Equivalency of Perceived Equivalent Psychometric Instruments, the Complexity of Technol ogical Innovations in Instrument Development: Evidences from Utme 2013 Mathematics Computer-Based and Paper-Pencil Tests

Anagbogu German Effa and Idajor, Clement Ojim
Department of Educational Foundations
Faculty of Education, University of Calabar, Nigeria
anagbogu@yahoo.com

Abstract
The recent technological innovations in the psychometric space in Nigeria has seen JAMB's introduction of CBT to replace the traditional PPT which has been their test form from the point of establishment. This study investigated the b-parameter (item difficulty) of (PPT) and (CBT) of UTME 2013 Mathematics examination, with their test information functions with the aim of determining their equivalence and comparability. Evaluation research design was adopted for the study and simple random sampling technique was used to select sample for the study. The sample for the study consisted of 12,991 candidates responses in Mathematics drawn from JAMB data base in Abuja being ten percent of the candidates for PPT and CBT. One research question was posed and answered using descriptive statistics of percentages, graphs, and histogram accordingly. The research question was translated to hypothesis, and tested using Chi-Square (X^2). Results of the analysis showed that, UTME Computer-Based Test differ from the Paper-Pencil Test in their difficulty indices when calibrated in their test information functions and the Chi-Square analysis showed that the b-parameter of PPT and CBT differ significantly. Based on the results of the findings, it was recommended among other things that examination bodies should avoid a situation where more than one test forms are administered for the same purpose because of the differences that may result from test mode effect. It was suggested that further research can also be carried out on other parameters to ascertain their differences in such parameters.

Keywords: Equivalent, Psychometric, Instruments, Complexity, Technological, Innovations, Development

1. Introduction

Technological innovations have been implemented in different spheres of education ranging from teaching, learning and research both in and out of the classroom setting. People can now access, store and process educational materials anywhere in the world. All credit to technological advancement in education and learning process. The wave of technological innovations in the educational sector is also being utilized for assessment process, hence the recent changes that have occurred in the field of educational assessment, measurement and evaluation. The shift from the traditional Paper–Pencil Based Test (PPT) to Computer-Based Test (CBT) is one of such adoption/use of new technology for educational assessment, which allows the use of computers to develop, process and deliver different types of assessment. Uysal and Kuzu, as cited in Janil and Shami (2012) reported that the rapid advancement of Information and Communication Technologies (ICTs) in teaching and learning has shifted the paradigm from PPT to CBT. There is also a pedagogical change among the people of the need for the inclusion of technology in the field of assessment and this has cumulated to the introduction of computer-based testing by JAMB in Nigeria.

When Joint Admission and Matriculation Board (JAMB) started in 1978, it conducts entrance examination for placement into universities in Nigeria by issuing examination booklets to candidates, the question for all the subjects were contained in a single booklet and four Optical Mark Readable (OMR) sheets serve as answer sheets for the four subjects attempted by each candidate. Later, each candidate was given single question paper booklet containing questions to the four subjects registered alongside one single OMR sheet as answer sheet (JAMB, 2014). In 2013, JAMB introduced three (3) examination modes in the conduct of its Unified Tertiary Matriculation Examination (UTME). The three modes were: Paper-Pencil Test (PPT), Dual Based Test (DBT) and Computer Based Test (CBT). Test takers were free to choose from the three modes of test suggesting that all the three test modes are equivalent, thus serve same purpose (JAMB report, 2014).

Computer-Based Testing (CBT) allows prospective test takers to take the JAMB UTME using a computer in the board's accredited centers. Just like PPT, the CBTs are taken in secured centers by the board and certified fit for such examination. It allows candidates to access questions only on the computer and at the same time take options directly from the computer as

against the PPT where candidates are provided with booklets containing questions and answer sheets to write on. To get the process of CBT started on smooth sail, Professional Testing Inc (PTI, 2006) had said that, it is important and needful to get CBT stakeholders involved in the transition processes of testing. They observed that one component of getting ready to move an examination programme from PPT to CBT is the development of a proper plan for orienting stakeholders concerning the major issues in the new test. Sad enough, observations revealed that, the suggestions by Professional Testing Incorporated were not adequately followed by JAMB in the introduction of CBT. We all woke to the sudden new innovation of testing and test takers make choices out of blatant ignorance, the mode of the test that was plausible to them.

