International Journal of Asian Social Science

journal homepage: http://www.aessweb.com/journal-detail.php?id=5007

USING IRT PSYCHOMETRIC ANALYSIS IN EXAMINING THE QUALITY OF JUNIOR CERTIFICATE MATHEMATICS MULTIPLE CHOICE EXAMINATION TEST ITEMS

O.O Adedoyin

University of Botswana, Botswana

T Mokobi

Department of Educational Planning and Research Services, Botswana

ABSTRACT

This study aims at providing the psychometric analysis of 2010 Botswana mathematics JC paper 1 in determining the quality of the junior certificate mathematics multiple choice examination test items. The mathematics paper 1 consisted of forty (40) multiple choice test items which was constructed using the three year JC mathematics curriculum. The population for the study was all the 36,940 students who sat for the JC mathematics examination in 2010, out of which a sample of 10,000 was selected randomly by the use of SPSS computer software. The students' responses were analysed using IRT (3PL) model to examine the psychometric parameter estimates of the forty test items which were: item difficulty, item discrimination, and the guessing value. The item characteristics curves were also generated for each test item that fitted the IRT (3PL) model. Twenty three (23) items fitted the 3PLM out of the forty (40) items, and were used in examining the psychometric qualities of the JC mathematics test paper 1.The findings from this study indicated that out of the twenty three (23) items that fitted the IRT model, twelve (12) items were classified as poor test items, ten (10) items were classified as fairly good test items which could be revised or improved and one (1) item was considered to be good test item. It was therefore recommended that examination bodies should consider improving the quality of their test items by conducting IRT psychometric analysis for validation purposes.

Keywords: IRT (Item response theory), CTT (Classical test theory), ICC (Item characteristics curve).

INTRODUCTION

The quality of test items in any public examinations is always examined through item analysis of examinees' responses. Item analysis is a process which examines students' responses to individual test items in order to assess the quality of those items and of the test as a whole. Traditionally, the proficiency of individual examinees is reported in terms of number-right scores (number of items answered correctly). One limitation or weakness with CTT approach, is that students with the same number-right score may have different response patterns (i.e., correct answers on different items) and, thus, may not have the same level of proficiency measured by the test. Reports related to the quality of test items, on the other side, are usually limited to indexes of *item difficulty* (proportion of correct answers on the item) and *item discrimination*. But a key problem with such indexes is that they depend on the group of examinees being tested and, therefore, do not adequately reflect the measurement quality of the test items.

Classical Test Theory

Classical Test Theory (CTT) is based on the assumption that every individual or person has a true score, T, and this true score can be obtained if and only if traits are constant and there are no random errors which can affect the result. Yu (2008) indicated that a person's true score is defined as the expected correct score over an infinite number of independent administrations of the test. That is, the random errors are expected to cancel over many repeated measurements e.g. when the test taker writes the test sometimes he/she would be so lucky to get a score greater than her/his true score and sometimes he/she would get a score less than his/her true score. The mean of these scores would then bring the test takers marks close to his/her true score. Unfortunately, in real life it is rather impossible to have such a situation where repeated measures are possible unless if one retakes a different examinations of the same level. "Furthermore, with most, if not all psychological constructs, learning and memory processes are involved that will have a systematic, but undesirable influence on performance if a test is repeatedly administered" (Klerk, 2008). For instance, people could recall their previous test session and answer in a similar or better way to improve their performance as testing is repeated

Test takers never observe a person's true score but an observed score, X. It is therefore important to note that CTT assumes that the true score plus the error gives the observed score.

$$X = T + E$$

X = the total score/observed score obtained
T = the true score and
E = the error component
This type of analysis is in the realm of Classical test theory (CTT) and problems that occur with CTT analysis of the examinees' proficiency and quality of test items are successfully addressed in the framework of item response theory (IRT).

Item Response Theory

Item Response Theory (IRT) is mostly used for modeling responses to items and scoring of educational tests. Using the appropriate IRT model, the ability level of an examinee is accurately estimated with any (sub) set of items that measure this ability. IRT is a powerful tool used in measurement of examinee ability, selection of test items and for equating tests. The concept of Item Characteristics Curve (ICC) is used in IRT, to show the relationship between examinee ability and performance on an item. In IRT, ability and item parameters are both estimated based on examinees' response patterns on the test. There are three (3) common IRT models based on the number of item parameters, the one-parameter model has only the item difficulty parameter *(b)*, while a three-parameter model has item difficulty *(b)*, item discrimination *(a)*, and guessing *(c)*. The number of item parameters to be estimated determines which IRT statistical model will be used, and the test item analysis of any examination is based on item discrimination, item difficulty and the guessing parameters.

The *a* parameter: Item discrimination

One characteristic of a good test item is that high-ability candidates will answer it correctly more frequently than lower-ability candidates. The *a* parameter expresses how well an item can differentiate among examinees with different ability levels. A test item has positive discrimination when higher ability students have a high probability of answering an item correctly and lower ability students have a low probability of answering the item correctly. A test item has negative discrimination when high ability students have a low probability of answering an item correctly and low ability students have a higher probability of answering an item correctly. The discrimination values (a-values) of good items ranges between 0.5 to 2, and the steeper the slope of an ICC, the higher an item's discrimination value. High discrimination level indicate that the item discriminates well between low and high skilled individuals. The *a* parameter is a measure that can be graphically expressed by the steepness of the ICC. If the values of the item discrimination *a* is above 1, this is normally desirable value for a good test item and values above 0.75 can also be acceptable.

The *b* parameter: Item difficulty

The difficulty of an item, known as the *b* parameter, is the point where the S-shaped curve has the steepest slope. The more difficult an item is, the higher an examinee's ability must be in order to answer the item correctly. Items with high *b* values are hard items, that is, values of *b* greater than 1 indicate a very difficult item and low-ability examinees are unlikely to answer it correctly. Items with low *b* values below -1 indicate easy items, which most examinees, including those with low ability, will have at least a moderate chance of answering correctly. When the values of *b* is between -0.5 to 0.5, then the test items with such difficulty indexes have medium difficulty level.

