Educational Research and Reviews Vol. 5 (7), pp. 385-399, July 2010
Available online at http://www.academicjournals.org/ERR2
ISSN 1990-3839 © 2010 Academic Journals

Full Length Research Paper

Using IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in Mathematics

O. O. Adedoyin

University of Botswana, Botswana. E-mail: omobola_adedoyin@yahoo.com. Tel: 002673555107 or 0026771429736.

Accepted 11 January, 2010

This is a quantitative study, which attempted to detect gender bias test items from the Botswana Junior Certificate Examination in mathematics. To detect gender bias test items, a randomly selected sample of 4000 students responses to mathematics paper 1 of the Botswana Junior Certificate examination were selected from 36,000 students who sat for the examination. Out of which 2,000 were males and 2000 were females. The examination paper consisted of 38 test items. To detect the gender biased test items, the study used 3PL (Multilog software) item response theory (IRT) statistical analysis. This generated the item characteristics curves (ICC for the two groups (male/female). The study compared the results generated from the ICC curves for the male and female groups, and found that, out of 16 test items that fitted the 3PL item response theory (IRT) statistical analysis, 5 items were gender biased.

Key words: IRT (item response theory), ICC (item characteristics curve).

INTRODUCTION

There has been a lot of research in educational measurement directed towards improving the fairness of tests/examinations across various subgroups of examinees, because important decisions are made based on test scores. A fair test is one that is comparably valid for all groups and individuals and that affords all examinees an equal opportunity to demonstrate the skills and knowledge which they have acquired and which are relevant to the test's purpose (Roever, 2005).

The presence of bias is a cause for concern because, tests are used as a gatekeeper for educational opportunities, and it is a very important issue that test items are fair for every examinee. Bias is the presence of some characteristic of an item that results in differential performance for individuals of the same ability but from different ethnic, sex, cultural or religious groups. Item bias can also be defined as invalidity or systematic error in how a test item measures a construct for the members of a particular group (Camilli and Shepard, 1994, p. 8). An examination item is considered biased if it functions differently for a specified subgroup of test-takers, in such a case, students who are equally able do not have an equal chance of success (Zumbo, 1999). A biased item according to (Williams, 1997), measures attributes irrelevant to the test construct.

An item may be biased if it contains content or language that is differentially familiar to subgroups of examinees, or if the item structure or format is differentially difficult for subgroups of examinees. Content bias refers to situations where knowledge and or skills are not part of the educational background of the examinee. Lack of familiarity with content in test items disadvantages individuals in their performance. The individual's responses to items are not based on the appropriate ability level but on other wrong premises. Language bias occurs where words in the items have different or unfamiliar meanings for different examinee subgroups. Item structure and format bias occurs where there is ambiguity in the instructions, item stem or options. The content, or clues and explanations given to successfully complete the tasks provided disadvantage to individuals in some subgroups (Hambleton and Rodgers, 1995; Perrone, 2006).

To ensure that tests are fair for all examinees, most

large testing boards, organisation, have a formal review, which is part of the test development process, where items are screened by content specialist for text that might be inappropriate or unfair to relevant subgroups. The items are reviewed before field testing by content specialists as well as after field testing. The screening and the test development processes involves test developers and practicing teachers. After field trial, statistical measures are employed to identify items that are biased against a certain group of examinees, such test items are removed before the final set of test items are compiled. Item bias is of particular concern on tests of mathematics achievement where differences between males and females are commonly found (Kimball, 1989; Scheuneman and Grima, 1997). The failure to understand and account for gender differences on any test may lead to misinterpretations of the examination results.

Previously, a variety of methods had been used for detecting item biasness, e.g. the chi-squared method, the transformed item difficulty method, but recent interest in item response theory (IRT) in the measurement community has helped to assess differences in subgroup performance at the item level. IRT statistical analysis produces parameter estimates and item characteristics curves for each test item. Item characteristic curve method is a kind of detecting biased test items, which is based on item response theory. That is, item characteristic curve is used to compare the item characteristic curve difference between different groups. Under an IRT framework, a test item is biased if the ICC is not the same for various different groups, e.g. gender groups (boys and girls), who are equal in level on the latent trait do not have the same probability of endorsing a test item (Embretson and Reise, 2000).

The purpose of this study is to detect gender test items that are biased from the Botswana Junior Certificate test paper 1 in mathematics, IRT method of gender item bias detection will primarily focus on if there is any difference between the item characteristic curves of the male / female sub-groups. The results will shed light on the effective use of IRT approach in detecting biased test items for sub-groups based on gender, ethnic, race or culture from the population of all examinees.

The Botswana Junior Certificate (JC) examination is a national examination completely managed by the Botswana Examination Council (BEC). The current structure of education in Botswana is seven years of primary education, three years of junior secondary education, two years of senior education, and four years of university education (7+3+2+4). The Botswana Junior Certificate (JC) examination is administered at the end of the third year of the Junior Certificate (JC) course to measure the achievement level of candidates at that point. The examination is used for two purposes; as a tool to select students who are to proceed to the next level of education, which is the senior secondary, also as an assessment mechanism that measures the extent to which basic competencies and skills have been acquired. For the three years of junior secondary school, students are required to take six core subjects and one optional subject. The core subjects are Setswana, English, Mathematics, Science, Social Studies and Agriculture. The optional subjects are Home Economics, Design and Technology, Religious and Moral education.

Item response theory

IRT is that the probability of answering an item correctly or of attaining a particular response level is modeled as a function of an individual's ability and the characteristics of the item. And a paramount goal of IRT is predicting the probability of an examinee of a given ability level responding correctly to an item of a particular difficulty. The latent traits can be measured on a transformable scale having a midpoint of zero, a unit measurement of one and arrange from negative infinity to positive infinity. While the theoretical range of ability is from negative infinity to positive infinity, practical considerations usually limit the range of values from -3 to +3 (Hambleton et al., 1991).

IRT begins with the proposition that an individual's response to a specific item or questions is determined by an unobserved mental attribute of the individual. Each of these underlying attributes, most often referred to as latent traits, is assumed to vary continuously along a single dimension usually designated by theta (θ) (Hambleton et al., 1991). There are traditionally three IRT mathematical equations termed; one, two, and three parameter models that are used to make predictions. The general, IRT framework encompasses a group of models and the applicability of each model in a particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items. These models relate the characteristics of individuals and the characteristics of the items to the probability of a person with a given characteristics or level of an attribute choosing a correct response. For test items that are dichotomously scored, there are three IRT models, known as three-, two- and one- parameter IRT models. A primary distinction among the models is the number of parameter used to describe items. Although the one- parameter model is the simplest of the three models, it may be better to start from the most complex, the three-parameter IRT model.