One of the gray area in assessment remains the comparability of scores across different test forms/modes of standardized assessment, and to the researcher, it should be a core objective of any examination body in Nigeria to accurately and fairly measure and compare educational skills using multiple test forms/modes from an educational assessment. Accordingly, the current study will seek to compare the two test modes under consideration bearing the difficulty indices, calibrating the responses of candidates on the Mathematics test in other to ascertain their comparability. It is worthy of note here that, such exercise is necessary in standardized educational assessment that uses numerous test forms/modes which tends to differ in difficulty even when there are built to the same specification or what you may call parallel test forms. Test forms are being designed based on the same specifications, to cover the same content, at the same ability level, but the score scales and alternate forms are not identical (Ojerinde, Popoola, Ojo&Onyeneho, 2012).

Thus, as psychometricians attempt to adopt technology for instrument development, the greater task for them remains the imbedded complexity of making the new instrument equivalent with that which technology tends to replace. In essence, a properly constructed instrument must test the appropriate knowledge, skills and abilities; it must do so consistently for all test-takers and must avoid any bias that could tamper with the result. If CBT and PPT are to be administered side-by-side to test-takers for the purpose of passing the same judgment, psychometricians must make every effort to adhere to the scientific approaches to analyze and review the test items and construction to ensure they possess the desired difficulty which will offer every test-taker, equal opportunity, a comparable testing experience and a degree of score

integrity.

ITEM DIFFICULTY

Item difficulty could be referred to as both p-value in classical test theory (CTT) and b-parameter in item response theory (IRT). When test-takers refer to difficulty of test, they are referring to the extent to which their ability or amount of success, in the area or domain tested is being challenged by the item(s)/test. However getting the right difficulty level for items is one of the challenging tasks for psychometricians and test practitioners in items writing. Professionally what is needful is to carryout pretesting of items on samples of test-takers almost similar to those in the target populating for gaining initial objective data on item difficulty level (Anderson & Morgan as cited in Greaney & Kellaghan, 2008.ed). This can help guide against common errors in item(s) writing or developing tests with items that prove to be much too difficult or too easy for the targeted examinees.According to Gronlund as cited in Joshua (2005), item difficulty refers to item easiness index. i.e the extent to which an item is cheap or hard for respondents. It is the proportion of test-takers that respond correctly to the item. Difficulty also known as p-value is relative to the ability level of the test-taker. It is concerned with the performance of the total sample, that is, the number of students' or test-takers that got the item correctly/right divided by the number of test-takers who responded to the item. The difficulty indices (p-value) in a classical test situation vary from zero (0) for a very difficult item (nobody got the item right) to one (1) for a very cheap item (everybody got the item right), hence, the higher the difficulty index of an item, the easier the item and the lower the difficulty index the harder the item (Joshua 2005). It is important to note here that, in the context of IRT; difficulty is referred to as b-parameter and is discussed in a similar way as the concept of p-value in ICT.

According to Brawn as cited in Anagbogu (2009), the procedure used in an item analysis depends on several factors, the type of items and test (i.e. if the test is multi-choice type, easy based or other type) the numbers of test takers, the available computational facilities and above all, the purpose of the analysis. The process involves counting all the number of examines who got each item correctly and converted into proportions or percentages in the CCT situation. The b-parameter is item difficulty parameter also called item locations by some psychometricians or item threshold. This is where the inflection on the ability scale is, a point on the scale wherein examinees have 50% probability of correctly answering the item (Ojerinde, Popoola, Ojo &

Onyenaho 2012). They noted, that theoretically difficulty values in IRT range from negative infinity to positive infinity, but in practice, value usually are in the range of (-3 to +3) when theta (θ) is scaled to have a mean 0 and standard deviation of 1.0.

Thus, they problem that prompted this research is the fact that, JAMB set the same cut-off make for all the test takers not minding the test modes taken by the candidates for an examination they could not satisfactorily establish their equivalency to those concern deserve further scrutiny to confirm or disprove the claim of equivalency by the board in terms of the item difficulty. The justification for this feeling by the researcher was further strengthened when in March 2015, the National Dailies were painted with headlines like: "protest marred JAMB CBT", "Candidates protest CBT inconveniences" and of particular note was the report in People's Daily of March 12, 2015, that over 50 JAMB candidates were arrested in Benue state over CBT protest. Notwithstanding the candidates' lack of confidence in the process, no further steps to address their grievances were put before the candidate. Because one would feel there exist some difference in the instrument and process. The current research is aimed at shedding light on the fuzzy picture painted above which are bugging.