The *c* parameter: Pseudo-guessing

Some IRT models include a pseudo-guessing parameter, the ***c*** parameter which expresses the likelihood that an examinee with very low ability can be able to guess the correct response to an item and therefore has a greater-than-zero probability of answering correctly. The item guessing parameter c, is the lowest value that an ICC curve attains. For example, an examinee who randomly select responses to items that have four response choices can answer these items correctly about 1 out of 4 times, meaning that the probability of guessing correctly is about .25.The purpose of this paper is to examine the quality of 2010 Mathematics Junior Certificate paper 1 multiple choice test items of a public examinations using IRT psychometric analysis. (Embretson, 1983; Embretson, 1994; Embretson, 1998) specified that most educational and psychological tests require examinees to engage in some form of cognitive problem solving and on these tests, the cognitive processes, strategies, and knowledge used by examinees to solve problems should be considered when attempting to validate the inferences made about the examinees.

Embretson (1983), also suggested that while cognitive theory can inform psychometric practice in many ways, so also the cognitive theory can enhance psychometric practice by illuminating the construct representation of a test. The construct or latent trait that underlies test performance is represented by the cognitive processes, strategies, and knowledge used by an examinee to respond to a set of test items. Once these cognitive requirements are sufficiently described, they can be assembled into cognitive models that are then used to develop items that elicit specific knowledge structures and cognitive processes. Test scores anchored to a cognitive model should be more interpretable and, perhaps, more meaningful to a diverse group of users because performance is described using a specific set of cognitive skills in a well-defined content area.

Embretson (1994) believed that test developers have been slow to integrate cognitive theory into psychometric practice because they lack a framework for using cognitive theory to develop tests.Embretson (1998) also argued that cognitive theory is not likely to impact testing practice until its role can be clearly established in test design. To try to overcome this impasse, Embretson (1995) developed the cognitive design system (CDS). The CDS is a framework where testdesign and examinee performance were explicitly linked to cognitive theory (also see (Embretson, 1994; Embretson, 1998; Embretson, 1999). The goal of such a link was to make both the test score and the construct underlying the score interpretable using cognitive theory. Embretson (1999) recently described the CDS as a three-stage process. In the first stage, the goals of measurement were described, in the second stage, construct representation was established and in the third stage, nomothetic span research (i.e., correlating the test score with other well-defined measures) was conducted.

LITERATURE REVIEW

Item Response Theory models

The three parameter IRT logistic model (3PLM) takes the following form:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}} \qquad (1)$$

where c_i is the guessing factor, a_i is the item discrimination 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 probability of answering the item correctly. At the point of the location parameter, the item discrimination parameter is the slope of the tangent line of the item characteristics curve (ICC).

When the guessing factor is assumed or constrained to be zero (c_i =0) the three-parameter logistic model is reduced to the two- parameter IRT logistic model (2PLM) for which only item location and item slope parameters need to be estimated.

$$P_i(\theta)=\frac{1}{1+e^{-Da_i(\theta-b_i)}} \tag{2}$$

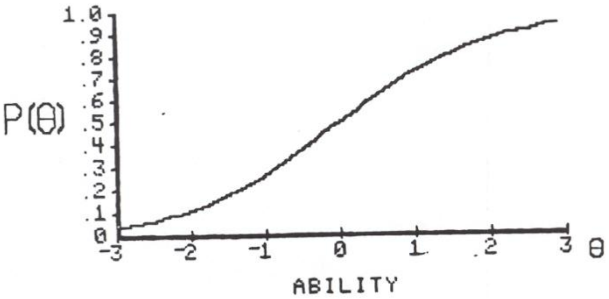
If another restriction is imposed that stipulates that all items have equal and fixed discrimination, then a_i becomes a constant rather than a variable, and as such, this parameter does not require estimation, and the IRT model is further reduced to

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

so, for the one- parameter IRT logistic model(IPLM), constraints have been imposed on two of the three possible item parameters, and item difficulty remains and item difficulty alone with θ remain the only parameters to be estimated. 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) (Fig. 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 o give ($\int$)

Fig-1. Example of Item Characteristics Curve (ICC)



The probability of the correct response is closer to 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). To describe the ICC, two technical properties are used, 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.

METHODOLOGY

Sample

The data for this study were the responses of 2010 JSS form three students in paper 1 mathematics multiple choice paper, these responses were obtained from Botswana Examinations Council. The examination mathematics paper 1 was administered at the end of form three to all students in JSS schools in Botswana. The population of this study was thirty six thousand, nine hundred and thirty-nine (36,939) JSS three students who sat for the mathematics paper one JC examination in government schools in Botswana, out of which,18271 were males and 18668 were females. A random sampling of ten thousand (10,000) was selected using the computer software for this study from the population of students.

Instrument

The researcher collected the data for this study from Botswana Examination Council (BEC). The data were the responses of students to 2010 mathematics paper one JC examination. The examination was a multiple choice paper which consisted of 40 items. The examination was administered for one and a half hours. Before the paper was administered, an assessment syllabus was used to create a scheme of assessment (test blue print). In the scheme of assessment the content area to be covered and the cognitive levels were shown to ensure a proper balance and emphasis of the syllabus.

The forty (40) multiple choice items of 2010 mathematics paper 1 were assessed for item fit and analyses were performed on only the items that fitted the 3PL model were subjected to IRT psychometric analysis to examine the quality of the test items, in terms of item difficulty, item discrimination and guessing parameter estimates using the examinees' responses. Item characteristic curves (ICC) were generated for the twenty four test items, using the Multi log 3.0 software. The most popular software packages for IRT model estimation, BILOG-MG (Zimowski *et al.*, 1996)and MULTILOG (Thissen, 1991), all IRT models estimable with BILOG-MG and MULTILOG are based on the three assumptions of local independence, monotonicity, and uni-dimensionality. The first assumption, local independence, states that the conditional probability of observing any response vector can be expressed as a product, across all items and examinees, of the probabilities of observing the individual response probabilities so that the response probabilities are independent at the local item level. BILOG (Mislevy and Bock, 1990) and MULTILOG (Thissen, 1991) are computer programme designed to facilitate item analysis and scoring of psychological tests within the framework of IRT. MULTILOG is for items with multiple alternatives and makes use of logistic response models, and commonly used for logistic models for binary item response data. MULTILOG provides Marginal Maximum Likelihood (MML) item parameter estimates for data in which the latent variable of IRT is random, as well as Maximum Likelihood (ML) estimates for the fixed effects case.

