seventeen items or 43% of the items revealed DIF (i.e., the items1, 5, 8, 21, 22, 23, 24 and 26 were revealed uniform DIF, whereas the items 9, 10, 11, 13, 16, 27, 28 and 40 were revealed non-uniform DIF). The items 1, 8, 10, 13, 16, 21, 24, 26, 28 and 40 were in favor of males, whereas the items 5, 9, 11, 22, 23 and 27 were in favor of females.

Table 1

Summary Result of the LR Analysis

Item	Variable	Statistical significance	Odds-Ratio	Type of DIF
1	Group	0.0146	2.6975	Uniform
	Interaction	0.1106	0.9361	
2	Group	0.1693	1.7383	
	Interaction	0.3355	1.0433	
3	Group	0.5994	0.8146	
	Interaction	0.1895	1.0547	
4	Group	0.6374	1.2300	
	Interaction	0.8349	1.0110	
5	Group	0.0001	0.1410	
	Interaction	0.0601	1.2605	Uniform
6	Group	0.1599	0.5481	
	Interaction	0.0883	1.0835	
7	Group	0.7014	0.8516	
	Interaction	0.5831	1.0257	
8	Group	0.0136	2.6964	Uniform
	Interaction	0.1103	0.9361	
9	Group	0.0046	0.2549	
	Interaction	0.0238	1.1367	Non-uniform
10	Group	0.0790	2.6214	
	Interaction	0.0072	0.8781	Non-uniform
11	Group	0.0800	0.1749	
	Interaction	0.0001	1.2538	Non-uniform
12	Group	0.1331	1.9886	
	Interaction	0.4770	0.9655	
13	Group	0.0972	2.6889	
	Interaction	0.0214	0.9166	Non-uniform
14	Group	0.7044	0.8767	
	Interaction	0.7403	0.9897	
15	Group	0.2468	0.5718	
	Interaction	0.6688	0.9830	
16	Group	0.0387	2.1889	
	Interaction	0.0234	0.9874	Non-uniform
17	Group	0.0559	3.1515	
	Interaction	0.0033	0.8967	Non-uniform
18	Group	0.4950	0.7738	
	Interaction	0.6900	0.9870	
19	Group	0.1244	0.4939	
	Interaction	0.7097	0.0139	
20	Group	0.7527	0.8720	
	Interaction	0.6213	0.9828	
21	Group	0.0068	1.7702	Uniform
	Interaction	0.5432	0.9951	
22	Group	0.0177	0.1328	Uniform
	Interaction	0.1285	1.0529	
23	Group	0.0303	0.3351	Uniform
	Interaction	0.1151	1.0351	

(to be continued)

24	Group	0.0490	1.2586	Uniform
	Interaction	0.1257	0.5649	
25	Group	0.4476	1.9074	
	Interaction	0.2477	0.9048	
26	Group	0.0490	1.2579	Uniform
	Interaction	0.1256	0.5648	
27	Group	0.1285	0.3759	Non-uniform
	Interaction	0.0266	1.0739	
28	Group	0.0673	3.5217	Non-uniform
	Interaction	0.0242	0.8809	
29	Group	0.4951	0.7740	
	Interaction	0.6901	0.9880	
30	Group	0.1693	1.7383	
	Interaction	0.3355	1.0434	
31	Group	0.5995	0.8148	
	Interaction	0.1897	1.0549	
32	Group	0.1798	1.7384	
	Interaction	0.3355	1.0439	
33	Group	0.5998	0.8147	
	Interaction	0.2895	1.0547	
34	Group	0.1696	1.7386	
	Interaction	0.3455	1.0433	
35	Group	0.8992	0.8145	
	Interaction	0.2894	1.0548	
36	Group	0.1694	1.7383	
	Interaction	0.3456	1.0433	
37	Group	0.5993	0.8145	
	Interaction	0.1891	1.0547	
38	Group	0.1694	1.7384	
	Interaction	0.3358	1.0433	
39	Group	0.5993	0.8146	
	Interaction	0.1893	1.0544	
40	Group	0.1842	2.1923	Non-uniform
	Interaction	0.0348	0.9362	

Table 2 shows the DIF statistic of the TID method for each of 40 items. The TID method flagged ten items at the significance level of 0.05 (the item 27 was in favor of female students, whereas the items 1, 14, 19, 20, 25, 33, 34, 37 and 39 were in favor of male students).

