

EXPLORING DIF: COMPARISON OF CTT AND IRT METHODS

Dr. Nabeel Abdelazeez

*Faculty of Education Department of Educational psychology and counseling, University of
Malaya, Kuala Lumpur, Malaysia*

e-mail: nabeelabdelazeez@yahoo.com

Mobile number: 006 0162501958

Office [TEL:0060379675171](tel:0060379675171)

DIF DIFFERENTIATION

- 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 [4].
- Uniform DIF, occurs when two ICC's differ, but are more or less parallel [6]. Uniform DIF is likely to occur when two ICC's have different b (difficulty) parameters and similar a (discrimination or slope) parameters [3].
- Nonuniform DIF, occurs when there is an interaction between ability level and subgroup membership [3], and the result is that the ICC's for the two subgroups cross at some ability value [6].

Gender differences in mathematics

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, Gierl & Khalaq, 2000; DeMars, 1998; Gamer & Engelhard, 1999; Scheuneman & Grima, 1997; Willingham & Cole, 1997). Presumably, because of the complexity of gender-related issues, results reported from a variety of studies are inconsistent and often even contradictory (Cleary, 1992; Hyde, 1991; Willingham & Cole, 1997).

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).

Research questions

The present study sought answers to the following questions: (1) To what extent do the four methods (i.e. area index for the two-parameter logistic model, transformed item difficulty, b-parameter difference, and Chi-square) agree or disagree in the identification of DIF? (2) What is the efficiency of the four methods in detecting DIF? (3) Are there gender differences in mathematical proficiency? (4) Are gender differences linked to content areas within mathematics?

Description of the Test Data and Examinees Samples

A mathematical proficiency test was developed in order to measure four components of the mathematical proficiency: Relations and functions, polynomial, Trigonometric functions, and triangles. The primary form of the scale (60 items) was tried out to a sample of 144 students-males and females, chosen from tenth grade. Accordingly, the final version of the scale compressed of 54 items.

The test of the mathematical proficiency was applied during the last quarter of the school – year 2009/2010 to samples of (1228) students- males and females- from the tenth grade (656 males, and 624 females). In Jordan.

Data about validity of the test were collected through four methods: Internal consistency, item analysis, Logical judgment, and Factor analysis. Cronbach alpha method was used to collect data about the reliability of the test ($\alpha = 0.91$). Confirmatory Factor Analysis reveals that the data obtain fits the models, and the test measures a single trait (unidimensionality).

DETECTING ITEM BIAS (DIF) METHODS

The various methods include techniques that examine (a) differences in relative item difficulty across different groups [25], (b) differences in item discrimination across groups [27], (c) differences in the item-characteristic curves for different groups [30], (d) differences in the distribution of incorrect responses for various groups [31], and, (e) differences in multivariate factor structures across groups [28].

Rudner (1977) and Scheuneman (1997) have noted the need to empirically compare the various methods.

Area Index for Two-Parameter Logistic Model

It measures the area between the two ICCs of the reference (males) and the focal (females) groups as an index of the difference between the performances of the two groups matched on ability. The larger the area, the larger the difference between the two curves.

In this study, a cut-off value (critical area= 0.220) was obtained by carrying out an analysis on two randomly equivalent groups. Because there is no DIF present, the largest area statistic obtained serves as an indicator of the greatest value of the statistic likely to occur by chance. This approach is not ideal; however, it does provide an approximate answer to the cut-off-score determination problem [6].

Raju [42] formula for the 2-parameter area index was used to find out the area between the two curves as follow:

$$Area = \left| 2 \frac{(a_2 - a_1)}{Da_1 a_2} \ln \left[1 + e^{\frac{Da_1 a_2 (b_2 - b_1)}{a_2 - a_1}} \right] - (b_2 - b_1) \right|$$

where:

a_1 :discrimination parameter for males (reference group).

a_2 : discrimination parameter for females (focal group).

b_1 : difficulty parameter for males (reference group).

b_2 : difficulty parameter for females (focal group).

D=1.7 (constant: scaling factor).

The item reveals DIF when the area between the two curves is greater than 0.220.

The direction of DIF is determine by testing ICC_S .

b-parameter difference

In the present study, the one-parameter logistic model was used to find out: the difficulty parameter for males and females by BILOG-MG program, and the difficulty difference was defined as follow:

$$\Delta b = b_F - b_R$$

Where:

b_F : Estimated difficulty parameter for males (reference group).

b_R : Estimated difficulty parameter for females (focal group).

Δb : Estimated difficulty parameter difference.

To test the significant of , the statistic d was defined as follow:

$$d = \frac{\Delta b}{S_{\Delta b}}$$

Where:

$$S_{\Delta b} = \sqrt{S_F^2 + S_R^2}$$

$S_{\Delta b}$: The standard error of b-difference.

