EXPLORING DIF: COMPARISON OF CTT AND IRT METHODS

Nabeel Abedlazeez [a]
[a] Faculty of Education, Department of Educational Psychology and Counseling
University of Malaya, Kuala Lumpur, Malaysia
[a] Corresponding author: nabeelabdelazeez@yahoo.com

 ISSN 1923-6654 (Print),
ISSN 1923-6662 (Online). Available at http://www.ssrn.com/link/OIDA-Intl-Journal-Sustainable-Dev.html

Abstract: 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 different groups have differing probabilities of success on an item, after controlling for overall ability level. The study was conducted to answer the following questions: To what extent do the four methods (i.e. area differential index procedure for the 2- parameter Logistic model, TID, b-difference, and Chi-square) agree or disagree in the identification of DIF? Are there gender differences in mathematical proficiency? What is the content or nature of those items identified as revealing DIF? Achievement test covering the following subjects: Relations and functions, polynomial, Trigonometric functions, and triangles was developed. The test was administered to a sample of 1228 tenth grade students (656 males and 624 females) in Jordan. 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%). The agreement among IRT based methods and CTT based methods were convergent. (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.

Keywords: Transformed Item Difficulty, Area index , Chi-square, b-parameter difference

I. INTRODUCTION

One of the important issues faced by counseling psychologists is that of responding to the diversity of clients. In particular, it is important that the tests used by counseling psychologists be free of systematic demographic subgroup bias. Item response theory (IRT) techniques provide a powerful means of testing items for bias, using what is known as differential item functioning (DIF), as well as assessing the cumulative effect of any item-level bias on the test's total score.

In contrast, classical test theory (CTT) methods of assessing bias are fundamentally limited, especially approaches that base their assessment of bias on the presence of group mean differences in total tests scores across demographic groups, or on differential item-passing/endorsement rates between subgroups [9]. In essence, such methods cannot distinguish between the situation in which (a) the subgroups have different means, and the test is biased, versus (b) the means differ, but the test is not biased (i.e., one group truly has a higher average on the test).

Methods bias have proliferated in recent years and have been reviewed. 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].

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]. 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 [5]. The tight relationship between the probability of correct responses and ability or trait levels is an explicit assumption of item response theory (IRT) [6] and an implicit assumption of classical test theory [7]. 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 several techniques have been promulgated for the statistical assessment of DIF. Several excellent reviews are available [4, 5, 8]. Most techniques for DIF assessment were developed in educational settings in which items are generally dichotomously scored as correct or incorrect.

Two distinct forms of DIF have been recognized. These have been called uniform and non-uniform DIF. 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]. Before the crossover point, the item is favoring one subgroup, and after the ICC's cross, the item starts to favor the other group, so the DIF could cancel themselves out, and the item shows no net DIF. Nonuniform DIF is likely to occur when the two ICC's have similar b parameters and different maximum slopes. IRT area-based statistics are powerful in detecting nonuniform DIF.

Gender differences in mathematics

Gender differences in mathematics have been a popular but complex issue in educational research [14, 20]. Since Sells [23] expressed the concern that mathematics is the critical filter for the differential representation of women and men in scientific and technical fields, there has been increased interest in research about gender and mathematics [1, 14, 16, 40]. In particular, researchers have focused on investigations of gender-related performance differences in mathematics and have provided different theoretical models to explain the gender differences in mathematics from various perspectives, such as biological, educational, and sociological [12, 13, 15, 18, 20, 41]. Although recent reviews of research on gender differences in mathematics by Friedman [17] and Hyde et al [19] suggest that gender differences in mathematical performance are declining, female students continue to show less confidence in their mathematical ability and a lower perception of the usefulness of mathematics to them in the future [37, 38]. Even among the mathematically gifted students, females have lower educational aspirations in mathematics and sciences than do males [12].

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 [18]. 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 [21] 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. In another study, Marshall and Smith [22] 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 [22], 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, but there were not significant differences on word problems and nontraditional problems.

The present study sought answers to the following questions: The first question of interest was: 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 DIF? A second question was: Are there gender differences in mathematical proficiency? A third question was: Are gender differences linked to content areas within mathematics?

II. METHOD

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 to make sure that the items of the test are clear and are understood by those who were tested, and to recognize the levels of difficulty and discrimination and the effectiveness of the detractors of the items. 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 sample of (1228) students- males and females- from the tenth grade (656 males, and 624 females). The item analysis revealed levels of difficulty from 0.16 to 0.96 and levels of discriminate ability from 0.19 to o 0.56. Besides, it revealed that the detractors were reversal to the item discriminate.

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 model, and the test measures a single trait (unidimensionality).

DIF Detection Procedures

Four methods were used to investigate DIF (area index for the two-parameter logistic model, transformed item difficulty , b-parameter difference, and Chi-square).

Item response theory-based methods for assessing differential item functioning

The principal conceptual unit of IRT is the item characteristic curve. An ICC is the function that relates the probability of a correct answer on an item to the "ability" measured by the test containing the item. If the unidimensional assumption of the test is met, an item response function or item characteristic curve defined by its item parameters will remain unchanged across subpopulation groups. An ICC estimated from any group will be equal to an ICC from another, and both will be equal to the ICC estimated from responses of all examinees.

Area Index for Two-Parameter Logistic Model

An example of the item characteristic curve approaches is the area index. It measures the area between the two ICCs of the reference and the focal 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.

The area is calculated over a specified ability interval, which in this study was from the lower group mean minus 3 SD to the upper group mean plus 3 SD. Because there is no known sampling distribution for the area statistic under the null hypothesis of no group difference, item are typically ranked according to the values of the statistic and those with the highest values flagged as revealing DIF. 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:

$$Area = \left| 2\frac{(a_{2-}a_1)}{Da_1a_2} Ln\left[1+e^{Da_1a_2}\frac{b_2-b_1}{a_2-a_1}\right] - (b_2-b_1) \right|$$

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).

Area index is powerful in detecting nonuniform DIF. Area-based statistics rest on the premise that when an item is not revealed DIF, the ICCs for two subgroups are identical, and the area between the curves is zero. However, when an item is revealed DIF, the ICCs are not the same, the area between the curves is not zero, and DIF is present [6]. The most important aspect of area index is also the most difficult to attain. In order to accurately calculate the area between two item characteristic curves, both curves must be on the same metric, otherwise, observed large areas may be due to scaling differences rather than actual DIF. This problem is referred to as the "linking" problem [2], and it arises whenever item parameters are estimated using data from two different subgroups (samples) of examinees [36]. Item DIF studies using area index method will always require subgroup parameters to be linked.

b-parameter difference

The simplest index available to reflect differences in item parameters is the differences in estimated b parameters for two groups; 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. The simple difference in b parameters for the two groups conveys the "size" rather than the statistical significance of the DIF [5].

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 Δb , 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].

Classical test theory based methods

Classical Test Theory (CTT) models state that an examinee's observed score consists of his/her true score plus error. IRT has a similar interest in determining an examinee's true score (latent trait score). However, CTT approaches are limited in that examinee ability is defined in terms of a particular test, and the difficulty of that test is determined by the ability of the examinees who take it. This circularity of item and examine characteristics in CTT branches into the estimation of reliability and validity as well because the test and item characteristics change as the examinee pool changes.

Transformed Item Difficulty (TID) Method

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 [25]. It will be referred to here as the transformed difficulty method TID).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) th 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 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]. This procedure has been used to study cultural differences in a wide variety of contexts [24, 25, 26, 33, 34, 43].

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 ($\Delta_i M$), *X* represents females delta values ($\Delta_i X$) , 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 ($\Delta_i F$)

μ_x: The mean of delta values for males ($\Delta_i M$) , 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.

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:

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

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 ± 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 *Di* less than minus one unit indicates DIF favoring males [35].

Chi Square Type DIF Methods

Analogous to, yet independent of the item characteristic curve, is Scheuneman's [30] *modified chi-square* DIF method. The ability dimension is divided into discrete categories with the probability of correct responses in each category assumed constant, while discrimination among items vary and the lower asymptote is typically not zero. Scheuneman [30] stated that "item characteristic curves for different ethnic groups can be very roughly approximated using relatively small samples…" (p. 145). 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:

$$2 = \sum \frac{[[(B_e] - B_o)^2]}{B_e} + \sum \frac{[[(W_e - W_o])^2]}{W_e}$$

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

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

where O is the observed frequency in a given category and E is the expected frequency in a given category.

When establishing ability intervals on the total score scale, several criteria need to be met. The probability of a correct response within each ability interval must be less than one, and intervals are made larger or smaller to insure that there are some incorrect responses included in each interval. Expected frequencies must be at least five and all other cells must have somewhat large counts, a minimum of ten to twenty observed correct responses, due to small cells producing spurious results [30].

The Chi-Square Method Instead of focusing on a single-item parameter, like difficulty or discrimination, other methods compare entire distributions of responses for the two groups in question. Scheuneman's chi-square procedure is one such technique. According to her definition "An item would be considered unbiased if for persons with the same ability in the area being measured, the probability of a correct response on the item is the same regardless of the population group membership of the individual." Operationally, this definition may be restated: "An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered".

III. RESULTS AND DISCUSSION

Appendix 1 shows the summary results of the TID method to identify Differential Item Functioning on the mathematics proficiency test for each of the fifty-four items. Nineteen or 35 percent of items revealed DIF (the items: 25, 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, 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.

Appendix 2 shows the summary results of the b-parameter difference method to identify Differential Item Functioning on the mathematics proficiency test for each of the fifty-four items at the 0.05 level of significance. 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, 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, 48 were in favor of females).

Appendix 3 shows the summary results of the area index to identify Differential Item Functioning on the mathematics proficiency test for each of fifty-four items. 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). In order to inspect the direction of DIF (i.e. uniform or nonuniform), item characteristic curve of each item for males and female were drawn (see appendix 5)

Appendix 4 shows that 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, 46 , 9, 37, 52, 53, 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, 46 were in favor of females and the items: 9, 37, 52, 53, 54 were in favor of males), whereas the items: 16, 18, 21, 25, 26, 43, 45, 49, 50, 51, 55, 56, 57 revealed nonuniform DIF. The area between the two curves for the items: 36, 38, 41 were closed to zero.