Table 3 summarizes the consistency in which the TID and LR methods flagged the items. The two methods were agreeable in allocating 14 items as revealing DIF, and ten items as not revealing DIF. As such, the percentage of agreement between TID and LR methods is 45% (i.e., 16 + 2/40 = 45%).

Discussion and Conclusions

In summary, the percentages of agreement among the two approaches in detecting DIF are relatively low. Not surprisingly, simulation studies from educational testing experts have found that LR-based DIF detection techniques enable the detection of both uniform and non-uniform DIF, whereas TID DIF detection techniques unable the detection of both non-uniform DIF.

Table 2

Summary Results From the TID Method to Identify Differential Items Functioning on a Mathematical Ability Test

Item	P_M	P_F	Z_M	Z_F	Δ_M	Δ_F	D_i
1	0.87	0.83	- 1.10	- 0.97	8.60	9.12	- 1.20*
2	0.62	0.62	- 0.31	- 0.29	11.76	11.84	- 0.45
3	0.16	0.12	1.01	1.17	17.04	17.68	- 0.01
4	0.88	0.84	- 1.20	1.00	8.20	17.00	- 6.44
5	0.88	0.89	- 1.18	- 1.23	8.28	8.08	- 0.80
6	0.63	0.68	- 0.32	- 0.47	11.72	11.12	- 0.04
7	0.70	0.64	- 0.53	- 0.36	10.88	11.56	- 0.96
8	0.67	0.55	- 0.45	- 0.38	11.20	11.48	- 0.66
9	0.58	0.64	- 0.21	- 0.36	12.16	11.56	0.03
10	0.38	0.20	0.86	0.86	16.44	16.44	0.30
11	0.23	0.13	1.12	1.12	17.48	17.48	0.45
12	0.64	0.62	- 0.36	- 0.31	11.56	11.76	- 0.56
13	0.67	0.60	- 0.43	- 0.25	11.28	12.00	- 0.93
14	0.67	0.50	- 0.44	0.01	11.24	13.04	- 1.61*
15	0.60	0.50	- 0.24	0.00	12.04	13.00	- 0.96
16	0.71	0.74	- 0.56	- 0.65	10.76	10.40	- 0.33
17	0.63	0.70	- 0.32	- 0.53	11.72	10.88	0.11
18	0.51	0.46	- 0.02	0.10	12.92	13.40	- 0.53
19	0.60	0.55	- 0.36	- 0.13	11.56	12.48	- 1.01*
20	0.82	0.71	- 0.91	- 0.56	9.36	10.76	- 1.64*
21	0.60	0.53	- 0.23	- 0.07	12.08	12.72	- 0.76
22	0.69	0.79	- 0.49	- 0.8	11.04	9.80	0.26
23	0.51	0.47	- 0.02	0.07	12.92	13.28	- 0.46
24	0.77	0.79	- 0.72	- 0.82	10.12	9.72	- 0.40
25	0.61	0.54	- 0.37	- 0.11	11.52	12.56	- 1.09*
26	0.64	0.68	- 0.37	- 0.48	11.52	11.08	- 0.17
27	0.54	0.84	- 0.10	- 1.00	12.60	9.00	1.97*
28	0.36	0.42	0.36	0.19	14.44	13.76	0.42
29	0.20	0.18	0.84	0.90	16.36	16.6	0.14
30	0.53	0.54	- 0.09	- 0.11	12.64	12.56	- 0.22
31	0.60	0.52	- 0.24	- 0.04	12.04	12.84	- 0.86
32	0.50	0.60	0.00	- 0.26	13.00	11.96	0.43
33	0.57	0.42	- 0.18	0.20	12.28	13.80	- 1.28*
34	0.67	0.57	- 0.44	- 0.19	11.24	12.24	- 1.11*
35	0.26	0.30	0.66	0.54	15.64	15.16	0.48
36	0.29	0.42	0.55	0.20	15.20	13.80	0.99
37	0.46	0.23	0.10	0.73	13.40	15.92	- 1.73*
38	0.57	0.48	- 0.18	0.05	12.28	13.20	- 0.90
39	0.34	0.19	0.41	0.87	14.64	16.48	- 1.12*
40	0.31	0.31	0.49	0.50	14.96	15.00	0.05