PRESENTATION OF RESULTS

Test for Uni-dimensionality

The method used to assess uni-dimensionality in this study was confirmatory factor analysis. It was performed to determine whether or not a dormant factor existed among all items as it was expected that the mathematics national examination would come up with one dominant factor. This factor would represent the construct underlining the mathematics skills measured by the examination. The exploratory factor analysis performed on the 40 items of the 2010 JC mathematics paper one yielded nine eigen values greater than one. The first eigen value was 5.909 greater than the next eight eigen values (1.492, 1.096, 1.088, 1.060, 1.029, 1.022, 1.017 and 1.010). The first factor explained 14.772% of the variance in the data set. The second factor explained 3.73% of the remaining variance. The rest of the variance was explained by the other 38 factors with 24 factors each having an percentage of variance between 2 and 3 and 14 factors each having a percentage of variance of between 1 and 2.

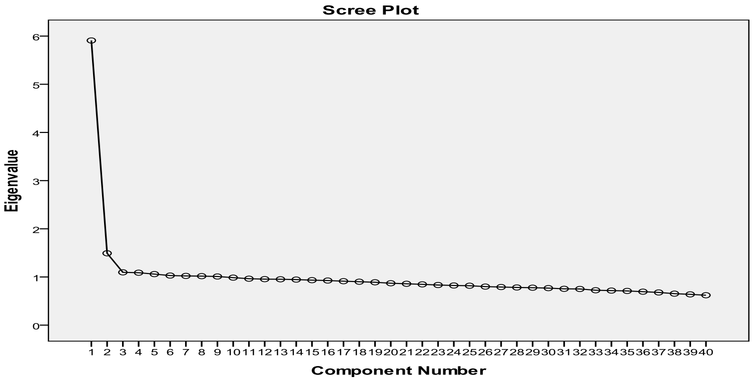
Table-1. Total Variance Explained by the result of factor analysis

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	5.909	14.772	14.772
2	1.492	3.730	18.502
3	1.096	2.740	21.242
4	1.088	2.720	23.963
5	1.060	2.649	26.612
6	1.029	2.573	29.185
7	1.022	2.554	31.739
8	1.017	2.543	34.282
9	1.010	2.524	36.806
10	.986	2.466	39.272
11	.964	2.410	41.682
12	.955	2.387	44.070
13	.952	2.381	46.450
14	.946	2.364	48.814
15	.934	2.334	51.149
16	.926	2.314	53.463
17	.912	2.281	55.744
18	.901	2.253	57.997
19	.890	2.224	60.222
20	.870	2.175	62.396
21	.859	2.147	64.543
22	.846	2.116	66.659

23	.833	2.082	68.742
24	.823	2.058	70.799
25	.818	2.045	72.845
26	.800	2.000	74.844
27	.791	1.978	76.822
28	.782	1.955	78.778
29	.777	1.943	80.721
30	.769	1.921	82.642
31	.754	1.885	84.527
32	.751	1.877	86.404
33	.724	1.811	88.215
34	.717	1.793	90.008
35	.710	1.775	91.783
36	.695	1.737	93.519
37	.679	1.697	95.216
38	.654	1.634	96.850
39	.639	1.598	98.448
40	.621	1.552	100.000
Extraction Method: Principal Component Analysis.			

A scree plot was produced to determine whether uni-dimensionality could be inferred. The scree plots provided a convenient way of visualising a dominant factor in principal component analysis.

Figure 2 Scree Plot for the eigenvalues



Test for model fit

The utility of the IRT model is dependent upon the extent to which the given responses reflect this model. To determine whether the test item fitted the model, a Chi-square test was run on the data set using Bilog-M to establish whether the items fitted the 1PL, 2PL and 3PL models. Table 2 showed the results of the chi-square statistics. The Chi-square goodness of fit analysis showed that only one item fitted the 1PL model, eleven items fitted the 2PL model and 23 items fitted the 3PL model. For the 2PL and 3PL model item 9 was omitted from the calibration as its initial slope was less than -0.15