Appendix 5 shows the summary results of the Chi-square method to identify Differential Item Functioning on the mathematics proficiency test for each of the fifty four items at the 0.05 level of significance. Twenty-seven or 50 percent of items revealed DIF (the items: 9, 23, 52, 53, 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, 47 were in favor of females).

In order to inspect the consistency between any two of the four methods in detecting DIF, the percentage of pair wise agreements among the four approaches were computed (i.e. the degree of correspondence among each of pair wise methods with respect to the items revealing or not revealing DIF for all items were computed).

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	19

Note. b-difference: Difficulty difference

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

Note. TID: Transformed Item Difficulty

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

Note. b-difference: Difficulty difference

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

Note. TID: Transformed Item Difficulty

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

Note. TID: Transformed Item Difficulty, b-difference: Difficulty difference,

IV. DISCSSION

In summary, the percentage of agreement among the four methods in detecting DIF were from 41% to 85%.. The moderate agreement was among b-parameter difference and TID methods (69%). The agreement among IRT based methods and CTT based methods were convergent. The highest agreement was among Chi-square and b-parameter difference methods (85%), however, this may due to: the two methods has a statistical test of significant (a standard norm to identify significant DIF). The lowest agreement was among Area index and TID methods (41%), however, this may due to: the two methods did not have a statistical test of significant.

The theoretical reasons for the lack agreement between certain pairs of methods in the identification of DIF of items are given by Hunter [29]. He discussed several factors which may cause an item to be labeled as revealed DIF when, in fact, no DIF exists. These are (a) non-unidimensional tests, (b) differences in ability distribution of the two groups, (c) differences in item quality, (d) guessing, and (e) 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 revealing DIF. The fact is that studies of convergence of methods for investigating DIF are influenced greatly by the unreliability of the statistics.

Hunter also claims that the chi-square approach has several problems also. The test must be unidimensional and very reliable in order for the total test score to be a valid measure of ability. In addition, this approach is very sensitive if there are differences in the total test score distributions of the groups. Finally, Hunter notes several flaws in the item characteristic curve method. Differences in the ability distributions will be reflected in instability at the ends of the curves for different groups and different displacements of the item-characteristic curves because of unreliability. He finally states that all methods fail if the test is not unidimensional.

To better understand lack of agreement between certain pairs of methods, a closer look was taken at those items identified as revealing DIF by one method but not another. Rudner [33] found that all the items identified as revealing DIF by both the transformed difficulty and b-parameter difference methods were items where the discrimination parameters were similar, but the location of the parameters differed.

In previous studies, after a statistical procedure had been used to identify potentially biased items, attempts were made to identify possible content sources of this bias. In general, this procedure has neither provided consistent nor easily generalizable results

Chi-square method flagged the items: 9, and 23 as revealing DIF in favor of males, whereas TID method flagged the items: 25, and 26 as revealing DIF in favor of males .The two CTT methods were disagreeable in allocating the items as revealing DIF in favor of males, whereas the two methods are relatively agreeable in allocating the items as revealing DIF in favor of females. In the contrast, IRT methods were agreeable in allocating the items: 9, 52, 53, and 54 as revealing DIF in favor of males.

Data analysis indicates that the four methods flagged most of items as revealing DIF in favor of females. Mathematics items indicating DIF in favor of males were found to involve triangles and items involving real world references. Mathematics items indicating DIF in favor of females tended to involve: Relations and functions, polynomial, Trigonometric functions , miscellaneous, and regular mathematics items. For mathematics items revealing DIF in favor of male student the content characteristics involved: solving triangles equations. The fact that this test was tied to a specific curriculum appeared to help females' performance.

Although men appear to have the advantage on mathematics achievement tests, women usually have higher average classroom grades than men [45]. To explain this discrepancy, Kimball [44] speculated that men and women have different learning styles; women rely more on routine use of rules learned in class, whereas men have a more autonomous style that allows them to generalize knowledge to unfamiliar problems. Then, it might be expected that women would do better on tests that are closely linked to classroom instruction. In support of this view, Smith and Walker [47] found that women performed slightly better than men on the ninth- and eleventh – grade New York State Regents Examination, whereas men did better on the tenth – grade paper. Although the authors explained these results by speculating that women may do better on curriculum – specific tests, it could be argued that Smith and Walker's results reflect a female advantage in algebra in the ninth grade –and eleventh grade and a male advantage in geometry in the tenth grade. Seegers an boekaerts [46] showed that eighth grade boys in the nethelands performed better than girl on a mathematics test, even though the test was specifically designed to reflect classroom tests.

Such a strong female advantage in Relations and functions, polynomial, Trigonometric functions, as reflected in DIF indexes has not been previously noted. 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 Relations and functions, polynomial, Trigonometric functions does indicate a reliance on algorithmic learning, females might benefit even more than males from an instructional strategy that relies less on teaching algorithms and more on teaching problem solving and effective means of approaching nonroutine problems.

REFERENCE

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

[2]. 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.

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

[4]. Clauser BE, Mazor KM. (1998). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice; 17:31– 44.

[5]. Camilli G, Shepard LA. (1994). Methods for Identifying Biased Test Items. Sage: Thousand Oaks.

[6]. Hambleton RK, Swaminathan H, Rogers HJ. (1991). Fundamentals of Item Response Theory. Sage: Newbury Park

[7]. McDonald RP. (1999).Test Theory: A Unified Treatment. Lawrence Erlbaum: Mahwah, NJ.

[8]. Millsap RE, Everson HT. (1993). Methodology review: statistical approaches for assessing measurement bias. Applied Psychological Measurement; 17(4):297–334.

[9]. Drasgow, F. (1987). A study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72, 19-29.

[10]. Lord, F. M.(1980). Applications of item response theory to practical testing problems. Hillsdale (NJ): Erlbaum.

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

[12]. Benbow, C. P. (1992). Academic achievement in mathematics and science between ages 13 and 23: Are there differences among students in the top one percent of mathematical ability? *Journal of Educational Psychology, 84*, 51-61.

[13]. Carr, M & Jessup, D. L. (1997). Gender differences in first-grade mathematics strategy use: Social and metacognitive influences. *Journal of Educational Psychology, 89*, 318-328.

[14]. Fennema, E. & Leder, G. C. (Eds.) (1990). *Mathematics and gender*. New York: Teachers College Press.

[15]. Fennema, E., & Peterson, P. (1985). Autonomous learning behavior: A possible explanation of gender-related differences in mathematics. In L. S. Wilkinson & C. B. Marrett (Eds.), *Gender influences in classroom*

interaction (pp. 17-36). New York: Academic Press.
[16]. Fennema, F., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational researcher, 27*(5), 6-11.
[17]. Friedman, L. (1989). Mathematics and gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research, 59*, 185-214.
[18]. Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences, 19*, 229-284.
[19]. Hyde, J. S., Fennema, E., Lamon, S. J. (1990). Gender differences in mathematics performance: A metaanalysis. *Psychological Bulletin, 107*(2), 139-155.
[20]. Leder, G. C. (1990). Gender differences in mathematics: An overview. In E. Fennema & G. C. Leder (Eds.), *Mathematics and gender* (pp. 10-26). New York: Teachers College Press.
[21]. Marshall, S. P. (1984). Sex differences in children's mathematics achievement: Solving computations and story problems. *Journal of Educational Psychology, 76*, 194-204.
[22]. 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.
[23]. Sells, L. W. (1973). High school mathematics as the critical filter in the job market. In R. T. Thomas (Ed.), *Developing opportunities for minorities in graduate education* (pp. 37-39). Berkeley, CA: University of California Press.
[24]. Angoff, W. H. (1975). The investigation of test bias in the absence of an outside criterion. Paper presented at the NIE Conference on Test Bias.
[25]. Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.
[26]. Angoff, W. H., & Sharon, A. L. (1972) Patterns of test and item difficulty for six foreign language groups on the Test of English as a Foreign language. Research Bulletin 72-2; CEEB RDR 71-72, No. 5. Princeton, New Jersey: Educational Testing Service.
[27]. Green, D. R., & Draper, J. F. (September 1972). Exploratory studies of bias in achievement tests. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu.
[28]. Green, D. R. (April 1976). Reducing bias in achievement tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.
[29]. Hunter, J. E. (December 1975). 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.
[30]. Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.
[31]. Veale, J. R., & Foreman, D. I. (April 1976). Cultural variation in criterion referenced tests: A global item analysis. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.
[32]. Subkoviak, M. j. Mack, J. S. Ironson, G. H. and Crag, R. D. (1987). Empirical Comparison of Selected Item bias Detection procedures with bias Manipulation. *Journal of education Measurement*, 21(1),209-223.
[33]. Rudner, L. M. (1976). Item and format bias and appropriateness. *Journal of Educational Measurement*, 17(1),143-165.
[34]. Gulliksen, H. (1964). Intercultural attitude comparisons and introductory remarks at the Princeton University Conference on Preference Analysis and Subjective Measurement. *Research Memo-random* 60-8. Princeton, New Jersey: Educational Testing Service.
[35]. Osterlind, S,J. (1983). *Test Item Bias*. Beverly Hills: Sage Publications.
[36]. Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.
[37]. Kenney, P. A. & Silver, E. A. (1997). *Results from the seventh mathematics assessment of the NAEP*. Reston, VA: National Council of Teachers of Mathematics.
[38]. Lindquist, M. M. (1989). *Results from the fourth mathematics assessment of the NAEP*. Reston, VA: National Council of Teachers of Mathematics.
[39]. Halpern, D. F. (1986). *Sex differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.
[40]. Chipman, S. F., Brush, L. R., Wilson, D. M. (Eds.) (1985). *Women and mathematics: Balancing the equation.* Hillsdale, NJ: Erlbaum.
[41]. Eccles, J. S. (1987). Gender roles and women's achievement-related decisions. *Psychology of Women Quarterly, 100*, 309-330.
[42]. Raju, N.S. (1988).The area between two item characteristic curves. Psychometrica,12, (6).
[43]. Breland, H. M., Stocking, M., Pinchak, B. M., & Abrams, N. (1974) The cross-cultural stability of mental test items: An investigation of response patterns for 10 sociocultural groups. Project Report 74-2. Princeton, New Jersey: Educational Testing Service.