Notes. * The item reveal DIF; P_M item difficulty for males; P_F item difficulty for females; Δ_M delta value for males group; Δ_F delta value for females group; Z_M *z* score for males; Z_F *z* score for females.

Table 3
Pair Wise Agreement Between TID and LR Methods

	Results from TID		
Results from LR	No. of non-flagged items	No. of flagged items	Marginal total
No. of non-flagged items	16	8	24
No. of flagged items	14	2	16
Marginal total	30	10	40

The theoretical reasons for the lack agreement between both methods in the identification of DIF of items are given by Hunter (1975) who discussed several factors which may cause an item to be labeled as revealed DIF when, in fact, no DIF exists. These are: (1) non-unidimensional tests; (2) differences in ability distribution of the two groups; (3) differences in item quality; (4) guessing; and (5) nonlinearity of regression. Finally, one should consider the fairness of an item in addition to its statistical index of bias. Also, this result helps to explain the low and moderate agreement reported in the measurement literature among DIF methods concerning items flagged as reveal DIF. The fact is that studies of convergence of methods for investigating DIF are influenced greatly by the unreliability of the statistics (Abedalaziz, 2010).

The DIF analysis pointed to the conclusion that females had an advantage over males on the numerical ability, whereas males had an advantage over females on items involving spatial ability and deductive ability. The tendency for males to perform better than females on spatial ability and inductive ability and women to perform better on numerical ability is consistent with previous findings (Willson, Fernandez, & Hadaway, 1993; Gallagher, DeLisi, Holst, McGillicuddy-DeLisi, Morely, & Cahalan, 2000).

In previous studies, however, females usually performed better on number and computation. The fact that this test was tied to a specific curriculum did not appear to help females' performance. The researchers consistently found that male students are superior in geometry and visualization (Geary, 1996). On the other hand, females show superiority in computation based on the data available. Gender differences in achievement of mathematics in favor of boys have been found in standardized tests and are most prominent at the very high levels of achievement (Leder, 1992). These differences are likely to both content and ability dependent. While males outperform females in scientific and mathematical tasks, females outperform males in tasks involving verbal abilities.

There are many studies that focus on differences between men and women in tests (Gallaghe et al., 2000; Kimball, 1994; Willingham & Cole, 1997). From the findings of earlier studies, one conclusion that can be drawn is that men have a better spatial ability than women (Geary, 1996). Men use this spatial more often than women when solving problems, which can give advantages while solving certain kinds of problems in geometry (Geary, 1996). Many studies indicated that women are better than men in verbal skills, which can give them advantages on items where communication is important. Women also score relatively higher on tests in mathematics that better match coursework. Men tend to outperform women in geometry and arithmetic and algebraic reasoning questions. Women tend to be better at intermediate algebra and arithmetic and algebraic operations (Willingham & Cole, 1997). Gallagher et al. (2000) found men outperformed women in all kind of problems, but that the differences were greater for problems requiring spatial skills or multiple solution paths than for problems requiring verbal skills or containing classroom-based content.