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

Items	1PL			2PL			3PL		
	Chi-square	**p**	**df**	**Chi-square**	**p**	**df**	**Chi-square**	**p**	**df**
1	60.3	0.0000	9.0	15.4	0.0812**	9.0	16.4	0.0585**	9.0
2	94.3	0.0000	9.0	34.2	0.0001	9.0	27.0	0.0014	9.0
3	361.1	0.0000	9.0	15.5	0.0000	9.0	13.7	0.0011	9.0
4	247.6	0.0000	9.0	25.0	0.0016	8.0	13.2	0.1556**	9.0
5	154.4	0.0000	9.0	78.8	0.0000	8.0	67.0	0.0000	8.0
6	82.3	0.0000	9.0	13.5	0.1409**	9.0	9.7	0.3741**	9.0
7	28.3	0.0008	9.0	15.5	0.0788**	9.0	5.7	0.773**	9.0
8	240.4	0.0000	9.0	61.8	0.0000	9.0	44.7	0.0000	9.0
9	428.4	0.0000	8.0						
10	16.4	0.0593**	9.0	12.3	0.1967**	9.0	7.9	0.5469**	9.0
11	63.3	0.0000	9.0	56.4	0.0000	9.0	41.6	0.0000	9.0
12	112.0	0.0000	9.0	21.8	0.0094	9.0	19.9	0.0182	9.0
13	329.6	0.0000	9.0	53.2	0.0000	9.0	67.4	0.0000	9.0
14	38.7	0.0000	9.0	8.6	0.4726**	9.0	6.7	0.6718**	9.0
15	45.0	0.0000	9.0	11.6	0.2356**	9.0	16.4	0.0596**	9.0
16	85.9	0.0000	9.0	58.1	0.0000	9.0	42.9	0.0000	9.0
17	429.8	0.0000	9.0	69.6	0.0000	7.0	35.0	0.0000	8.0
18	76.6	0.0000	9.0	24.8	0.0032	9.0	19.1	0.0242	9.0
19	222.5	0.0000	9.0	29.3	0.0003	8.0	28.7	0.0007	9.0
20	387.5	0.0000	8.0	20.6	0.0083	8.0	20.6	0.0146	9.0
21	21.7	0.0099	9.0	21.1	0.0122	9.0	6.6	0.6812**	9.0
22	405.9	0.0000	8.0	58.6	0.0000	8.0	39.2	0.0000	8.0
23	57.8	0.0000	9.0	15.3	0.0840**	9.0	12.8	0.1736**	9.0
24	93.6	0.0000	9.0	60.9	0.0000	9.0	12.7	0.1747**	9.0
25	140.3	0.0000	9.0	40.3	0.0000	9.0	14.1	0.1196**	9.0
26	57.7	0.0000	9.0	64.9	0.0000	9.0	5.8	0.7557**	9.0
27	46.8	0.0000	9.0	14.4	0.1080**	9.0	15.6	0.0749**	9.0
28	252.5	0.0000	9.0	39.3	0.0000	9.0	27.7	0.0011	9.0
29	47.6	0.0000	9.0	18.2	0.0330	9.0	6.0	0.7404**	9.0
30	143.8	0.0000	9.0	31.5	0.0002	9.0	8.3	0.4999**	9.0
31	180.5	0.0000	9.0	20.9	0.0075	8.0	10.3	0.3234**	9.0
32	194.3	0.0000	9.0	28.8	0.0007	9.0	9.0	0.4404**	9.0
33	47.9	0.0000	9.0	10.3	0.3294**	9.0	4.6	0.8708**	9.0
34	128.6	0.0000	9.0	98.4	0.0000	9.0	12.8	0.1723**	9.0
35	171.7	0.0000	9.0	23.5	0.0028	8.0	14.3	0.1126**	9.0
36	103.9	0.0000	9.0	41.7	0.0000	9.0	40.8	0.0000	9.0
37	146.5	0.0000	9.0	29.6	0.0003	8.0	24.8	0.0032	9.0
38	68.7	0.0000	9.0	19.4	0.0222	9.0	14.1	0.1204**	9.0
39	97.8	0.0000	9.0	4.0	0.9111**	9.0	8.4	0.4980**	9.0
40	136.1	0.0000	9.0	17.0	0.0298	8.0	15.1	0.0893**	9.0

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

Table-3. The number of items fitting each model

IRT model	**1PL**	**2PL**	**3PL**
Items fitting the model	10	1, 6, 7, 10, 14, 15, 23, 27, 33, 39	1, 4, 6, 7, 10, 14, 15, 21, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 38, 39, 40
Number of items	1	10	23

fitting the model

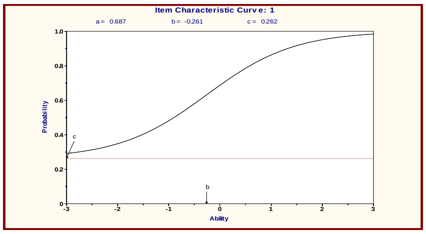
Since 23 items fitted the IRT (3PL) model, the 3PL IRT model was used to estimate the item parameters and to generate item characteristics curves (ICC).

DISCUSSION

Using IRT (3PL)model to examine the quality of test items, three parameter estimates considered were that the item must be able to discriminate very well among the examinees i.e must have a high value of item discrimination (a-value) greater or than 1 (a value of greater or equal to 1 is desirable), the item difficulty value (b value) between -0.5 to +0.5 (or values very near to – 1 to + 1), any value greater than 1, the item has a high difficulty value and since all test items should have a minimum difficulty of b=0.000 and a low guessing value (c-value) very close to 0.000 .

One useful feature about the item characteristics curve is that if the test is made up of test items that are relatively difficult, the ICC curve shifts to the right. But if the test items are very easy the ICC curve shifts to the left. 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.

For this study, in order to classify the test items into good, fairly good or poor items, the following criteria was used; for good test items, the discrimination parameter value *a* must be greater or equal to 1. The value of the difficulty parameter *b* should be from 0.5 to +1, any test item with value above +1 would be considered as difficult. Any test item with a *b* value less than 0.5 was considered as easy item. For the ***c*** value, it should be between 0.00 to 0.25, test items with *c* values greater than 0.25 was considered as an item with a high probability of guessing the answer correctly and such test items would be classified as not good test items.

23 Items that fitted the 3PL	Determining the quality of each item using the ICC curves and IRT item parameter estimates
Item 1 Which class of numbers is listed below? 2, 3, 5, 7,… a) Rectangle numbers b) Even numbers c) Prime numbers d) Odd numbers **Answer is C**	 (a=0.26,b=0.69,c=0.26) This is a poor item of medium ability the b value of 0.69, but the item has a discrimination low value of 0.26.

