Journal of Emerging Trends in Educational Research and Policy Studies (JETERAPS) 12(6): 253-259

scholarlinkinstitute.org/jeteraps

Post-Test Estimates of Item Parameters of Mathematics Multiple-Choice Test of a Public Examination in Nigeria

Dibu Ojerinde and Taiwo Oluwafemi Ajeigbe
Department of Educational Foundations and Counselling
Faculty of Education
Obafemi Awolowo University, Ile-Ife, Nigeria.
Corresponding Author: Dibu Ojerinde

Abstract
Post-test estimates of item parameters of multiple-choice tests are usually ignored in the assessment industry, except when it is necessary for research purposes. The study estimated item parameters using responses of candidates in mathematics multiple-choice items for three years. This is a longitudinal study which adopted the ex-post facto research design. The population comprised all the candidates who sat for mathematics examination during the years under consideration. Sample consisted of 10% of those candidates. Candidates' result in the Senior School Certificate examination conducted by a public examination body constituted data for the study. Data was analysed using Factor Analysis and X-Caliber IRT software. The results of the study showed variations in the item parameter estimates. Also, some of the items calibrated for three years showed avoidable clerical, concept and construct errors which could only be detected using statistical estimate of the parameters. The study therefore concluded that manual development of marking scheme for multiple-choice items is neither sufficient nor efficient enough to ignore statistical estimates. It is recommended that before scoring, statistical post-test estimates of multiple-choice test parameters ought to be carried out to ascertain the characteristics of the test items.

Keywords: Post-test, Item Parameters, Mathematics, Multiple-Choice Tests, Examination

INTRODUCTION

Examination procedure follows item generation/development, pre-test, administration, scoring and reporting. Experts from different subject specialisations are employed at every stage to enhance quality of the examination items. One major area of omission in the procedure of testing which may strengthen and authenticate the validity of scores usually reported by examinations bodies in Nigeria is post-test estimates of item parameters. The extent to which the parameter estimates of pre-test and post-test are comparable is yet to be empirically ascertained. This may suggest that examination bodies in Nigeria rely on the results of the pre-test to establish the psychometric characteristics of the items, even after test administration. The tendencies to introduce error in in students' scores if post-test estimate is not carried out is eminent which may consequently result into misinformation, as well as passing wrong judgement about students' ability. The study is apt in preventing or correcting error that could be avoidable.

A good innovation in a good direction in the field of psychometrics is that attention should be focused on how individual item behave in a real examination condition/situation. When items parameters are estimated after the examination, it is very likely that such items will behave in a similar situation over time will be guaranteed. This may also strengthen the item bank of the examination body. At the end of conduct of any examination, marking schemes are usually moderated, especially in essay items to ascertain acceptable responses for uniform award of marks. The use of this procedure for mathematics multiple-choice items is the focus of this study.

Mathematics is one of the major subjects that candidates should pass to enable them study any science-related discipline in tertiary institutions of learning. Its importance cannot be overstressed as the basic principles of mathematics are imperative to technology advancement and development. To this end, test developers need to pay adequate attention to the items used in assessing students' knowledge of the concept of mathematics to ensure that reliable and valid decisions emanate from such assessment. In otherwords, to screen the items for possible errors, a modern theory (item response theory) can be employed.

The ultimate goal of this paper is to serve as an eye opener for examination bodies and test developer on the actual estimate of item parameters of students' responses to items and also provide feedback on the item quality, as well as quality assurance before students' grades are released for public consumption. Most of the past researches do focus on parameter

estimates without calibrating for key problems most especially after the examination might have been taken place. Hence, this study aim to provide empirical evidence by analysing the parameter estimates of multiple-choice mathematics items and identifying items that flay key (i.e key problem), with a view to establishing the need for key calibration before making students' scripts to avoid measurement error(s)

Thus, IRT models are mathematical models that permit prediction of examinee's test performance from an individual's standing on an attribute or trait and the characteristics of the items that make up a test [4]. The application of the theory as a modern statistical screening tool requires that the assumptions of unidimensionality, local independence and Item Characteristic Curve (ICC) need to be established. Many researchers have advocated for the use of IRT for item analysis in recent time.