The study is a detailed and step-by-step evaluation of the difficulty parameter of PPT and CBT of UTME in 2013 Mathematics in order to make meaningful comparison using the test information function, and ascertain the appropriateness of JAMB conducting the two modes of test in one year. Specifically, the study established if the UTME 2013 Mathematics Computer-Based Test and Paper-Pencils Test differ in the frequencies of the test items that belong to the various intervals of the calibration of the b-parameter (item difficulty) in their test information function among candidates.

Research questions

One research question guides the study:

1. What are the differences among the frequencies of the test items that belong to the various intervals of the calibrations of the b-parameter for the CBT and PPT groups of candidates?

Research hypothesis

H_o: The mean of all calibrated b-parameter for PPT group does not differ significantly from the mean of all calibrated b-parameter for CBT group ($P<0.05$)

2. Research Methodology

The study adopted evaluation research design. The researcher adopted this design because the research investigated, assessed and appraised the equivalence of the two testing mode (PPT and CBT) instruments to ascertain some level of certainty concerning their difficulty indices (b-parameters). Evaluation research according to Odinko (2014) is a structured process of measuring and assessing the success of a project in meeting its goals and to reflect on the lessons learned. The area of study comprises of all the areas covered by JAMB as an examination body, (JAMB-UTME is conducted nationwide and in some selected diasporas centers in countries where JAMB has partnership).The population of the study comprises of the entire candidates who sat for 2013 JAMB Mathematics examination in the two test modes (CBT & PPT). A total of one-hundred and twenty-nine thousand, nine hundred and ten (129,910) sat for the two forms of test for Mathematics in 2013. Ninety-nine thousand, nine hundred and thirty sat for PPT and twenty-nine thousand nine hundred and eighty sat for CBT that same year. The study adopted the stratified random sampling techniques for sample selection. The test was first stratified into modes of administration (CBT and PPT) and ten per cent of each test form was randomly selected to form the sample with the help of a computer program (X-calibre4.2) for adequate representation. The program gave equal opportunity to the candidate's responses on JAMB data base for Mathematics to be picked for the study. The sample was made up of twelve thousand, nine hundred and ninety-one (12,991) candidates' responses who sat for UTME Mathematics2013 in the CBT and PPT modes. Ten percent of the candidates' responses in each test modes were selected for the study. Further breakdown shows that 9993 were selected from PPT and 2998 were selected from CBT being the total ten per cent.

Instrumentation

The instrument for the study was the UTME question type Q for both PPT and CBT which was taken by majority of the candidates. The instrument is made up of 50 items drown from JAMB Mathematics syllabus, constructed and validated by experts in the board which makes it a standardized instrument. For PPT, the instrument is hard-copy (i.e. produced on paper) but the CBT was delivered online via computer program and thus, only the responses to the instrument recorded in the JAMB data based that was obtained and used. The instruments are standardize tests, thus, the validity of the instruments were established by the board and assumed

valid given the known testing practice in JANB. It is a standard practice in JAMB to try-test all test items and calibrate for item parameters and their appropriateness for the category of test-takers as part of their quality assurance. Recently, JAMB relied on IRT software for their item calibration for reliability of their instruments. However, a confirmatory reliability coefficient from the data calibrated for CBT was .84 and that of PPT was 0.94 using the same software as used by JAMB (Xcalibre).

3. Results and Discussion

To answer the research question, the b parameter of the 50 calibrated items for PPT and CBT were used to determine the extent to which the two tests defer using descriptive statistics. The various mean and standard deviation for the b-parameter and that of the total items are presented in summary from the calibrated statistics.