<table>
<tr>
<td>
Item 4

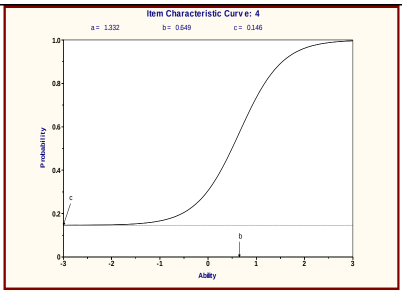
Work out $\begin{pmatrix}-5\\2\end{pmatrix}+\begin{pmatrix}2\\-4\end{pmatrix}$

a) $\begin{pmatrix}7\\-2\end{pmatrix}$

b) $\begin{pmatrix}-7\\-6\end{pmatrix}$

c) $\begin{pmatrix}-3\\2\end{pmatrix}$

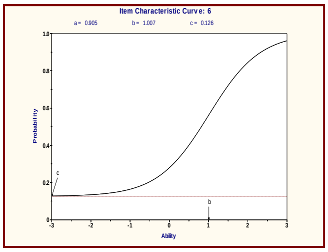
d) $\begin{pmatrix}-3\\-2\end{pmatrix}$

Answer is D
</td>
<td>


(a=0.65,b=1.33,c=0.15)
This item can be classified as poor since the item discrimination value is less than 1. The item difficulty parameter shows that the item is a bit difficulty (b>+1), the ICC curve shifts to the right, although the c-value is less than 0.25.
</td>
</tr>
<tr>
<td>
Item 6

Mpho’s bank balance id P300. He withdraws P310 and was charged P20 for overdrawing. What is his balance, in Pula after the withdrawal and the charge?

a) -40
b) -30
c) 10
d) 30

Answer is B
</td>
<td>


(a=1.01, b=0.9,c=0.13)
This item is a fairly good item because the item discrimination parameter (a>1) indicates that it differentiates well between the high ability and low ability examinees. The item difficulty parameter shows that the item is of medium difficulty because the b-value is lower than 1 and the value for the guessing parameter is lower than 0.25.
</td>
</tr>
</table>

Item 7

The table below shows how Kaelo spends 70 minutes of her study time. Use it to answer question 7.

English	Mathematics	Setswana
20 minutes	q minutes	Twice the time spent on Mathematics

Which of the following shows an equation for the total time Kaelo spent on studying the tree subject?

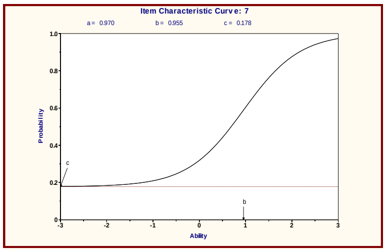
a) $q + 20 = 70$

b) $2q + 20 = 70$

c) $3q + 20 = 70$

d) $4q + 20 = 70$

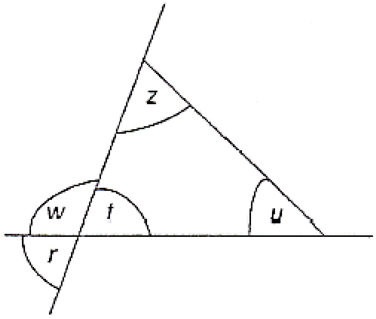
Answer is C



(a=0.95,b=0.97,c=0.18)
This item is fairly good item because the item discrimination parameter is very close to +1 indicates that it differentiates fairly well between the high ability and low ability examinees (***a*** very close to +1). The item difficulty parameter shows that the item difficulty level is not greater than +1 and the value of the guessing parameter is very low (c<0.25).

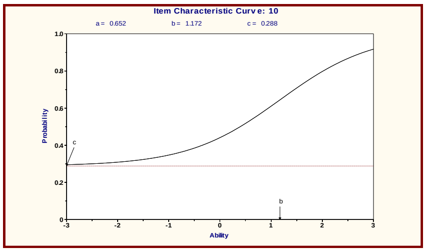
Item 10

Which angle is vertically opposite to t in the diagram below?



a) w
b) z
c) r
d) u

Answer is C



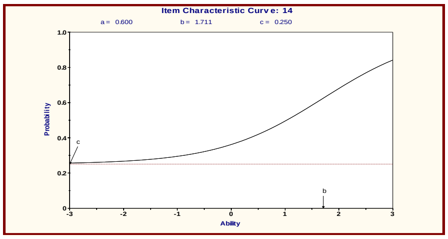
(a=1.17,b=0.65, c=0.29)
This item is a fairly good item because the item discrimination parameter indicates that it differentiates well between the high ability and low ability examinees. The item difficulty parameter shows that the item is of medium difficulty, although the value of the guessing parameter is a bit higher than 0.25. This item can be modified or improved.

Item 14

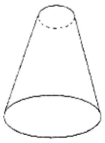
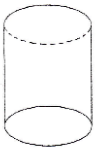
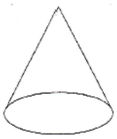
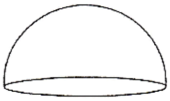
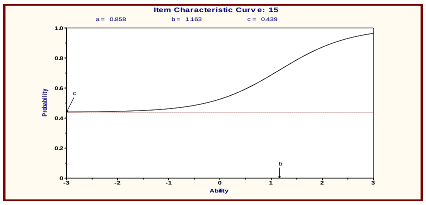
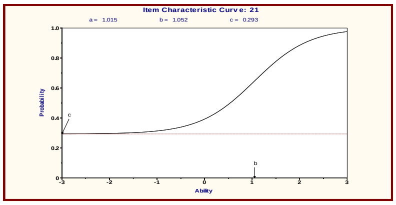
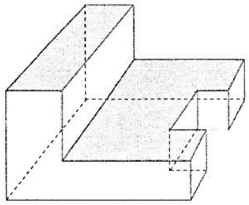
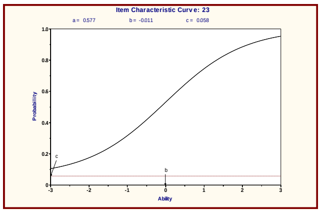
The time in Australia is 6 hours ahead of that in Botswana. A live broadcast of a game starts at 0815 in Australia. At what time will the same live broadcast start in Botswana.

