British Journal of Psychology Research

Vol.4, No.1, pp.1-32, May 2016

Published by European Centre for Research Training and Development UK (www.eajournals.org)

ITEM RESPONSE THEORY VALIDATION OF ADVANCED PROGRESSIVE MATRICES IN NIGERIA

Kpolovie, P. J[*] and Emekene C. O

Academic Planning, Research & Control Unit; Vice-Chancellor's Office; University of Port Harcourt; PMB 5323; Choba; Port Harcourt; Nigeria.

ABSTRACT: *Raven's Advanced Progressive Matrices (APM) is a leading global non-verbal mental ability test for identification of individuals with clear thinking skills who can handle rigorous study programmes and cope with complexity and ambiguity of the contemporary workplace. The test is popularly used in America, Europe and Asia but has never been validated for use in Nigeria. A validation sample of 2100 in Nigeria was randomly drawn for this study. Triangulation research design, adopting Item Response Theory (IRT), guided this validation of APM. Results revealed that all items of the test yield favourable statistics under 3-Parameter Logistic IRT Model with regards to discrimination, difficulty and guessing. Item Response Function showed preponderance of APM's reliability (0.948) and construct cum concurrent validity (0.701) with Culture Fair Intelligence Test (CFIT). X-Calibre analysis confirmed suitable difficulty indexes (-2.595 to 2.133 b parameter) of APM. The APM is bias-free and very suitable for use in Nigeria.*

KEYWORDS: Advanced Progressive Matrices, Item Response Theory, X-Calibre, Test Bias, 3-Parameter Logistic Model, CFIT, Nigeria

INTRODUCTION

Psychometrics is the field of study concerned with the theories and techniques of psychological measurement of knowledge, abilities, attitudes, personality traits, academic achievement, and educational attainment. While the core theories include Classical Test Theory (CTT), Item Response Theory (IRT), and theories of various constructs; the central techniques include test development, validation and standardization as well as statistical analysis and programme evaluation (Kpolovie, 2016; 2011; 2014). Psychometrics can be defined as the branch of psychology concerned with the design and use of psychological tests and the application of statistical and mathematical techniques to psychological testing (Michell, 1999) and programme evaluation (Kpolovie, 2012).

Raven's Progressive Matrices tests developed by John C. Raven in 1936 are examples of psychological testing tools. Raven's tests exist in three different forms that are progressively more difficult in contents and are intended for different populations (Verguts and De Boeck, 2002). They are the Standard Progressive Matrices (SPM), the Coloured Progressive Matrices (CPM) and the Advanced Progressive Matrices (APM). The APM scale is the most difficult of the three and it is the main instrument of this study. APM test is a leading global non-verbal measure of mental ability, helping to identify individuals with advanced observation, high-level imagination including the domain of duty and clear thinking skills who can handle rigorous study programmes as well as the complexity and ambiguity of the modern workplace. APM test offers information about someone's capacity for analysing and solving problems, abstract reasoning, logical reasoning, quick recognition of differences and similarities, intellectual capacity and the ability to learn (Raven & Raven, 2008). The APM assesses the

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

ability or capacity to detect a certain order or structure in a chaos or chaotic situation and the ability to find meaning of apparently randomly compiled elements (Raven, Raven & Court, 1998; Pearson, 2011).

The APM reduces cultural biases with a nonverbal approach. The APM was developed to reduce cultural biases in a manner similar to Culture Fair Intelligence Test that was developed by Cattell and Cattell in 1963 to measure fluid intelligence in accordance with the Theory of Fluid (gf) and Crystalized Intelligence (gc) by Cattell (1963). The APM is said to be very suitable for individuals whose native language is not English (Bors & Stokes, 1998). The test when administered untimed, differentiates between people at the high end of intellectual ability (Chiesi, Ciancaleoni, Galli, & Primi, 2012, March 26).

When administered under timed conditions, the APM can be used to assess intellectual efficiency - quick and accurate high-level intellectual work and the ability to be sharp and quick at decision making. Items on all forms ask the examinee to identify the missing component in a series of figural patterns. Grouped in sets, the items graduates in the difficulty index from very easy items to very difficult items (Raven, 1962; Raven, Raven, & Court, 2003; Raven, Court & Raven, 1983; Sefcek, 2007). Therefore the items require increasingly greater skills in encoding, analysing, recognizing patterns and identifying the right answers (Van de Ark, 2010). The Raven’s APM produces a single raw score as well as percentile rank to indicate the candidate’s educative ability or the ability to make sense of complex situations, compared to a norm group (Raven, Raven and Court, 2012).

Presenting the report of a large survey conducted in nineteen European countries by several members of the International Test Commission (ITC) at the 12th European Congress of Psychology that held in Istanbul in the month of July 2011, Evers (2011) asserted that “the Raven's Matrices are in the fourth position among the ten most used tests in Europe.” The report further stated that “among them the Advanced Progressive Matrices are widely employed for assessing fluid ability in adolescents and adults.” Evers (2011) also reported that “the Raven’s Advanced Progressive Matrices (APM) scale has also been recommended as a useful measure for identifying academic potential.” Thus the APM is in high demand as an instrument of choice among researchers in America, Europe and Asia because of its utility value in psychological research works (Herrnstein & Murray, 1994). Meanwhile, the instrument is hardly known let alone effectively employed in psychological research works particularly Nigeria, and in some other African countries (Kpolovie, 2015; Rushton, Skuy and Bons, 2004).

The problem of this study can be categorized into three. First, there exists a dearth of well validated, standardized and normed instruments for psychological testing in Nigeria. According to Kpolovie (2012), this has severely affected decisions being made about people's capacity for analysing and solving problems, abstract and logical reasoning, quick recognition of differences and similarities, intellectual capacity and the ability to learn within and outside the education sector in the country (Ololube, Emejuru, Kpolovie, Amaele, & Uzorka (2012). To fill the existing great knowledge gap, this study was designed to validate the APM for use in Nigeria by establishing its temporal consistency, consistency of equivalence, internal consistency, criterion-related validity as well as construct validity. The APM can only be correctly used in Nigeria after the test has duly been validated and/or standardized locally in Nigeria in line with the golden rule of the International Test Commission (2000) in its *international guidelines for test use* that a measuring instrument that is developed in one country should by necessity be validated in another country before it is adopted there for use.

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Second, the International Test Commission recommended that IRT be used for the proper description and evaluation of existing and widely used psychological instruments (Muñiz, 2011). The APM scale is yet to be examined in Nigeria with IRT. Therefore IRT will be applied on the APM scale. From the review of literature, there does not exist a single study on record that has carried out operational validation of the psychometric properties of the Advanced Progressive Matrices in Nigeria, using the framework of either CTT, or IRT, or both. Such knowledge gap, if not filled, continues to breed what Thompson (2004) described as the "five methodology errors in educational research, the pantheon of statistical significance and other faux pas" that Spearman (1927) and Rathus (1990) hah much earlier enjoined psychometricians to guide against. Hence the need exists for the gap to be filled; and this current study was carefully designed to achieve.

Thirdly, it is feared that with the type of cultural diversity in Nigeria, a test may not suitably measure mental ability without bias (Gilovich, Griffin & Kahneman, 2002). This investigation is therefore also aimed at ascertaining whether the AMP could actually be biased in measuring the attribute in Nigeria or not.

The purpose of this study therefore, is to solve the identified and categorized problem, using multiple perspectives that satisfactorily establish the reliability and validity of APM in addition to empirically determining whether the test can suitably be used in Nigeria without biases. Consequently, nine research questions as follows were posed and answered in this study because they all help in inferring test reliability, validity and bias in IRT.

1. Which is the most suitable Item Response Theory (IRT) Parameter Logistics Model (PLM) for Raven's Advanced Progressive Matrices (APM) in Nigeria?
2. What is the Overall Model Fit of APM using Nigerian validation sample?
3. What is the person separation reliability of APM that can be inferred from the contribution of each of the items to the Test Response Function (TRF)?
4. What is the Item Response Function (item-by item) evidence of reliability of APM in Nigeria?
5. What is the evidence of unidimensionality, if any, of APM in Nigeria?
6. What is the b-parameter index (item difficulty parameter) for each APM item in Nigeria?
7. What is the range of Differential Item Function (discriminatory index) popularly referred to in IRT as a-parameter of the APM in the country?
8. What is the c-parameter (probability of guessing) for each APM item as evidence of bias culturally (ethnic group, school type/age, and sex)?
9. What is the concurrent validity of APM with Culture Fair Intelligence Test in Nigeria?

METHODOLOGY

Multiple triangulation research design which Kpolovie (2016) described as the highest and most comprehensive and all-embracing form of triangulation research was employed in this study. This design was used because "it allows for multi-method approach in studying psychometric properties of an instrument and some aspects of human behaviour. It helps to map out or explain more fully, the richness and complexity of a psychometrical instrument and/or human characteristics by studying it from more than one stand point" (Kpolovie, 2010). This research design enabled the researchers to apply various methods including the IRT

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

logistic models (1-PLM, 2-PLM, 3-PLM), DIF, TIF, Factor analyses using the data reduction option, Test of Fitness of Good Statistics, CTT, t-test, Correlations, analysis of variance, (ANOVA) Kuder-Richardson's estimates (KR_{20} and KR_{21}), normalized standard score and percentile ranks, among many other statistical procedures in a single investigation for a much more robust results.

The study was carried out in Nigeria. The population of the study comprised all the university undergraduates (1,794,989) and all the senior secondary school students in Nigeria (4,758,739); making a total population of 6,553,728 (FRN National Population Commission, 2014; Federal Ministry of Education, 2014; NEEDS Assessment of Nigerian Universities, 2013). With the use of Table of Random Numbers, a total sample of 2,100 (1,000 undergraduates and 1,100 secondary school students) was randomly drawn, using disproportional stratified random sampling technique (Kpolovie, 2011) as participants in the study. The sample had males and females between the age range of 12 and 40 years; spread over four main cultural groups (Hausa, Igbo, Yoruba and Minorities) that adequately cover the four geopolitical zones in Nigeria.

The main instrument of this study was the APM test that was under validation. Another instrument used was the Culture Fair Intelligence Test (CFIT) that has since been validated and standardized for use in Nigeria (Kpolovie, 2015) to allow for establishment of concurrent validity of the APM. The CFIT has test-retest reliability of 0.92, equivalent forms reliability of 0.91, split-half reliability of 0.93, internal consistency reliability of 0.91 via KR_{20} and 0.87 via KR_{21}; and a construct validity of 0.83 through subtest-total correlation; in addition to satisfactory developmental changes evidence as scores increased significantly from age 9 to 15 and flattened out thereafter strictly in accordance with the Fluid and Crystalized Theory of Intelligence (Kpolovie, 2015; Cattell, 1962). Further construct validity evidence of CFIT showed no significant difference in score across the four cultural groups in Nigeria (Igbo, Hausa, Yoruba and Minority) indicating that the test is not culturally biased; in addition to overwhelmingly significant difference between mentally retarded students (MRS), normal students (NS) and gifted students (GS) with the GS significantly higher than the NS, and the NS significantly higher than the MRS (Kpolovie, 2015).

Eight research assistants (two from each geopolitical region) were engaged, trained to administer the tests and accompanied the researchers to the four geopolitical zones to conduct the tests to the various participants. Administration of the APM test strictly lasted for 40 minutes, while administration of the CFIT lasted for 25 minutes in each centre. The APM test has two subsections (Sets 1 and 2). The Set 1 contains 12 items, ant the Set 2 contains 36 items. The Set 1 items were used as practice test. The Set 2 was the main scale of this study. Each item of APM has eight options from which the participant is expected to select one option. Each correct option chosen by an examinee was scored 1 point while 0 point was given for each question the examinee marked a wrong option.