The three parameter IRT model takes the following form:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-D_{a_i}(\theta - b_i)}} \quad (1)$$

Where c_i is the guessing factor, a_i is the item discrimination

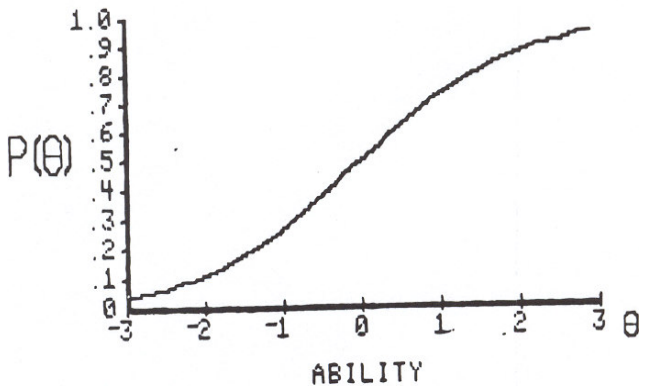


Figure 1. Example of item characteristics curve (ICC).

parameter commonly known as item slope, b_i is the item difficulty parameter commonly known as the item location parameter, D is the arbitrary constant (normally D = 1.7) and θ is the ability level of a particular examinee. The item location parameter is on the same scale of ability, Θ, and takes the value of Θ at the point at which an examinee with the ability-level θ has a 50/50 probability of answering the item correctly. The item discrimination parameter is the slope of the tangent line of the item characteristics curve at the point of the location parameter.

When the guessing factor is assumed or constrained to be zero (c_i =0) the three-parameter model is reduced to the two- parameter model for which only item location and item slope parameters need to be estimated.

$$P_i(\theta) = \frac{1}{1 + e^{-D_{a_i}(\theta - b_i)}} \quad (2)$$

If another restriction is imposed which stipulates that all items have equal and fixed discrimination, then a_i becomes a constant rather than a variable, and as such, this parameter does not require estimation, and the IRT model is further reduced to:

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}} \quad (3)$$

so, for the one- parameter IRT model, constraints have been imposed on two of the three possible item parameters, and item difficulty remains the only item parameter that needs to be estimated. The three-parameter model is the most general model, and the other two IRT models (two- and one-parameter models) can be considered as models nested or subsumed under the three- parameter model (Lord, 1980; Hambleton and Swaminathan, 1985; Hambleton et al., 1991). The three IRT models are based on the logistic (cumulative) distribution function (Hambleton et al., 1991).

These logistic equations when graphed, produce plots that are called item characteristic curves (ICCs) (Figure 1). When ICCs are plotted the ability of the examinee is denoted by theta (θ) on the x-axis, while the probability of an examinee correctly answering the question is denoted by P (θ) on the y-axis. ICCs typically take the shape of an S – shaped curve called ogive (∫).

The probability of the correct response is near zero at the lowest levels of the trait and it increases to the highest levels of the traits where the probability of correct response approaches 1 (Hambleton et al., 1991). There are two technical properties that are used to describe ICC, the values of item difficulty and item discrimination. The value of item difficulty denoted by (b) is a location parameter, indicating the position of the item characteristics curve in relation to the ability that is required for an examinee to have a 50% chance of getting the item right. The item discrimination provides information on how well an item separates people with high and low ability levels.

The flatter the ICCs curve, the less the item is able to discriminate since the probability of correct response at the low ability levels is nearly the same as it is at high ability levels. The steeper the curve, the better the item can discriminate. The strongest utilization of IRT models have been in education, psychology and statistics fields primarily in instrument development and computerised adaptive testing. The growth in psychometrics, and computer adaptive testing in particular, has supported the growing interest in the use of IRT (Embretson and Reise, 2000). The backbone of IRT is the item characteristics curves produced for each test item. In using IRT to detect item bias, different ICC curves for each items / subgroups are produced for comparison.

In the context of large scale testing like the JC examinations, which is a national examination, the analysis of the scores of students is very essential in the production of student scores and grades, and in monitoring and evaluation of the quality of the test items for fairness within the country. In most of the African countries, Botswana inclusive, national examination are still analysed and interpreted using the classical test theory, which involves the use of the basic item analysis, like item difficulty, item discrimination and reliability coefficients. Apart from the common parameter estimates for analyzing tests/ examination scores, it is also necessary to assess differences in subgroup performance at the item level, commonly referred to as differential item functioning (DIF).

METHOD

The research population for this present study consisted of all students who sat for the Paper 1 of 2004 Botswana Junior secondary

school examinations in mathematics. The population of all students who sat for the Junior secondary school examination was thirty five thousand two hundred and sixty two (35, 262). Out of which 4000 students (2000 males and 2000 females) were randomly selected.

IRT (3PL) statistical method was used to analyse the responses from the different sub-groups (males/ females). The IRT statistical analyses produced parameter estimates for the two sub-groups and their corresponding Item characteristics curves. These item characteristics curves for the males and females groups were compared for gender item biased analysis.

The IRT model assumption

Unidimensionality is the most important assumption common for all IRT models. This assumption is sometimes empirically assessed by investigating whether or not a dormant factor exists among all the items of the test (Hambleton et al., 1991). The method used in this study for assessing the unidimensionality was performing exploratory factor analysis, principal component analysis with varimax rotation on the responses to the 38 items of Paper 1 Botswana JSS Certificate Examination in Mathematics using a sample size of 5000 examinees. This yielded five eigenvalues greater than 1. The first eigenvalue (5.718) was greater than the next four eigenvalues (1.781, 1.143. 1.012 and 1.004) (Table 1). The scree plot was plotted to guide in the determination of whether unidimensionality could be inferred (Figure 2). Unidimensionality was inferred because of the presence of a dominating factor, that is, a single underlying factor.

Determining the appropriate model for this study's data (the model FIT)

All applications of IRT assume that the model is correct. The utility of the IRT model is dependent upon the extent to which the model accurately reflects the data. The overall fit of the model to the data was examined using goodness of fit statistics to test if the items fitted the given IRT model. The one parameter model (1PL), two parameter model (2PL) and three parameter model (3PL) were used for the overall model fit. But the resulting approximate chi-square statistics for the goodness of fit, (Table 2),however, showed that only two items out of 38 items fitted the one parameter model, eleven items (11) fitted the two parameter model (2PL) and sixteen items (16) fitted the three parameter model. The results of goodness of fit indicated that the data fitted the two and three parameter IRT models.

With the result of the goodness of fit analysis, the 3PL model was used for the item parameter estimates of the 16 items and for the item characteristics curves. The items are numbers (1, 2, 3, 4, 5, 12, 15, 16, 21, 22, 23, 30, 31, 34, 37, 38).

RESULTS

Table 3, summaries the results for male/female students item parameter estimates from the generated item characteristics curves, for the following items (1, 2, 3, 4, 5, 12, 15, 16, 21, 22, 23, 30, 31, 34, 37 and 38). Out of the sixteen items the following eleven (11) items were non significant (1,4, 5, 16, 21, 22, 23, 30, 34 37 and 38), and five items were significant (2, 3, 12, 15, 31) because the item characteristics curves for both the male and female were different, which shows that these items were gender biased towards a particular group.