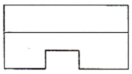
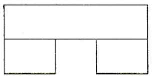
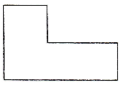
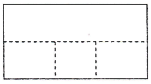
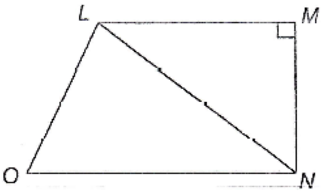
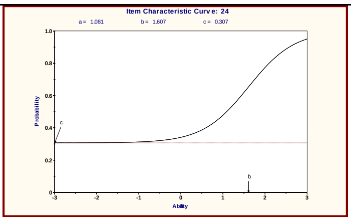
a) 0215
b) 0315
c) 1315
d) 1415

Answer is A

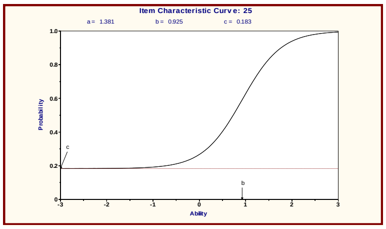


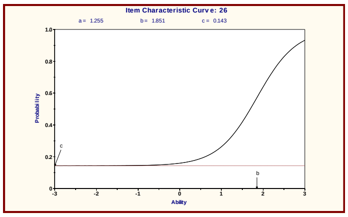
(a=1.71, b=0.60,c=0.25)
This item is a good item because the item discrimination value of +1.71 indicates that it differentiates well between the high ability and low ability

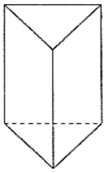
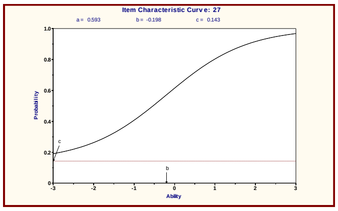
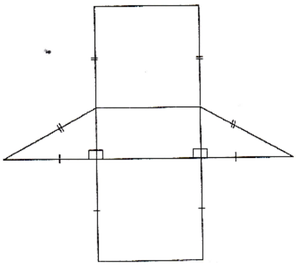
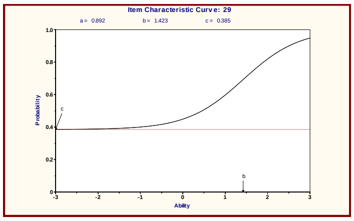
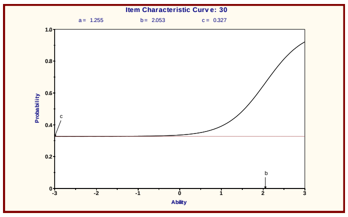
	students. The item difficulty value of 0.60 shows that the item is of medium difficulty and the value of the guessing parameter is 0.25.
Item 15 Which of the following is a prism?  A  B  C  D **Answer is B**	 (a=1.16,b=0.86,c=0.44) This item has a difficulty value of 0.86 and can discriminate between the high and low ability examinees, with a-value of +1.16. But it has a high guessing value of 0.44. The item can be classified a very poor test item.
Item 21 Expand (a+x)(a+2) a) 4ax+3ax b) 2a+ax+2x c) $a^2+3ax+2x$ d) $a^2+2a+ax+2x$ **Answer is D**	 (a=1.05,b=1.02,c=0.29) This item is a fairly good item because the item discrimination value of a=1.05 indicates that it differentiates between the high ability and low ability examinees. The item difficulty parameter (b= 1.02) shows that the item is a bit difficult, but the b-value is close to 1 and the value of the guessing parameter is very close to the specified c value of 0.25 for good items.
Item 23 The figure below, the shaded part represents the plan of the figure. Use the figure to answer question 23  Which of the following diagrams shows the plan of the figure?	 (a=-0.01, b=0.58,c=0.06) This item has a negative discrimination value (a=-0.01), which shows that the test item cannot discriminate well between the high and low ability examinees, although it is of medium

<table>
<tr>
<td>A 
B 
C 
D 
<b>Answer is A</b></td>
<td>difficulty (b=0.58) and with a low value of c the guessing parameter, this is a very poor item.</td>
</tr>
<tr>
<td>Item 24
Which side is adjacent to angle MNL in the figure below?

a) LM
b) MN
c) NO
d) LO
<b>Answer is B</b></td>
<td>
(a=1.61,b=1.08,c=0.31)
This item is a fairly good item because the item discrimination value indicates that it differentiates between the high ability and low ability examinees. The item difficulty value(b= 1.08) is very close to 1 and the value of the guessing parameter is a bit high.</td>
</tr>
<tr>
<td>Item 25

The number of students in a school decreased by 5%. After the decrease there are 475 students in the school. Calculate the number of students before the decrease.

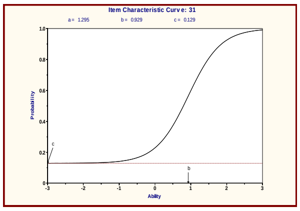
a) 570
b) 500
c) 480
d) 740
<b>Answer is B</b></td>
<td>
(a=0.92,b=1.38,c=0.18)
This item is a difficult item because the item difficulty value is greater than 1, the discrimination value indicates that it can differentiate to a little extent between the high ability and low ability examinees. The ICC curve is shifted to the right due to high value of b parameter. Although the c-value is less than 0.25. This test item can be classified as a poor item.</td>
</tr>
<tr>
<td>Item 26

The frequency table below shows the marks obtained by 12 pupils in a mathematics quiz. Use it to answer question 26.
<table>
<tr><th>Marks</th><th>Frequency</th></tr>
<tr><td>3</td><td>5</td></tr>
<tr><td>4</td><td>3</td></tr>
<tr><td>5</td><td>3</td></tr>
<tr><td>6</td><td>1</td></tr>
</table>
</td>
<td>
(a=1.85,b=1.25. c=0.14)
Although the ICC curve is shifted to the right, which shows that the item is a bit</td>
</tr>
</table>