S_F^2 : The variance for estimating b-parameter for females group.

S_R^2 : The variance for estimating b-parameter for males group.

Since d with normal distribution and similar to z scores, the normal probability distribution tables can be used to reference the level of significance under the null hypothesis $H_0: \Delta b = 0$ [10].

A positive value of the difference indicates DIF favoring the reference group, whereas a negative value of the difference indicates DIF favoring the focal group. In the present study, a significant value of d greater than or equal 1.96 indicates DIF favoring female students at 0.05 level, whereas a significant value of d less than or equal - 1.96 indicates DIF favoring male students at the 0.05 level [10].

Transformed Item Difficulty (TID) Method

The method involves computing the difficulty or p-value for each item separately for each group. The normal deviate z is obtained corresponding to the $(1-p)$ th percentile of the distribution, Then to eliminate negative z -values, a delta value is calculated from the z -value by the equation $A = 4z + 13$. 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. A line can be fitted to the plot of points; and the deviation of a given point from the line is taken as measure of that item's bias, large deviations indicating much bias [32].

In the present study, 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)

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

$$A = \frac{(\sigma_y^2 - \sigma_x^2) \pm \sqrt{(\sigma_y^2 - \sigma_x^2)^2 + 4r_{xy}\sigma_y^2\sigma_x^2}}{2r_{xy}\sigma_y^2\sigma_x^2}$$

Where:

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

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

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

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

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

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 \pm one unit reveals DIF. 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. In the present study a value of D_i greater than one unit indicates DIF favoring females, whereas a value of D_i less than minus one unit indicates DIF favoring males [35].

Chi Square DIF Method

The ability dimension is divided into discrete categories with the probability of correct responses in each category assumed constant. Scheuneman's version of the chi square method is concerned not only with frequencies of persons in each category as the usual chi square is, but with the number of correct responses made by persons in each group (or subpopulation) of interest. This is evident in the degrees of freedom for this method, which is $(k - 1)(r - 1)$ where k is number of subpopulations and r is the number of score groups, or categories.

Scheuneman's [30] modified χ^2 formula is:

$$\chi^2 = \sum [(B_s - B_o)^2 / B_s] + \sum [(W_s - W_o)^2 / W_s]$$

where B stands for subpopulation one (B_s : expected frequencies, B_o : observed frequencies) and W stands for subpopulation two (W_s : expected frequencies, W_o : observed frequencies). For comparison purposes the usual χ^2 formula is:

$$\chi^2 = \sum [(O - E)^2 / E]$$

where O is the observed frequency in a given category and E is the expected frequency in a given category. In the present study, the total score divided into eleven classes, and the values of chi square were computed for each item. The item reveals DIF at $\alpha=0.05$

Results

-Results from TID

Nineteen or 35 percent of items revealed DIF (the items: 25, and 26 were in favor of males and the items: 2, 10, 11, 12, 13, 16, 22, 23, 24, 28, 31, 32, 33, 39, 40, 44, and 47 were in favor of females). The range of D signifies DIF in favor of males were from -1.03 to -1.02, whereas the significant value of D for female students were from 1.01 to 1.91. Item difficulty (p) for each item indicates that the test is easier for females.

-Results from b-parameter difference

Forty-one or 75 percent of items were easier for females (i.e. the lowest value of b-parameter for one group indicates that the item is easy for this group), as such, the test is easier for females. The range of b-parameter difference signifies DIF in favor of males were from 0.189 to 1,708, whereas the range of b-parameter difference signifies DIF in favor of females were from -0.717 to -0.163. Thirty-two or 56 percent of fifty-four items revealed DIF (the items: 9, 18, 25, 52, 53, and 54 were in favor of males and the items: 7, 10, 11, 12, 13, 16, 20, 22, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34, 39, 40, 42, 43, 44, 46, 47, and 48 were in favor of females).

-Results from area index

Forty-four or 77 percent of items revealed DIF (i.e. the area between the two curves were greater than a critical value; the critical value was 0.222). The items: 7, 8, 9, 10, 11, 12, 14, 17, 19, 20, 22, 23, 24, 27, 28, 30, 31, 32, 34, 35, 37, 39, 40, 42, 44, 46, 52, 53, and 54 revealed uniform DIF (i.e. the items: 7, 8, 10, 11, 12, 14, 17, 19, 20, 22, 23, 24, 27, 28, 30, 31, 32, 34, 35, 39, 40, 42, 44, and 46 were in favor of females and the items: 9, 37, 52, 53, and 54 were in favor of males), whereas the items: 16, 18, 21, 25, 26, 43, 45, 49, 50, 51, 55, 56, and 57 revealed nonuniform DIF.