DISCUSSIONS

From the item characteristics curves for both male and female groups, the most obvious test items that exhibited gender test items bias were five (item numbers 2, 3, 12, 15, 31). The item characteristics curves for the five identified test items were not the same for both male and female groups, which means that the five test items were gender biased.

Item 2

Both the male and female item characteristics curves (ICC) shifted vertically up due to high guessing factor (male c= 0.438 and female c= 0.569) but it is more pronounced in the female group than the male group. This represents an easy item because of the probability of correct response is high for the low ability examinees due to guessing factor. The two item characteristics curves are not identical, which shows that item 2 is biased towards a male group, due to the fact that the female group can answered the question by guessing (Figure 3).

Item 3

For item 3, the two item characteristics curves are not identical, since the ICC for the female group shifted towards the right, and the guessing value c =0.199 is higher than the male group of c=0.093. This is a difficult item for both the male / female examinees, but more difficult for the female examinees due to the fact that the probability of correct response is low for most of the ability scale and it increases only at the high ability levels. This item was biased towards the female group (Figure 4).

Item 12

The ICC curves for the male and female groups are not the same, the two curves shifted up, more pronounced in the female than the male group. The test item was a bit easier for the female group than the male group. Looking at the two ICC curves, the item was biased towards the male group. While this item is easy for the female groups, but a bit difficult for the male group (Figure 5).

Item 15

Item 15 was easier for the males than the females. The low ability group in the male group has a high probability of correct response than the low ability female group. The female item characteristics curve shifted towards the right

Table 1. Total variance explained by the result of factor analysis.

Component	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	5.718	15.046	15.046	5.718	15.046	15.046	3.756	9.884	9.884
2	1.781	4.688	19.734	1.781	4.688	19.734	3.216	8.462	18.347
3	1.143	3.007	22.741	1.143	3.007	22.741	1.458	3.837	22.183
4	1.012	2.662	25.404	1.012	2.662	25.404	1.151	3.030	25.213
5	1.004	2.643	28.047	1.004	2.643	28.047	1.077	2.834	28.047
6	0.997	2.623	30.670						
7	0.986	2.595	33.266						
8	0.955	2.514	35.780						
9	0.950	2.499	38.279						
10	0.935	2.460	40.739						
11	0.925	2.434	43.172						
12	0.919	2.418	45.590						
13	0.910	2.395	47.985						
14	0.896	2.359	50.344						
15	0.887	2.333	52.677						
16	0.882	2.321	54.998						
17	0.875	2.302	57.299						
18	0.867	2.282	59.581						
19	0.855	2.251	61.833						
20	0.845	2.224	64.056						
21	0.838	2.205	66.261						
22	0.834	2.195	68.456						
23	0.822	2.164	70.620						
24	0.816	2.146	72.766						
25	0.813	2.140	74.906						
26	0.810	2.133	77.039						
27	0.805	2.118	79.157						
28	0.791	2.080	81.237						
29	0.782	2.057	83.294						
30	0.777	2.044	85.338						
31	0.755	1.987	87.325						
32	0.742	1.953	89.278						
33	0.726	1.912	91.190						
34	0.706	1.857	93.047						

Table 1. Contd.

35	0.679	1.786	94.833
36	0.675	1.777	96.610
37	0.672	1.768	98.378
38	0.617	1.622	100.000

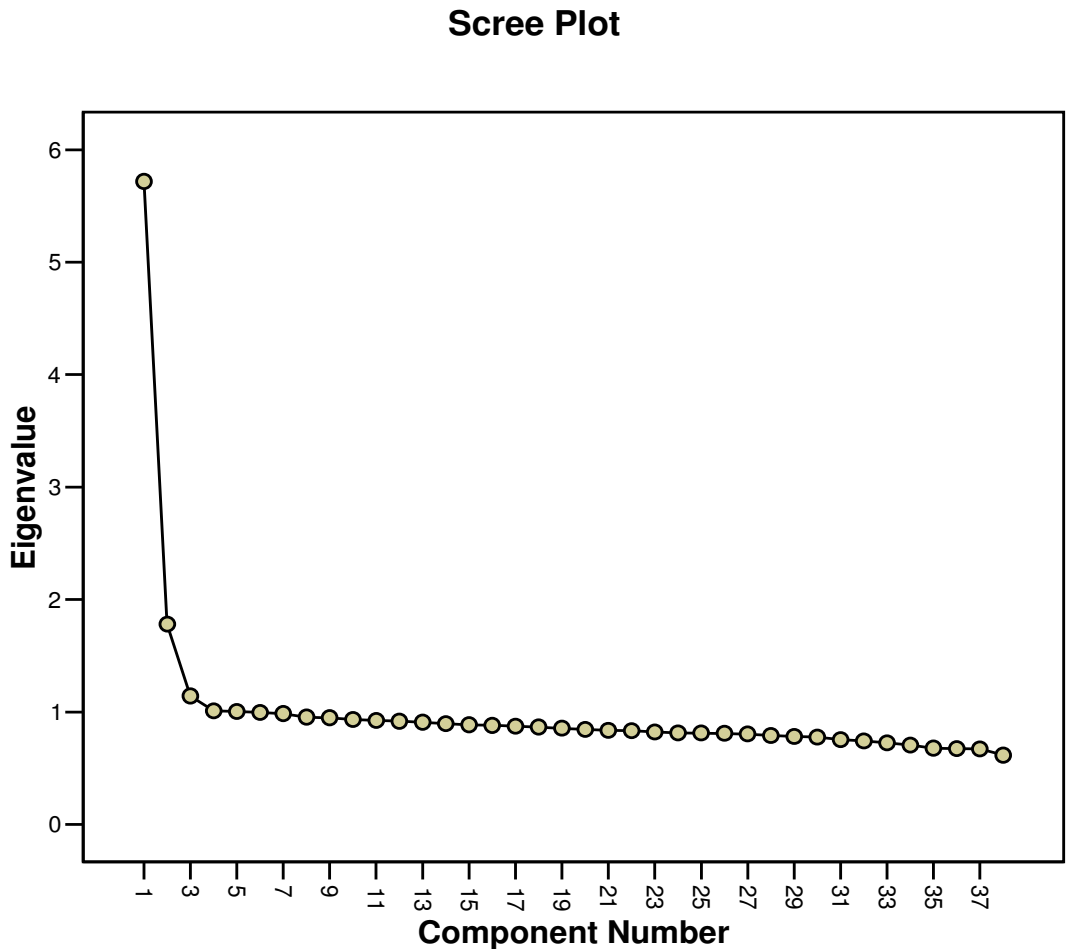


Figure 2. Scree plot of eigenvalue.

Table 2. Results of chi-square statistics for 1PL, 2PL and 3PL IRT models.