Spatial abilities were reported to have relationship with mathematics test scores (Casey, Nuttall, Pezaris, & Benbow, 1995; Geary, Saults, Liu, & Hoard, 2000; Nuttall, Casey, & Pezaris, 2005). This relationship indicates

that gender differences in spatial abilities may contribute to gender differences in mathematical problem-solving.

The study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even when content is closely tied to curriculum. The presence of a gender related DIF in mathematical ability test can be attributed to: (1) the unfamiliar with the content of the items which caused the examinees to be attracted to the incorrect options; (2) the ambiguities in the item stem, keyed response, or distracter; (3) the disparities in the matched examinees' exposure to concepts or skills reflected on the items; and (4) the inability of the matched examinees to understand the concepts reflected on the items (Pedrajita, 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.

Angoff, W. H. (1972, May). A technique for the investigation of cultural differences. Paper presented at *the Annual Meeting of the American Psychological Association*, Honolulu.

Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education, 21*(1), 47-60.

Camilli, G., & Shepard, L. A. (1999). *Methods for identifying biased test items*. Sage: Thousand Oaks.

Caplan, J. B., & Caplan, P. J. (2005). The perseverative search for sex differences in mathematics abilities. In A. M. Gallagher, & J. C. Kaufman (Eds.), Gender differences in mathematics: An integrative psychological approach. Cambridge: Cambridge University Press.

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

Clauser B. E., & Mazor, K. M. (1998). Using statistical procedures to identify deferentially functioning test items. *Educational Measurement: Issues and Practic, 17,* 31-44.

Fennema, E., & Carpenter, T. P. (1981). Sex-related differences in mathematics: Results from national assessment. *The Mathematics Teacher, 74,* 554-559.

Gallagher, A. M. (1990). *Sex differences in the performance of high-scoring examinees on the SAT-M.* New York: College Entrance Examination Board.

Gallagher, A. M. (1992). *Sex differences in problem-solving strategies used by high-scoring examinees on the SAT-M.* Princeton, N. J.: Educational Testing Service.

Gallagher, A. M. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record, 100,* 297-314.

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

Gallagher, A. M., De lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced problem-solving. *Journal of Experimental Child Psychology, 75,* 165-190.

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Science, 19,* 229-284.

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency and arithmetical reasoning. *Journal of Experimental Child Psychology, 77,* 337-353.

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, N. J.: Lawrence Erlbaum Associates.

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

Hunter, J. E. (1975, December). A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the *National Institute of Education conference on test Bias*. Maryland.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139-155.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an eect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.

Kessel, C., & Linn, M. C. (1996). Grades or scores: Predicting future college mathematics performance. *Educational Measurement: Issues and Practice, 15,* 10-14.

Kimball, M. M. (1994). It is only a myth that girls are poorer in mathematics. *Kvinnovetenskaplig tidskrift, 15*(4), 39-53.

Leder, G. C. (1992). Mathematics and gender: changing perspectives. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning*. New York: Macmillan.

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

McDonald, R. P. (1999). *Test theory: A united treatment*. Lawrence Erlbaum: Mahwah, N. J..

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

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

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

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

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, V. A.: National Council of Teachers of Mathematics. Retrieved October 14, 2006, from http://www.NCTM.ORG/

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

Osterlind, S. J. (1983). *Test item bias*. Beverly Hills: Sage Publications.

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

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

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

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*(13), 109-131.

Stumpf, H., & Stanley, J. C. (1998). Standardized tests: Still gender biased? *Current Directions in Psychological Science, 7,* 335-344.

Swaminathan, H., & Rogers, H. J. (1990). Detecting dierential item functioning using logistic regression procedures. *Journal of Educational Measurement , 27*(4), 361-370.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A metaanalysis and consideration of critical variables. *Psychological Bulletin, 117,* 250-270.

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

Willson, J. W., Fernandez, M. L., & Hadaway, N. (1993). Mathematical problem-solving. In P. S. Wilson (Ed.), *Research ideas for the classroom: High school mathematics.* New York: MacMillan.