ITEM	PPT	CBT	DIFF
MEAN TOTAL (X_T)	15.46	13.21	2.25
STANDARDDEVIATION TOTAL (SD_T)	11.17	7.18	3.99
MEAN b (X_b)	0.94	1.62	0.68
SD_b	0.61	0.52	0.09
X_θ	0.02	0.07	0.05
SD_θ	1.03	1.02	0.01
MAXIMUM θ	7.00	2.66	4.34
MINIMUMθ	-7.00	-7.00	0.00
MAXIMUMb	2.46	2.44	0.02
MINIMUM b	0.22	0.19	0.03

DIFF: line showed the differences in the mean of the total items (2.25) and b-mean of (0.68). Same cases appeared in the minimum and maximum values

TABLE 1

A comparative look at the b parameter of PPT and CBT with the associated properties

CBT b PARAMETER FOR ALL ITEMS

Seq.	Item ID	P	R	b	Flag(s)
1	1	0.185	0.462	1.816	K, F
2	2	0.386	0.102	1.274	
3	3	0.278	0.214	2.283	K, F
4	4	0.226	0.392	1.663	F
5	5	0.146	0.211	2.331	K, F
6	6	0.232	0.294	1.871	K
7	7	0.381	0.251	1.778	K, F
8	8	0.216	0.345	1.691	K
9	9	0.219	0.386	1.659	K
10	10	0.294	0.319	1.645	K
11	11	0.171	0.377	1.968	K
12	12	0.291	0.259	1.998	K, F
13	13	0.213	0.381	1.655	K, F
14	14	0.282	0.315	1.599	
15	15	0.217	0.304	1.010	F
16	16	0.199	0.401	1.655	F
17	17	0.251	0.321	1.583	
18	18	0.222	0.364	1.656	F
19	19	0.249	0.301	1.266	
20	20	0.208	0.389	1.574	F
21	21	0.284	0.266	1.216	
22	22	0.289	0.323	1.417	
23	23	0.257	0.110	2.405	
24	24	0.070	0.186	2.406	K, F
25	25	0.412	0.241	1.578	F
26	26	0.272	0.317	1.641	K, F
27	27	0.214	0.377	1.662	K, F
28	28	0.507	0.192	0.382	
29	29	0.346	0.235	0.906	
30	30	0.182	0.407	1.632	F
31	31	0.229	0.374	1.675	K, F
32	32	0.266	0.315	1.632	F
33	33	0.318	0.293	1.509	F
34	34	0.320	0.286	1.620	F
35	35	0.419	0.224	0.950	
36	36	0.176	0.101	2.434	K, F
37	37	0.185	0.070	2.437	K, F
38	38	0.461	0.131	0.955	
39	39	0.162	0.315	1.761	
40	40	0.576	0.098	0.185	
41	41	0.155	0.398	1.813	K, F
42	42	0.217	0.370	1.746	K, F
43	43	0.154	0.183	2.434	K
44	44	0.348	0.228	0.728	
45	45	0.151	0.480	1.663	K, F
46	46	0.255	0.337	1.667	
47	47	0.295	0.102	2.424	K
48	48	0.221	0.393	1.647	
49	49	0.224	0.281	1.971	K
50	50	0.378	0.204	0.651	

PPT b PARAMETERS FOR ALLITEMS

Seq.	Item ID	P	R	b	Flag(s)
2	2	0.443	0.616	-0.017	F
3	3	0.371	0.559	0.376	F
4	4	0.522	0.500	-0.266	F
5	5	0.391	0.596	0.204	F
6	6	0.288	0.597	0.644	F
7	7	0.309	0.593	0.526	F
8	8	0.346	0.558	0.498	
9	9	0.574	0.129	1.066	K, F
10	10	0.251	0.460	1.047	F
11	11	0.195	0.356	1.629	F
12	12	0.307	0.600	0.528	F
13	13	0.284	0.602	0.636	F
14	14	0.237	0.497	1.041	F
15	15	0.308	0.573	0.634	F
16	16	0.314	0.520	0.708	F
17	17	0.337	0.447	0.811	F
18	18	0.217	0.150	1.720	K, F
19	19	0.180	0.361	1.559	F
20	20	0.242	0.449	1.154	F
21	21	0.358	0.588	0.328	F
22	22	0.394	0.637	0.135	F
23	23	0.289	0.547	0.720	F
24	24	0.282	0.558	0.831	F
25	25	0.397	0.569	0.205	F
26	26	0.348	0.606	0.388	F
27	27	0.287	0.493	0.918	F
28	28	0.288	0.437	1.057	F
29	29	0.203	0.203	2.459	F
30	30	0.189	0.082	2.459	K, F
31	31	0.201	0.492	1.256	F
32	32	0.206	0.129	2.459	F
33	33	0.228	0.430	1.296	F
34	34	0.340	0.573	0.516	F
35	35	0.261	0.484	1.069	F
36	36	0.254	0.466	1.095	F
37	37	0.228	0.419	1.332	F
38	38	0.220	0.387	1.440	F
39	39	0.220	0.334	1.657	F
40	40	0.347	0.587	0.366	F
41	41	0.198	0.335	1.704	F
42	42	0.348	0.629	0.334	F
43	43	0.304	0.609	0.545	F
44	44	0.320	0.541	0.569	F
45	45	0.262	0.552	0.824	F
46	46	0.392	0.546	0.270	F
47	47	0.289	0.352	1.266	K, F
48	48	0.260	0.398	1.264	F
49	49	0.245	0.551	0.933	F
50	50	0.270	0.489	0.909	F