The theory is a mathematical model of three facets namely; 1-parameter model (1PL), 2-parameter Logistic Model (2PL) and 3-parameter Logistic Model (3PL). The Three parameter model (3PL) has *a*, *b*, and *c* parameters where *a*, discrimination, *b*,difficulty and *c*, guessing are prominent characteristics of the model. When *c* is zero (0), we refer to 2-PL; 'c'=0, and 'a'=0, the 1PL stands. However, for the purpose of this study, the 3-PL was adopted. Theoretically, *c* ranges from 0.0 to 1.0, but is typically < 0.3 for 4 or 5 option lengths according to [6].This model is considered appropriate to guarantee necessary improvement in the item parameter estimates [1].The 3-parameter is presented in the equation proposed by Lord [5].

Where: Pi(Θ) = the probability that an examinee with ability level Θ answers item correctly; *b* = the item difficulty parameter, *a* = the item discrimination

$$Pi(\Theta) = ci + \frac{1- ci}{1+ \exp [-1.7a (\Theta-b)]} \quad (i)$$

parameter, 1.7= scaling factor (D) and *c*= the lower asymptote parameter. The ICC is presented in Figure 1.

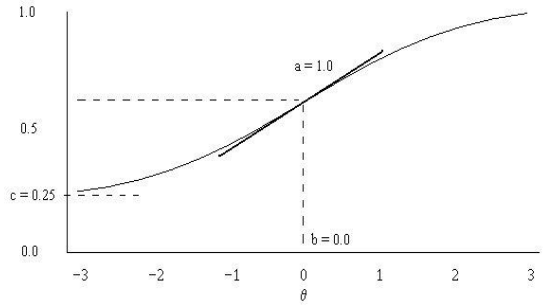

Figure 1: ICC for a 3PL Model

From the above, many questions readily come to mind such as: Are the Multiple-choice mathematics questions administered to the target population compliant with the pre-test item parameters? Is there any proof that the item parameters of the final examination taken by the target population is still acceptable? In fact, one may ask if the avoidable errors during the test development were finally rectified in the final version of the test?

Determination of final marking scheme is usually done manually. Is there a guarantee for the usability of item parameters of the final examination grades? The development of marking scheme in Mathematics can be very challenging. Because of its importance, one can ask a number of questions in respect of mathematics' post-test estimated parameters. These concerns should be statistically handled at the post-test level before scoring the test. In this study, the pre-test estimated parameters were not available for easy comparison in this study.

However, the following questions can be asked in respect of mathematics post-test estimated parameters.

Mathematics as a Subject

(1) Are the mathematics multiple-choice items unidimensional in nature? In otherwords, is there any evidence that the administered test items were unidimensional?
(2) What are the item parameter estimates (a, b, and c) of the multiple-choice items? What are the final item parameter estimates (a, b, c) of the administered test?
(3) How many of the multiple-choice items have key problems?

The purpose of the study therefore is restricted to the three questions above. However, the study is clear departure from other studies that are restricted to pre-test analysis of items before the actual administration to the tests. As such, it will provide scientific evidence on the true parameter estimates of the items and afford the opportunity to make comparison.

METHODOLOGY

The calibration procedure was carried out using X-Calibre 4.2 that generated coefficients for the a, b, and c parameters adopting the three parameter logistic model. The study adopted the ex-post facto research design. The population comprised candidates who sat for the June/July Mathematics Senior School Certificate Examination (SSCE) conducted by a public examination body. Paper I during years 2011, 2012 and 2013 examinations in Osun State. These consisted of 35,792 candidates (male-18487, female-17305) for 2011, 23,936 candidate (male- 12293, female- 11,643) for 2012 and 24,754 candidates (male-12,390, female-12,364). Three different sample sizes were randomly selected, using proportional sampling procedure 10% of each

population cohort of three years. As a result, the sample used for the study consisted of 3,579 candidates (male- female); 2,394 candidates (male female) and 2,475 candidates (male female) for the three years respectively.

Three instruments were adopted for the study. These were responses of the candidates to 60 multiple-choice Mathematics items Paper I of 2011, 2012 and 2013 from a Public Examination in Nigeria. The instruments covered contents areas such as; Algebra, Arithmetic, Geometry, Trigonometry, and Statistics. The data collected for the study were generated from the data base of the examination body with respect to Osun State.