Items	1PL			2PL			3PL		
	Chi-square	df	p	Chi-square	df	p	Chi-square	df	P
1	201.4	8.0	0.0000	12.6	9.0	0.1796**	8.1	9.0	0.5271**
2	179.5	8.0	0.0000	19.3	9.0	0.2310**	11.4	9.0	0.2480**
3	215.5	8.0	0.0000	15.4	9.0	0.0817**	14.0	9.0	0.1220**
4	121.5	9.0	0.0000	16.6	9.0	0.0554**	18.4	9.0	0.0308**
5	23.3	9.0	0.0057	16.7	9.0	0.0537**	13.6	9.0	0.1370**
6	324.6	9.0	0.0000	57.5	8.0	0.0000	52.9	9.0	0.0000
7	296.4	9.0	0.0000	80.6	8.0	0.0000	76.7	8.0	0.0000
8	227	9.0	0.0000	53.5	9.0	0.0000	36.3	9.0	0.0000
9	96.2	9.0	0.0000	29.0	9.0	0.0006	29.7	9.0	0.0005
10	528.4	9.0	0.0000	78.2	9.0	0.0000	50.2	8.0	0.0000
11	310.1	9.0	0.0000	28.5	8.0	0.0004	24.6	9.0	0.0034
12	80.7	9.0	0.0000	12.2	9.0	0.2017**	14.1	9.0	0.1178**
13	78.9	9.0	0.0000	21.3	9.0	0.0115	7.7	9.0	0.0017
14	257.7	8.0	0.0000	57.3	7.0	0.0000	51.3	8.0	0.0000
15	6.9	9.0	0.6435**	6.3	9.0	0.7136**	8.7	9.0	0.4644**
16	75.1	9.0	0.0000	32.4	9.0	0.0002	8.4	9.0	0.4985**
17	31.8	8..0	0.0001	30.8	8.0	0.0002	31.5	9.0	0.0002
18	308.2	9.0	0.0000	79.3	7.0	0.0000	81.6	8.0	0.0000
19	94.1	9	0.0000	62.8	9.0	0.0000	62.0	9.0	0.0000
20	86.0	9.0	0.0000	73.0	9.0	0.0000	29.1	9.0	0.0006
21	34.4	8.0	0.0000	4.7	9.0	0.8612**	14.4	9.0	0.1084**
22	43.8	9.0	0.0000	7.7	9.0	0.5691**	13.8	9.0	0.1312**
23	20.0	9.0	0.0178	30.8	9.0	0.0003	10.9	9.0	0.2843**
24	105.6	9.0	0.0000	47.1	9.0	0.0000	34.7	9.0	0.0001
25	107.0	8.0	0.0000	125.0	9.0	0.0000	25.1	9.0	0.0028
26	27.2	9.0	0.0013	12.2	9.0	0.0003	21.2	9.0	0.0005
27	132.6	9.0	0.0000	111.4	8.0	0.0000	14.7	9.0	0.0001
28	24.7	9.0	0.0033	7.1	9.0	0.0002	8.5	9.0	0.0004
29	223.9	8.0	0.0000	81.6	9.0	0.0000	35.4	9.0	0.0001
30	17.3	9.0	0.0446	22.1	9.0	0.0087	13.3	9.0	0.1507**
31	59.4	9.0	0.0000	33.7	9.0	0.0001	15.5	9.0	0.07770**
32	366.6	9.0	0.0000	59.5	8.0	0.0000	34.5	9.0	0.0001
33	154.2	9.0	0.0000	62.0	9.0	0.0000	74.3	9.0	0.0000
34	440.2	9.0	0.0000	17.8	9.0	0.0373**	10.6	9.0	0.3068**
35	151.4	9.0	0.0000	105.0	9.0	0.0000	47.6	9.0	0.0000
36	259.5	9.0	0.0000	234.8	9.0	0.0000	104.0	9.0	0.0000
37	11.6	9.0	0.2353**	14.2	9.0	0.1156**	6.0	9.0	0.7436**
38	146.8	8.0	0.0000	27.9	8.0	0.0005	9.7	9.0	0.3773**

** The items selected with probability greater than the alpha level of 0.05 significant level.

Table 3. Summary of all non significant results for male /female students

items	Item discrimination (a)	Item difficulty (b)	Guessing parameter (c)
Item 1			
Males	1.053	0.220	0.216
Females	0.780	0.294	0.183
Item 2			
Males	0.887	-0.028	0.438 ***
Females	0.883	0.141	0.569
Item 3			
Males	1.345	0.320	0.093 ***
Females	1.495	0.501	0.199
Item 4			
Males	1.123	1.310	0.000
Females	1.217	1.213	0.104
Item 5			
Males	1.151	0.946	0.208
Females	1.028	1.253	0.149
Item 12			
Males	0.755	-0.689	0.244 ***
Females	0.859	-0.986	0.289
Item 15			
Males	1.015	-1.341	0.001 ***
Females	0.861	-1.586	0.000
Item 16			
Males	0.681	-1.954	0.001
Females	0.697	-2.329	0.000
Item 21			
Males	0.676	-0.415	0.000
Females	0.606	-0.914	0.000

Table 3. Contd.

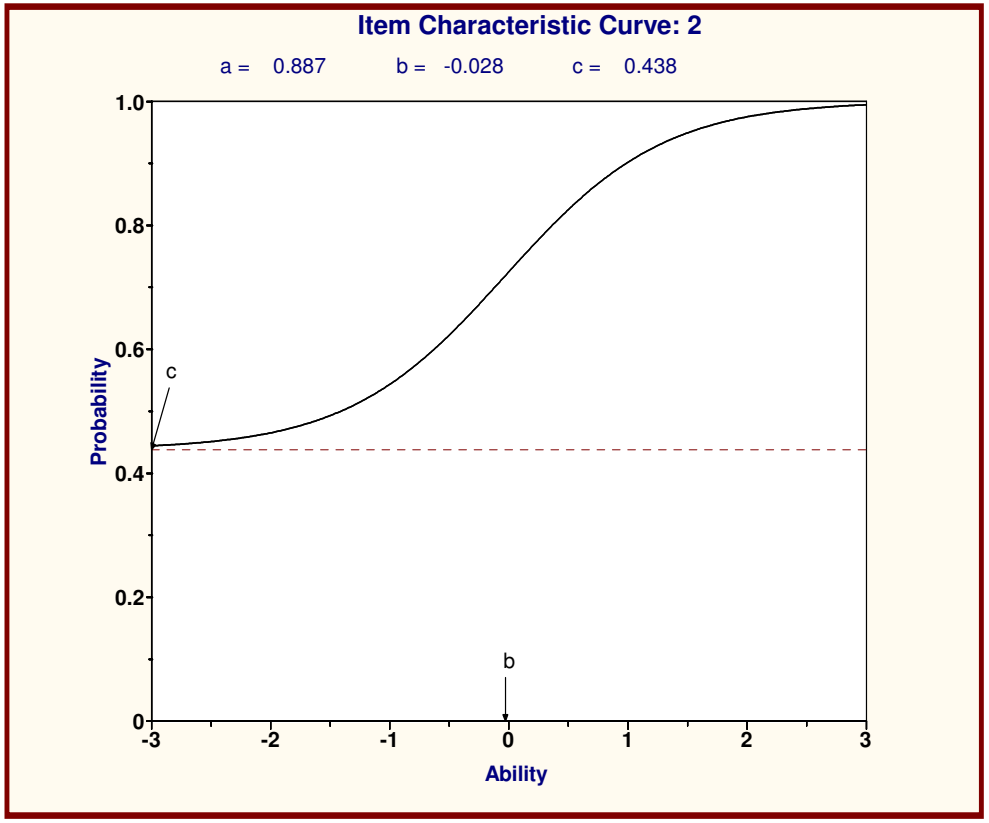
Item 22			
Males	2.049	1.506	0.168
Females	1.770	1.580	0.131
Item 23			
Males	0.839	0.155	0.311
Females	0.844	0.417	0.426
Item 30			
Males	1.906	1.448	0.578
Females	1.547	1.474	0.550
Item 31			
Males	1.793	2.186	0.424***
Females	-0.026	-24.300	0.252
Item 34			
Males	0.859	0.779	0.250
Females	1.062	0.938	0.304
Item 37			
Males	0.752	-0.634	0.000
Females	0.692	-0.255	0.018
Item 38			
Males	2.243	2.050	0.185
Females	1.562	2.228	0.142