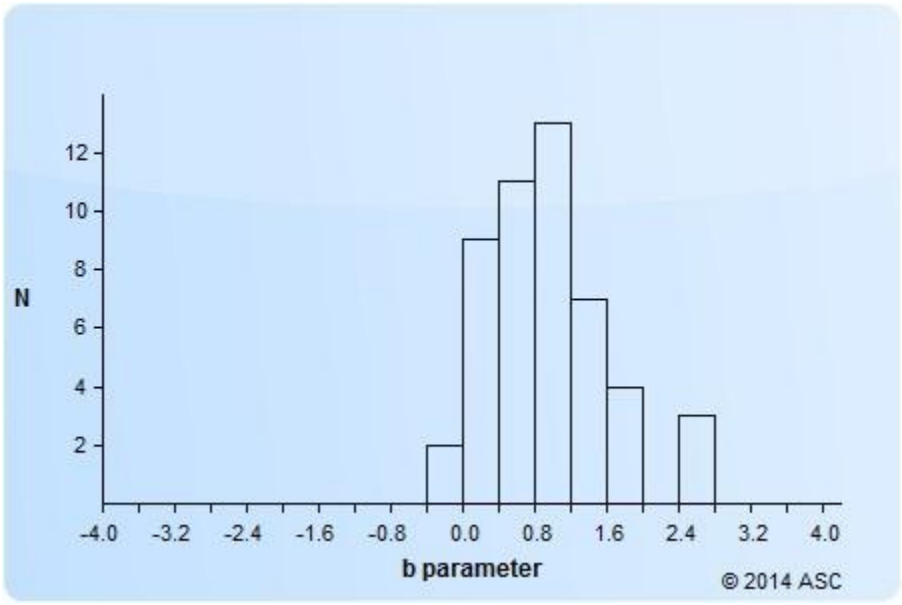

Fig: 1A: Histogram of the b Parameters of PPT

Table 1 presents the classical statistics, the item parameters, and any flags for each calibrated item. The K flag indicates that the keyed alternative (correct option) did not have the highest correlation with total score. The F flag indicates the item fit statistics. The Lb, flag indicate that the b parameter was lower than the minimum acceptable value. The Hb, flag indicate that the b parameter was higher than the maximum acceptable value. From the result presented above, the software did not report an item with Lb or Hb which means all calibrated items are within acceptable range exception of item one in the PPT test form which was not included in the calibration because the item has no variance and so did not qualify for calibration. A further analysis shows that only six items representing 12% of the entire 50 items in the CBT form produce b of less than one, while 28 items representing 56% of the 50 items in PPT produce a b of less than one. Contrarily, 35 items which is 70% of the CBT items produced b parameter between one and two, while 18 items representing 36% of the PPT items produced b parameter of between one and two. Eight CBT items had b parameter of two and above which is 16% of the total items and three items of the PPT form which is 6% had a b parameter of two and above

Figure 1 displays the distribution of the b parameters. Corroborating the earlier position in the table 1, the histogram revealed that majority of b of the items in the PPT form cluster

between 0.8 to 1.2 with the frequency of 13of the b parameter, 0.4 to 0.8 with frequency of 11 and 1.2 to1.6 with frequency of seven respectively while that of the CBT form has it highest cluster of b within 1.6 to 2.0 with a high frequency of 25, followed by 1.2 to 1.6 with the frequency of nine, 2.4 to 2.8 with the frequency of six and 0.8 to 1.2 with the frequency of four respectively which is almost progressive as against that of PPT which was retrogressive and largely within the lower bs, while that of the CBT was largely within the higher bs. This revealed that the bs for the items of the two test forms differ in values and range of clusters. This question is completely answered in the observed contingency table frequencies of calibrated b parameter and values within specified intervals that are in Table 2 which revealed the standing differences in the two test modes.

