Detecting A gender-Related Differential Item Functioning Using Transformed Item Difficulty

Nabeel Abedalaziz [1], Chin Hai Leng [2], Ahlam Alahmadi [3]

[1] nabilabedalaziz@yahoo.com
Department of Educational Psychology & Counseling, Faculty of Education University of Malaya, Kuala Lumpur, Malaysia

[2] chin@um.edu.my
Department of Curriculum & Instructional Technology, Faculty of Education,University of Malaya,
50603 Kuala Lumpur, Malaysia

[3] Hamed.303@yahoo.com
Faculty of Education University of Malaya, Kuala Lumpur, Malaysia

ABSTRACT

The purpose of the study was to examine gender differences in performance on multiple-choice mathematical ability test, administered within the context of high school graduation test that was designed to match eleventh grade curriculum. The transformed item difficulty (TID) was used to detect a gender related DIF. A random sample of 1400 eleventh graders in Kuala Lumpur was selected. In DIF indexes, females showed a statistically significant and consistent advantage over males on items involving algebra, whereas males showed a less consistent advantage on items involving geometry and measurement, number and computation, data analysis, and proportional reasoning. However it was concluded that gender differences in mathematics may well be linked to content.

Keywords: *Differential Item Functioning, Transformed Item Difficulty, Item Bias*

INTRODUCTION

Educational or psychological tests may include items that operate differently for certain groups. It is important to identify these items because they may lead to unfair results for groups being compared. The reason for such items to operate differently may be gender, culture or language differences between groups. In the measurement literature the analyses to determine such items are called differential item functioning (DIF) analyses. And the items detected trough these analyses are called items functioning differentially among the groups, or shortly DIF items.

Assessment of test bias is important to establish the construct validity of tests. Assessment of differential item functioning (DIF) is an important first step in this process. DIF is present when examinees from deferent groups have differing probabilities of success on an item, after controlling for overall ability level. Here, we present analysis of DIF Bias is a serious problem in psychometric tests. Differential item functioning (DIF) is said to be present when examinees from different groups have differing probabilities of success on an item, after controlling for overall ability. 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 factor as well as the level of the underlying trait. The tight relationship between the probability of correct responses and ability or trait levels is an explicit assumption of Item Response Theory (IRT), and an implicit assumption of classical test theory. 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.

Gender Differences in Mathematics

There are many studies that focus on differences between men and women in tests (Benbow & Stanley, 1980, 1983; Hedges & Nowell, 1995; Stumpf & Stanley, 1996; Bielinski & Davison, 2001; Boughton, Gierl, & Khalaq, 2000; DeMars, 1998; Gamer & Engelhard, 1999; Scheuneman & Grima, 1997; Willingham & Cole, 1997; Gallagher De Lisi, Holset, Mc Gillicuddy-De Lisi, Morly & Cahalan, 2000; Kimball, 1994).

Researchers consistently found that male students are superior in geometry and visualization (Geary, 1996).

On the other hand, female show superiority in computation based on the data available. Gender differences in achievement in 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 be both content and ability dependent. While males outperform females in scientific and mathematical tasks, females outperform males in tasks involving verbal abilities.

From the findings of earlier studies, one conclusion 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 indicate 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 in 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 kinds 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.

PURPOSE

The purpose of this study was to analyze gender differences in performance on multiple-choice items mathematical ability test. Because there is now a movement away from exclusive reliance on multiple-choice assessment, led by the National Council of Teachers of Mathematics (NCTM, 1994), the ability meaningfully to compare performance on different item types is important.

This study sought answers to the following questions: Are there gender differences in mathematical ability? Are gender differences linked to content areas within mathematics?

METHOD

Participants

A total of 1400 (690 males and 710 females) eleventh grade students in Kuala Lumpur were targeted as participants in this study, during the ending period of the First semester, school year 2009- 2010.

Materials

The mathematical ability test was developed as part of this study. The 40-item instrument consists of four components (basic arithmetic, verbal arithmetic, elementary algebra, and geometry). Psychometric properties of the test reveal some items needing revision. Nonetheless, reliability is reported KR-20 indices to be 0.87. Spearman-Brown Correction on split-half reliabilities for odd even comparison also show similar results $r = .89$. Validity of the instrument was shown using inter-correlation of the sub scales (0.19 to 0.855). Confirmatory FactorAnalysis reveals that the test measure one trait (unidimensionality).

Data Analysis

Two sections of analysis were done to establish psychometric properties. First is using theclassical test theory steps which include the item analysis. Microsoft Excel was used for the analyses and computations involved in the CTT analysis. SPSS software was also used todetermine reliability of the test. The second is detecting DIF.

Detecting DIF

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.

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 described in Angoff and Ford (1973).

Angoff (1972) offered the *delta-plot* or transformed item-difficulty (TID) method. The method 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 Δ = 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. ΔLine 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_X), 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)

μ_x: The mean of delta values for males (Δ_M), and

$$A = \frac{(\sigma_Y^2 - \sigma_X^2)^2 \pm \sqrt{(\sigma_Y^2 - \sigma_X^2)^2 + 4r_{XY}\sigma_Y^2\sigma_X^2}}{2r_{XY}\sigma_Y^2\sigma_X^2} \qquad (1)$$

Where:

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

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

r_{XY}:The correlation between deltas for males and females.