It was assumed that before the test was administered to the candidates, the items were trial-tested and found suitable in terms of coverage, construct and usability by the examination body.The keys adopted for scoring and calibrating the 60 multiple-choice mathematics items were obtained from the matrix scores for each candidate as contained in the database made available for the researchers by the examination body. The scoring was dichotomously done. Hence, the possible maximum and minimum obtainable raw scores were sixty (60) and zero (0) respectively. The scores were subjected to factor analysis as well as item analysis using *X*-Caliber software.

RESULTS

Research Question One: Are the mathematics multiple-choice items of 2011, 2012, and 2013 unidimensional in nature?

One of the major assumptions in the use item response theory is unidimensionality. In this case, it is primarily important to test the assumption to ascertain whether the mathematics multiple-choice items is measuring a distinct trait or construct in mathematics across the three years under consideration.

As such, to answer this question, method of scree plot was used. The theory behind the acceptance or otherwise of the scree are that:

1. The eigenvalue of the first factor must be greater than one (1)
2. The value of the eigenvalue of the first factor must be at least twice the value of the second factor.
3. The Cronbach alpha, reliability, of each item (i.e. relationship between the eigenvalue and factors) must be minimum of 0.7.

These are the benchmark values for estimation of unidimensionality of tests according to Lord, and Novick [6] and Cook, Kallen and Amtmann[2].

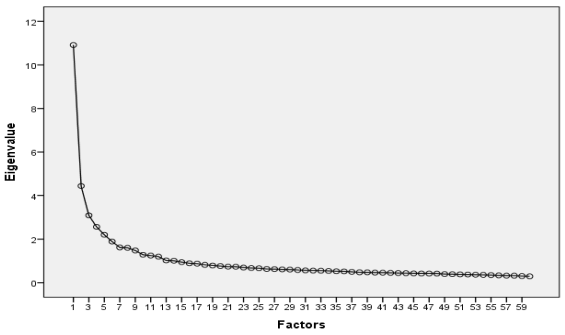

Fig.1a.: Scree Plot for 2011 Mathamatics Multiple-Choice Items

coefficients (F_1& F_2) = 10.07: 3.44
σ =0.90

Coefficients (F_1& F_2) = 9.09: 4.32

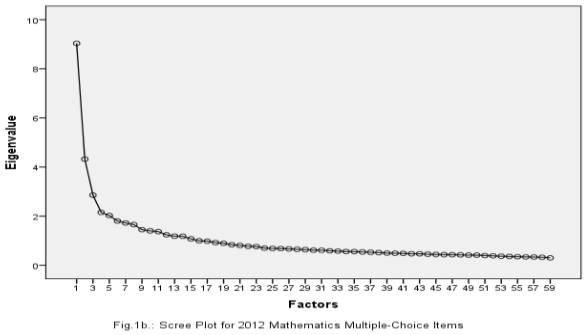

Fig.1b.: Scree Plot for 2012 Mathematics Multiple-Choice Items

σ =0.82

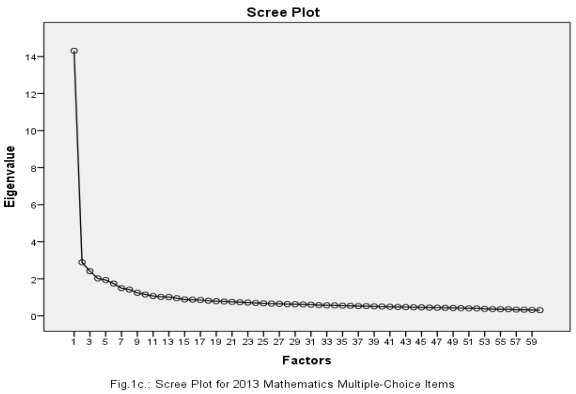

Fig.1c.: Scree Plot for 2013 Mathematics Multiple-Choice Items

Coefficients (F_1& F_2) = 14.31: 2.83
σ =0.94

The results in Figure 1a, 1b, and 1c showed 14, 15, and 13 factors with 59.26%, 59.36%, and 56.19% variance explained across the three years respectively. A cursory look at the three scree plots for the three years show some differences in the degree of unidimensionality. The plots obviously passed acceptable standard. However, the degree of pass varied from 2013 through 2011, and 2012 in order of merit. The order of merit demonstrated that there are improvement in the test development of the public examination in Nigeria.

Research Question Two: What are the item parameter estimates (a, b, and c) of the mathematics multiple-choice items during the years under consideration?