The APM scores were painstakingly subjected to IRT analysis (Arthur & Day, 1994; Chiesi, Ciancaleoni, Galli, Morsanyi & Primi, 2011), using the X-Calibre 4.2 software (Cikrikci-Demirtasli, 2000; Field, 2005; Gallini, 1983). The X-Calibre 4.2 IRT analysis was performed adopting the three different Parameter Logistic Models, that is, the 1- Parameter Logistic Model (1-PLM), the 2-Parameter Logistic Model (2-PLM), and the 3-Parameter Logistic Model (3-PLM). Other statistical analysis performed with the X-Calibre software includes:

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

i. The Differential Information Function (DIF) for group comparisons [gender (males and females), age as inferred from the school categories (university undergraduates within the age range of 16-40 years old and senior secondary school students within the age range of 12-20 years old) ethnicity (Hausa, Igbo, Yoruba and Minorities)].

ii. Test Information Function, Item Information Function, Item-by-Item Analysis that included the Item characteristic curve also known as Item Response Function.

SPSS was also used to perform Factor Analysis for determination of unidimensionality of the APM; dimension reduction analysis, reliability analysis, correlation analysis, one-way and two-way analysis of variance, etc. The Microsoft Excel software was used for scoring the responses of the tests, normalized scores analysis, and percentile ranks analysis as suggested by Ololube and Kpolovie (2013; 2012).

Statistical packages that the researchers employed in this study were:

1. X-Caliber 4.2
2. SPSS version 22 and
3. Microsoft Excel software.

These statistical packages were used in accordance with the statistical triangulation demands of Multiple Triangulation research design (Kpolovie, 2016; 2015; Verguts & De Boeck, 2002). Ololube, Kpolovie and Makewa (2015); Ojerinde, Popoola, Ojo and Onyeneho (2012); and Ojerinde, Popoola, Ojo and Ariyo (2014) have equally called for use of these three statistical packages for analysis of data in an investigation of this nature. Furthermore, Guyer & Thompson (2011) posited that “Item response theory (IRT) presents a powerful psychometric paradigm for developing, delivering, analysing, and scoring assessments, and that in order to utilize IRT with the aim of obtaining accurate results, assessment data must be calibrated with sophisticated software designed for that purpose.” Similar calls have also been made by Muniz (2009), Orluwene (2012), Raven, Raven and Court (1993), and Vigneau and Bors (2015).

RESULTS

The findings of this investigation are presented herein in accordance with the research question and briefly explained.

The most suitable parameter logistic model for APM

In order to determine the most suitable Parameter Logistic Model for the main instrument of this research, the APM scale, the X-Calibre IRT analysis was performed for each of the three different Parameter Logistic Models, namely, the 1-Parameter Logistic Model (1-PLM), the 2-Parameter Logistic Model (2-PLM) and the 3-Parameter Logistic Model (3-PLM).

Table 1:1-PLM Item Parameters for All Calibrated Items

Seq.	Item ID	P	R	a	Flag(s)
1	1	0.746	0.050	1.000	F
2	2	0.760	0.158	1.000	F
3	3	0.758	0.045	1.000	F

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

4	4	0.754	0.060	1.000	F
5	5	0.735	0.089	1.000	F
6	6	0.765	0.042	1.000	F
7	7	0.721	0.158	1.000	F
8	8	0.750	0.044	1.000	F
9	9	0.734	0.035	1.000	F
10	10	0.753	0.085	1.000	F
11	11	0.730	0.049	1.000	F
12	12	0.734	0.108	1.000	F
13	13	0.730	0.129	1.000	F
14	14	0.697	0.104	1.000	F
15	15	0.714	0.136	1.000	F
16	16	0.740	0.090	1.000	F
17	17	0.740	0.155	1.000	F
18	18	0.743	0.124	1.000	F
19	19	0.733	0.093	1.000	F
20	20	0.709	0.116	1.000	F
21	21	0.719	0.207	1.000	F
22	22	0.685	0.128	1.000	F
23	23	0.716	0.232	1.000	F
24	24	0.709	0.199	1.000	F
25	25	0.662	0.232	1.000	F
26	26	0.716	0.189	1.000	F
27	27	0.717	0.221	1.000	F
28	28	0.746	0.290	1.000	F
29	29	0.726	0.273	1.000	F
30	30	0.690	0.220	1.000	F
31	31	0.016	0.016	1.000	K, FHb
32	32	0.633	0.264	1.000	F
33	33	0.607	0.269	1.000	F
34	34	0.530	0.239	1.000	F
35	35	0.500	0.271	1.000	F
36	36	0.177	0.076	1.000	

Table 1 present the classical statistics, the item parameters, and flags (if any) for each calibrated item. The K flag indicates that the keyed alternative did not have the highest correlation with total score. The F flag indicates that the item fit statistic (z Resid for dichotomous) was significant, and the item did not fit the IRT model. The La, Lb, and Lc flags indicate that the a/b/c parameters (that is the discriminatory, difficulty and guessing indices) were lower than the minimum acceptable value. The Ha, Hb, and Hc flags indicate that the a/b/c parameters were higher than the maximum acceptable value. As shown above in **Table 1**, all the 36 items of the APM Set 2 under the 1-PLM were flagged which indicates that all the

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

items did not fit the 1-Parameter Logistic IRT Model. Item 31 in addition to the F-flag, further showed the K-flag which indicates that the keyed alternative for item 31 did not have the highest required correlation with total score. Item 31 also showed the Hb Flag which indicates that the b-parameter for item 31 was higher than the maximum acceptable value. Therefore the 1-PLM is not suitable for the APM scale.

Table 2:2-PLM Item Parameters for All Calibrated Items

Seq.	Item ID	P	R	a	b	Flag(s)
1	1	0.746	0.050	0.415	-4.000	F, Lb
2	2	0.760	0.158	0.656	-4.000	Lb
3	3	0.758	0.045	0.520	-4.000	F, Lb
4	4	0.754	0.060	0.590	-4.000	F, Lb
5	5	0.735	0.089	0.504	-4.000	F, Lb
6	6	0.765	0.042	0.599	-4.000	F, Lb
7	7	0.721	0.158	0.682	-3.828	Lb
8	8	0.750	0.044	0.541	-4.000	F, Lb
9	9	0.734	0.035	0.450	-4.000	F, Lb
10	10	0.753	0.085	0.603	-4.000	F, Lb
11	11	0.730	0.049	0.484	-4.000	F, Lb
12	12	0.734	0.108	0.547	-4.000	Lb
13	13	0.730	0.129	0.615	-4.000	Lb
14	14	0.697	0.104	0.523	-4.000	Lb
15	15	0.714	0.136	0.566	-4.000	Lb
16	16	0.740	0.090	0.555	-4.000	F, Lb
17	17	0.740	0.155	0.625	-4.000	Lb
18	18	0.743	0.124	0.636	-4.000	Lb
19	19	0.733	0.093	0.469	-4.000	F, Lb
20	20	0.709	0.116	0.465	-4.000	Lb
21	21	0.719	0.207	0.516	-4.000	Lb
22	22	0.685	0.128	0.427	-4.000	Lb
23	23	0.616	0.232	0.569	-4.000	Lb
24	24	0.709	0.199	0.536	-4.000	Lb
25	25	0.662	0.232	0.512	-3.615	Lb
26	26	0.716	0.189	0.433	-4.000	F, Lb
27	27	0.717	0.221	0.447	-4.000	F, Lb
28	28	0.746	0.290	0.535	-4.000	F, Lb
29	29	0.726	0.273	0.498	-4.000	F, Lb
30	30	0.690	0.220	0.408	-4.000	F, Lb
31	31	0.016	0.016	0.705	4.000	K, F, Hb
32	32	0.633	0.264	0.361	-4.000	Lb

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

33	33	0.607	0.269	0.317	-4.000	Lb
34	34	0.530	0.239	0.255	-3.694	La, Lb
35	35	0.500	0.271	0.270	-2.946	La
36	36	0.177	0.076	0.179	3.022	La, Hb

In **Table 2** above, 17 out of the 36 items of the APM Set 2 under the 2-PLM were flagged. This indicates that 17 items did not fit the 2-Parameter Logistic IRT Model. Consequently the result indicates a partial fit since most of the items did not fit the 2-Parameter Logistic IRT Model. In addition to the F-flag, 31 items further showed the Lb-flag (that is low difficulty index) and items 34, 35 and 36 also showed the La-flag (that is low discriminatory index). These indicate that the *a*- and *b*-parameters for those items were lower than the minimum acceptable value. Item 31 also showed the K and Hb flags (that is high difficulty index) which indicate that the keyed alternative for item 31 did not have the highest correlation with total score while the b-parameter for item 31 was higher than the maximum acceptable value. Therefore the 2-PLM is also not suitable for the APM scale.

Table 3: 3-PLM Item Parameters for All Calibrated Items

Seq.	Item ID	P	R	a	b	c	Flag(s)
1	1	0.746	0.159	0.640	-2.595	0.499	
2	2	0.760	0.158	0.756	-2.574	0.189	
3	3	0.758	0.045	0.663	-2.557	0.331	
4	4	0.754	0.060	0.772	-2.546	0.448	
5	5	0.735	0.089	0.607	-2.436	0.252	
6	6	0.765	0.042	0.746	-2.394	0.252	
7	7	0.721	0.158	0.834	-2.234	0.249	
8	8	0.750	0.044	0.674	-2.170	0.253	
9	9	0.734	0.035	0.550	-2.092	0.253	
10	10	0.753	0.085	0.752	-2.076	0.251	
11	11	0.730	0.049	0.608	-2.049	0.253	
12	12	0.734	0.108	0.677	-2.001	0.252	
13	13	0.730	0.129	0.760	-1.919	0.251	
14	14	0.797	0.104	0.704	-1.891	0.255	
15	15	0.714	0.136	0.748	-1.505	0.252	
16	16	0.740	0.090	0.703	-1.347	0.251	
17	17	0.740	0.155	0.803	-1.145	0.251	
18	18	0.743	0.124	0.772	-1.076	0.250	
19	19	0.733	0.093	0.590	-1.041	0.253	
20	20	0.709	0.116	0.599	-1.007	0.253	
21	21	0.719	0.207	0.616	0.085	0.252	

22	22	0.685	0.128	0.557	0.105	0.256	
23	23	0.716	0.232	0.707	0.112	0.252	
24	24	0.709	0.199	0.717	0.313	0.254	
25	25	0.662	0.232	0.636	0.425	0.253	
26	26	0.716	0.189	0.512	0.523	0.255	
27	27	0.717	0.221	0.536	0.645	0.255	
28	28	0.746	0.290	0.645	0.775	0.253	
29	29	0.726	0.273	0.599	0.838	0.254	
30	30	0.790	0.220	0.494	0.850	0.256	
31	31	0.016	0.283	0.784	0.993	0.617	
32	32	0.633	0.264	0.431	1.076	0.258	
33	33	0.607	0.269	0.367	1.086	0.260	
34	34	0.530	0.239	0.313	1.505	0.263	
35	35	0.500	0.271	0.338	2.060	0.263	
36	36	0.177	0.076	1.180	2.133	0.449	

In **Table 3** above no item out of the 36 items of the APM Set 2 under the 3-PLM was flagged either for F, K, La/b/c or Ha/b/c. Consequently the result indicates a perfect fit since all the items fit the 3-Parameter Logistic IRT Model. Therefore within the framework of IRT, the 3-Parameter Logistic IRT Model is the most suitable for examining the Advance Progressive Matrices (APM) scale.