*** Significant items that were gender biased.

that is the probability of correct response is low for most of the ability scale and it increases only at the high ability levels. The guessing parameter for the female group was also higher than the male group. This item was biased towards the female group (Figure 6).

Item 31

The most obvious test item that exhibited test bias was item number 31, the item discrimination value for the male sub-group a= 1.793, item difficulty value b=2.186 and the guessing value c= 0.424, whereas the item discrimination value for the female sub-group a = -0.026, item difficulty value b= -24.300, and the guessing value c= 0.252. It shows that this test item was very difficult for both groups (male/female) but more difficult for the female examinees than the male examinees. The



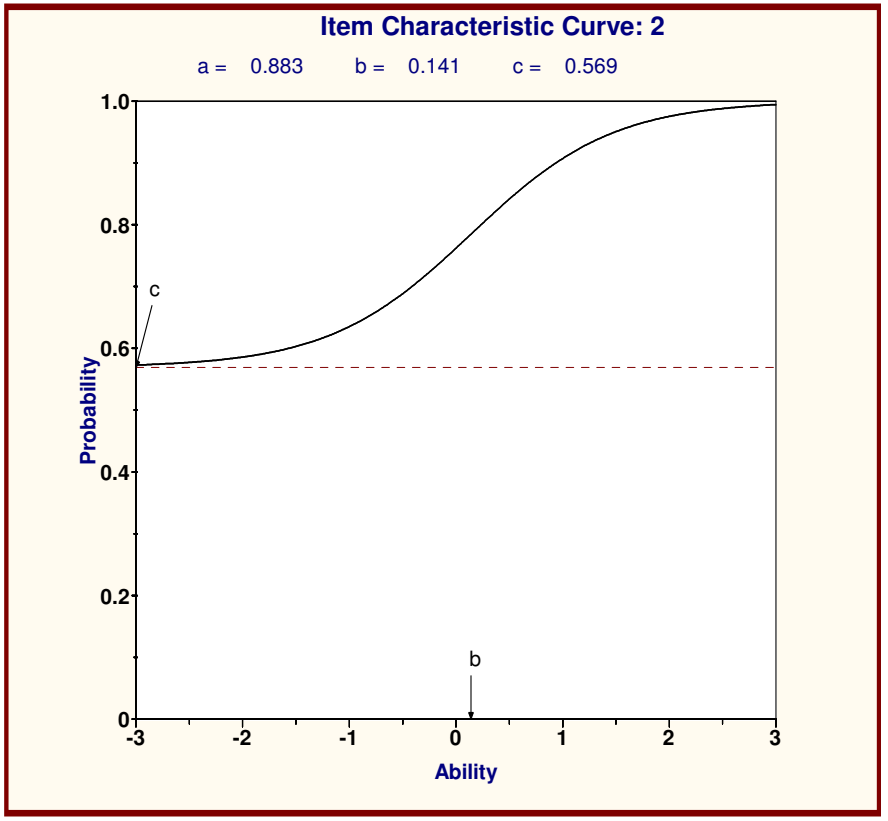
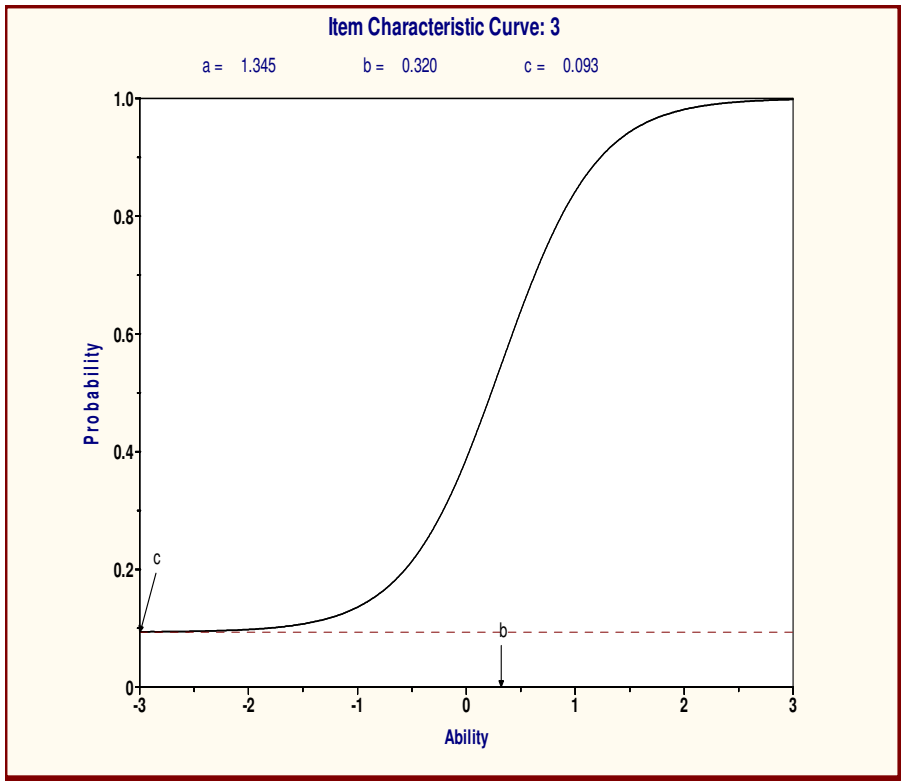


Figure 3. The two item characteristics curves. M (a=0.887, b=-0.028, c=0.438); F(a=0.883, b=0.141,c=0.569).

female item characteristics curve for item 31 was more or less flat, that is a straight horizontal line, which cannot discriminate well among the female group. This item was biased towards the female group. This item is not fit to be part of a public examination (Figure 7).