Measures of item parameter estimate is gamine to this study to provide statistical evidence to establish how individual item behave across different test takers and ascertain the stability of the parameter estimates across the three years under consideration. Hence, to provide answer this question, the item parameter estimates were obtained from the calibrated 60 Mathematics multiple-choice items for 2011, 2012 and 2013 using 3-parameter model of item response theory, using *X*-Caliber software. The results produced item parameters (*a*, *b*, and *c*) as shown in Table 1 and 2 respectively.

Table 1: Item Parameters for 2011, 2012 and 2013 Mathematics Multiple-choice Calibrated Items

Items	**2011**			**2012**			**2013**		
	a	b	c	a	b	c	a	b	c
1	3.17	3.58	0.04	3.62	2.11	0.08	3.66	2.89	0.06
2	0.73	-2.08	**0.25***	3.80	3.02	0.12	3.79	3.30	0.05
3	1.80	3.46	0.06	0.98	-0.75	**0.22***	2.37	3.30	0.05
4	0.48	1.40	**0.37***	1.73	3.35	**0.27***	1.90	3.30	0.08
5	2.32	2.76	0.08	1.12	-0.68	**0.22***	1.59	2.90	**0.35***
6	1.68	3.28	0.10	0.64	0.25	**0.23***	1.76	3.30	0.07
7	1.74	3.35	0.08	1.65	3.36	0.12	.89	-1.31	**0.25***
8	1.67	3.49	0.08	3.94	1.45	0.11	1.74	3.26	0.07
9	1.76	3.35	0.05	1.67	3.32	0.17	1.08	-0.73	**0.24***
10	1.72	3.36	0.08	1.23	-0.46	0.20	1.64	3.20	0.09
11	2.93	1.85	0.07	1.33	-0.22	**0.22***	1.13	-0.60	**0.23***
12	1.76	3.32	0.06	1.20	-0.92	**0.24***	1.73	3.27	0.09
13	0.61	-1.79	**0.25***	1.68	3.28	0.11	0.57	-1.29	**0.25***
14	1.70	3.31	0.08	1.71	2.72	0.17	1.74	3.27	0.09
15	1.68	3.47	0.06	1.67	3.28	0.09	1.73	3.26	0.07
16	1.65	3.26	0.08	0.45	1.61	**0.22***	1.70	3.24	0.10
17	1.65	3.42	0.08	0.86	-0.12	**0.22***	1.47	3.00	0.17
18	1.71	3.32	0.07	1.70	3.21	0.10	1.83	3.04	0.07
19	2.47	2.61	0.07	1.66	3.30	0.09	1.81	3.03	0.08
20	0.74	-0.86	**0.23***	2.88	0.94	**0.25***	1.39	-0.70	**0.24***
21	2.46	2.68	0.07	1.67	3.24	0.09	1.81	2.99	0.08
22	2.47	2.63	0.06	1.77	2.74	0.09	1.79	2.94	0.15
23	2.17	2.77	0.06	3.76	1.25	**0.29***	1.14	-0.91	**0.24***
24	2.30	2.66	0.06	0.97	-0.89	**0.26***	1.84	3.02	0.07
25	2.29	2.63	0.10	0.90	-0.92	**0.28***	1.77	3.03	0.09
26	2.07	2.73	0.16	0.77	0.58	**0.32***	1.64	2.89	0.17
27	2.08	2.91	0.07	1.29	-1.04	**0.24***	1.72	2.98	0.10
28	0.77	-1.88	**0.25***	1.92	2.50	0.11	1.85	3.04	0.06
29	1.19	0.02	**0.25***	1.64	3.32	0.13	1.74	2.98	0.14
30	1.01	-1.22	**0.25***	0.20	1.08	**0.26***	.71	-1.73	**0.25***
31	1.48	-0.09	**0.26***	1.04	-0.52	0.20	2.01	2.82	0.09
32	0.68	-1.22	**0.24***	1.57	3.35	**0.35**	2.04	2.61	0.08
33	1.85	3.10	0.18	1.67	3.34	0.10	2.03	2.71	0.06
34	0.52	-1.66	**0.25***	1.67	3.36	0.11	1.94	2.76	0.17
35	1.65	3.58	0.07	4.29	1.16	0.13	2.04	2.63	0.08
36	1.65	3.58	0.07	1.68	3.34	0.09	1.77	2.78	0.12
37	1.62	3.06	**0.37***	1.46	2.74	0.14	1.35	-.61	**0.23***
38	0.88	-1.61	**0.23***	1.65	3.32	0.12	2.02	2.75	0.09
39	1.62	3.58	0.06	1.67	3.33	0.08	1.88	2.79	0.16
40	1.62	3.58	0.12	1.58	3.24	0.18	2.00	2.77	0.10
41	1.65	3.58	0.06	1.71	3.04	0.07	2.02	2.66	0.07
42	1.50	3.58	**0.30***	1.59	3.26	0.16	1.85	2.95	0.11
43	1.54	3.57	0.09	1.70	2.73	0.13	1.86	2.92	0.10
44	1.59	3.58	0.09	0.80	-0.20	**0.21***	1.55	-0.43	**0.22***
45	0.52	-0.42	0.24	0.91	-0.93	**0.24***	1.74	2.92	0.10
46	1.60	3.58	0.15	1.67	3.33	0.08	1.70	3.17	0.09
47	1.60	3.58	0.17	1.63	3.30	0.14	0.69	1.86	**0.27***
48	1.61	3.58	0.07	.47	1.80	**0.22***	1.82	2.59	0.10
49	0.43	-0.62	**0.25***	1.63	3.28	0.16	1.86	2.91	0.10
50	1.61	3.58	0.06	1.63	3.30	0.11	1.86	3.02	0.11
51	1.09	-1.97	**0.24***	1.70	3.36	0.06	1.75	2.64	0.14
52	1.63	3.58	0.06	1.65	3.20	0.13	1.91	2.87	0.06
53	0.89	-1.59	**0.24***	1.70	3.36	0.06	1.69	3.22	0.12
54	2.26	1.59	0.10	1.65	2.88	0.10	1.90	2.84	0.07
55	1.61	3.58	0.07	3.02	1.24	0.13	1.86	2.87	0.13
56	1.60	3.58	0.09	1.55	3.29	0.14	1.81	-.55	**0.24***
57	1.64	3.51	0.06	1.66	3.32	0.10	1.84	2.90	0.08
58	1.59	3.58	0.10	0.32	-0.70	**0.23***	1.84	2.95	0.08
59	0.37	-0.83	**0.25***	1.56	3.05	0.09	1.84	3.02	0.11
60	3.17	3.58	0.04	3.62	2.11	0.08	3.66	2.89	0.06