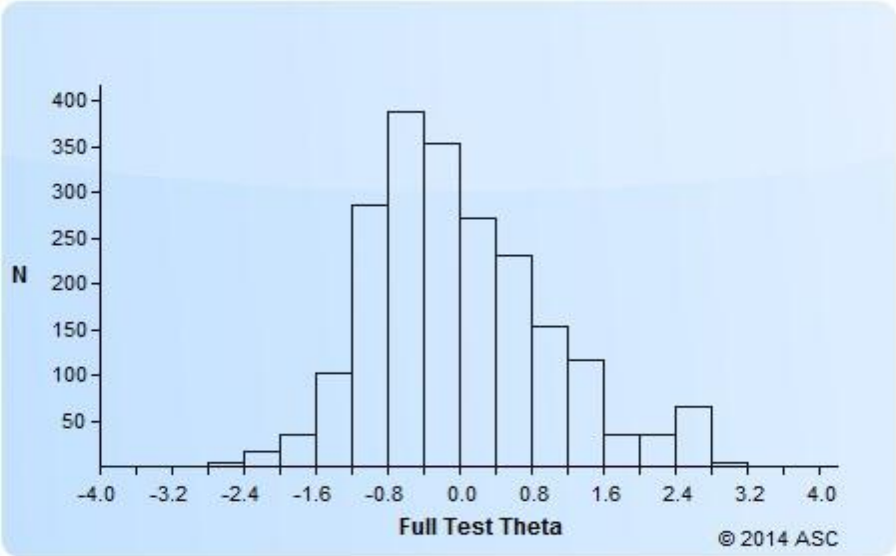
The Overall Model Fit of APM

Table 4: Overall Model Fit

Test	Items	Chi-square	df	p	-2LL
Full Test	36	1120.993	432	0.000	40593

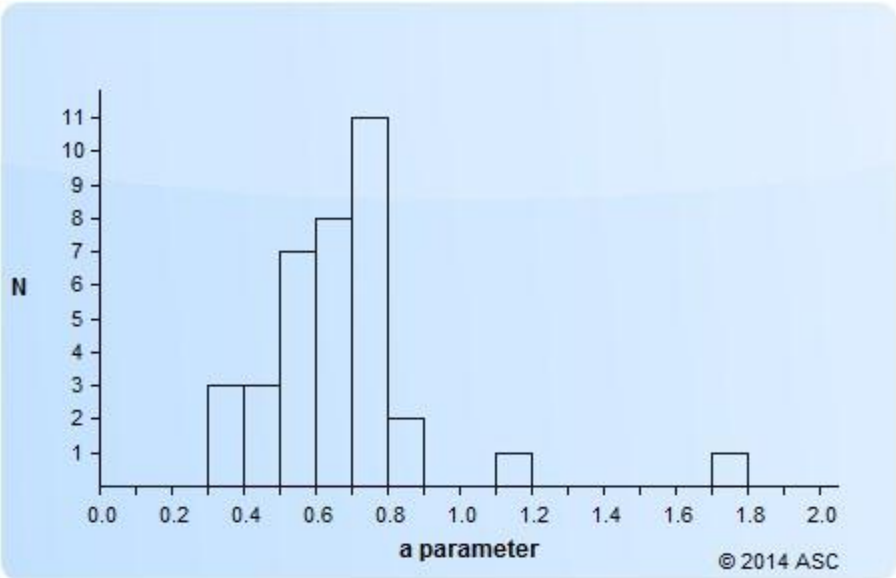
TABLE 4 above presents the Overall Model Fit with a Chi-Square value of 1120.993, degrees of freedom (df) of 432, a probability of 0.000 and -2 logistic likelihood of 40593. To further appreciate the Overall Model fit, the distribution of the theta estimates for all calibrated items, frequency distribution for the theta estimates, the distribution of the a-, b- and c-parameters are presented below: **Figure 1** displays the distribution of the theta estimates for all calibrated items.

Figure 1: Theta Estimates for All Calibrated Items



Additional information about the fit statistics are contained in the item-by-item results of the analysis (contained in the full work). Each scored item has four tables and a plot of the item response function (IRF). The chi-square fit statistic and its degrees of freedom are reported for each item. All the items indicated good fit statistics under the 3-PLM. **Figure 2** below displays the distribution of the a-parameters.

Figure 2: Histogram of the a-Parameters



ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Figure 3 displays the distribution of the b parameters.

Figure 3: Histogram of the b Parameters

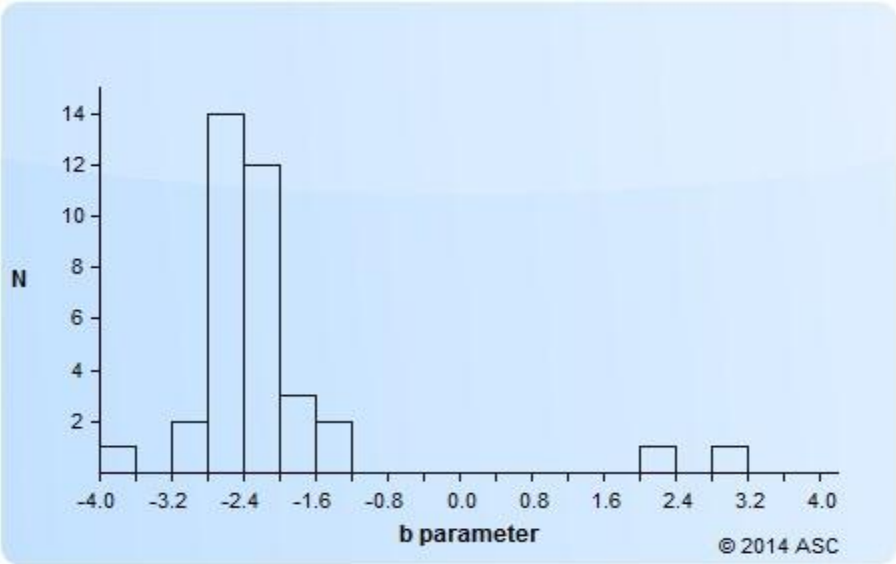


Figure 4 displays the distribution of the c parameters.

Figure 4: Histogram of the c Parameters

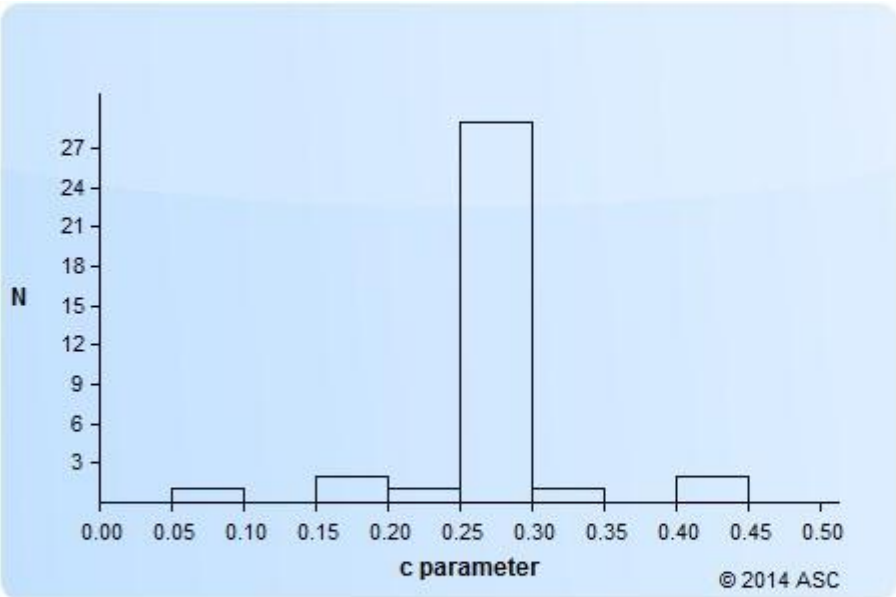


Figure 5 displays the scatter plot of the b parameter (difficulty) by the a-parameter (discrimination) for all calibrated items.

Figure 5: b-Parameter by a-Parameter

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

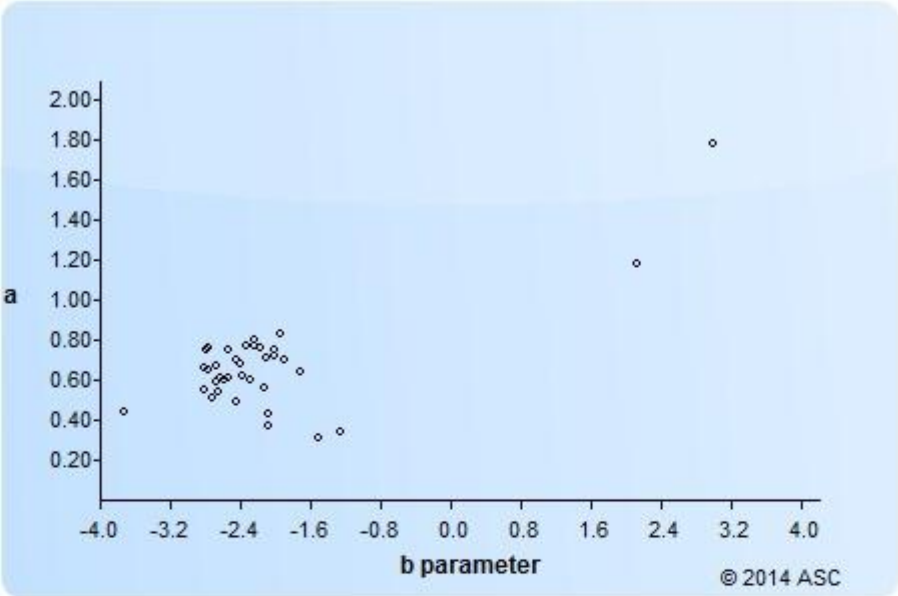
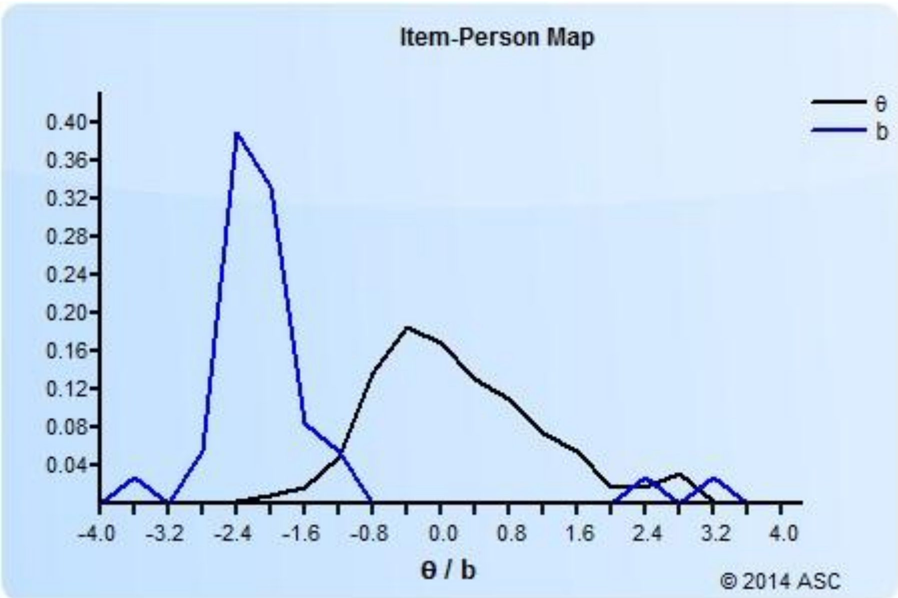


Figure 6 displays the joint distribution of the b parameter by Theta.