The nature of the identified five (5) items that demonstrated DIF were such that the questions leaned towards ideas that depict inherent interests depending on gender. Such interest could be in the area of questions that relate to sports that normally interest male rather than female, or



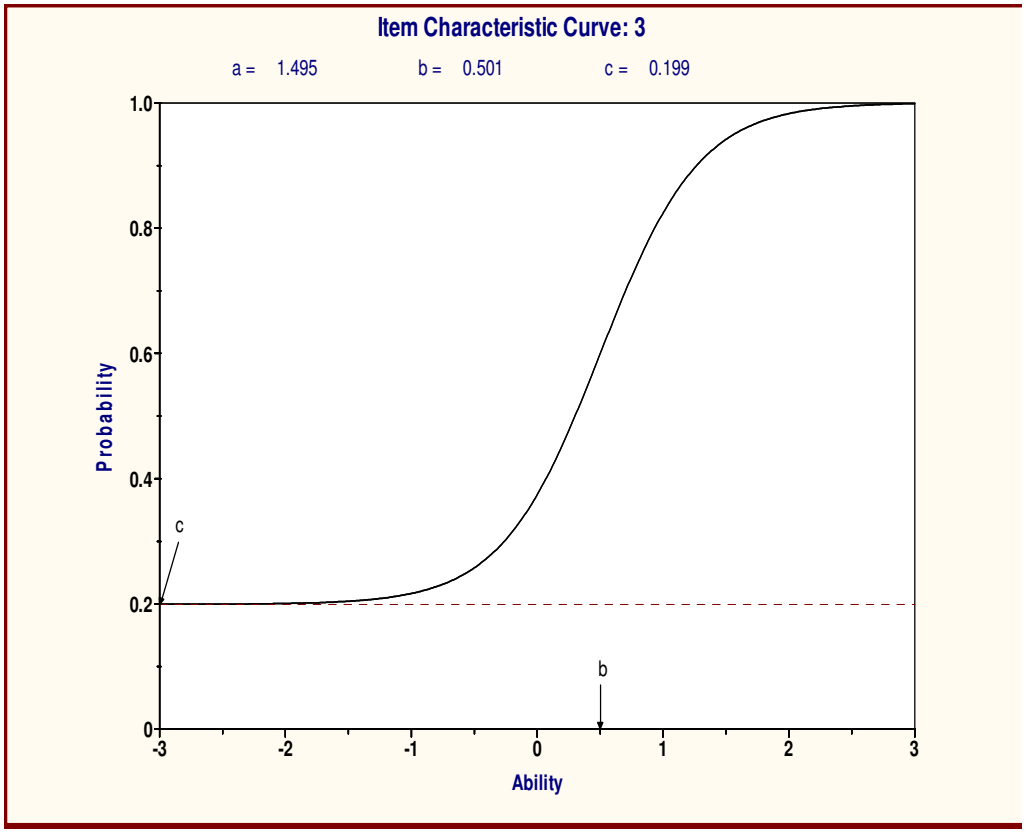
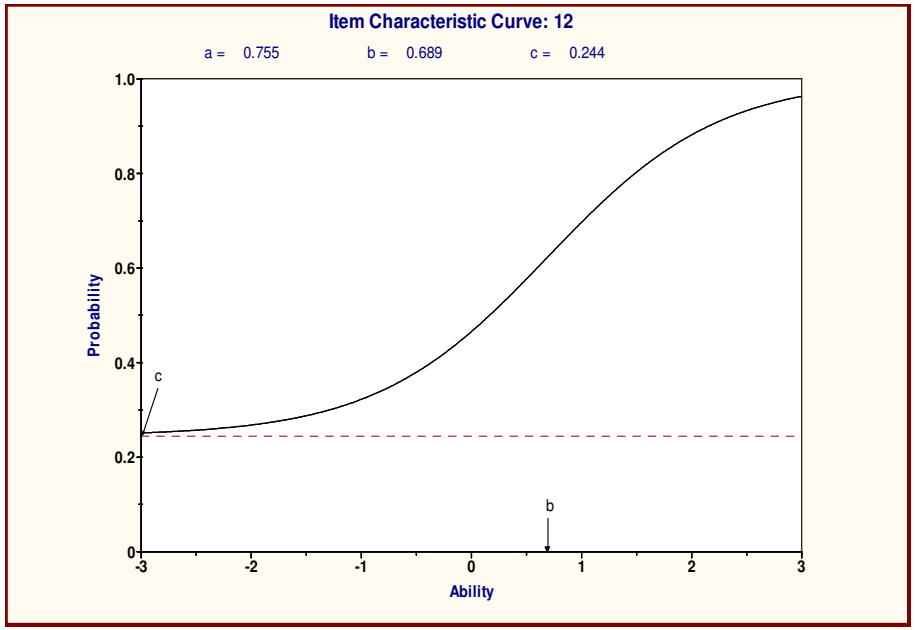


Figure 4. The three item characteristics curves M(a=1.345, b=0.320, c=0. 093); F(a=1.495, b=0.501, c=0.199).

questions on everyday domestic activities which will interest the female rather than the male.

For example, item 2 was biased towards the male group because the question was on purchasing of items using the knowledge of cost price, selling price and calculation of discounts. This item favoured the female group because they were more familiar with buying and selling of goods.

Item 3 was a technical question on measurements, the diagram given showed some potatoes on a measuring scale, the question was “what is