[44]. Kimball, M. (1989). A new perspective on women's math achievement. Psychological Bulletin, 105, 198-214.
[45]. Linn, M. C., & Kessel, C. (1995, April). Participation in mathematics courses and careers: Climate, grades, and entrance examination scores. Paper presented the annual meeting of the American Educational Research Association, San Francisco.
[46]. Seegers, G., & Boekaerts, M. (1996). Gender – related differences in self-referenced cognitions in relation to mathematics. Journal for Research in Mathematics Education, 27, 215-240.
[47]. Smith, S. E., & Walker, W. J. (1988). Sex differences on New York State Regents Examinations: Support for the differential course taking hypothesis. Journal for Research in Mathematics Education, 19, 81-85.

Appendix 1: Summary Result of the TID Method to Identify Differential Item Functioning on the Mathematics Proficiency Test.

	Male		Female		
Item	**P**	**Δ**	**P**	**Δ**	**D_i**
1.	0.90	7.88	0.91	7.64	0.32
2.	0.68	11.12	0.84	9.04	1.70*
3.	0.72	9.32	0.77	8.48	0.78
4.	0.90	7.88	0.96	5.96	1.48*
5.	0.81	9.48	0.87	8.44	1.01*
6.	0.69	11.00	0.73	10.56	0.57
7.	0.88	8.32	0.92	7.36	0.83
8.	0.76	10.16	0.81	9.48	0.70
9.	0.61	10.80	0.53	11.68	-0.35
10.	0.72	10.52	0.85	8.84	1.40*
11.	0.64	11.56	0.73	9.20	1.91*
12.	0.49	13.08	0.55	11.44	1.46*
13.	0.70	10.92	0.76	9.48	1.25*
14.	0.54	12.60	0.57	12.28	0.54
15.	0.41	13.92	0.38	14.20	0.17
16.	0.69	11.00	0.74	9.040	1.61*
17.	0.44	13.60	0.46	13.40	0.49
18.	0.29	15.20	0.27	15.44	0.24
19.	0.75	10.32	0.78	9.92	0.51
20.	0.54	12.60	0.60	12.00	0.73
21.	0.43	13.72	0.41	13.92	0.22
22.	0.36	14.44	0.50	13.00	1.37*
23.	0.19	16.52	0.33	14.76	1.67*
24.	0.16	16.96	0.23	15.96	1.18*
25.	0.75	10.32	0.72	9.32	-1.02*
26.	0.60	12.00	0.59	13.92	-1.03*

27.	0.77	8.48	0.83	9.20	-0.32
28.	0.61	10.80	0.79	9.76	1.03*
29.	0.58	12.20	0.58	12.20	0.30
30.	0.52	12.80	0.57	12.28	0.68
31.	0.46	13.40	0.64	11.56	1.61*
32.	0.32	14.88	0.50	13.00	1.69*
33.	0.19	16.52	0.30	15.08	1.45*
34.	0.14	16.24	0.22	16.08	0.55
35.	0.46	13.40	0.49	13.08	0.57
36.	0.32	14.88	0.32	14.88	0.40
37.	0.18	15.32	0.22	16.08	-0.11
38.	0.46	13.40	0.48	13.20	0.48
39.	0.37	14.32	0.46	13.40	1.01*
40.	0.32	14.88	0.39	14.12	1.04*
41.	0.50	13.00	0.47	13.32	0.11
42.	0.35	14.56	0.43	13.72	0.96
43.	0.28	15.32	0.33	14.76	0.80
44.	0.23	15.96	0.32	14.88	1.18*
45.	0.51	12.20	0.56	12.4	0.17
46.	0.41	13.08	0.49	13.92	-0.25
47.	0.27	15.44	0.34	14.64	1.05*
48.	0.22	16.08	0.26	15.56	0.80
49.	0.47	13.32	0.46	13.4	0.29
50.	0.39	14.12	0.41	13.92	0.51
51.	0.35	14.56	0.34	14.64	0.33
52.	0.49	13.92	0.39	14.12	0.22
53.	0.41	13.92	0.27	15.44	-0.69
54.	0.27	15.44	0.17	16.8	-0.52
55.	0.29	15.2	0.29	15.2	0.41
56.	0.2	16.36	0.21	16.24	0.53

57.	17	16.8	0.17	16.8	0.46
58.	0.39	14.12	0.38	14.24	0.29
59.	0.32	14.88	0.32	14.88	0.40
60.	0.23	15.96	0.28	15.32	0.87

Note. * item indicating DIF. P: item difficulty. Δ: delta value. SD: standard deviation.

Appendix 2: Summary Results of the b-parameter Difference to Identify Differential Item Functioning on the Mathematics Proficiency Test

Item	Females b_F	Males b_R	$\Delta b = b_F - b_R$	d Statistic
1.	-1.876	-1.980	0.104	0.849
2.	-1.441	-1.444	0.003	0.030
3.	-1.077	-0.925	-0.152	-1.853
7.	-2.021	-1.782	-0.239	-1.960*
8.	-1.197	-1.116	-0.081	-0.738
9.	-0.102	-0.441	0.339	3.961*
10.	-1.545	-1.029	-0.516	-5.309*
11.	-0.998	-0.555	-0.443	-4.779*
12.	-0.231	0.083	-0.314	-3.828*
13.	-1.048	-0.813	-0.235	-2.724*
14.	-0.302	-0.181	-0.121	-1.378
15.	0.422	0.297	0.125	1.359
16.	-0.956	-0.793	-0.163	-1.967*
17.	0.116	0.201	-0.085	-0.976
18.	0.949	-0.759	1.708	17.753*
19.	-1.144	-1.029	-0.115	-1.221
20.	-0.414	-0.181	-0.233	-2.574*
21.	0.310	0.207	0.103	1.1746
22.	-0.043	0.547	-0.590	-7.859*
23.	0.590	1.298	-0.708	-7.106*
24.	1.035	1.515	-0.480	-4.545*
25.	-0.869	-1.058	0.189	1.965*
26.	-0.391	-0.393	0.002	0.024
27.	-1.405	-1.109	-0.296	-2.867*

28.	-1.198	-0.685	-0.513	-5.139*
29.	-0.516	-0.346	-0.170	-2.003*
30.	-0.296	-0.117	-0.179	-2.220*
31.	-0.564	0.153	-0.717	-8.506*
32.	-0.014	0.677	-0.691	-7.576*
33.	0.746	1.325	-0.579	-5.501*
34.	1.165	1.675	-0.510	-4.578*
35.	0.009	0.130	-0.121	-1.391
36.	0.636	0.704	-0.068	-0.697
37.	1.288	1.371	-0.083	-0.776
38.	0.045	0.088	-0.043	-0.486
39.	0.128	0.451	-0.323	-3.510*
40.	0.384	0.691	-0.307	-3.312*
41.	-0.009	-0.047	0.038	0.395
42.	0.206	0.533	-0.327	-3.342*
43.	0.583	0.842	-0.259	-2.599*
44.	0.705	1.097	-0.392	-4.148*
45.	-0.243	-0.077	-0.166	-1.719
46.	0.009	0.333	-0.324	-3.294*
47.	0.557	0.922	-0.365	-3.935*
48.	0.943	1.204	-0.261	-2.971*
49.	0.039	0.088	-0.049	-0.531
50.	0.322	0.370	-0.048	-0.499
51.	0.570	0.552	0.018	0.190
52.	0.397	0.041	0.356	3.721*
53.	0.860	0.309	0.551	5.729*
54.	1.352	0.915	0.437	4.313*
55.	0.795	0.821	-0.026	-0.265

56.	1.139	1.246	-0.107	-0.934
57.	1.399	1.427	-0.028	-0.231

Note:*significant at level α=0.5, b_F: Estimated difficulty parameter for males. b_R: Estimated difficulty parameter for females (focal group). Δb: Estimated difficulty parameter difference.

Appendix 3: Summary Result of the Area index to Identify Differential Item Functioning on the Mathematics Proficiency Test

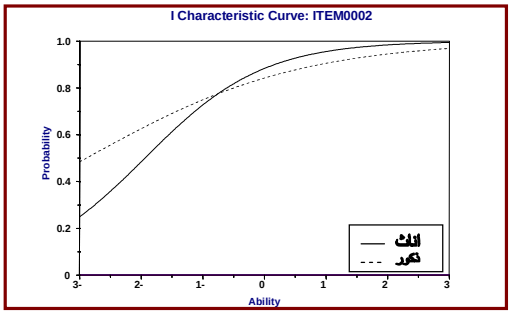
Item	Females		Males		Area
	Discrimination A	**Difficulty B**	**Discrimination a**	**Difficulty b**	
1.	0.696	-3.542	0.525	-4.425	0.028
2.	1.044	-1.939	0.579	-2.880	0.158
3.	0.650	-2.028	0.518	-1.962	0.038
7.	0.881	-3.154	1.212	-2.059	0.795*
8.	1.454	-1.300	1.258	-1.226	0.298*
9.	1.660	-0.086	1.245	-0.454	0.788*
10.	1.390	-1.930	0.803	-1.530	0.517*
11.	1.892	-0.954	1.241	-0.585	0.793*
12.	1.439	-0.235	1.830	0.142	1.624*
13.	1.118	-1.307	1.008	-1.007	0.119
14.	1.904	-0.286	1.643	-0.136	0.638*
15.	1.895	0.408	1.825	0.307	0.189
16.	1.241	-1.121	0.754	-1.214	0.366*
17.	1.709	0.126	1.587	0.234	0.279*
18.	1.9897	0.905	1.493	0.787	1.365*
19.	1.993	-1.075	0.834	-1.486	2.870*
20.	2.182	-0.370	1.666	-0.135	1.034*
21.	1.992	0.295	1.434	0.250	1.373*
22.	0.913	-0.021	0.583	1.112	0.241*
23.	1.556	0.621	1.402	1.352	0.341*
24.	1.412	1.133	1.509	1.507	0.238*
25.	1.493	-0.924	0.664	-1.833	0.860*
26.	1.910	-0.367	0.833	-0.626	1.958*
27.	1.758	-1.389	1.248	-1.222	0.817*