Figure 6: b-parameter by Theta



Inferred reliability of APM from Test Response Function

The focus of IRT is the contribution of each item to the overall fit of any given instrument. Therefore discussion on reliability of instrument is usually inferred from the Test Response Function (TRF) since the concept of test response function is analogous to the concept of reliability in Classical Test Theory. The TRF is pictorially displayed in **Figure 7**. Reliability in this case is conceived as the person separation reliability or item separation reliability. The person separation reliability is analogous to Cronbach's α. This is the degree to which the APM scale differentiates persons in the test's outcome. The range of course is 0 – 1. Item separation reliability on the other hand is the degree to which item difficulties are differentiated. Again the range of course is 0 – 1. Now due to the sophistication of the X-Calibre software, a power

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

tool designed for analysing IRT, it is now possible to estimate the reliability of an instrument under the IRT models.

Table 5: Summary Statistics for the Total Scores indicating person separation reliability of APM

Test	Items	Alpha	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	36	0.948	33.005	4.74	-2.361	7	31.00	33.0	34.00	36	3.00

Table 5 above shows the Alpha value of the full test. The Alpha value is 0.948 which tends towards 1 and it indicates a strong reliability. Therefore the reliability of the APM scale within the framework of IRT as indicated by the Alpha value is 0.948. **Figure 7** displays a graph of the Test Response Function (TRF) for all calibrated items. The TRF predicts the proportion or number of items that an examinee would answer correctly as a function of theta. The left Y-axis is in proportion correct units while the right Y-axis is in number-correct units. In this case TRF will predict 94.8% or its equivalent of the score of each examinee on the APM in Nigeria.

Figure 7: Test Response Function

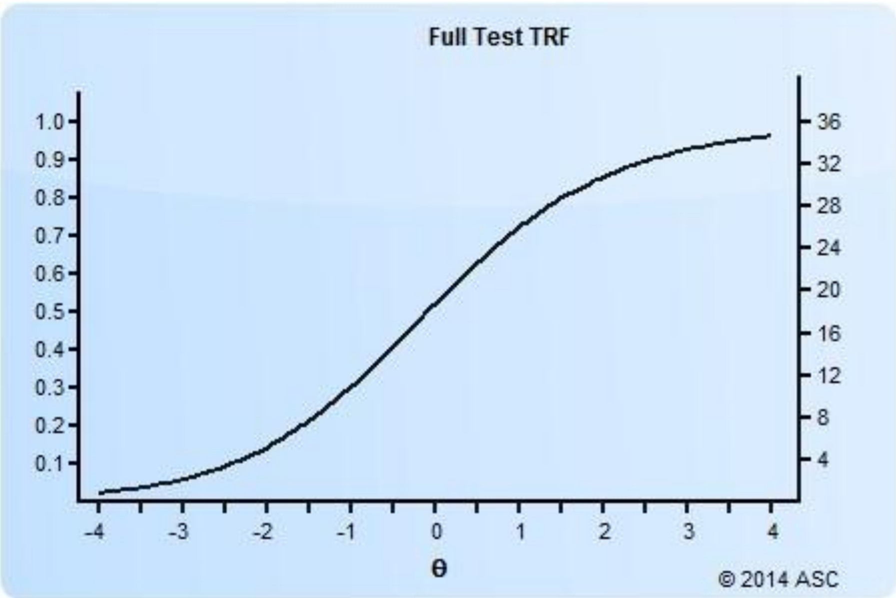


Figure 8 displays a graph of the Test Information Function for all calibrated items. The TIF is a graphical representation of how much information the test is providing at each level of theta. Maximum information was 5.825 at theta = -2.050.

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Figure 8: Test Information Function

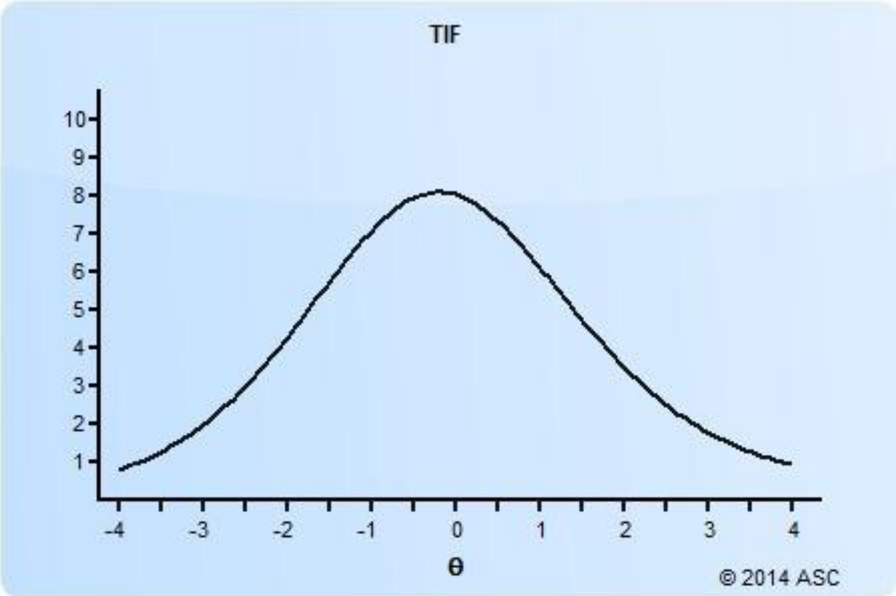
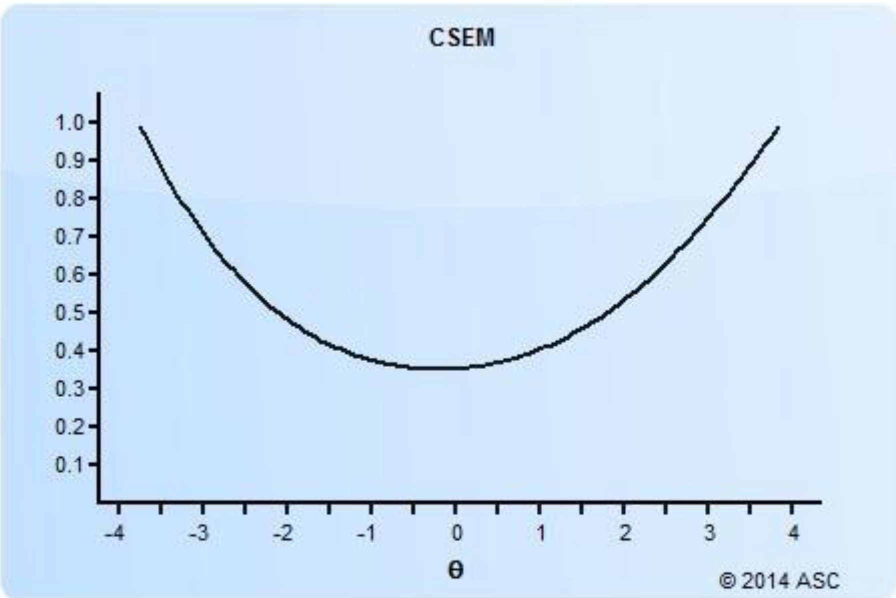


Figure 9 displays a graph of the Conditional Standard Error of Measurement (CSEM) Function. The CSEM is an inverted function of the TIF, and estimates the amount of error in theta estimation for each level of theta. The minimum CSEM was 0.414 at theta = -2.050.

Figure 9: CSEM Function



Item Response Function (item-by- item) reliability of APM

The item-by-item results of the analysis shows that each scored item has four tables and a plot of the item response function (IRF). The item-by-item analysis report is a sequel to the Test Information Function (TIF). The red line (fit line) represents the observed proportion correct

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

conditional on theta. In almost all the items of the APM scale, there were no large deviations of the red line from the IRF which are suggestive of good item fit. Thus, the fit line further identifies why and how the particular item fits the chosen 3-Parameter Logistic IRT model. There are four tables presented for each item by the X-Calibre analysis, but due to secrecy and confidentiality of the items, only those for the 1st, 18th and 36th items are presented here for illustration.

1. Item information table: records the information supplied by the control file (or Classic Data Header) for this item.
2. Classical statistics table: classical statistics for the item.
3. IRT parameters table: item parameter estimates for the item.
4. Option/Category statistics: detailed statistics for each item, which helps diagnose issues in items with poor statistics. In the case of the APM, there are no items with poor statistics under the chosen 3-PLM.

The classical statistics presents classical summary statistics for the item. For multiple choice items instrument like the instrument of this study: APM, the P value and the point-biserial correlations are presented in the first three columns of the table. The P value is the proportion of examinees that answered an item in the keyed direction and ranges from 0 to 1. The S-Rpbis and T-Rpbis are the point-biserial correlations of an item with total score and theta, respectively. The Alpha w/o is Cronbach's alpha computed with the current item excluded. The item-total correlation is a measure of the discriminating power of the item and is related to the IRT discrimination parameter. The IRT parameters table presents the IRT item parameters and the fit statistics. The latent trait theta is expressed on a standardized scale, so a one unit change equals a one standard deviation change. The "a" parameter indexes the discrimination of the item, as larger values for "a" will result in a greater steepness of the slope of the IRF or Item Characteristic Curve (ICC) and indicate the item differentiates examinees well. The "b" parameter is the item difficulty parameter and equals the location on the theta continuum where the probability of a correct response equals .50 + (c/2). It follows that multiple choice items with more positive "b" parameters are more difficult for examinees, as a higher trait level is required to endorse the keyed response 50% of the time. The "c" parameter equals the probability of an examinee of infinitely low theta obtaining a correct response due to guessing. Thus, "c" is also the lower asymptote of the IRF or ICC. The standard errors (SE) for each item parameter estimate are also presented in the item parameter table. A large SE for an item parameter (compared to the other items) indicates that the item parameter was poorly estimated. The IRT standardized (z) residual is the last entry in the item parameter table. It indexes the fit of the data to the Item Response Function. For dichotomous items, the p-value for rejecting the item as poor fit was computed using the z residual with the standard normal distribution as its sampling distribution. The chi-square fit statistic and its degrees of freedom are reported for each item.

Presented below are three examples of the item-by-item report.