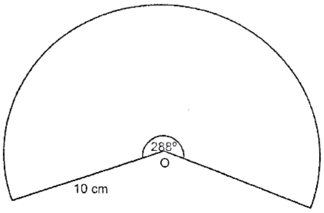
Calculate the median mark. a) 3 b) 3.5 c) 4 d) 4.5 **Answer is C**	difficult for the examinees, the test item can discriminate to a large extent between the high and low ability examinees. The item has a low guessing value of 0.14. This item can be classified as fairly good test item.
Item 27 What is the name of the shape that forms the cross-section of the prism shown below?  a) Parallelogram b) Rectangle c) Triangle d) Square **Answer is C**	 (a=-0.20,b=0.59.c=0.14) This item has a negative discrimination value (a=-0.20), which shows that the test item cannot discriminate well between the high and low ability examinees, although it is of medium difficulty (b=0.59) and with a low value of c the guessing parameter, this is a very poor item.
Item 29 The diagram below shows the net of a solid figure. Use it to answer question 29  What is the name of the solid figure a) Cube b) Cuboid c) Triangular Prism d) Triangular pyramid **Answer is C**	 (a=1.42,b=0.89,c=0.38) This item has a medium difficulty level and can discriminate to an extent between the high and low ability examinees, but it has a high guessing value of 0.38. The item can be classified as poor test item.
Item 30 Which of the dimensions below forms the sides of a right angled triangle? a) 8, 10, 12 b) 6, 10, 14 c) 6, 8, 16	 (a=2.05,b=1.25,c=0.33)

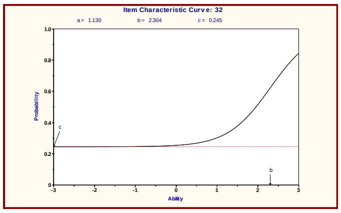
<table>
<tr>
<td>d) 6, 8, 10
Answer is D</td>
<td>This item discriminates very well between the high and low ability examinees, but it is a bit difficult since the b value is greater than +1. The guessing value ***c*** is high, the ICC curve shifted to the right. The item is a poor item.</td>
</tr>
<tr>
<td>Item 31

Simplify $2ax - 5bd - 4ax + 3bd$

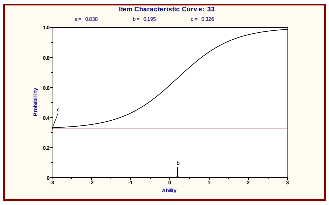
a) $-2ax - 2bd$
b) $2ax - 2bd$
c) $2ax + 8bd$
d) $6ax - 2bd$
Answer is A</td>
<td>

(a=0.93,b=1.30,c=0.13)
The item discrimination value indicates that it differentiates fairly well between the high ability and low ability students. The item difficulty parameter shows that the item is a bit difficult and the value of the guessing parameter is very low. This item can be classified as poor test item.</td>
</tr>
<tr>
<td>Item 32

The diagram below shows a sector of a circle with center O. Use it to answer question 32.


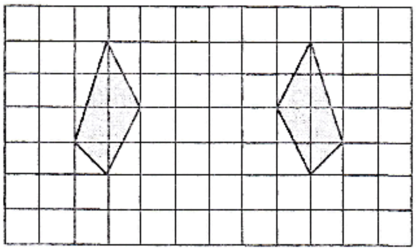
Calculate the length of the major arc. Take π as 3.14
a) 12.56 cm
b) 25.12 cm
c) 50.24 cm
d) 251.2 cm
Answer is C</td>
<td>

(a=2.30,b=1.13,c=0.25)
This item discriminates very well between the high and low ability examinees, but it is a bit difficult with a b-value of 1.13, greater than +1. The ICC curve is shifted to the right, and the c-value is of 0.25. This can be classified as fairly good test item.</td>
</tr>
<tr>
<td>Item 33

Which transformation is shown in the figure below?</td>
<td>

(a=0.19,b=0.84, c=0.33)
This item has a high guessing value of 0.33, and has a low value of the</td>
</tr>
</table>

<table>
<tr>
<td>



a) Enlargement
b) Translation
c) Reflection
d) Rotation

Answer is C

</td>
<td>

discrimination value a=0.19, with a medium difficulty level for the examinees. This is a poor item because the item cannot discriminate between the examinees.

</td>
</tr>
<tr>
<td>

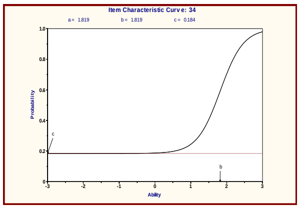
Item 34

Katso invests P1000 at a compound interest of 10% per annum. Calculate the total amount of money he will have after two years.

a) P800
b) P1020
c) P1200
d) P1210

Answer is D

</td>
<td>



(a=1.82,b=1.82,c=0.18)
This item is too difficult because of the value of b which is 1.82, the ICC curve has shifted to the right, although it can discriminate well among the ability groups and the guessing value is low. This item can be classified as poor item.

</td>
</tr>
<tr>
<td>

Item 35

What is the probability of selecting a spade from a pack of 52 playing cards

a) $\frac{26}{52}$

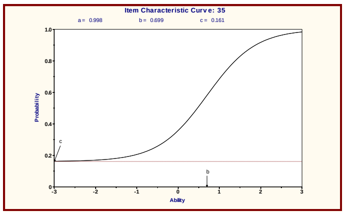
b) $\frac{13}{52}$

c) $\frac{10}{52}$

d) $\frac{1}{52}$

Answer is B

</td>
<td>



(a=0.70,b=1.00,c=0.16
Although the item discrimination value is less than +1, and for any item to be good, it must have a value greater than 1.The item difficulty value is +1 and the value of the guessing parameter is 0.16 which is less than 0.25. This is a fairly good item.

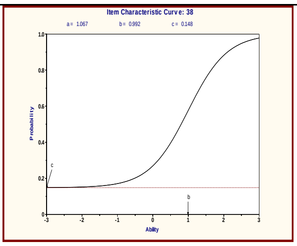
</td>
</tr>
</table>

Item 38

A fridge can be bought cash for P2400. It can also be bought on hire purchase by paying a deposit of P400 followed by 24 monthly payments of P150 each. Calculate the difference between the hire purchase price and the cash price.

a) P1200
b) P1600
c) P1850
d) P2950

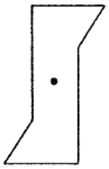
Answer is B



(a=0.99,b=1.07,c=0.15)
This item is a fairly good item because the item discrimination value of 0.99 is very close to +1.00 indicates that it differentiates to an extent between the high ability and low ability students. The item difficulty parameter shows that the item is a bit difficulty and the value of the guessing value of 0.15 is lower than 0.25.