*** c > 0.2**

Table one shows the three parameter estimates (a, b, c) estimates of the calibrated 60 mathematics multiple-choice items for 2011, 2012 and 2013 respectively. The three parameter estimates is found useable here because the information capture in the data base of the examination body did not include the upper asymptote (carelessness), hence, the three parameter model is considered appropriate. It could be seen that 17 items (28%), 20 items (33%) and 12 items (20%) are with guessing parameter (c) greater than 0.2 (acceptable standard for 5-option length) in the three years respectively. The remaining two parameter estimates (a, b) are summarised in Table 2.

Table 2: Summary of Item Parameters for 2011, 2012 and 2013 Mathematics Multiple-Choice Calibrated Items

Years	b-parameter	N/% Difficulty	a-parameter	N/% Discrimination
2011	Easy (-3.00≤-1.00)	14 (23.33%)	Excellent (a≥1.70)	24 (40%)
	Moderate (-1.00≤1.00)	1 (1.67%)	Good (1.35 - 1.69)	22 (36.67%)
	Difficult (1.00≥2.00)	45 (75%)	Moderate (0.65 - 1.34)	9 (15%)
			Marginal (0.35 - 0.64)	5 (8.33%)
			Poor (0.1 - 0.34)	0 (0.00%)
2012	Easy (-3.00≤-1.00)	13 (21.67%)	Excellent (a≥1.70)	18 (30%)
	Moderate (-1.00≤1.00)	3 (5%)	Good (1.35 - 1.69)	24 (40%)
	Difficult (1.00≥2.00)	44 (73.33%)	Moderate (0.65 - 1.34)	11 (18.33%)
			Marginal (0.35 - 0.64)	5 (8.33%)
			Poor (0.1 - 0.34)	2 (3.33%)
2013	Easy (-3.00≤-1.00)	9 (15%)	Excellent (a≥1.70)	46 (76.76%)
	Moderate (-1.00≤1.00)	1 (1.67%)	Good (1.35 - 1.69)	8 (13.33%)
	Difficult (1.00≥2.00)	50 (83.33%)	Moderate (0.65 - 1.34)	5 (8.33%)
			Marginal (0.35 - 0.64)	1 (1.67%)
			Poor (0.1 - 0.34)	0 (0.00%)