ThePerpendicular distance(D_i) that each point deviates from the major axis was calculated from the formula:

$$D_i = \frac{AX_{i-}Y_{i+}B}{\sqrt{A^{2+}\mathrm{I}}} \qquad (2)$$

Where:

$X_{i:}$Represents males delta value for item *i*.

$Y_{i:}$Represents females delta value for item *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.

RESULTS

Table 1 shows the DIF statistic of the TID method for each of 40 items. The TID method flagged ten items at the .05 level of significance (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 1 provides the p-value (item difficulty) of test items for male and female students, which were relatively convergent.

From the analysis it appears that the effect of DIF in a well construct test was not very large. Neither group was greatly affected across all items since some items were revealed DIF in favor for each group.

Table 1: Summary Results from the TID method to Identify Differential Item 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

Note. * 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_Mz score for males. Z_F z score for females.

DISCUSSION

Both mean raw scores and DIF index point to the conclusion that females had an advantage over males on Algebra (item number 27), whereas males had an advantage on items involving proportional reasoning, Number and Computation, Data Analysis, and Geometry and Measurement. The tendency for males to perform better than females on Geometry and Measurement and females to perform better on Algebra is consistent with previous finding. 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 (for example, Halpern, 2000; Fennema & Carpenter, 1981; Maccoby & Jacklin, 1974).

Such a strong female advantage in algebra, as reflected in DIF indexes has not been previously noted. The advantage could perhaps be explained by noting that the algebra items were very abstract and algorithmic, unlike the items in the other component, the algebra component included no real-world situation, no geometry, no word problems. As already noted, females tend to do better on such problems. Perhaps these gender related differences in performance are a result of both a reliance on routines learned in class as proposed by Bohlin (1994), Hyde and Linn (1989), Kimball (1989), and Gallaghher (1992), and a lack of confidence on nonroutine tasks as suggested by Seegers and Boekaerts (1996).

For those five mathematics items identified by Agoff's technique as revealing DIF in favor of females, the p values ranged from .2922 to .5480 with a mean p value of .4293. The mean p value for males for all 90 mathematics items was .4937. The point biserial correlations for these five items tended to be in the moderate to high range, with only one item having a discrimination index below .30. Of the eight mathematics items identified as revealing DIF in favor of males, four were verbal application items. The content of these items dealt with such concepts as determining: (a) the number of ounces of chemical to be added to each 100 gallons of water, (b) the relative altitude of two towns in a desert, (c) the number of miles per gallon of gasoline averaged by a plane, and (d) the number of gallons of orange juice remaining in a tank if a certain percentage was lost by leakage. Two other of the night mathematics items involved geometry or referred to graphs. Of the remaining two items, one item dealt with finding the sum of two times expressed in hours and minutes and the other item involved the determination of a percent. Of the seven mathematics items identified as revealing DIF in favor of females, one item involved the multiplication of decimal fractions.The other six items required knowledge of the more abstract concepts of mathematics, such as, algebraic concepts, mathematical definitions such as prime numbers, associative property, and the expression of a number in expanded notation.

This 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. Furthermore, assuming that females' better performance on algebra 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 and effective means of approaching nonroutine problems.

REFERENCES

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.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95-105.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Facts or Artifact? *Science*, *210*(12), 1262-1264.

Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science*, *222*, 1029-1031.

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

Bohlin, C. F. (1994). Learning Style factors and mathematics performance: Sex-related differences. *International Journal of Educational Researchers, 21,* 387-397.

Boughton, K., Gierl, M. J., & Khaliq, S. (2000, May). *Differential bundle functioning on mathematics and science achievement tests*. Paper presented at the annual meeting of the Canadian Society in Education, Edmonton, Alberta, Canada.

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,* 279-299.

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

Gallagher, A. M. (1992). *Sex differences in problem-solving strategies used by high-scoring examinees on the SAT-M*. (College Board Rep. No. 92-2; ETS RR No. 92-33). New York, NY: College Entrance Examination Board.

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.

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

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

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

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41-45.

Hyde, J. S., & Linn, M. C. (1989). Gender difference in verbal ability.A meta-analysis. *Psychological Bulletin, 104,* 53-69.

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

Kimball, M. M. (1994). It is only a myth that girls are poorer in mathematics. *Kvinnovetenskapligtidskrift*, *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, NY: Macmillan.

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

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

Osterlind, S. J. (1983). *Test Item Bias*. Beverly Hills, CA: Sage.

Scheuneman, J. D., & Grima, A. (1997). Characteristic of quantitative word items associated with differential item functioning for female and black examinees. *Applied Measurement in Education, 4,* 299-320.

Seegers, G., & Boekaerts, M. (1996). Gender-related differences in self-referenced congnitions in relation to mathematics. *Journal for Research in Mathematics Education, 27*(3), 215-240.

Stumpf, H., & Stanley, J. C. (1996). Standardized tests: Still gender biased? *Current Directions in Psychological Science, 7,* 335-344.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.