Item1 information

Seq.	*ID*	*Model*	*Scored*	*Num Options*	*Domain*	*Flags*
1	1	3PL	Yes	8	1	Lb

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Classical statistics for item1

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
2100	0.946	0.050	0.147	0.589

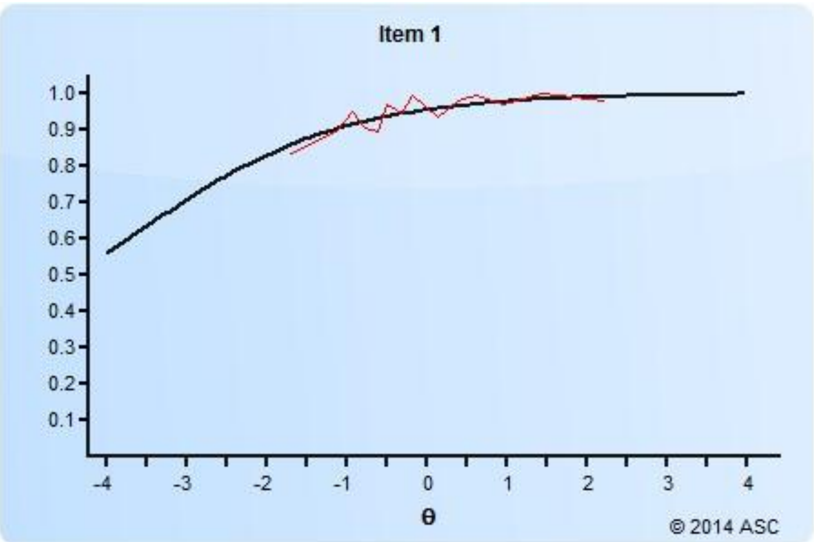
IRT parameters for item1

a	*B*	*c*	*a SE*	*b SE*	*c SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.540	-2.525	0.399	0.039	0.130	0.281	25.162	12	0.014	0.316	0.752

Option statistics for item1

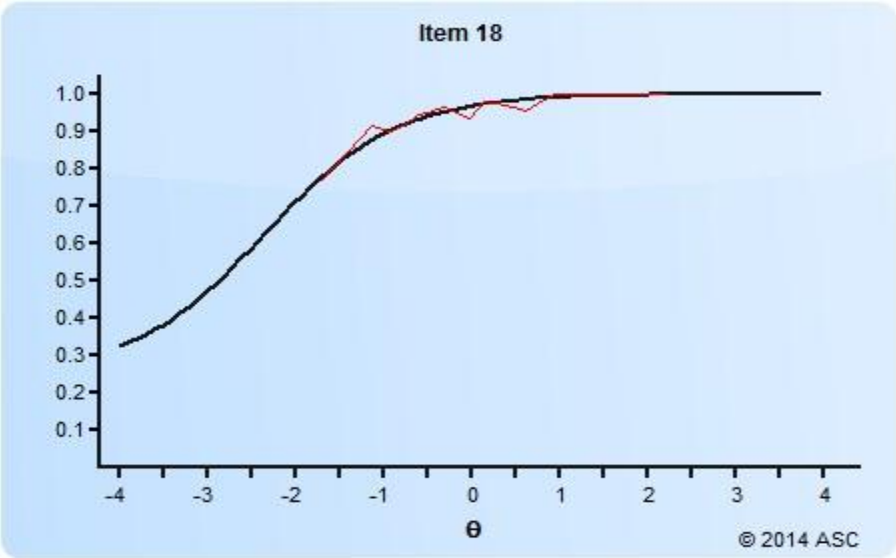
Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
A	36	0.017	-0.006	-0.078	-0.613	0.717	
B	16	0.008	-0.031	-0.058	-0.680	0.947	
C	2	0.001	0.002	-0.017	-0.565	0.836	
D	30	0.014	-0.021	-0.073	-0.629	0.746	
E	1987	0.946	0.050	0.147	0.033	1.027	
F	5	0.002	-0.063	-0.057	-1.199	0.409	
G	15	0.007	-0.001	-0.046	-0.562	0.655	
H	7	0.003	-0.008	-0.016	-0.283	1.084	
Omit	2	0.001	-0.058	-0.034	-1.133	1.341	
Not Admin	0						

For the first item a total of 1987 selected the right key, indicating that this particular item was an easy item. It is however within the normal difficulty index bound of -3 to +3 making the item reasonably fit for the APM instrument. Below is the graphical display of item 1:



Similar information for Item 18 is given here. The interpretation or explanation is just as given for Item 1.

ISSN 2055-0863(Print), ISSN 2055-0871(Online)



Item information

Seq.	*ID*	*Model*	*Scored*	*Num Options*	*Domain*	*Flags*
18	18	3PL	Yes	8	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
2100	0.943	0.124	0.213	0.584

IRT parameters

a	*b*	*c*	*a SE*	*b SE*	*c SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.772	-2.326	0.250	0.044	0.077	0.252	18.123	12	0.112	1.363	0.173

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
A	39	0.019	-0.095	-0.124	-0.929	1.316	
B	13	0.006	-0.060	-0.087	-1.134	0.493	
C	0	0.000	--	--	--	--	
D	31	0.015	-0.056	-0.120	-1.009	0.754	
E	2	0.001	0.020	-0.001	-0.023	0.185	
F	10	0.005	0.031	-0.015	-0.219	0.527	
G	1981	0.943	0.124	0.213	0.051	1.005	
H	24	0.011	-0.067	-0.096	-0.916	0.774	
Omit	0						
Not Admin	0						

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

The third illustration presented here is for Item 36.

Item36 information

Seq.	*ID*	*Model*	*Scored*	*Num Options*	*Domain*	*Flags*
36	36	3PL	Yes	8	1	Hc

Classical statistics for item36

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
2100	0.377	0.076	0.086	0.597

IRT parameters for item36

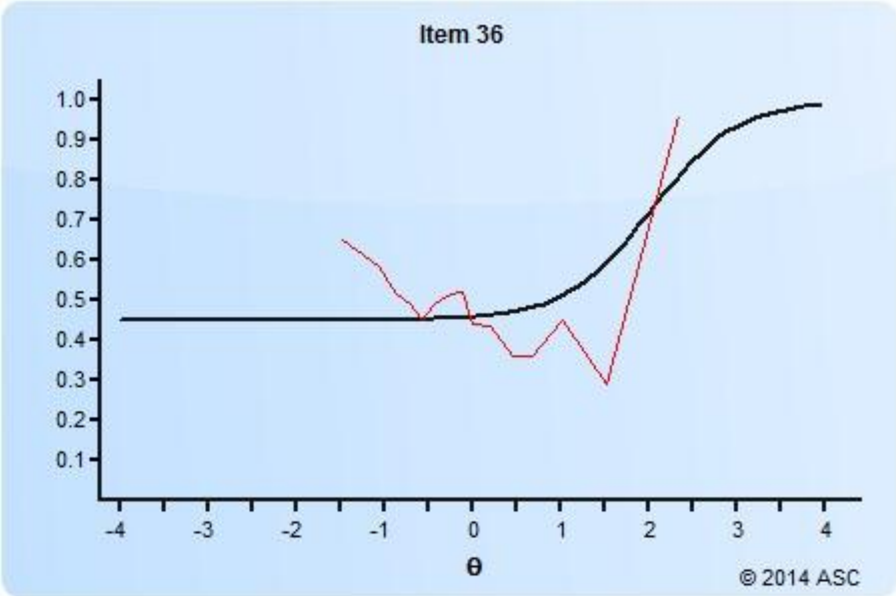
a	*B*	*c*	*a SE*	*b SE*	*c SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
2.180	2.733	0.749	0.195	0.114	0.048	97.697	12	0.000	1.277	0.201

Option statistics for item36

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
A	197	0.094	0.052	-0.015	-0.052	0.770	
B	492	0.377	0.076	0.086	0.111	1.166	
C	322	0.058	0.058	0.006	0.023	0.897	
D	234	0.111	0.103	0.040	0.114	0.849	
E	134	0.016	0.013	-0.007	-0.062	0.766	
F	76	0.036	0.049	0.005	0.023	0.797	
G	75	0.036	0.048	0.023	0.119	0.910	
H	53	0.025	0.058	0.026	0.160	0.781	
Omit	517	0.246	-0.294	-0.139	-0.252	1.026	
Not Admin	0						

The right key was got by 492 examinees selected this option. Again the b-parameter is near the end of the rung in the b-continuum confirming that the item was in the difficulty side of the scale. In any case the APM is constructed and calibrated in such a way that the b-parameter graduates in the difficulty index from very easy items to very difficult items. This report confirms the aim and design of the APM tool. In the overall report for guessing, guessing was not significant but was for item 36. This is the beauty, power and potency of IRT. The graphical display for item 36 is shown below:

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

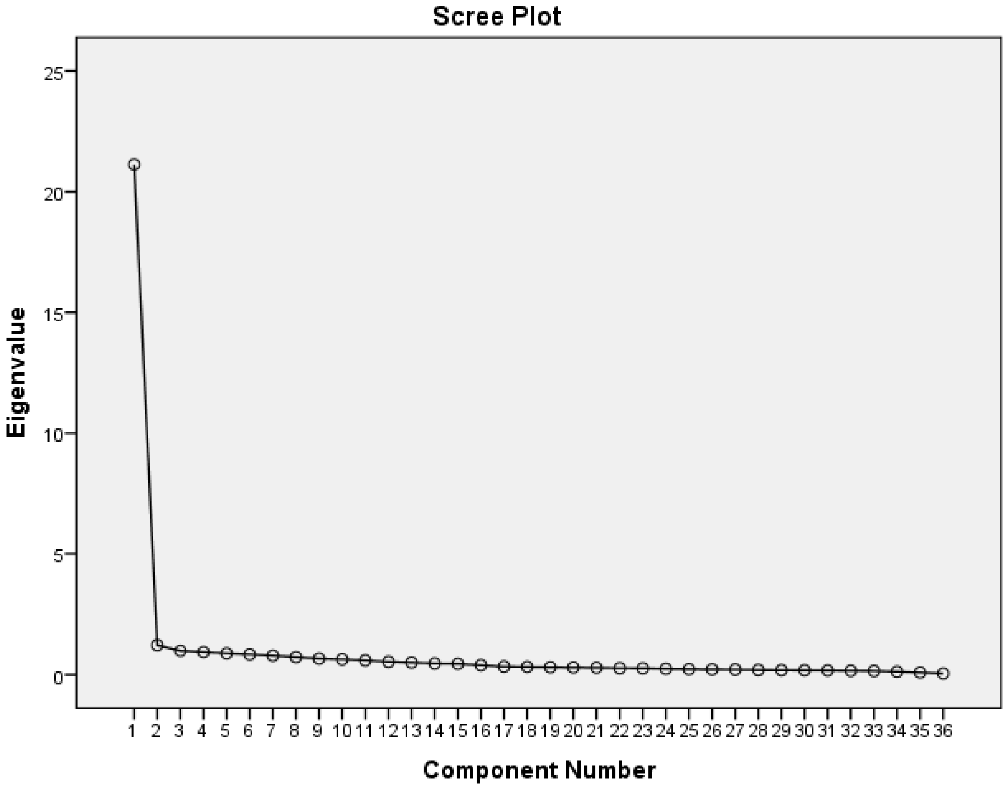


Unidimensionality evidence of APM in Nigeria

Unidimensional models require a single trait (ability or domain) dimension, $\boldsymbol{\theta}$. (Table 4.8 contained in the full report of this work), part of the X-Calibre report indicated that the scale essentially examined one trait, construct or domain. However in order to extrapolate the unidimensionality of the scale, factor analysis was performed for the scores generated from the APM test administered to the 2100 university undergraduates and senior secondary school students. Field (2005) wrote that "factors or traits or underlining constructs can be extrapolated or established through the use of eigenvalues and variance, scree plot and communalities." Georgiev (2008), Morsanyi, Primi, Handley (2009), Raven (2000), Van der Ven and Ellis (2000), Raven, Raven and Court (1997) as well as the WPS (20015) stressed the need for extrapolation of unidimensionality of an instrument that is indeed measuring only one domain or construct; or even more than one factor with the use of eigenvalues. Using Guttman-Kaiser rule, "all factors with eigenvalues greater than 1 should be retained as the factors that the scale measures." Guttman-Kaiser also suggested that "factors which account for 70% and above of the variance should be accepted as the underlining construct." Analysis of the scree plot is another way to determine the underlining construct or unidimensionality of a scale. The rule of thumb in analysing the scree plot is very simple. Traits or constructs or factors before the breaking point or elbow joint in the scree plot graph is assumed to be the main construct under examination. Furthermore it is also important to check the communalities after construct extraction. If the communalities are low, the extracted constructs account for only a little part of the variance, and therefore more constructs might be deemed to be in view which might provide better account for the total variance. Dimension reduction analysis was utilized to determine significant unidimensionality extraction at greater than 0.50. The choice of 0.50 was made by the researcher because according to Thomson (2004) "determining the number of factors or construct to be extract or extrapolated requires judgment." In this analysis, promax rotation was utilized to maximize the establishment of the construct under examination. The choice of promax rotation was made because orthogonality is not assumed in this case and therefore the items of the construct to be examined are expected to correlate. A careful examination of the scree plot shown below shows that there is only one construct before the breaking point or elbow joint. This therefore succinctly shows the unidimensionality of the underlining construct of the APM scale, namely intelligence or fluid ability. All the 36 items

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

measure one construct, the fluid ability of the test taker as confirmed by the scree plot.