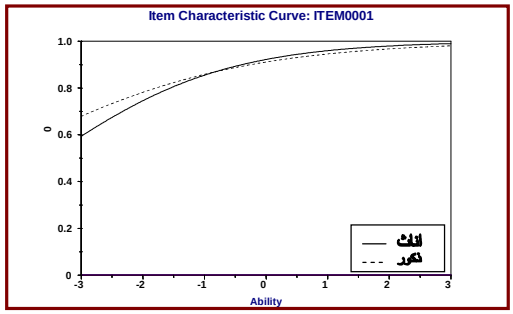
28.	2.330	-1.061	1.464	-0.669	0.653*
29.	1.341	-0.567	1.205	-0.353	0.197
30.	1.257	-0.325	1.042	-0.102	0.249*
31.	1.300	0.626-	1.446	0.196	0.474*
32.	1.426	0.001	1.962	0.628	4.107*
33.	1.900	0.713	1.936	1.180	0.159
34.	1.821	1.127	1.500	1.665	0.588*
35.	1.918	0.016	1.569	0.165	0.786*
36.	2.126	0.584	2.071	0.639	0.202
37.	1.529	1.358	1.891	1.230	0.654*
38.	1.956	0.054	1.854	0.119	0.307*
39.	2.101	0.127	2.020	0.431	0.307*
40.	2.035	0.361	1.770	0.667	0.676*
41.	2.532	-0.008	2.464	0.004	0.347*
42.	2.199	0.189	2.586	0.466	4.017*
43.	1.850	0.565	2.706	0.695	6.908*
44.	1.518	0.748	1.748	1.032	0.735*
45.	2.040	-0.216	3.129	-0.011	22.058*
46.	2.813	0.004	2.856	0.305	0.390*
47.	1.733	0.557	1.768	0.872	0.113
48.	1.181	1.147	1.179	1.394	0.003
49.	1.817	0.052	2.806	0.115	6.166*
50.	2.342	0.287	2.725	0.336	2.379*
51.	1.962	0.540	2.55	0.488	1.901*
52.	2.050	0.372	2.935	0.079	0.321*
53.	1.997	0.805	2.348	0.297	0.263*
54.	1.777	1.319	1.567	0.915	0.476*
55.	2.104	0.730	1.771	0.783	0.970*

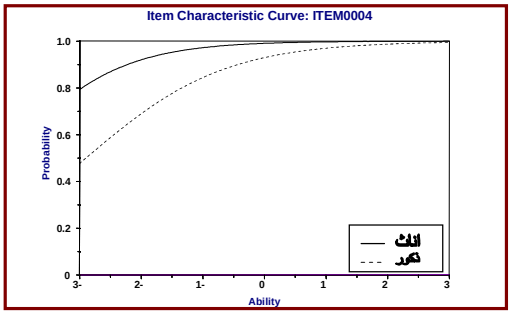
56.	2.807	0.957	2.126	1.074	1.941*
57.	3.007	1.153	2.103	1.224	2.897*

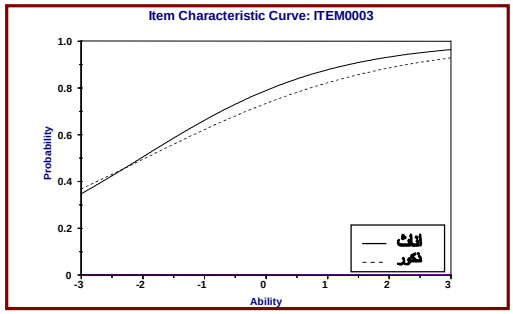
Note. * item indicating DIF, a: item discrimination, b:item difficulty

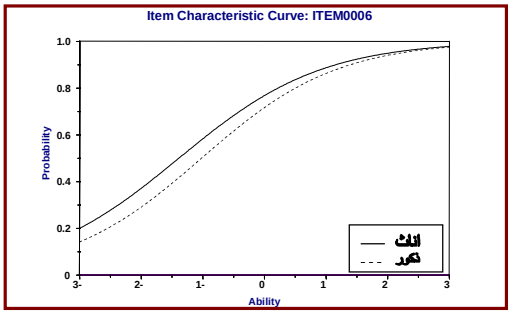
Appendix 4: Item Characteristics Curves for 2-parameter logistic model