TABLE 2

Observed contingency table frequencies of calibrated b parameter and values within specified intervals

Group	**Intervals**			
	-0.260 to 0.645	**0.646 to 1.550**	**1.551 – 2.459**	**Total**
PPT	20	21	8	49
CBT	2	11	37	50
Total	**22**	**32**	**45**	**99**

The table showed the differences in the range of b parameter for the two test modes

H_o1: The mean of all calibrated b-parameter for PPT group does not differ significantly from the mean of all calibrated b-parameter for CBT group ($P<0.05$).The independent variable is the test modes (CBT and PPT) while the dependent variable is the difficulty parameter (b-parameter).To test the hypothesis, the contingency Chi-Square (X^2) was utilized for the analysis. The detail is presented in Table 3.

TABLE3

Observed contingency table frequencies of calibrated b parameter and values within specified intervals

Group	Intervals			
	-0.260 to 0.645	**0.646 to 1.550**	**1.551 – 2.459**	**Total**
PPT	20	21	8	49
CBT	2	11	37	50
Total	22	32	45	99

Chi-Square table of observe and expected frequency for the PPT and CBT b-parameter

Cell	Observed	Expected	O-E	$(O-E)^2$	$(O-E)^2/E$
1	20	10.9	9.1	82.81	7.6
2	21	15.8	5.2	27.04	1.7
3	8	22.3	-14.3	204.49	9.2
4	2	11.1	-9.1	82.81	7.5
5	11	16.2	-5.2	27.04	1.7
6	37	22.7	14.3	204.49	9.0
Total	**99**	**99**			**X^2=36.7**

***Significant at P<0.05; DF=2; X^2- critical = 5.99; X^2-Calculated = 36.70**

From the result of the analysis, the calculated X^2-value of 36.70 was found to be more than the X^2-critical of 5.99. Because the calculated X^2-value was higher than the critical X^2-value, the null hypothesis which states that, the mean of all calibrated b-parameter for PPT group does not differ significantly from the mean of all calibrated b-parameter for CBT group was rejected and thus, an alternate hypothesis formulated: the mean of all calibrated b-parameter for PPT group differ significantly from the mean of all calibrated b-parameter for CBT group and was upheld. By this result therefore, PPT and CBT UTME Mathematics of 2013 differ significantly in terms of difficulty (b-parameter).

4. Discussion

Item difficulty of PPT and CBT of 2013 UTME Mathematics Test Items. This parameter was investigated using the calibrated b-values of the 50 items for the two tests forms, histogram of the b-values and percentages. The findings on this parameter is in line with the submission of Baker (2001), Natarajan (2009) and Ojerinde et.al (2012), that the value of b in practice ranges

between -3 and +3 which confirms why no b-value in either of the test forms was >3. However, it was discovered that, the CBT test form has a higher mean of b-value of 1.65 as against PPT b-value mean of 0.94 which implies that the CBT have higher b-values than the PPT form. This means that the CBT items proved to be more difficult than PPT. This is also seen in the range of the b-values for the two tests forms as they cluster.

Six items in the CBT which is 12% have difficulty b-value of < 1(low) while that of PPT is 28 items representing 56% of items with low difficulty level implying clearly their disparities in this parameter. Also, the findings indicate that 35 (70%) CBT items has b-value between 1 – 2 (moderate) and only 18 (36%) of PPT were in the same moderate range. Finally, 8 items of CBT falls in the high difficulty range with b-values of > 2 while only 3 PPT items were at this range, indicating another disparity. This means that majority of the items for PPT have low difficulty parameter, and majority of the CBT form have moderate difficulty parameter. More CBT items were in the moderate and high difficulty level than the number of PPT items. This means that, the CBT items proved to be more difficult for the examinees than the PPT. The findings agrees to Harris, cited in Ojerinde et.al (2012) which says that, items with high values of b are difficult and items with low-values of b are easy with most examinees, even those with low-ability values having at least a moderate probability of answering the item correctly.