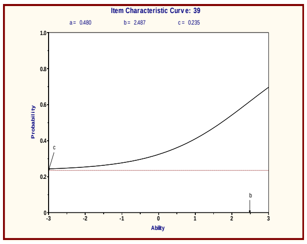
Item 39

What is the order of rotational symmetry of the shape given below?



a) 0
b) 1
c) 2
d) 3

Answer is B



(a=2.49,b=0.48,c=0.23)
This item discriminates very well between the high and low ability groups of examinees, it also has a medium difficulty level and a low guessing value. The ICC curve shifts to the left, which makes the item a bit easy. This is a fairly good item.

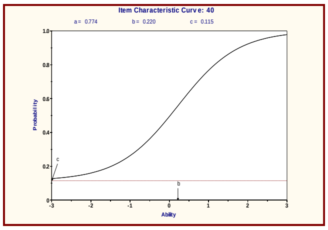
Item 40

The table below shows the names of four chiefs and the years in which each one of them became a chief.

Name of chief	**Year of becoming a chief**
Dintho	1918
Tsie	1924
Thotobolo	1938
Tau	1974

Who became chief during a leap year
a) Tau
b) Tsie
c) Dintsho
d) Thotobolo

Answer is B



(a=0.22,b=0.77,c=0.11)
Although the item difficulty parameter shows that the item is of medium difficulty and the value of the guessing parameter is very low, this item cannot discriminate among the ability groups of examinees and it can be classify as a poor item.

CONCLUSIONS

The findings from this study indicated that out of the twenty three (23) items that fitted the IRT model, twelve (12) items were classified as poor test items, ten (10) items were classified as fairly good test items which can be revised or improved and only one (1) item was considered to be good test item. This shows that test developers or examining bodies, especially in Africa should be concerned about the quality of test items and how examinees respond to them when constructing tests. Nenty (2004) emphasized that, in educational practice, one of the principal tasks is the development of tests that measure the facets of learning with the greatest precision and accuracy, and this is associated with the quality of test items. The growth in psychometrics, and computer adaptive testing in particular, have supported the growing interest in the use of IRT (Embretson and Reise, 2000). According to Hays *et al.* (2000), IRT has a number of potential advantages over CTT in assessing learning, in developing better measures and in assessing change over time. Its models yield invariant item and latent trait estimates.IRT psychometric methodologies have been used to solve assessment challenges as identified by Aiken (2003), Cook *et al.* (2003). The use of IRT to identify differential item functioning (DIF) or to distinguish between bias and real differences in an ability or trait among groups was also addressed by Fayers and Machin (2000) and Hahn and Cella (2003). Educational tests are a main source of information about student achievement in schools and in the context of large-scale testing the analysis of test data is essential in determining the quality of the test and the information the test generates. The worth of any educational assessment endeavor depends on the instruments i.e. the tools and techniques used, if these instruments are poorly designed, the assessment can be a waste of time and money.

RECOMMENDATIONS

It is therefore recommended that examination bodies especially in Africa should consider improving the quality of their test items by conducting an item analysis using IRT psychometric analysis to examine the quality of their constructed test items for validation purposes. It is now time for educational measurement experts in Africa to rise to the challenges pose by the measurement community and be fully aware of the usefulness of IRT constructing quality test items for examination purposes. There must be a paradigm shift from CTT to IRT for the construction and analysis of items in a test or examination especially public examinations in Africa.

REFERENCES

Aiken, L.R., 2003. Psychological testing and assessment 11th Edn., Boston: Allen & Bacon.

Cook, K.F., P.O. Monahan and C.A. McHorney, 2003. Delicate balance between theory and practice: Health status assessment and item response theory. Medical Care, 41(5): 571-571.

Embretson, S.E., 1983. Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 83: 179-197.

Embretson, S.E., 1994. Application of cognitive design systems to test development. New York: Plenum Press.

Embretson, S.E., 1995. A measurement model for linking individual learning to process and knowledge: Application to mathematical reasoning. Journal of Educational Measurement, 32: 277-294.

Embretson, S.E., 1998. A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods, 3: 300-396.

Embretson, S.E., 1999. Generating items during testing: Psychometric issues and models. Psychometrika, 64: 407-433.

Embretson, S.E. and S.P. Reise, 2000. Item response theory for psychologists. Mahwal, NJ:Erlbaum.

Fayers, P.M. and D. Machin, 2000. Quality of life: Assessment, analysis, and interpretation

West Sussex, England:Wiley.

Hahn, E.A. and D. Cella, 2003. Health outcomes assessment in vulnerable populations: Measurement challenges and recommendations. Archives of physical medicine and rehabilitation, 84(suppl.2), s35-s42.

Hambleton, R.K., H. Swaminathan and H.J. Rogers, 1991. Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Hays, R.D., L.S. Morales and S.P. Reise, 2000. Item response theory and health outcomes measurement in the 21st century. . Medical Care, 38 (9, SUPPL.2), II28-II42.

Klerk, G., 2008. Classical test theory (ctt)

Mislevy, R.J. and R.D. Bock, 1990. Bilog 3. 2nd Edn.: Mooresville IN .Scientific Sofware.

Nenty, H.J., 2004. The application of item response theory in strengthening assessment’s role on the implementation of national education policy

Thissen, D., 1991. Multilog: Multiple category item analysis and test scoring using item response theory [computer software]. Chicago: Scientific Software International.

Yu, C., 2008. True score model and item response theory.

Zimowski, M.F., E. Muraki, R.J. Mislevy and R.D. Bock, 1996. Bilog-mg: Multiple-group irt analysis and test maintenance for binary items [computer software]. Chicago: Scientific Software International.