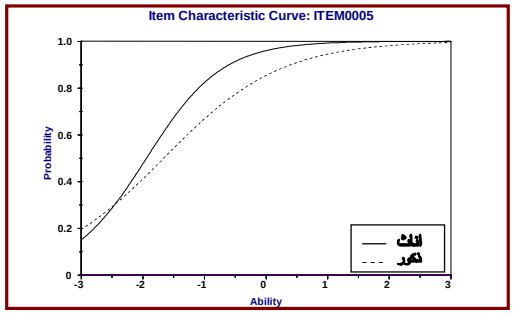


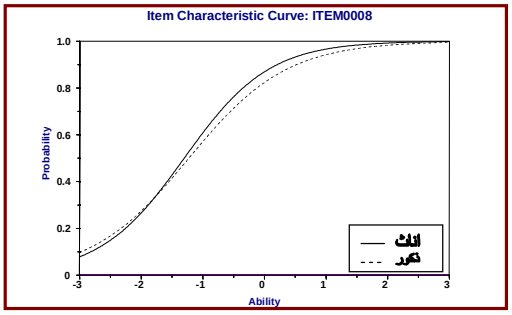


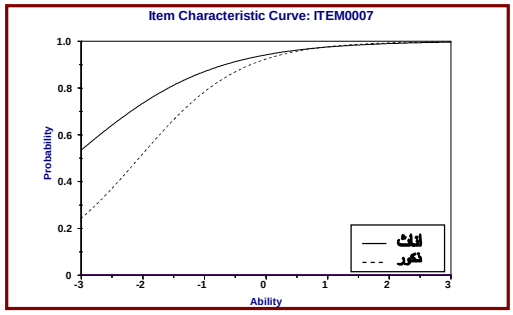


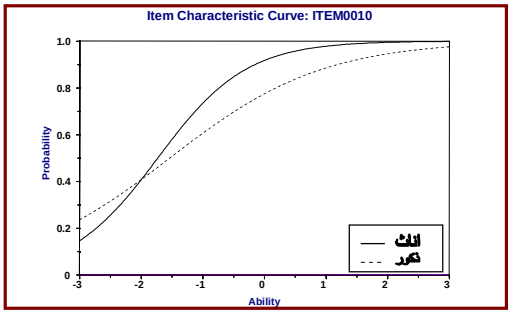


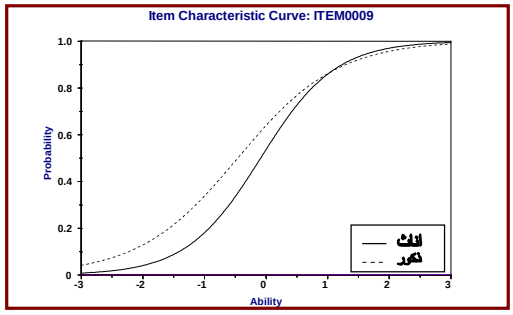

Item Characteristic Curve: ITEM0006
Probability
Ability
اناث
ذكور

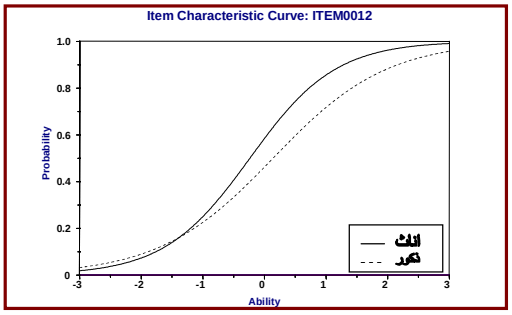

Item Characteristic Curve: ITEM0005
Probability
Ability
اناث
ذكور

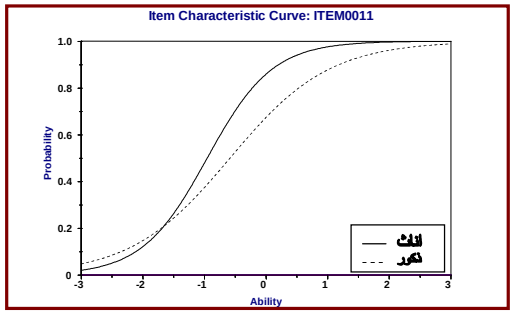

Item Characteristic Curve: ITEM0008
Probability
Ability
اناث
ذكور

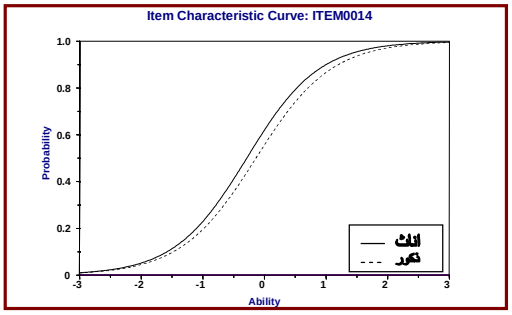

Item Characteristic Curve: ITEM0007
Probability
Ability
اناث
ذكور

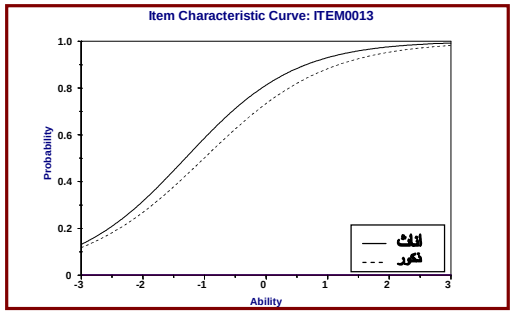

Item Characteristic Curve: ITEM0010
Probability
Ability
اناث
ذكور

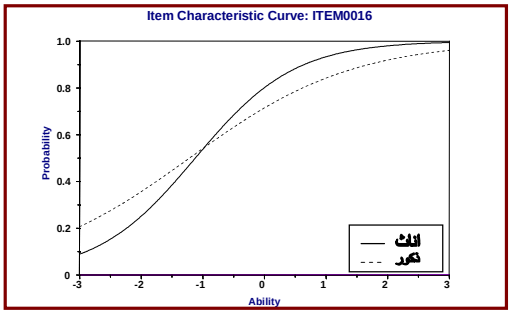

Item Characteristic Curve: ITEM0009
Probability
Ability
اناث
ذكور

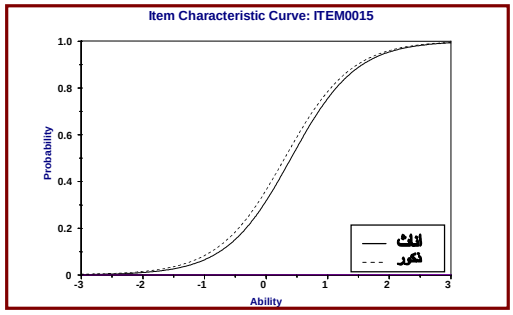

Item Characteristic Curve: ITEM0012
Probability
Ability
اناث
ذكور

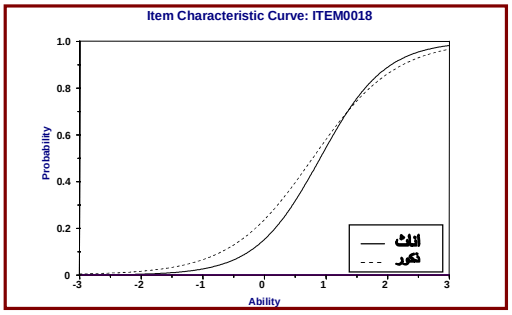

Item Characteristic Curve: ITEM0011
Probability
Ability
اناث
ذكور

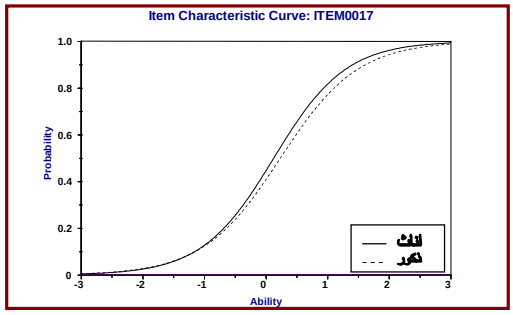

Item Characteristic Curve: ITEM0014
Probability
Ability
اناث
ذكور

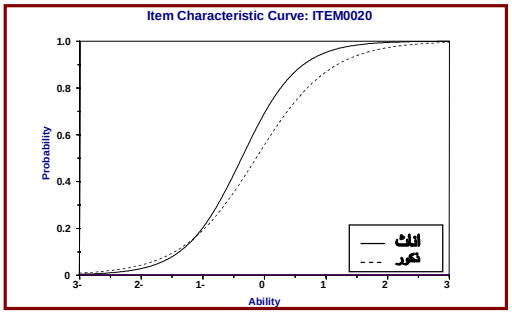

Item Characteristic Curve: ITEM0013
Probability
Ability
اناث
ذكور

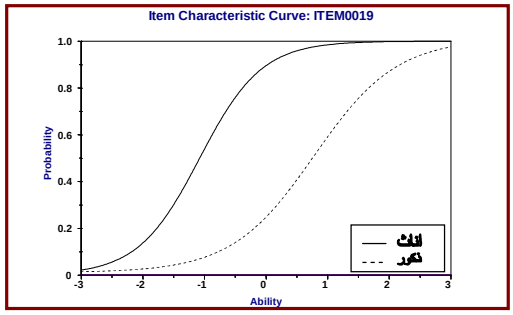

Item Characteristic Curve: ITEM0016
Probability
Ability
اناث
ذكور

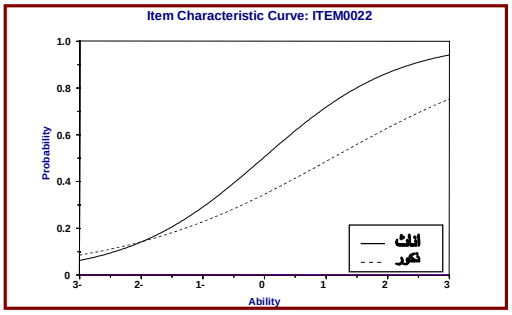

Item Characteristic Curve: ITEM0015
Probability
Ability
اناث
ذكور

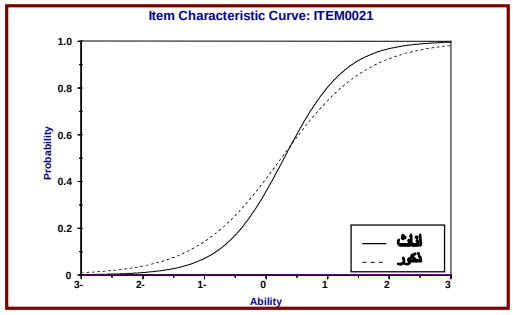

Item Characteristic Curve: ITEM0018
Probability
Ability
اناث
ذكور

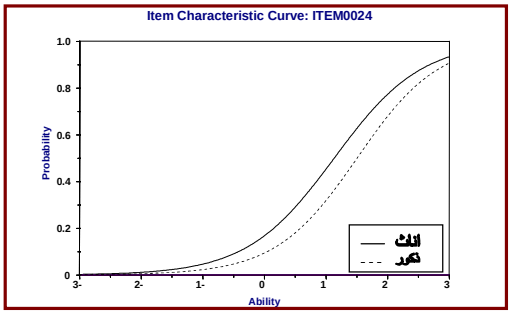

Item Characteristic Curve: ITEM0017
Probability
Ability
اناث
ذكور

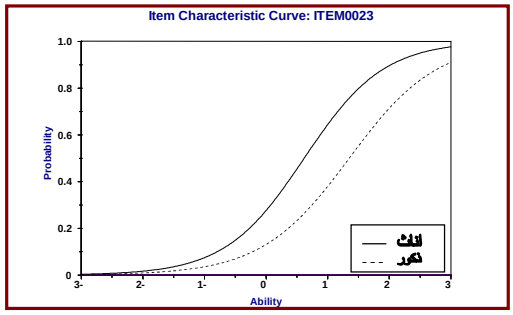

Item Characteristic Curve: ITEM0020
Probability
Ability
اناث
ذكور

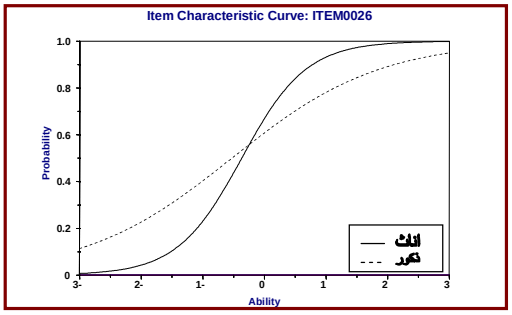

Item Characteristic Curve: ITEM0019
Probability
Ability
اناث
ذكور

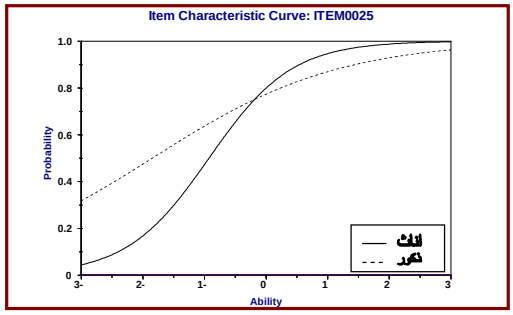

Item Characteristic Curve: ITEM0022
Probability
Ability
اناث