Table 2 shows the summary of the item parameter estimates for the three years as proposed by Georgiev [3]. In terms of the difficulty (b-parameter) of the items, 45 (75%), 44 (73.33%), and 50 (83.33%) of the items were found difficult in 2011, 2012, and 2013 respectively. It also showed that 1 (1.67%), 3 (5%), 1 (1.67%) items were moderately difficult. In addition, 14 (23.33%), 13 (21.67%) and 9 (15%) of the items were found to be easy respectively for the three years. In terms of the discrimination (a-parameter) of the items, 24 (40%), 18 (30%), and 46 (76.76%) of the items were found to discriminate excellently in 2011, 2012, and 2013 respectively. It also showed that 22 (36.67%), 24(40%), 8(13.33%) items had good discrimination. In addition, 9 (15%), 11(18.33%) and 5(8.33%) of the items discriminated moderately across the three years respectively. Furthermore, 5 (8.33%), 5(8.33%), and 1(1.67%) items discriminated marginally across the three years under consideration. Finally, 2(3.33%) of the items in 2012 were found to discriminate poorly, while no item discriminated poorly in 2011 and 2013 respectively. From the results, it could be seen that most of the items were difficult as the percentage difficulty ranged from 73.33% to 83.33%, while percentage discrimination ranged from 40% to 76.76% across the three years respectively.

Research Question Three: How many of the mathematics multiple-choice items have key problems?To answer this question, the results in Table 1 showed items whose keys the candidates found it difficult to recognise as correct option. Table 3 shows the summary of those items across the three years under review.

Table 3: Item Parameters for 2011, 2012 and 2013 Mathematics Multiple-Choice Calibrated Items

Items	2011				2012				2013			
	P	R	b	Flag	P	R	b	Flag	P	R	b	Flag
1	.01	-.04	3.58	K, F, Hb	.05	.02	2.11	K, F	.03	.05	2.89	K, F
2	.92	.02	-2.08		.10	-.19	3.02	K, F, Hb	.01	.03	3.30	K, F, Hb
3	.04	.01	3.46	K, F, Hb	.74	.32	-.75		.02	.03	3.30	K, F, Hb
4	.52	.06	1.40		.26	-.04	3.35	K, F, Hb	.05	-.03	3.30	K, F, Hb
5	.06	.09	2.76	K, F	.74	.29	-.68		.36	.04	2.90	K, F
6	.09	.02	3.28	K, F, Hb	.55	.21	.25	F	.02	.03	3.30	K, F, Hb
7	.05	.04	3.35	K, F, Hb	.09	-.07	3.36	K, F, Hb	.85	-.05	-1.31	
8	.06	.05	3.49	K, F, Hb	.12	.07	1.45	K, F	.04	.09	3.26	K, F, Hb
9	.03	.07	3.35	K, F, Hb	.15	-.17	3.32	K, F, Hb	.77	-.14	-.73	
10	.06	-.02	3.36	K, F, Hb	.68	.44	-.46		.06	.11	3.20	K, F, Hb
11	.06	.08	1.85	F	.64	.39	-.22		.73	-.04	-.60	
12	.04	-.02	3.32	K, F, Hb	.80	.34	-.92		.06	.07	3.27	K, F, Hb
13	.87	.05	-1.79		.08	-.13	3.28	K, F, Hb	.79	-.02	-1.29	
14	.06	-.03	3.31	K, F, Hb	.16	-.13	2.72	K, F	.06	.11	3.27	K, F, Hb