The eigenvalue associated with each construct represents the variance explained by that particular linear or unidimensional component and SPSS factor analysis of the scores of the APM test administered to 2100 university undergraduates and senior secondary school students displays the eigenvalue in terms of the percentage of the variance explained. In this case, the one construct examined by the scale, which is intelligence explains 33.005 or 85.686% of the total variance. This is incontrovertibly a sizable chunk of the model. Therefore the underlining construct is effectively examined by the scale and it ensures its unidimensionality. Since the assumption of unidimensionality is met by this model, it invariably means that local independence holds. Thus all the items APM unquestionably measure just one general intelligence factor in Nigeria just as it does in all other countries that the test is actively in use. So, it should be used in Nigeria to validly and reliably measure the construct without bias.

Item Difficulty Parameter (b-parameter) of APM

The answer to the sixth research question is glaringly obvious in **Table 3** (already presented above). The *b*-parameter is the item difficulty parameter and indicates the location on the theta (θ) continuum where the probability of a correct response equals $c/2 + .50$. Thus, the *b*-parameter is the centre of the IRF and is where the slope steeps most to show the discriminating power of the item maximally. Since the APM scale is centred on the examinees drawn from the university undergraduates and senior secondary school students, the *b* parameter shows the examinee's θ value for which the item is appropriate. Higher *b*-parameters (> 1.0) indicate that the item is more difficult; a value below -1.0 indicates that the item is very easy. The purpose of the APM test is the measure of fluid ability, the test is conceived and designed by its constructor in such a way that the difficulty index graduates from very easy item to very

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

difficult item. According to X-Calibre manual, the difficulty index "ranges in theory from negative to positive infinity, but in practice from -3.0 (very easy) to +3.0 (very difficult)." A careful examination of the *b*-parameter column shows that the values of *b* for item 1 is -2.595, item 2 is -2.574, items 15 and 21 have *b* values of -1.505 and 0.085 respectively. The *b*-parameter kept graduating in difficulty until the last item which has a *b* value of 2.133. The *b*-parameter is related to the classical *P* statistic, as items with low *P* values will tend to have higher (more positive) *b*-parameters and items with high *P* values will tend to have lower (more negative) *b*-parameters.

Differential Item Function (Discriminatory Index) or a parameter of APM in Nigeria

Differential item functioning (DIF) occurs when the performance of an item differs across groups of examinees with equal latent trait as an evidence of item bias which leads to test bias. In this study, the university undergraduates and senior secondary school students' responses to the APM test were examined for DIF across gender (i.e. males and females), age (i.e. university undergraduates Vs senior secondary school students), and ethnic groups in Nigeria (i.e. Hausa, Igbo, Yoruba and Minority). The goal of this analysis was to flag items that are potentially biased against one group in favour of another. The X-Calibre's Mantel-Haenszel statistical analysis tool, where each group is split into several ability levels, and the probability of a correct response compared between the groups for each level was applied to the results of the respondents to the APM scale test. The Mantel-Haenszel (M-H) coefficient is reported for each item as an odds ratio. The coefficient is a weighted average of the odds ratios for each θ level. If the odds ratio is less than 1.0, then the item is more likely to be correctly endorsed by one group than the other group(s). Likewise, odds ratios greater than 1.0 indicate that one group was more likely to correctly endorse the item than other group(s). According to Brouwers, Van de Vijver, & Van Hhemert (2009), 'the M-H coefficient is standardized through a log transformation, which is referred to as M-H DIF. The transformed value less than 0 indicates a reference group advantage whereas a value greater than 0 indicates the item is more likely to be correctly endorsed by a particular group than the other group or groups.' These ratios were used to determine if the DIF present in the responses to the APM scale was constant for all abilities (uniform DIF) or varied conditional on θ (crossing DIF). The M-H coefficient is not sensitive to crossing DIF, so null results were checked to confirm that crossing DIF was present or not present (Dorans & Holland 1993). Subsequently the X-Calibre ***z-test Statistic*** was also applied so that the negative of the natural logarithm of the M-H odds ratio was divided by its standard error to obtain the *z*-test statistic used to test the significance of the M-H against a null of zero DIF (odds ratio of 1.0). The two-tailed *p* value associated with the *z* test for DIF was then prorated. Items with *p* values less than .05 were flagged as having significant DIF. Thus the group that the item or items of the scale is/are ***Bias Against*** are flagged. This then is the group the item or items is/are disfavouring, or "biased against" when the *p* value is less than .05. In the context of the M-H test for DIF, the group that the item is disfavouring has a lower probability of a correct response than the other group, controlling for ability level. Below are the results of the analysis for each of the identified groups:

GENDER: MALES AND FEMALES

Table 6: Subgroup statistics for the Full Test

Subgroup	Examinees	Mean Theta	SD Theta
MALES	908	1.876	0.469
FEMALES	1092	1.891	0.472

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Table 6 above shows the gender subgroup statistics for the full test. The Meanθ and SDθ values for male with a total number of 908 are 1.876 and 0.469 respectively, while the Meanθ and SDθ values for female with a total number of 1092 are 1.891and 0.472 respectively. Both values are not far apart indicating no element of bias towards any gender. The APM scale was equally weighted and rated for both genders.

Table 7: Mantel-Haenszel's Item Parameters for All Calibrated Items for 2 Groups: Male Vs Female

Seq.	Item ID	P	R	a	B	c	Flag(s)
1	1	0.002	0.008	0.700	4.000	0.051	K, F, Hb
2	2	0.001	0.021	0.700	4.000	0.051	K, F, Hb
3	3	0.003	0.012	0.487	4.000	0.052	K, F, Hb
4	4	0.021	-0.050	0.392	4.000	0.062	K, F, Hb
5	5	0.002	-0.003	0.378	4.000	0.056	F, Hb
6	6	0.010	0.007	0.366	4.000	0.066	K, F, Hb
7	7	0.004	0.031	0.366	4.000	0.064	F, Hb
8	8	0.010	0.010	0.360	4.000	0.056	F, Hb
9	9	0.002	-0.003	0.365	4.000	0.051	K, F, Hb
10	10	0.009	0.004	0.359	4.000	0.056	K, F, Hb
11	11	0.008	0.036	0.360	4.000	0.055	K, F, Hb
12	12	0.003	-0.013	0.362	4.000	0.052	K, F, Hb
13	13	0.037	0.109	0.339	4.000	0.075	F, Hb
14	14	0.050	0.075	0.332	4.000	0.084	F, Hb
15	15	0.001	0.040	0.363	4.000	0.051	F, Hb
16	16	0.005	-0.040	0.360	4.000	0.053	K, F, Hb
17	17	0.005	-0.020	0.360	4.000	0.053	F, Hb
18	18	0.001	-0.014	0.362	4.000	0.051	K, F, Hb
19	19	0.001	-0.014	0.362	4.000	0.050	K, F, Hb
20	20	0.919	-0.046	0.120	-4.000	0.253	F, La, Lb
21	21	0.015	0.024	0.352	4.000	0.060	K, F, Hb
22	22	0.019	0.039	0.352	4.000	0.062	K, F, Hb
24	24	0.002	-0.003	0.361	4.000	0.051	K, F, Hb
25	25	0.006	0.024	0.359	4.000	0.054	K, F, Hb
26	26	0.012	-0.047	0.353	4.000	0.058	K, F, Hb
27	27	0.002	-0.021	0.360	4.000	0.052	K, F, Hb
28	28	0.009	-0.011	0.357	4.000	0.056	K, F, Hb
29	29	0.009	0.004	0.356	4.000	0.056	K, F, Hb
30	30	0.016	0.028	0.349	4.000	0.061	F, Hb
32	32	0.008	-0.013	0.354	4.000	0.057	K, F, Hb
33	33	0.019	0.000	0.344	4.000	0.067	F, Hb
34	34	0.026	-0.008	0.335	4.000	0.076	F, Hb

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

35	35	0.058	-0.044	0.313	4.000	0.110	K, F, Hb
36	36	0.061	-0.052	0.358	4.000	0.187	K, F, Hb

Again, as shown in Table 7 above, the Mantel-Haenszel's item parameter for all calibrated items assigned equal *b* parameter values to all items of the APM scale from the analysis of the scores generated from both genders. Similarly, the discriminatory, *a*, parameter did not show much discrepancies. Apart from items 1, 2, 3 and 20 which had *a* parameter values of 0.700, 0.700, 0.487 and 0.120 respectively, all the other *parameters* ranged from 0.313 to 0.392. This is an indication that the APM did not discriminate the test outcome in terms of group. These conclusively show that the items of the APM were bias free towards the gender groups of males and females. In other words the items are gender blind. The items are not biased towards any gender.

AGE: (UNIVERSITY UNDERGRADUATES VS SENIOR SECONDARY SCHOOL STUDENTS)

Table 8: Subgroup statistics for the Full Test

Subgroup	Examinees	Mean Theta	SD Theta
UG	1000	0.841	0.115
SS	1100	0.839	0.113

Table 8 above shows the school category subgroup statistics for the full test. The Meanθ and SDθ values for undergraduates (UG) with a total number of 1000 are 0.841 and 0.115 respectively, while the Meanθ and SDθ values for senior secondary (SS) with a total number of 1100 are 0.839 and 0.113 respectively. Both values are not far apart indicating no element of bias towards any age. The APM scale was equally weighted and rated for the two age groups.

Table 9: Mantel-Haenszel's Item Parameters for All Calibrated Items for 2 Groups: UG Vs *SS*

Seq.	Item ID	P	R	a	B	c	Flag(s)
1	1	0.002	0.008	0.594	4.000	0.051	K, F, Hb
2	2	0.001	0.021	0.594	4.000	0.051	F, Hb
3	3	0.003	0.012	0.454	4.000	0.052	F, Hb
4	4	0.021	-0.050	0.391	4.000	0.062	F, Hb
5	5	0.002	-0.003	0.381	4.000	0.056	F, Hb
6	6	0.010	0.007	0.374	4.000	0.066	F, Hb
7	7	0.004	0.031	0.374	4.000	0.064	F, Hb
8	8	0.010	0.010	0.370	4.000	0.056	F, Hb
9	9	0.002	-0.003	0.373	4.000	0.051	F, Hb
10	10	0.009	0.004	0.369	4.000	0.056	F, Hb
11	11	0.008	0.036	0.370	4.000	0.055	F, Hb
12	12	0.003	-0.013	0.371	4.000	0.052	F, Hb

13	13	0.037	0.109	0.356	4.000	0.075	F, Hb
14	14	0.050	0.075	0.352	4.000	0.084	F, Hb
15	15	0.001	0.040	0.372	4.000	0.051	F, Hb
16	16	0.005	-0.040	0.370	4.000	0.053	F, Hb
17	17	0.005	-0.020	0.370	4.000	0.053	F, Hb
18	18	0.001	-0.014	0.371	4.000	0.051	K, F, Hb
19	19	0.001	-0.014	0.371	4.000	0.050	K, F, Hb
20	20	0.919	-0.046	0.145	-4.000	0.253	F, La, Lb
21	21	0.015	0.024	0.364	4.000	0.060	F, Hb
22	22	0.019	0.039	0.364	4.000	0.062	F, Hb
24	24	0.002	-0.003	0.370	4.000	0.051	F, Hb
25	25	0.006	0.024	0.369	4.000	0.054	F, Hb
26	26	0.012	-0.047	0.365	4.000	0.058	F, Hb
27	27	0.002	-0.021	0.370	4.000	0.052	F, Hb
28	28	0.009	-0.011	0.367	4.000	0.056	F, Hb
29	29	0.009	0.004	0.367	4.000	0.056	F, Hb
30	30	0.016	0.028	0.363	4.000	0.061	F, Hb
32	32	0.008	-0.013	0.366	4.000	0.057	F, Hb
33	33	0.019	0.000	0.359	4.000	0.067	F, Hb
34	34	0.026	-0.008	0.353	4.000	0.076	F, Hb
35	35	0.058	-0.044	0.339	4.000	0.110	F, Hb
36	36	0.062	-0.049	0.405	4.000	0.152	F, Hb

Again, as shown in **Table 9**, the Mantel-Haenszel's item parameter for all calibrated items assigned equal *b*-parameter values to all items of the APM scale from the analysis of the scores generated from both age groups. Similarly, the discriminatory, *a*, parameter did not show much discrepancies. Apart from items 1, 2, 3, 20 and 36 which had *a* parameter values of 0.594, 0.594, 0.454, 0.145, and 0.405, respectively, all the other *a*, *parameters* ranged from 0.313 to 0.392. This is an indication that the APM did not discriminate the test outcome in terms of group. These conclusively show that the items of the APM were bias free towards the age groups of undergraduates of ages 16 to 40 years and senior secondary school students of ages 11 to 20 years. The items are not biased towards any age group on the account of age.