ذكور

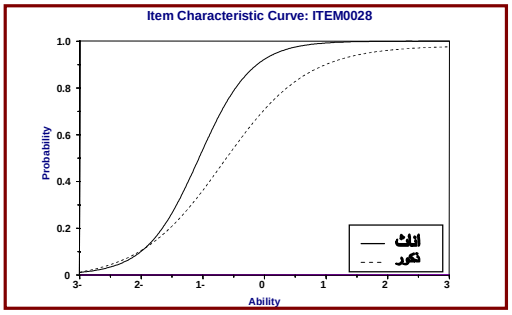

Item Characteristic Curve: ITEM0021
Probability
Ability
اناث
ذكور

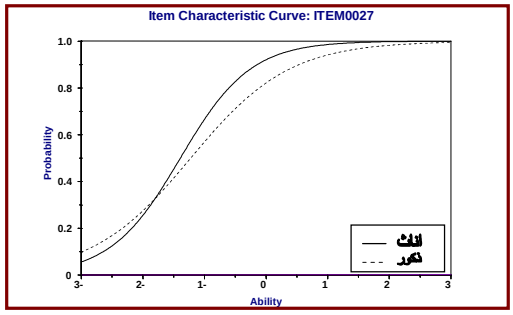

Item Characteristic Curve: ITEM0024
Probability
Ability
اناث
ذكور

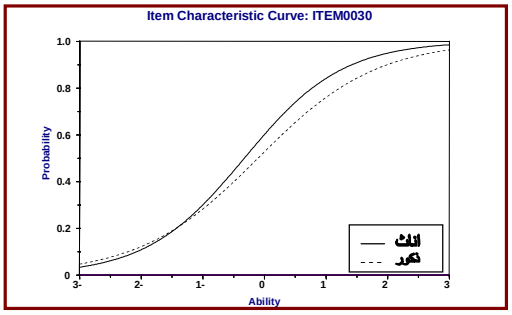

Item Characteristic Curve: ITEM0023
Probability
Ability
اناث
ذكور

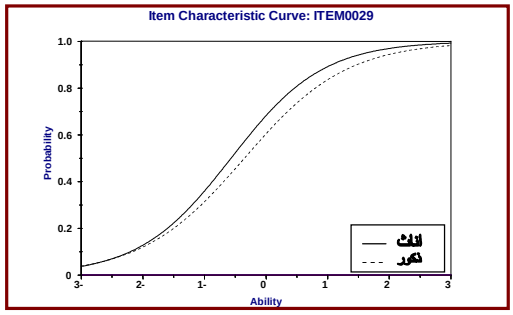

Item Characteristic Curve: ITEM0026
Probability
Ability
اناث
ذكور

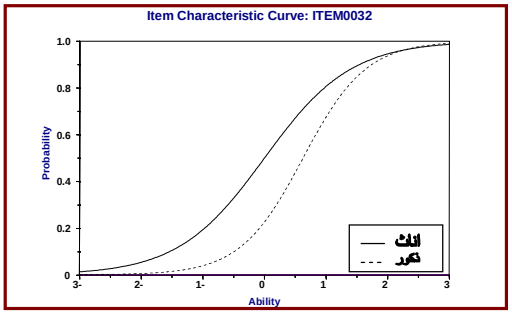

Item Characteristic Curve: ITEM0025
Probability
Ability
اناث
ذكور

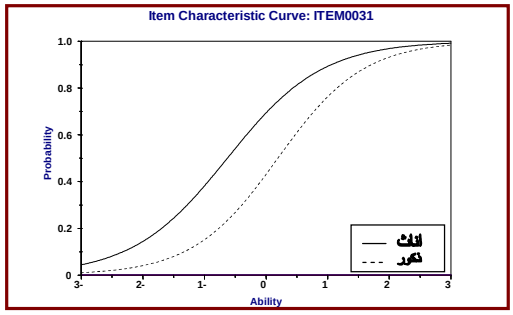

Item Characteristic Curve: ITEM0028
Probability
Ability
اناث
ذكور

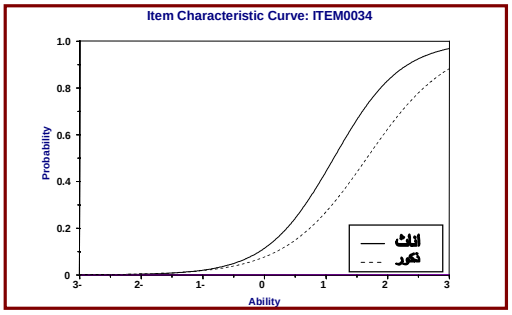

Item Characteristic Curve: ITEM0027
Probability
Ability
اناث
ذكور

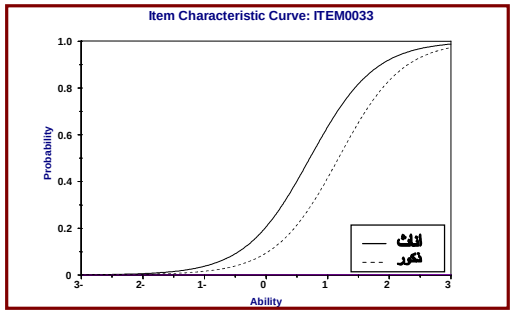

Item Characteristic Curve: ITEM0030
Probability
Ability
اناث
ذكور

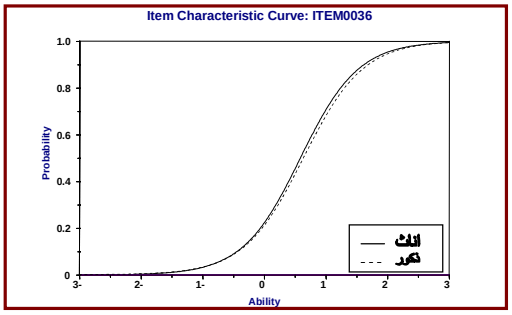

Item Characteristic Curve: ITEM0029
Probability
Ability
اناث
ذكور

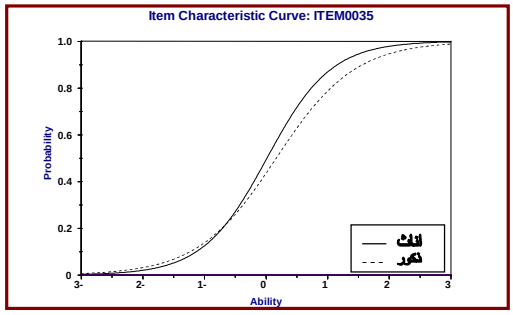

Item Characteristic Curve: ITEM0032
Probability
Ability
اناث
ذكور

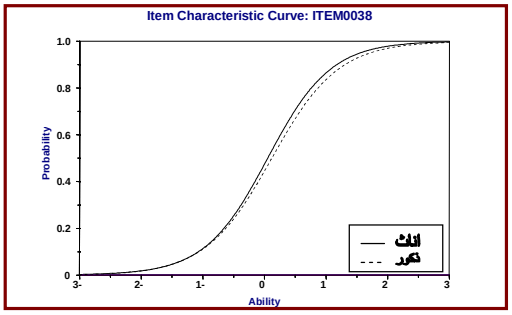

Item Characteristic Curve: ITEM0031
Probability
Ability
اناث
ذكور

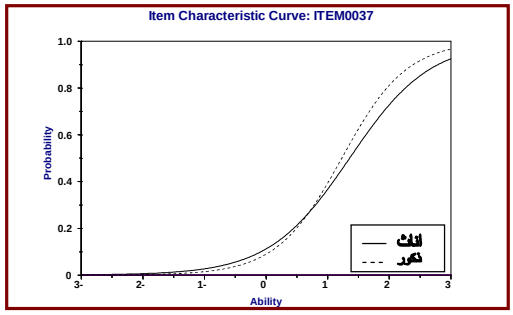

Item Characteristic Curve: ITEM0034
Probability
Ability
اناث
ذكور

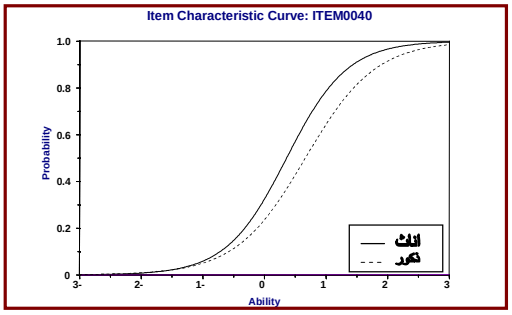

Item Characteristic Curve: ITEM0033
Probability
Ability
اناث
ذكور

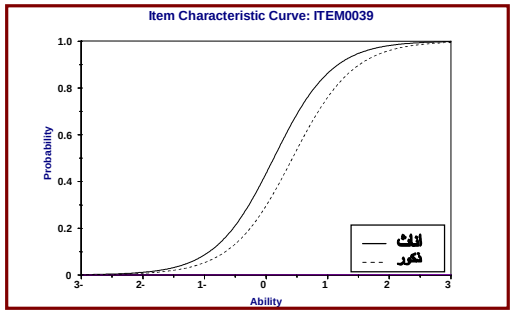

Item Characteristic Curve: ITEM0036
Probability
Ability
اناث
ذكور

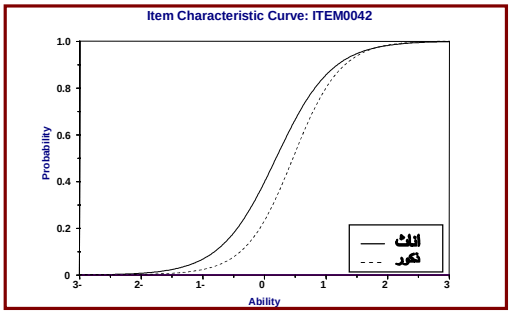

Item Characteristic Curve: ITEM0035
Probability
Ability
اناث
ذكور

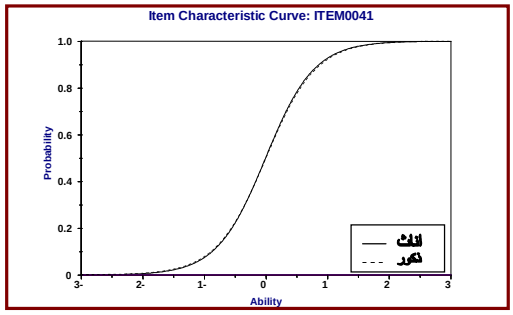

Item Characteristic Curve: ITEM0038
Probability
Ability
اناث
ذكور

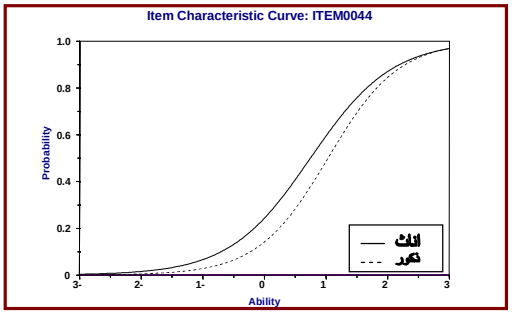

Item Characteristic Curve: ITEM0037
Probability
Ability
اناث
ذكور

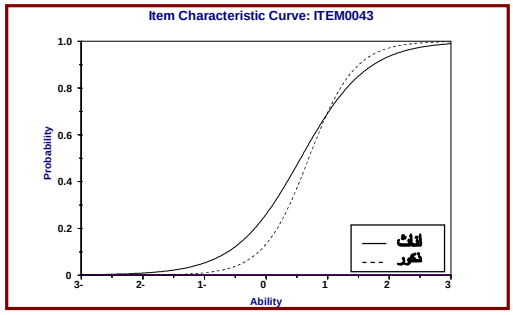

Item Characteristic Curve: ITEM0040
Probability
Ability
اناث
ذكور

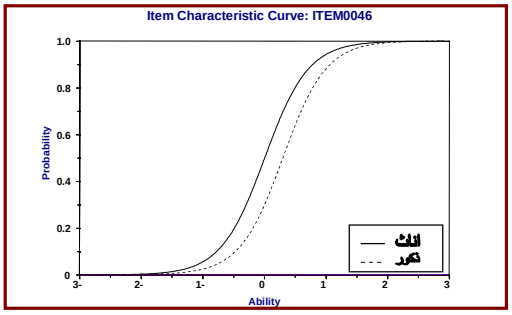