15	.04	.04	3.47	K, F, Hb	.05	.00	3.28	K, F, Hb	.04	.15	3.26	K, F, Hb
16	.06	.04	3.26	K, F, Hb	.40	.20	1.61	F	.07	.04	3.24	K, F, Hb
17	.06	.02	3.42	K, F, Hb	.61	.20	-.12		.16	.11	3.00	K
18	.05	.02	3.32	K, F, Hb	.07	-.04	3.21	K, F, Hb	.04	.17	3.04	K, F, Hb
19	.05	.02	2.61	K, F	.06	-.06	3.30	K, F, Hb	.05	.12	3.03	K, F, Hb
20	.76	-.03	-.86		.35	.24	.94	F	.77	-.09	-.70	
21	.05	.06	2.68	K, F	.06	-.06	3.24	K, F, Hb	.06	.16	2.99	K, F
22	.04	.07	2.63	K, F	.09	-.04	2.74	K, F	.13	.11	2.94	K, F
23	.04	.01	2.77	K, F	.33	.02	1.25	K, F	.80	-.09	-.91	
24	.04	.09	2.66	K, F	.78	.28	-.89		.03	.10	3.02	K, F, Hb
25	.08	.04	2.63	K, F	.79	.24	-.92		.06	.05	3.03	K, F, Hb
26	.15	.10	2.73	K	.55	.17	.58	F	.16	.18	2.89	K
27	.05	.03	2.91	K, F	.82	.35	-1.04		.07	.16	2.98	K
28	.90	.04	-1.88		.09	-.18	2.50	K, F	.02	.05	3.04	K, F, Hb
29	.61	.14	.02		.11	-.09	3.32	K, F, Hb	.12	.08	2.98	K, F
30	.85	.12	-1.22		.55	.01	1.08	F, La	.88	-.12	-1.73	
31	.64	.11	-.09		.69	.32	-.52	F	.06	.05	2.82	K, F
32	.81	.06	-1.22		.36	-.08	3.35	K, F, Hb	.05	.11	2.61	K, F
33	.17	.02	3.10	K, F, Hb	.07	-.09	3.34	K, F, Hb	.03	.15	2.71	K, F
34	.83	-.12	-1.66		.26	-.14	3.36	K, F, Hb	.16	.02	2.76	K, F
35	.04	-.04	3.58	K, F, Hb	.18	.19	1.16	F, Ha	.05	.17	2.63	K, F
36	.05	-.04	3.58	K, F, Hb	.06	-.12	3.34	K, F, Hb	.09	.15	2.78	K, F
37	.37	.01	3.06	K, Hb	.13	.08	2.74	K	.75	-.12	-.61	
38	.89	.07	-1.61		.10	-.03	3.32	K, F, Hb	.06	.12	2.75	K, F
39	.04	.07	3.58	K, F, Hb	.04	-.04	3.33	K, F, Hb	.15	.08	2.79	K, F
40	.10	.11	3.58	K, F, Hb	.17	-.04	3.24	K, Hb	.07	.08	2.77	K, F
41	.03	-.06	3.58	K, F, Hb	.04	-.01	3.04	K, F, Hb	.04	.07	2.66	K, F
42	.29	-.02	3.58	K, Hb	.14	-.03	3.26	K, F, Hb	.08	.09	2.95	K, F
43	.06	.02	3.57	K, F, Hb	.12	-.04	2.73	K, F	.07	.15	2.92	K, F
44	.07	.08	3.58	K, F, Hb	.61	.26	-.20		.71	-.15	-.43	F
45	.67	.03	-.42		.78	.32	-.93		.07	.11	2.92	K, F
46	.14	.06	3.58	K, F, Hb	.04	-.05	3.33	K, F, Hb	.07	.13	3.17	K, F, Hb
47	.16	.01	3.58	K, F, Hb	.20	-.02	3.30	K, F, Hb	.35	-.01	1.86	K
48	.05	.00	3.58	K, F, Hb	.37	.16	1.80	F	.08	.06	2.59	K, F
49	.69	-.02	-.62		.14	-.15	3.28	K, F, Hb	.07	.04	2.91	K, F
50	.03	.01	3.58	K, F, Hb	.08	-.07	3.30	K, F, Hb	.08	.05	3.02	K, F, Hb
51	.94	.11	-1.97		.02	-.07	3.36	K, F, Hb	.13	.15	2.64	K, F
52	.03	.04	3.58	K, F, Hb	.11	-.14	3.20	K, F, Hb	.03	.10	2.87	K, F
52	.89	.09	-1.59		.02	-.06	3.36	K, F, Hb	.09	.03	3.22	K, F, Hb
54	.12	.13	1.59	F	.08	-.09	2.88	K, F	.04	.13	2.84	K, F
55	.04	.01	3.58	K, F, Hb	.17	.18	1.24	F	.11	.06	2.87	K, F
56	.07	-.06	3.58	K, F, Hb	.12	-.05	3.29	K, F, Hb	.75	-.09	-.55	
57	.04	.09	3.51	K, F, Hb	.07	-.13	3.32	K, F, Hb	.05	.09	2.90	K, F
58	.08	-.02	3.58	K, F, Hb	.67	.12	-.70	F	.04	.10	2.95	K, F
59	.71	.05	-.83		.06	-.06	3.05	K, F, Hb	.08	.08	3.02	K, F, Hb
60	.71	.05	-.83		.06	-.06	3.05	K, F, Hb	.08	.08	3.02	K, F, Hb