ETHNIC GROUPS: HAUSA, IGBO, YORUBA AND MINORITY

Table 10: Subgroup statistics for the Full Test

Subgroup	Examinees	Mean Theta	SD Theta
HAUSA	235	1.276	0.169
IGBO	650	1.491	0.272
YORUBA	461	1.331	0.209
MINORITY	754	1.587	0.367

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Table 10 shows the school category subgroup statistics for the full test. The Meanθ and SDθ values for Hausa with a total number of 235 participants are 1.276 and 0.169 respectively. The Meanθ and SDθ values for Igbo with a total number of 650 are 1.491 and 0.272 respectively. The Meanθ and SDθ values for Yoruba with a total number of 461 participants are 1.331 and 0.209 respectively. The Meanθ and SDθ values for Minority with a total number of 754 participants are 1.587 and 0.367 respectively. What can be reasonably deduced from the above iterations is that there is a close parity between the values of Meanθ and SDθ and therefore the claim of bias towards any of the ethnic groups is not sustainable. The APM scale was equally weighted and rated for all the ethnic groups.

Table 11: Mantel-Haenszel's Item Parameters for All Calibrated Items for 4 Groups: Hau, Igb, Yor & Min

Seq.	Item ID	P	R	a	B	c	Flag(s)
1	1	0.002	0.008	0.594	4.000	0.051	K, F, Hb
2	2	0.001	0.021	0.594	4.000	0.051	F, Hb
3	3	0.003	0.012	0.454	4.000	0.052	F, Hb
4	4	0.021	-0.050	0.391	4.000	0.062	F, Hb
5	5	0.002	-0.003	0.381	4.000	0.056	F, Hb
6	6	0.010	0.007	0.374	4.000	0.066	F, Hb
7	7	0.004	0.031	0.374	4.000	0.064	F, Hb
8	8	0.010	0.010	0.370	4.000	0.056	F, Hb
9	9	0.002	-0.003	0.373	4.000	0.051	F, Hb
10	10	0.009	0.004	0.369	4.000	0.056	F, Hb
11	11	0.008	0.036	0.370	4.000	0.055	F, Hb
12	12	0.003	-0.013	0.371	4.000	0.052	F, Hb
13	13	0.037	0.109	0.356	4.000	0.075	F, Hb
14	14	0.050	0.075	0.352	4.000	0.084	F, Hb
15	15	0.001	0.040	0.372	4.000	0.051	F, Hb
16	16	0.005	-0.040	0.370	4.000	0.053	F, Hb
17	17	0.005	-0.020	0.370	4.000	0.053	F, Hb
18	18	0.001	-0.014	0.371	4.000	0.051	K, F, Hb
19	19	0.001	-0.014	0.371	4.000	0.050	K, F, Hb
20	20	0.919	-0.046	0.145	-4.000	0.253	F, La, Lb
21	21	0.015	0.024	0.364	4.000	0.060	F, Hb
22	22	0.019	0.039	0.364	4.000	0.062	F, Hb
24	24	0.002	-0.003	0.370	4.000	0.051	F, Hb
25	25	0.006	0.024	0.369	4.000	0.054	F, Hb
26	26	0.012	-0.047	0.365	4.000	0.058	F, Hb
27	27	0.002	-0.021	0.370	4.000	0.052	F, Hb
28	28	0.009	-0.011	0.367	4.000	0.056	F, Hb

29	29	0.009	0.004	0.367	4.000	0.056	F, Hb
30	30	0.016	0.028	0.363	4.000	0.061	F, Hb
32	32	0.008	-0.013	0.366	4.000	0.057	F, Hb
33	33	0.019	0.000	0.359	4.000	0.067	F, Hb
34	34	0.026	-0.008	0.353	4.000	0.076	F, Hb
35	35	0.058	-0.044	0.339	4.000	0.110	F, Hb
36	36	0.062	-0.049	0.405	4.000	0.152	F, Hb

Again, as shown in **Table 11** above, the Mantel-Haenszel's item parameter for all calibrated items assigned equal *b* parameter values to all items of the APM scale from the analysis of the scores generated from all ethnic groups. Similarly, the discriminatory, *a*, parameter did not show much discrepancies. Apart from items 1, 2, 3, 20 and 36 which had *a* parameter values of 0.594, 0.594, 0.454, 0.145, and 0.405, respectively, all the other *a, parameters* ranged from 0.313 to 0.392. This is an indication that the APM did not discriminate the test outcome in terms of ethnic group. These conclusively show that the items of the APM were bias free towards the ethnic groups in Nigeria: Hausa, Igbo, Yoruba and Minority. The items are not biased towards any group on the account of ethnicity or culture.

Probability of Guessing (c-parameter) of APM items in Nigeria

The *c* parameter equals the probability of an examinee of infinitely low θ obtaining a correct response due to guessing. Thus, *c* is also the lower asymptote of the IRF. The inclusion of a non-zero *c* parameter affects the location of *a* and *b* on the θ scale. The *c* parameter is expected to equal approximately 1 divided by the number of alternatives for multiple-choice tests. Therefore, for the APM that has 8 alternatives, a low examinee should have 1/8 = 0.125 chance of guessing the correct answer. Since *c* = 0.125 for this 8-alternative item, once the right key is isolated, the examinees will be guessing among the remaining seven options. Therefore where guessing is pronounced, the value of *c* will be much lower than 0.125. Higher value will mean that guessing is not strongly evidenced. From Table 3 already presented above, *c parameter* values range from 0.189 to 0.617 indicating limited or no guessing. Therefore the degree of guessing can be said to be low amongst the undergraduates as well as the senior secondary school students.

Concurrent validity of APM, using Culture Fair Intelligence as the criterion.

Concurrent validity of a test under validation or under development is simply established by correlating it with another test that validly and reliably measures the same trait or domain in the population that the test is being validated or developed for. It was for this reason that the CFIT that validly and reliably measures Fluid General Intelligence in Nigeria and (Kpolovie, 2015; 2005; 2003; 1999) was simultaneously administered with the APM to the subjects of the current study. The correlation coefficient between the APM and CFIT is the concurrent validity of the APM.

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Table 12: Concurrent validity of APM

Correlations

		CFIT	APM
CFIT	Pearson Correlation	1	.701**
	Sig. (2-tailed)		.000
	N	1509	1509
APM	Pearson Correlation	.701**	1
	Sig. (2-tailed)	.000	
	N	1509	1509

**. Correlation is significant at the 0.01 level (2-tailed).

The results have shown that the correlation coefficient between the two tests (CFIT and APM) is 0.701, and therefore the Advanced Progressive Matrices has a concurrent validity of 0.701 in Nigeria. Out of the 2100 subjects of the study, only 1509 completed the two tests. This accounts for why the number of cases (N) in the output is 1509 rather than 2100. A correlation coefficient of 0.701 for a sample that is as high as 1509 is a very strong correlation and a stunning indication of concurrent very high concurrent validity of the APM that was under validation in this investigation.

CONCLUSION

This investigation has successfully provided empirically verifiable and replicable answers to the posed nine research questions. The findings will be of immense significance to the validation population in a number of ways. For instance, the use of the APM in Nigeria will enhance easy identification, placement, acceleration and enrichment of the gifted/talented students in Nigerian; and in ability selection and placement within and outside the education management climes. The study will with a little or no doubt serve as an epoch in the annals of measurement practices in this part of the world as it successfully used IRT to validate the APM in a population that the test has hitherto never been employed for decision-making. The validation of APM in Nigeria has localized the appropriate use of the test in the country; and thus freed the test from being ‘a foreign instrument’ (Carlson, Geisinger, & Jonson, 2014) for the research community, the behavioral sciences, medical practitioners and decision makers in the most populous black nation where education has received the least attention (OECD, 2015; FRN National Population Commission, 2014; Kpolovie & Obilor, 2013 a; b; c). Findings of this study have clearly showed the complementary role of the IRT and CTT. The findings about the suitability of 3-Parameter Logistics Model is an eye opener to the fact that a perfect fit statistic can be achieved for the enhancement of the credibility of APM as a suitable measuring instrument in Nigeria; clearly showing the directions and dimensions of the items and how well the items are performing the intended objectives that they were designed to serve. The APM has appropriate difficulty, discrimination and guessing indexes; and of suitable validity and reliability; in addition to being bias-free in Nigeria.

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

REFERENCES

Abad, F. J., Colom R., Rebollo I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences*, 36 (2004), 1459–1470

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30–60

Arthur, W. & Day D. (1994) Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, 54 (1994), 395–403

Bennett, C.K. (1969). *Bennett mechanical comprehension test*, San Antonio: The Psychological Corporation.

Bors, D.A. Stokes T.L. (1998). Raven's Advanced Progressive Matrices: Norms for the first-year university students and the development of a short form Educational and Psychological Measurement, 58 (3) (1998), pp. 382–398

Brouwers, S. A., Van de Vijver, F. J. R., & Van Hemert, D. A. (2009). Variation in Raven's Progressive Matrices scores across time and place. Learn and Individual Differences. 19, 330-338.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (2014). The ninethenth mental measurements yearbook. Nebraska: The Buros Center for Testing, The University of Nebraska Press.

Dorans, N. J., & Holland, P. W. (1993). Detection and Description: Mantel-Haenszel and standardization. In P. W. Holland & Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum. Pp. 35-66.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54, 1-22.*

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012, March 26). Using the Advanced Progressive Matrices (Set I) to Assess Fluid Ability in a Short Time Frame: An Item Response Theory–Based Analysis. *Psychological Assessment*. Retrieved July 15, 2012, from http://my.apa.org/portal.