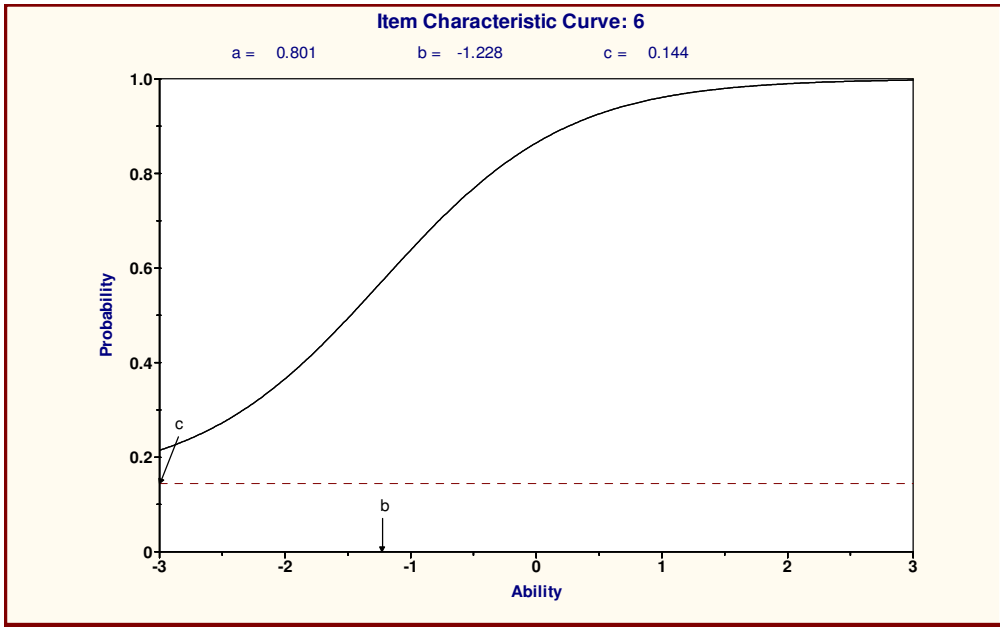
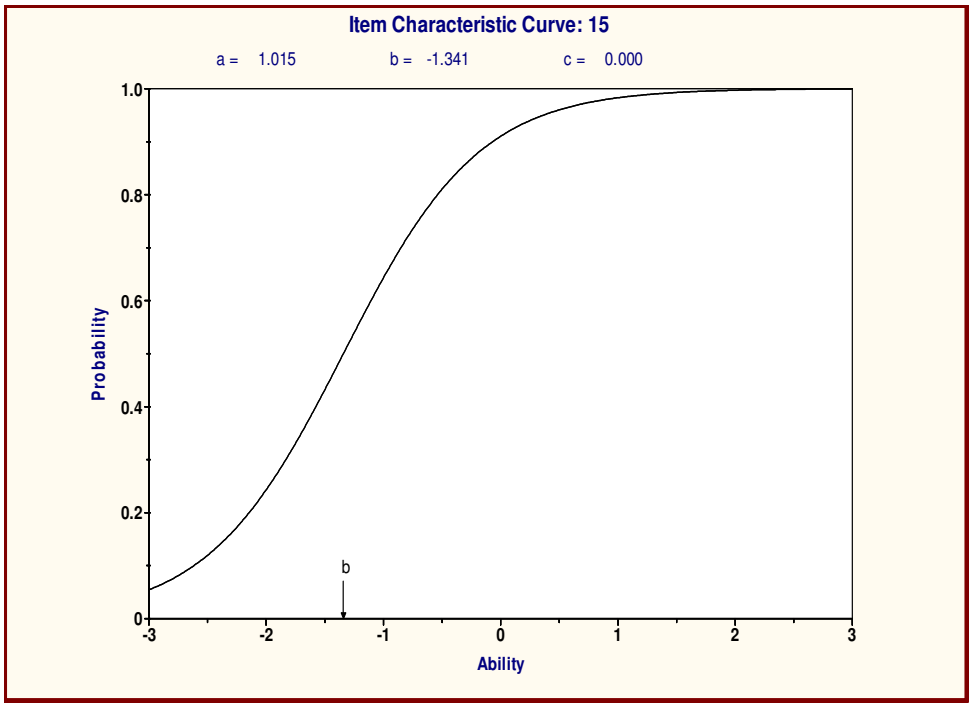


Figure 5. The twelve item characteristics curves M(a=0.755, b=-0.689, c=0.244); F(a=0.859, b=-0.986, c=0.289).

the mass of the potatoes" reading of the measuring scale could be a problem to the female group. Item 12 was biased towards the male group because of the language used, students were asked to give the name of the shape of "a living room". The word living room gave the female group more advantage to answer the question than the male group, because domestically, female are more familiar to the word 'living room' than their male counterpart. Item 15 was on comparing performances of students given a bar chart showing marks for four students in a test and the names of these students were left out. This item could be biased towards the female group, because it involved the use of diagrams and interpretation of the bar charts. Item 31, students were asked to give the number of lines of symmetry of a design which was inform of a football shape. Using a design like the football could be biased towards the female group. The findings of this study confirm the views of Scheuneman (1982a) on item bias, who stressed that an item may be biased if it contains content or language that is differentially familiar to subgroups of examinees, or if the item structure or format is differentially difficult for subgroups of examinees. According to Scheuneman (1982a) an example of content bias against girls would be one in which students are asked to compare the weights of several objects, including a football. Since girls are less likely to have handled a football, they might find the item more difficult than boys, even though they have mastered the