-Results from chi-square

Twenty-seven or 50 percent of items revealed DIF (the items: 9, 23, 52, 53, and 54 were in favor of males and the items: 4, 5, 7, 10, 11, 12, 13, 20, 22, 24, 27, 28, 31, 32, 33, 34, 39, 40, 42, 44, 46, and 47 were in favor of females).

-The efficiency of DIF method

The efficiency of the four approaches, were 59% for the area index , 69% for TID, 81% for b-difference, and 89% for chi square).

-The agreement among DIF methods

Table 1 summarizes the consistency in which Area index and Chi-square methods flagged the items. The two methods were agreeable in allocating twenty-three items as revealing DIF, and seven items as not revealing DIF. As such, the percentage of agreement between Area index and Chi-square methods is 56% (i.e. $7 + 23/54 = 56\%$).

Table 1 : Pair wise agreement between Chi-square and Area index methods.

	Results From Area index		
Results From Chi-square	No. of Nonflagged Items	No. of flagged Items	Marginal Total
No. of nonflagged items	7	21	28
No. of flagged items	3	23	26
Marginal total	10	44	54

Table 2 summarizes the consistency in which b-difference and Chi-square methods flagged the items. The two methods were agreeable in allocating twenty-five items as revealing DIF, and twenty-one items as not revealing DIF. As such, the percentage of agreement between b-difference and Chi-square methods is 85% (i.e. $21 + 25 / 54 = 85\%$).

Table 2: Pair wise agreement between Chi-square and b-difference methods.

	Results From Chi-square		
Results From b-difference	No. of Nonflagged Items	No. of flagged Items	Marginal Total
No. of nonflagged items	21	1	22
No. of flagged items	7	25	32
Marginal total	28	26	54

Table 3 summarizes the consistency in which TID and Chi-square methods flagged the items. The two methods were agreeable in allocating sixteen items as revealing DIF, and twenty-three items as not revealing DIF. As such, the percentage of agreement between TID and Chi-square methods is 72% (i.e. $23 + 16 / 54 = 56\%$).

Table 3: Pair wise agreement between Chi-square and TID methods.

	Results From Chi-square		
Results From TID	No. of Nonflagged Items	No. of flagged Items	Marginal Total
No. of nonflagged items	23	11	34
No. of flagged items	4	16	20
Marginal total	27	27	54

Table 4 summarizes the consistency in which Area index and b-difference methods flagged the items. The two methods were agreeable in allocating twenty-seven items as revealing DIF, and five items as not revealing DIF. As such, the percentage of agreement between Area index and b-difference methods is 59% (i.e. $5 + 27/54 = 59\%$)

Table 4: Pair wise agreement between b- difference and Area index methods.

	Results From Area index		
Results From b-difference	No. of Nonflagged Items	No. of flagged Items	Marginal Total
No. of nonflagged items	5	17	22
No. of flagged items	5	27	32
Marginal total	10	44	54

Table 5 summarizes the consistency in which Area index and TID methods flagged the items. The two methods were agreeable in allocating sixteen items as revealing DIF, and six items as not revealing DIF. As such, the percentage of agreement between Area index and TID methods is 41% (i.e. $6 + 16/54 = 41\%$).

Table 5: Pair wise agreement between TID and Area index methods.

	Results From Area index		
Results From TID	No. of Nonflagged Items	No. of flagged Items	Marginal Total
No. of nonflagged items	6	28	34
No. of flagged items	4	16	20
Marginal total	10	44	54

Table 6 summarizes the consistency in which TID and b-difference methods flagged the items. The two methods were agreeable in allocating seventeen items as revealing DIF, and twenty items as not revealing DIF. As such, the percentage of agreement between TID and b-difference methods is 69% (i.e. $20 + 17 / 54 = 69\%$).

Table 6: Pair wise agreement between TID and b-difference methods.

	Results From b-difference		
Results From TID	No. of Nonflagged Items	No. of flagged Items	Marginal Total
No. of nonflagged items	20	15	35
No. of flagged items	2	17	19
Marginal total	22	32	54

Conclusion

The study pointed out: (1) the percentage of agreement among the four methods in detecting DIF were from 41% to 85%. The highest agreement was between Chi-square and b-parameter difference methods (85%), whereas the lowest agreement was between Area index and TID methods (41%). (3) females showed a statistically significant and consistent advantage over males on items involving Relations and functions, polynomial, Trigonometric functions, whereas men showed a less consistent advantage on items involving triangles, however It was concluded that gender differences in mathematics may well be linked to content. The highest efficacy was for chi-square method, whereas the lowest efficacy was for area index.