Item Characteristic Curve: ITEM0039
Probability
Ability
اناث
ذكور

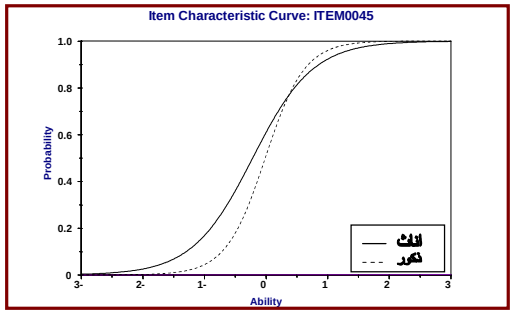

Item Characteristic Curve: ITEM0042
Probability
Ability
اناث
ذكور

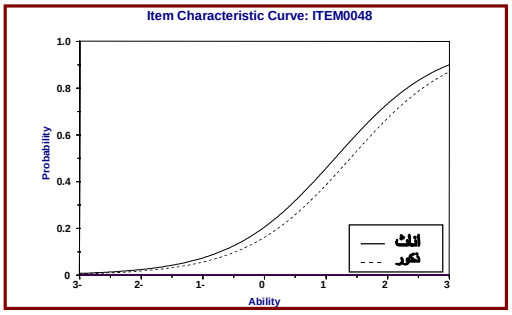

Item Characteristic Curve: ITEM0041
Probability
Ability
اناث
ذكور

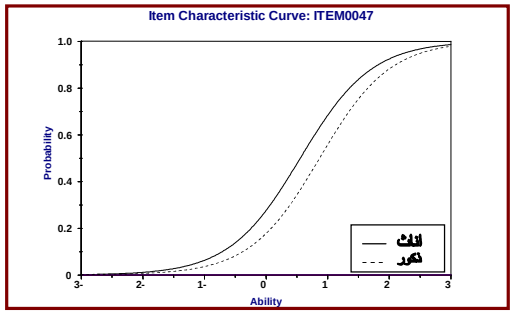

Item Characteristic Curve: ITEM0044
Probability
Ability
اناث
ذكور

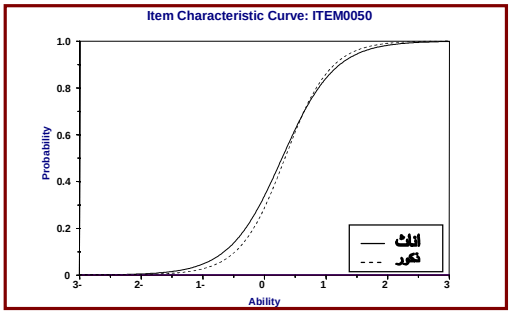

Item Characteristic Curve: ITEM0043
Probability
Ability
اناث
ذكور

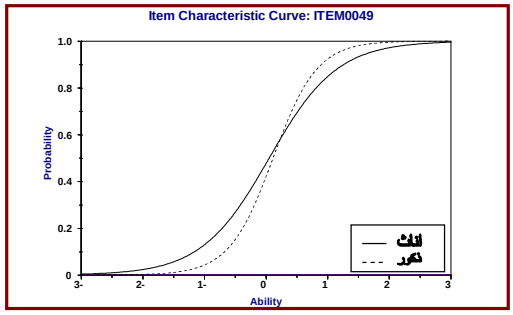

Item Characteristic Curve: ITEM0046
Probability
Ability
اناث
ذكور

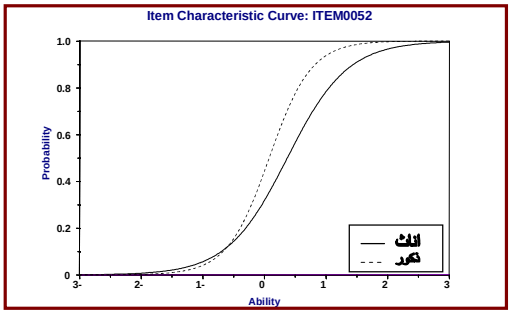

Item Characteristic Curve: ITEM0045
Probability
Ability
اناث
ذكور

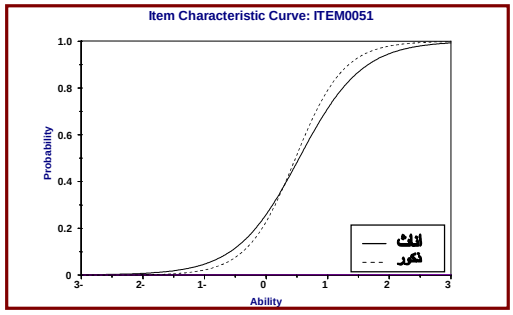

Item Characteristic Curve: ITEM0048
Probability
Ability
اناث
ذكور

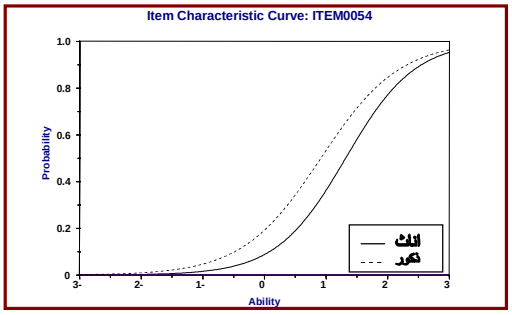

Item Characteristic Curve: ITEM0047
Probability
Ability
اناث
ذكور

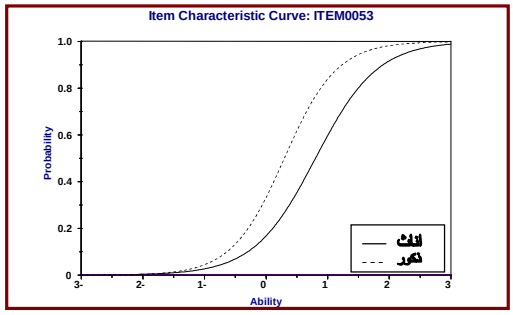

Item Characteristic Curve: ITEM0050
Probability
Ability
اناث
ذكور

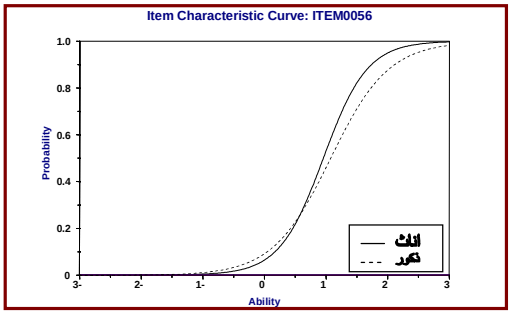

Item Characteristic Curve: ITEM0049
Probability
Ability
اناث
ذكور

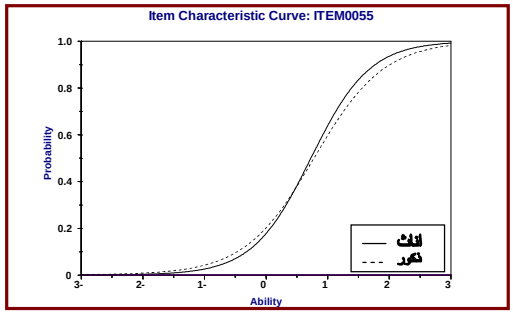

Item Characteristic Curve: ITEM0052
Probability
Ability
اناث
ذكور

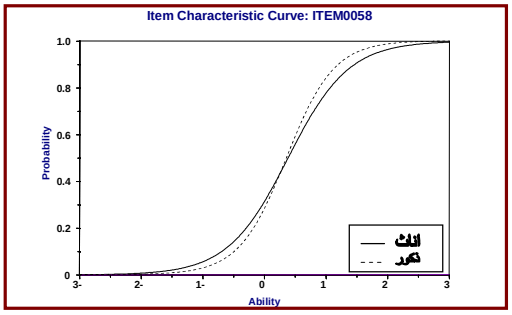

Item Characteristic Curve: ITEM0051
Probability
Ability
اناث
ذكور

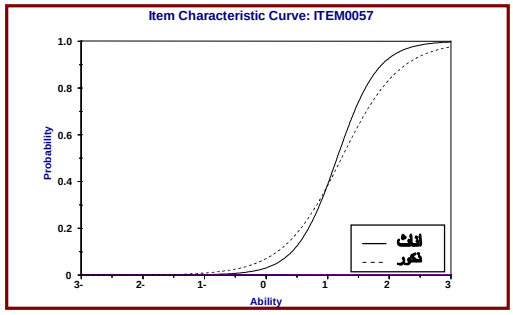

Item Characteristic Curve: ITEM0054
Probability
Ability
اناث
ذكور

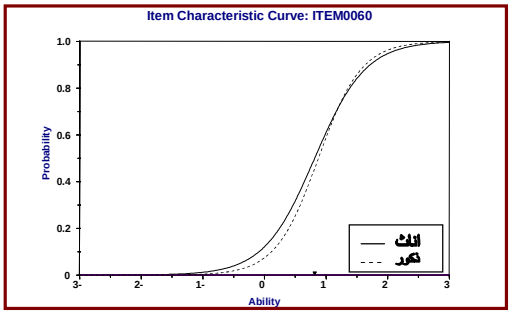

Item Characteristic Curve: ITEM0053
Probability
Ability
اناث
ذكور

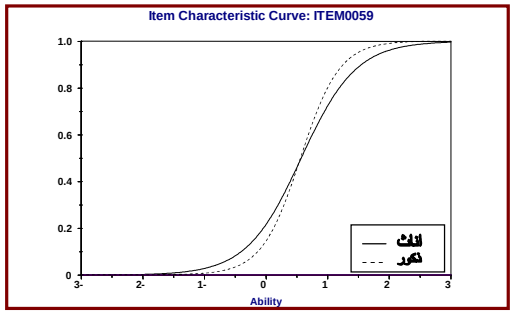

Item Characteristic Curve: ITEM0056
Probability
Ability
اناث
ذكور


Item Characteristic Curve: ITEM0055
Probability
Ability
اناث
ذكور


Item Characteristic Curve: ITEM0058
Probability
Ability
اناث
ذكور


Item Characteristic Curve: ITEM0057
Probability
Ability
اناث
ذكور


Item Characteristic Curve: ITEM0060
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
3-
2-
1-
0
1
2
3
اناث
ذكور