Contrarily to the above position of interpreting the values of difficulty in term of being low, moderate or high, Natarajan (2009) states that, the proper way to interpret a numerical value of the item difficulty parameter is in terms of where the item functions on the ability scale. The discrimination parameter can be used to add meaning to this interpretation. The slope of the ICC is at a maximum at an ability level corresponding to the item difficulty. Thus, the item is doing its best in distinguishing between test takers in the neighborhood of this ability level. Because of this, one can speak of the item functioning at this ability level. He noted for instance that, ‘an item whose difficulty is -1 function among the lower ability test takers. A value of +1 denotes an item that functions among higher ability test takers. The underlying principle being that the item difficulty is a location parameter so should be described using ‘where’ and not ‘how’. Thus, the findings still aligned that the two tests forms differ because the CBT functions more among the high-ability examinees than the low-ability examinees. Indicating, that more high-ability

examinees took the CBT form than the PPT form, and the same goes that, more low-ability examinees took the PPT form than the CBT form.

Again, the PPT b curve indicates a better spread of the b along the ability scale compared to that of CBT. The PPT curve is flatter ranging from -0.4 to 3.2 level of ability, while that of CBT is more steep which ranges only between 0.0 to 3.2 levels of ability. Therefore, from difficulty and theta plot of person-item, the two test forms are not the same.

5. Conclusion and Recommendations

Based on the findings of this study, the researcher concluded that, the two test forms as administered by JAMB in 2013 differs in terms of the difficulty parameter in the way the items functions among test takers. The researcher therefore recommended the following:

1. That, Psychometricians should always have at the back of their mind, the fact that even though parallel test are constructed to meet the same specifications in terms of their test/item difficulty, they will not necessarily scale up equal indices or function in the same way and thus should mind their usage in testing.
2. Examination bodies should avoid the use of technology that is not effective for test delivery.
3. The researcher strongly recommend and cautioned, that examination bodies should in the best interest of their candidate avoid administering two modes of parallel test for the same purpose, but for balance can administer a form of two or more parallel test to candidates to check test mode effect.

References

Anagbogu, G. E. (2005). Analysis of the psychometric properties of NECO and WACE Mathematics instruments and students' performance in Cross River State. Unpublished Ph.D thesis, Faculty of Education, University of Calabar, Nigeria.

Baker, F. B. (2001). The basics of item response theory. Wisconsin: Eric Clearing House on Assessment and Evaluation.

Greaney, V. & Kellanghan, T(ed 2012) Implementing a national assessment of education achievement.London Retrieved from http://web.worldbank.org/WBSITE/EXTERNAL/....

Jamil, M. & Shami, A. P (2012) Computer-based vs paper-based examination: perception of university teachers. Turkish online Journal of Educational Technology-TOJET/, ii(4), 173-381-381-381. Retrieved from files.eric.ed.gov/fulltext/E0546893.pdf on 03/12/2014

Joshua. M.T (2005), Fundamentals of test and measurement in Education, Calabar.Antia Press

Natarajan, V. (2009). Basic principles of IRT and application to practical testing &assessmentmen t. Merit Tracers. India

Odinko M. N (2014), Evaluation research, theory and practices. Ibadan, Giraffe Books

Ojerinde. D, Popoola. K, Ojo. F, and Onyeneho. P (2012), Introduction to Items Response Theory, Parameter Models, Estimation & application. Abuja, Marvelouse Mike Press Ltd.

ASSEREN

RESEARCH PAPER REVIEW FORM

1. Name(s):
2. Review Date:
3. Paper Title:

Aspect	**Expectations**					**Comment**
	Excellent	Very good	Good	Fair	poor	
(a) Topic			X			
(b) Abstract			X			
(c) Background/Introduction			X			
(d) Purpose/Objective/Research question			X			
(e) Theoretical/Conceptual Framework			X			

(f) Statement of the Problem			X			
(g) Methods/Procedures			X			
(h) Findings/Results and discussion			X			
(i) Conclusions/Implications/ Recommendations			X			
(j) References			X			

Oxverall Decision

a. Excellent: Publishable as it is
b. Very good: Publishable with very minor correction
c. Good: Publishable with minor correction √
d. Publishable with major correction
e. To be revised for reassessment

NB: In the case of (e) clearly identify what revision is required