Results in Table 3 show that most of the items flagged K are those that have key problems. The results show that 19 items (32%), 21 items (35%) and 9 items (15%) are completely free of being flagged in 2011, 2012, and 2013 respectively. High b-parameter is ≥ 3.00, denoted by high b-parameter estimate (H_b) and this is however prominent for most of the items, and items with F flag show that the items fit statistic (z Resid for dichotomous/chi-square for polytomous) was significant, only that the items did not fit the 3-parameter of IRT model.

These are the benchmark values for estimation of unidimensionality of tests according to Lordand Novick[6] and Cook, Kallen and Amtmann[2].

DISCUSSION

The study focused on the parameter estimates of 2011, 2012, and 2013 Mathematics multiple-choice items respectively. The results obtained from research question one of the study showed that the mathematics multiple-choice items comply with the assumption of unidimensionality. A single factor was found to be clearly distinct, as the eigenvalue of the first factor was more than twice the second factor, as shown in the scree plots for the three years under review. The scree plot of mathematics 2013 was found to be outstanding when compare to that of 2011 and 2012 respectively.

This result is in congruence with the finding of the study conducted by [2]. Also, the second research question revealed most of the mathematics multiple-choice items were found to be very difficult based on the classification of [3] across the three years under review. Many reasons may be adduced for the difficult items, such as content ambiguity, keys similarity, clerical errors and so on. Similarly, most of the mathematics multiple-choice items were found to discriminate excellently. A very wide gap was noticed between brilliant and poor candidates across the three years. This may be traceable to the fact that those items were extremely difficult for the candidates during the years under consideration. In

the same vein, the research question three revealed that more than half of the mathematics multiple-choice items had key problems. This calls for urgent attention so that candidates who have studied day and night are not frustrated and subsequently denied their deserved marks. A greater proportion of the mathematics multiple-choice items were found to be of high difficulty level, but had good discrimination across the three years under consideration

CONCLUSION

The study therefore concluded that manual development of marking scheme for multiple-choice items is neither sufficient nor efficient enough to ignore statistical estimates. It is recommended that before scoring, statistical post-test estimates of multiple-choice test be carried out to ascertain the parameters of the test items. Examination bodies should endeavour to carry out post-item-analysis using reasonable sample size from the original candidates who participated in the examination. This will enable them to identify problematic items before grading of candidates' responses. As such, candidates would not be unduly favoured or disfavoured because of errors never committed during the examination. Examination bodies should try to make use of items with moderate level of difficulty to ensure that all the candidates are adequately put on the scale.

REFERENCES

[1]. Baker, F. (2001).The basics of item response theory (nd). (2nd ed.) ERIC Clearinghouse on assessment and evaluation. MD: College Park.

[2]. Cook, K. F., Kallen, M. A. and Amtmann, D. (2009). Having a Fit: Impact of Number ofItems and Distribution of Data on Traditional Criteria for Assessing IRT's Unidimensionality Assumption. Published Online 2009 March 18. Doi: 10.1007/s11136-009-9464-4.

[3]. Geogiev, N. (2008). Item Analysis of C, D, and E Series from Raven's Standard Progressive Matrice with Item Response Theory Two Parameter Model. Europe's Journal of Psychology, 4(3).

[4]. Hambleton, R. K., andSwaminathan, H. (1985). Item Response Theory: Principles andApplications. Boston: Kluwer-Nijhoff.

[5]. Lord, F. (1952). A Theory of Test Scores (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf

[6]. Lord, F. M. and Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

[7]. Ojerinde, D, Popoola, K., Ojo, F., and Onyeneho, P. (2012). Introduction to Item Response Theory (2nd ed.). Lagos (Nigeria): Goshen Print Media Limited.