Item Characteristic Curve: ITEM0059
Ability
1.0
0.8
0.6
0.4
0.2
0
3-
2-
1-
0
1
2
3
اناث
ذكور

Appendix 5 : Summary Result of the Chi-square Method to Identify Differential Item Functioning on the Mathematics Proficiency Test

Item	Gender	Scores classes											Chi-square	P-value
		61-65	56-60	51-55	46-50	41-45	31-35	26-30	21-25	16-20	11-15	6-10		
1.	M	23	29	51	37	46	69	80	58	65	97	35	0.031	0.923
	F	32	47	54	48	41	66	48	63	69	67	30		
2.	M	23	29	51	36	45	66	75	43	59	86	30	0.511	0.496
	F	32	47	54	46	41	65	48	59	61	57	16		
3.	M	21	29	48	32	45	61	63	27	46	77	26	2.601	0.062
	F	32	41	47	41	35	62	45	57	56	52	13		
4.	M	23	29	51	44	61	63	70	60	66	90	34	37.057	0.000
	F	38	47	58	51	44	66	53	74	72	79	36		
5.	M	23	29	50	44	61	55	68	51	57	68	23	22.828	0.000
	F	32	47	58	51	43	64	51	73	61	64	18		
6.	M	23	29	47	41	58	48	48	41	46	52	18	3.019	0.084
	F	31	44	43	45	39	56	47	55	42	42	13		
7.	M	23	29	51	44	61	63	77	57	59	79	32	6.258	0.014
	F	32	47	58	50	44	56	49	67	71	73	26		
8.	M	23	29	51	43	59	63	62	45	51	59	16	1.403	0.251
	F	31	47	58	50	40	51	47	62	56	46	17		
9.	M	23	27	45	39	48	58	52	33	35	34	5	11.001	0.001
	F	31	45	53	42	32	28	25	34	21	10	1		
10.	M	23	29	50	43	54	50	57	47	51	58	29	24.349	0.000
	F	32	47	58	51	43	59	47	63	62	54	18		
11.	M	23	29	50	42	54	44	53	38	41	33	11	18.777	0.000
	F	21	47	57	51	41	58	46	56	50	25	4		
12.	M	21	23	44	41	43	24	39	28	34	16	7	7.947	0.005
	F	31	44	45	43	33	45	23	40	32	7	0		
13.	M	23	28	46	38	58	62	60	39	26	45	31	7.584	0.007
	F	32	47	57	46	38	53	42	51	45	51	13		
14.	M	23	28	44	35	54	59	43	28	13	14	10	1.645	0.216
	F	32	47	53	43	32	48	29	32	18	21	2		
15.	M	20	27	42	27	47	49	28	20	6	2	2	1.821	0.188
	F	32	43	40	33	21	32	11	15	4	2	1		
16.	M	23	29	45	37	45	58	59	38	39	51	28	3.510	0.064
	F	32	47	54	45	40	59	44	43	46	35	14		
17.	M	23	29	42	35	38	40	30	25	10	12	3	1.315	0.518
	F	32	46	46	35	24	37	19	23	14	5	3		
18.	M	23	18	31	26	30	28	19	14	6	4	1	5.388	0.068
	F	31	36	32	17	10	10	10	8	2	0	1		
19.	M	23	29	48	40	48	61	67	45	37	66	25	3.275	0.194
	F	32	47	58	49	42	64	49	50	53	30	13		
20.	M	23	28	46	37	47	43	56	36	18	15	2	5.704	0.018
	F	32	47	58	47	33	47	36	35	28	11	1		
21.	M	22	25	41	29	28	41	49	33	10	7	0	1.246	0.282
	F	32	39	50	36	24	34	12	14	10	1	0		
22.	M	33	25	35	18	18	11	22	15	15	29	18	27.936	0.000
	F	31	39	41	28	29	37	24	26	24	24	8		
23.	M	22	22	26	12	10	6	11	6	4	5	3	31.179	0.000
	F	30	34	39	27	27	19	10	10	6	5	1		

24.	M	22	20	18	8	8	6	9	5	3	4	2	10.781	0.001
	F	26	22	29	18	21	13	8	6	2	1	0		
25.	M	22	28	47	36	40	59	72	54	47	68	22	2.122	0.162
	F	32	46	57	45	37	59	42	49	40	32	9		
26.	M	21	28	45	36	38	41	48	41	39	44	10	0.002	1.000
	F	31	46	56	46	36	48	33	38	22	13	2		
27.	M	21	29	44	43	53	65	79	51	53	50	14	9.983	0.002
	F	31	45	57	51	43	64	50	65	58	44	13		
28.	M	23	29	43	43	54	56	68	45	42	28	8	25.309	0.000
	F	32	46	58	52	44	62	51	62	51	30	8		
29.	M	22	25	40	41	40	51	62	41	34	22	5	2.879	0.097
	F	29	36	55	48	40	5	40	48	27	16	4		
30.	M	19	21	38	38	40	44	50	40	33	19	2	2.769	0.103
	F	28	34	53	44	39	46	34	39	25	11	3		
31.	M	23	29	46	31	48	26	39	20	18	21	4	48.089	0.000
	F	32	47	53	40	33	46	37	44	36	27	6		
32.	M	23	29	42	28	34	12	13	12	11	5	1	39.805	0.000
	F	32	46	49	34	24	36	25	29	17	15	2		
33.	M	21	23	30	13	13	7	9	5	3	1	0	19.229	0.000
	F	32	36	36	22	18	24	6	3	7	1	0		
34.	M	14	13	23	10	10	5	6	4	4	1	0	11.019	0.001
	F	36	26	26	13	9	18	4	1	0	2	0		
35.	M	23	23	43	40	36	33	45	27	15	13	1	1.266	0.263
	F	32	46	50	45	29	34	19	20	16	11	2		
36.	M	23	23	41	34	25	26	17	8	7	3	0	0.086	0.811
	F	32	41	36	31	20	18	10	6	4	4	0		
37.	M	19	16	29	18	10	12	11	3	2	1	0	0.009	0.943
	F	28	34	16	13	10	13	6	0	1	6	0		
38.	M	23	26	46	33	41	38	47	26	13	10	2	0.203	0.655
	F	32	45	53	41	28	23	25	25	16	9	1		
39.	M	23	26	42	29	39	29	28	17	7	5	0	8.740	0.003
	F	32	45	54	41	25	20	23	22	11	10	1		
40.	M	21	19	39	29	32	25	22	15	3	3	0	6.931	0.010
	F	31	43	46	36	20	18	20	17	6	5	0		
41.	M	23	28	50	40	54	43	41	24	16	8	2	0.207	0.655
	F	31	37	54	43	35	37	27	15	9	4	3		
42.	M	23	27	47	30	46	22	17	9	7	2	2	7.982	0.005
	F	31	46	47	36	28	29	29	14	6	1	2		
43.	M	23	27	44	26	28	16	11	4	5	0	2	3.869	0.053
	F	31	35	40	25	20	23	22	9	4	0	0		
44.	M	14	18	38	24	22	19	8	4	4	0	2	8.462	0.004
	F	27	28	39	26	27	22	19	10	2	1	0		
45.	M	21	29	49	42	55	55	48	27	5	2	4	2.324	0.130
	F	31	46	55	45	34	50	29	27	18	5	7		
46.	M	22	28	46	35	46	35	33	20	2	1	0	7.974	0.005
	F	32	44	57	44	32	46	18	23	4	4	0		
47.	M	23	20	26	16	26	27	27	12	1	0	0	7.922	0.005
	F	27	33	42	36	23	28	10	10	3	3	0		

	M	14	9	21	14	24	25	24	12	0	0	0	2.417	0.101
48.	F	15	19	37	31	18	24	9	4	2	2	0		
	M	23	29	50	36	44	52	44	13	8	4	3	0.261	0.615
49.	F	21	47	52	38	27	34	21	20	16	10	4		
	M	22	29	45	31	42	45	31	7	3	3	1	0.143	0.732
50.	F	31	46	50	39	21	24	15	18	3	6	0		
	M	22	29	42	30	31	41	27	4	0	3	1	0.092	0.769
51.	F	31	44	40	27	16	19	10	18	5	4	0		
	M	23	29	45	37	48	54	53	16	6	2	0	11.164	0.001
52.	F	32	39	49	34	22	26	10	15	6	7	1		
	M	22	25	41	34	38	50	43	12	2	2	0	26.376	0.000
53.	F	31	33	34	25	18	15	2	5	3	5	0		
	M	22	20	25	16	20	33	31	6	2	2	0	16.931	0.000
54.	F	28	26	21	13	3	8	2	0	5	3	0		
	M	23	25	31	23	25	25	16	10	4	5	1	0.019	0.902
55.	F	32	44	37	17	13	10	10	4	6	4	1		
	M	23	22	24	17	21	11	7	6	1	1	0	0.124	0.832
56.	F	32	41	30	12	5	4	4	1	1	2	1		
	M	19	20	20	17	17	11	7	3	0	0	0	0.148	0.711
57.	F	30	36	22	8	3	1	1	1	1	0	1		
	M	23	28	46	32	42	27	31	15	5	4	0	0.117	0.774
58.	F	32	44	43	30	23	23	14	16	7	5	0		
	M	23	28	42	32	39	15	18	8	2	1	0	0.030	0.905
59.	F	32	43	38	27	17	19	11	9	3	2	0		
	M	22	23	33	26	24	11	7	4	0	1	0	0.990	0.329
60.	F	31	41	31	22	11	17	2	4	2	1	0		