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2011).Item response theory analysis and differential item functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learning and Individual Differences.* Retrieved June 3, 2012, from http://www.sciencedirect.com/science/article/

Chiesi, F., & Primi, C. (2009). Recency effects in primary-age children and college students using a gaming situation. *International Electronic Journal of Mathematics Education, 4, 3,* 259 –274. Retrieved June 3, 2012 from www.iejme.com

Chiesi, F., Primi, C., & Morsanyi, K. (2011). Developmental changes in probabilistic reasoning: The role of cognitive capacity, instructions, thinking styles and relevant knowledge. *Thinking & Reasoning, 17,* 315–350 Retrieved June 3, 2012, from http://www.sciencedirect.com/science/article/

.Çikrikçi-Demįrtaşli, N. (2000). A study of Raven Standard Progressive Matrices Test's item measures under Classical and Item Response Model Paper presented at 31st European Mathematical Psychology Congress, Austria: Graz. Retrieved June 3, 2012, from http://www.sciencedirect.com/science/article/

Colom, R., Florez-Mendoza, C., & Rebollo, I. (2003). Working memory and intelligence. *Personality and Individual Differences, 34,* 33–39

Evers, A. (2011) Testing practices and attitude towards tests and testing: The results of a global survey Paper presented at the 12th European Congress of Psychology, Istanbul (2011, July). Retrieved July 15, 2012, from http://my.apa.org/portal

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

FRN National Population Commission (2014). *Nigeria Demographic and Health Survey 2013*. USA: ICF International.

Federal Ministry of Education (2014). *Staff and students data in our educational institutions*. Abuja: FME press.

Field, A. (2005). *Discovering statistics using SPSS for windows*. London: Sage Publication.

Gallini, J.K., (1983). A Rasch analysis of Raven item data The Journal of Experimental Education, 52 (1) (1983), pp. 27–32.Retrieved July 15, 2012, from http://my.apa.org/portal

Georgiev, N., (2008) Item analysis of C, D and E series form Raven's Standard Progressive Matrices with Item Response Theory two-parameter logistic model Europe's Journal of Psychology (2008) http://www.ejop.org/archives/2008/08/item_analysis_o.html

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment.* New York, NY: CambridgeUniversity Press.

Herrnstein, R. J & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: The Free Press.

International Test Commission (2000). International guidelines for test use. Downloaded electronically from www.intestcom.org/itc-projects.htm

Joe, I. A. (2000). *Fundamental Statistics for Education and the behavioural sciences*. Ibadan: Kraft Books Ltd.

Kpolovie, P. J. (2012). *Education Reforms without Evaluation Designs: Nigeria at Risk,* New Owerri: Springfield Publishers Ltd.

Kpolovie, P. J. (2011), *Statistical Techniques for Advanced Research*. New Owerri: Springfield Publishers Ltd.

Kpolovie, P. J. (2010). *Advanced Research Methods*, New Owerri: Springfield Publishers Ltd.

Kpolovie, P.J. & Obilor, I. E. (2013a). Higher education for all in need through the National Open University of Nigeria: A paradox in policy practice. *Merit Research Journal of Education and Review (MRJER).* http://meritresearchjournals.org/er/content/2013/September/Kpolovie%20and%20Obilor.pdf ISSN: 2350-2282. 1(8), 172-180.

Kpolovie, P.J. & Obilor, I. E. (2013b). Nigerian universities bag ludicrous ranks in world rankings of universities. *Universal Journal of Education and General Studies*. http://www.universalresearchjournals.org/ujegs 2(9), 303-323.

Kpolovie, P. J. & Obilor, I. E. (2013c). Adequacy-inadequacy: Education funding in Nigeria. Universal Journal of Education and General Studies. http://www.universalresearchjournals.org/ujegs ISSN: 2277-0984. 2(8), 239-254.

Kpolovie, P. J. (2014). *Tests, Measurement and Evaluation in Education. Second Edition*. Owerri: Springfield Publishers Ltd.

Kpolovie, P. J. (2016). *Excellent research methods*. Indiana: Partridge Publishing.

Kpolovie, P. J. (2012). *Education Reforms without Evaluation Designs: Nigeria at Risk*. Owerri: Springfield Publishers Ltd.

Kpolovie, P. J. (2015). Indispensability of intelligence testing in the repositioning and revitalization of Nigerian education. *Multidisciplinary Journal of Research Development*. 6(4), 1-11. http://www.globalacademicgroup.com/journals/nard/Kpolovie.pdf

Kpolovie, P. J. (1999). Reliability of the Culture Fair Intelligence Test in Nigeria. *Journal of Education in Developing Areas (JEDA).* A Journal of the Faculty of Education, University of Port Harcourt. Nigeria. XIII, 18-24.

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Kpolovie, P. J. (2003). Construct validity of Culture Fair Intelligence Test (CFIT) in Nigeria. *Nigerian Journal of Research and Production.* A Journal of the Nigerian Research Forum. 2(5), 130-139.

Kpolovie, P. J. (2005). Nigerian norms of Culture Fair Intelligence Test. *International Journal of Communication and Humanistic Studies.* ISSN 1597-1716. 2(1), 1-16.

Michell, J. (1999). *Measurement in Psychology*. Cambridge: Cambridge University Press.

Mogle, J. A., Lovett, B. J., Stawski, R. S., & Sliwinski, M. J. (2008). What's so special about working memory? An examination of the relationship among working memory, secondary memory, and fluid intelligence. *Psychological Science*, 19, 1071-1077.

Morsanyi, K., Primi, C., Chiesi, F. & Handley, S.J., (2009).The effects and side-effects of statistics education. Psychology students' (mis-)conceptions of probability *Contemporary Educational Psychology*, 34 (2009), pp. 210–220

Muñiz J., (2011), International strategies to improve tests and testing. Paper presented at the 12th European Congress of Psychology, Istanbul (2011, July)

Muñiz, J., (2009), The role of EFPA in setting standards for tests and test use Paper presented at the 11th European Congress of Psychology, Oslo (2009, July). Retrieved June 3, 2012, from http://www.sciencedirect.com/science/article/

NEEDS Assessment of Nigerian Universities Committee (2013). *NEEDS Assessment of Nigerian Universities Reports*. Abuja: FME Press.

OECD (2015). *Education at a glance 2015 OECD indicators*. UK: OECD Publishing. http://dx.doi.org/10.1787/eag-2015-en

Ojerinde, D.; Popoola, K; Ojo, F.; and Ariyo, A. (2014). *Practical applications of item response theory in large-scale assessment*. Nigeria: Marvelous Mike Press Limited.

Ojerinde, D.; Popoola, K.; Ojo, F.; and Onyeneho, P. (2012). *Introduction to item response theory: Parameter models, estimation and application*. Nigeria: Marvelous Mike Press Limited.

Ololube, N. P. and Kpolovie, P. J. (2013). Literature and focus group analysis of the approaches and obstacle to effective educational planning in higher education in an emerging economy. *International Journal of Scientific Research in Education (IJSRE).* http://www.ijsre.com/Vol.,%206 3-Ololube%20&%20 Kpolovie.pdf 6(3): 233–254.

Ololube, N. P. and Kpolovie, P. J. (2012). Approaches to conducting scientific research in education, arts and the social sciences. *Online Journal of Education Research* 1(3): 44–56. http://onlineresearchjournals.org/IJER/pdf/2012/june/ Ololube%20and%20Kpolovie.pdf

Ololube, N. P., Emejuru, P. I., Kpolovie, P. J., Amaele, A., and Uzorka, M. C. (2012). Excellence in higher education: a powerful instrument for social and economic progress. In, N. P. Ololube and P. J. Kpolovie. *Educational Management in Developing Economies: Cases 'n' School Effectiveness and Quality Improvement*. Saarbucken, Germany: LAP LAMBERT Academic Publishing. ISBN 978-3-8465-8931-1. http://www.amazon.com/EducationalManagement-Developing-Economies-Effectiveness/dp/3846589314 93–112.

Ololube, N. P., Kpolovie, P. J. and Makewa, L. N. (2015). Handbook of Research on Enhancing Teacher Education with Advanced Instructional Technology. PA, USA: *Information Science Reference (an imprint of IGI Global)*. ISBN 13: 978146668162; EISBN 13: 9781466681637; DOI: 10.4018/978-1-4666-8162-0 http://www.igi-global.com/book/handbook-research-enhancing-teacher-education/120264

Orluwene, G. W. (2012). *Introduction to Test Theory and Development Process*. Port Harcourt: Lucky Press Ltd.

Pearson NCS (2011). *Raven's Advanced Progressive Matrices International Technical Manual*. UK: NCS Pearson, Inc.

Pearson NCS (2007). *Raven's Advanced Progressive Matrices (APM) evidence of reliability and validity*. UK: Pearson NCS Inc.

Rathus, S. A. (1990). *Psychology: Annotated instructor's edition*. Chicago: Hoilt, Rinehart and Winston.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1-48. Retrieved June 15, 2012, from http://www.wpspublish.com

Raven, J.C., (1962). *Advanced progressive matrices*, London: Lewis & Co. Ltd.

Raven, J. C., Court, J. H., & Raven, J. *(1983). Manual for Raven's Progressive Mamas and vocabulary scales, section* 4. *Advanced Progressive Matrices, Sets I and II.* London: H. K. Lewis.

Raven, J.C. Raven J. (Eds.), (2008) *Uses and abuses of intelligence: Studies advancing Spearman and Raven's quest for non-arbitrary metrics*, Royal Fireworks Press, Unionville, New York.

Raven, J., Raven, J. C., & Court, J. H. (2003). *Raven manual section 1: General overview*. Oxford: Oxford Psychologists Press.

Raven, J., Raven, J. C., & Court, J. H. (2012). *Raven manual section 1: General overview*. Oxford: Oxford Psychologists Press.

Raven, J, Raven, J.C. & Court, J. H. (1997). *Mill Hill Vocabulary Scale: 1998 Edition.* Oxford: Oxford Psychologists Press.

Raven, J., Raven, J. C., & Court, J. H. (1993). *Raven manual section 1: Generaloverview*. Oxford: Oxford Psychologists Press.

Raven, J., Raven, J.C. & Court, J.H. (1998). *Advanced Progressive Matrices*. Oxford: Oxford Psychologists Press.

Rushton, J. P. Skuy, M & Bons, T. A. (2004). Construct Validity of Raven's Advanced Progressive Matrices for African and Non-African Engineering Students in South Africa. International Journal of selection and assessment volume 12 number 3 September 2004, Retrieved May 23, 2012, from *www.charlesdarwinresearch.org/Ravens%20inIJSA04.pdf*

Sefcek, J. A., (2007). Development and of an 18-item short form of the Ravens Advanced Progressive Matrices (RAPM- 18). Retrieved June 9, 2012, fromhttp://www.highbeam.com/doc/1G1-108790099.html

Spearman, C. (1927).*The abilities of man*. London: Macmillan.

Thompson, B. (2004). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas. In B. Thompson (Ed.), *Advances in social science methodology* (pp. 23-86). Stamford, CT: JAI Press.

Van der Ark, A. A. (2010). Using item response theory to score the Advanced Progressive Matrices: Rationale and research findings. *Journal of Psychological Type*, 37, 16-60.

Van der Ven A.H.G.S. & Ellis, J.L., (2000). A Rasch analysis of Raven's Standard Progressive Matrices Personality and Individual Differences, 29 (2000), pp. 45–64.

Verguts, T. and De Boeck, P. (2002). The induction of solution rules in Raven's Progressive Matrices Test. *European Journal of Cognitive Psychology*. 14(4), 521-547.

Vigneau, F. Bors D.A. (2005) Items in context: Assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement*, 65 (1) (2005), pp. 109–123.

ISSN 2055-0863(Print), ISSN 2055-0871(Online)

Wechsler, D. (2013). *WISC–III.Wechsler Intelligence Scale for Children-Terza Edizione-Manuale* [*WISC–III. Wechsler Intelligence Scale for Children-Third edition-Handbook*]. Florence, Italy: Organizzazioni Speciali.
WPS (2015). *Western Psychological Services California*. Retrieved from www.wpspublish.com
Wikipedia, (2012). Item Response theory. *Retrieved June 3, 2012 from www.**wikipedia**.org/*

ISSN 2055-0863(Print), ISSN 2055-0871(Online)