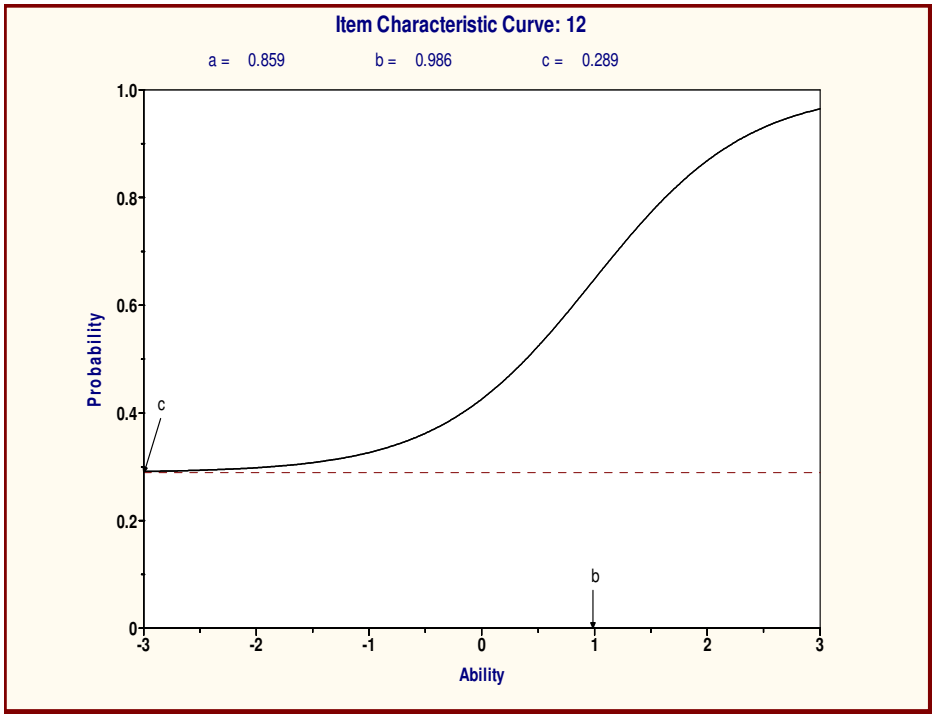


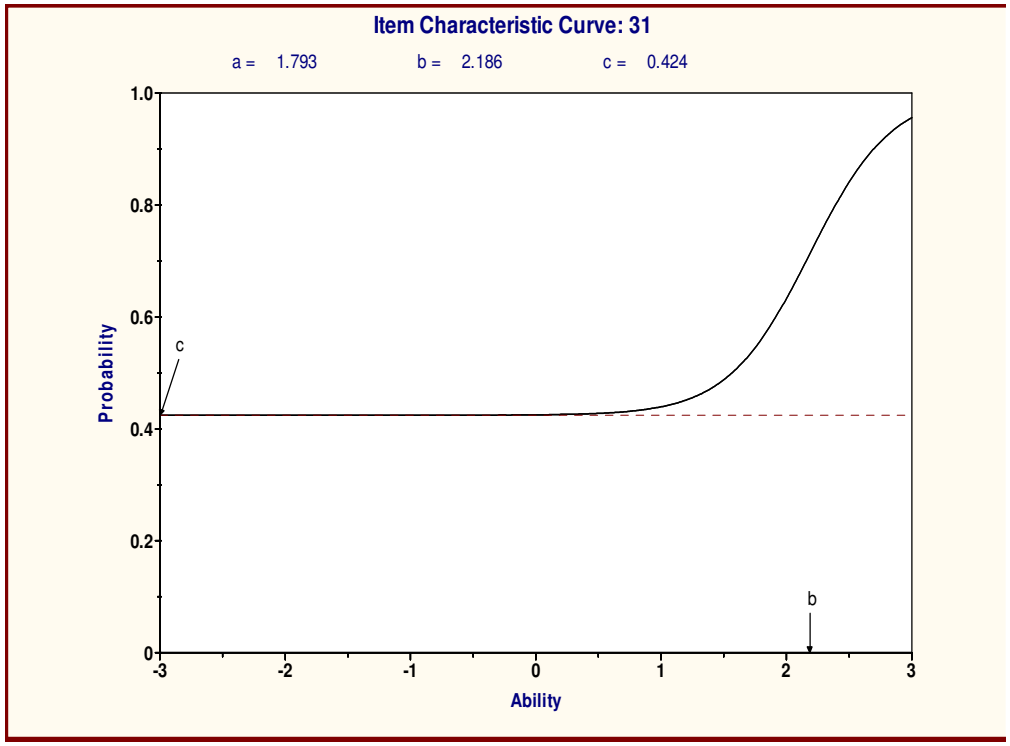
Figure 6. The fifteen item characteristics curves. M (a=1.015, b=-1.341, c=0.018); F(a=0.861, b=-1.586, c=0.283).

concept measured by the item.

Conclusion

The reality of item bias is a phenomenon that must be acknowledged and appropriately dealt with in examinations and tests designed for heterogeneous groups. In accordance with the findings of this study, it is suggested that item bias need not be a limitation to ensuring gender fairness, provided that the bias does not cause a recognizable difference in the total test scores of different groups.

Through the application of IRT methodology (ICC), it was clear that the biased items that were identified in the 2004 Botswana mathematics paper 1 examination would definitely have caused a recognizable difference in test scores for the male and female groups. This study only tried to identify the test item bias for the 2004 Botswana mathematics paper 1, there is need to detect gender bias test items from other subjects in any public examinations, through the use of item



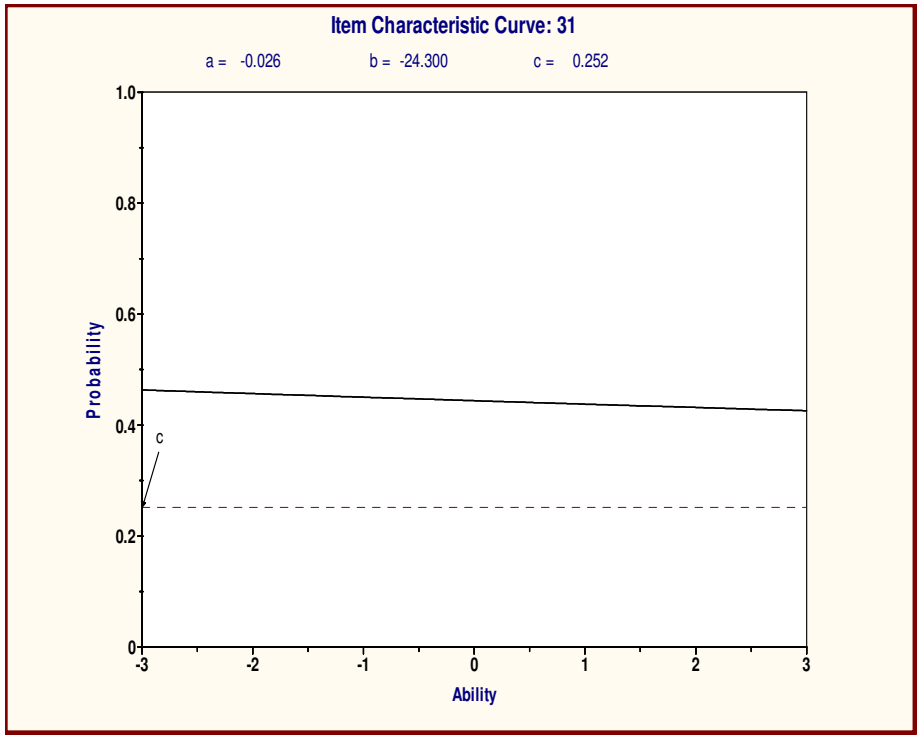


Figure 7. The thirty one item characteristic curve, M (a=1.793, b=2.186, c=0.424); F(a=-0.026, b=-24.300, c=0.252).

response approach (ICC curves).

RECOMMENDATIONS

1. For more objective, educational measurement IRT theoretical framework should be incorporated by Examination Boards into educational measurement practices, tests or examinations in Africa.
2. The construction and analysis of public examinations in Africa should utilise item response theory approach.
3. The item characteristic curves should be used to detect for gender bias test items. Test items should be free from bias towards a particular group.
4. The issue of IRT parameter estimates is still new in Africa, therefore, workshops, seminars and conferences should be organised for researchers in educational testing.

5. It is high time for experts in educational measurement in Africa to rise to the challenges pose by the measurement community and be fully aware of the usefulness of IRT in constructing and scoring of tests or examinations.

REFERENCES

Camilli G, Shepard LA (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.

Embretson S, Reise SP (2000). Item response theory for psychologists. Mahwah New Jersey: Lawrence Erlbaum Associates Publishers.

Hambleton R, Rodgers J (1995). Item bias review. Practical Assessment, Research, and Evaluation, 4(6). Retrieved March 18, 2009, from http://PAREonline.net/getvn.asp?v=4andn=6.

Hambleton RK, Swaminathan H, Rogers HJ (1991). Fundamentals of item response theory. Newbury Park, California: Sage Publications.

Hambleton RK, Swaminathan H (1995). Item response theory: Principles and application. Boston: Kluwer.

Kimball MM (1989). A new perspective on women's math achievement. Psychol. Bull. 105: 198-214.

Lord FM (1980). Applications of item response theory to practical testing problems. Mahwah, NJ: Erlbaum.

Perrone M (2006). Differential item functioning and item bias: Critical consideration in test fairness. Applied Linguistics, 6(2): 1-3.

Roever C (2005). "That's not fair!" Fairness, bias, and differential item functioning in language testing. Retrieved March 18, 2009 , from the University of Hawai'i System Web site: http://www2.hawaii.edu/~roever/brownbag.pdf.

Scheuneman JD (1982). A new look at bias in aptitude tests. In P. Merrifield (Ed.), New directions for testing and measurement: Measuring human abilities, No. 12. San Francisco: Jossey-Bass.

Scheuneman JD, Grima A (1997). Characteristics of quantitative word items associatedwith differential item functioning for female and black examinees. Applied Measurement in Education, 4: 299-320.

Thissen.D (1991). Multilog version7.0. Multiple categorical item analysis and test scoring using Item Response Theory. Chicago; Scientific Software International.

Zumbo BD (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-like (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation.