Comparative Analysis of Candidates' Performances in the Pre and Post IRT Eras in JAMB: Case study of the Use of English in the 2012 and 2013 UTME

Prof 'DibuOjerinde, Grace O. Onoja and Chidinma B. Ifewulu
Joint Admissions and Matriculation Board (JAMB)
National Headquarters, Bwari, Abuja,
Nigeria.
dibu65ojerinde@yahoo.com,geeonoja@yahoo.com& chyfewulu@yahoo.com

Abstract

Two major test theories categorized under traditional and modern test theories exist for educational test and measurement. These are Classical Test Theory (CTT) and Item Response Theory (IRT). CTT focuses primarily on test as a whole while IRT focuses on the item-level as the name implies. The Joint Admissions and Matriculation Board (JAMB), mandated by law to conduct the UTME, a high-stake examination for entry into the tertiary institutions in Nigeria used the CTT in its Universities Matriculation Examination (UME) until 2010. Thereafter, it embraced the IRT for its Unified Tertiary Matriculation Examination (UTME) and began actual application in 2013 with the introduction of the Computer Based Test (CBT) as a mode of examination delivery. The purpose of this paper is to examine the empirical relationship between CTT and IRT eras in JAMB to ascertain impact of both theories on performances of repeaters and on JAMB assessment practices in the 2012 and 2013 UTME Use of English (UOE) Language paper. The study employed the ex-post facto design. The Population was made up of candidates who sat for the 2012 UTME and repeated in 2013. A sample size of 100,000 candidates who had scores in both years was randomly selected and analysis carried out using ANOVA, Chi-Square, Microsoft Excel software and X-Calibre. Results revealed a significant improvement on the performances of candidates who repeated the examination in 2013 over those of the 2012 in the UTME UOE. The paper, therefore, recommends the application of IRT for making valid judgments on quality of items to be used by the examinees.

Keywords: CTT, IRT, UTME, Comparability, UOE

Background of the Study

As assessment becomes increasingly global, the field of education stands to benefit from theories. Therefore, two major tests categorized under traditional and modern test theories exist. These are: Classical Test Theory (CTT) and Item Response Theory (IRT). CTT and IRT albeit their consistencies and complementary attributes have a number of points of differences. CTT and IRT are widely perceived as representing two very different measurement frameworks. Whereas, CTT is the older, IRT is relatively newer. IRT is seen as the most significant and popular development in psychometrics to overcome the shortcomings of CTT, and maximize objectivity in measurement (Hambleton and Jones, 1993). The name IRT derives its meaning from the focus of the theory on items whereas the CTT analysis is on the test (not the items). It models examinee responses of a given ability to each item in the test. The application of IRT enables the assembling of equivalent test forms ensuring test security. That is to say that test forms are constructed to be equivalent in psychometric characteristics, as well as in non-psychometric properties. In the application of the CTT, the ability of the examinee is dependent on the type of test items used and the parameters of the items are dependent on the samples of test used by the examinees. Prior to the introduction of IRT in JAMB, the process of test development was in line with the traditional method of ascertaining reliability of test items by considering the item difficulty and the discrimination index of the test items. This approach as useful as it appears only focused on the test as a whole without any consideration given to the individual items and the test takers.

Therefore, the introduction of Computer Based Test (CBT) by the Board made the application of IRT in the Board's test development expedient. One of the underlining reasons for this decision was

based on the need for an item bank for the Board since item banks are a repository of test items stored with known item parameters and characteristics. It became necessary therefore to generate these parameters using a test theory which defines items in terms of its qualities and the person (ability). Over the years, the Test Development Department of the Board generated test items whose quality were judged using face and content validity without conducting trial test on them. In 2010, the Board introduced trial testing of test items and used the CTT approach in the analysis of the items. By this development, quantitative approach in test development came to the fore in 2013 as all trial tested items of the Board were subjected to analysis using IRT. The result of the application of IRT in item analysis was quite revealing and interesting. In carrying out the analysis, the Board employed the use of XCALIBRE 4.1 in the caliberation of its test items. The introduction of any new concept comes along with its problem of acceptability and adaptability. The Board thus, adopted the use of the XCALIBRE software to suit its needs after initial analysis rejected a number of trial tested items based on item low point biserial correlation and a number of other problems. The information generated from the Board's item analysis was forwarded to the Test Development Department for reconstruction, revalidation and re-trial testing of items initially rejected.

The repeaters used in this study are candidates who sat for the Board's examination in the 2012 UTME and repeated in 2013 UTME. The 2013 UTME was different from others in the sense that the Board expanded its mode of examination delivery to include the Dual Based Test (DBT), Computer Based Test (CBT) as well as the traditional Paper-and-Pencil Test (PPT).

Statement of the Problem

Addressing measurement problems with the right application of statistical test framework is a herculean task. This is so because if the wrong framework is applied, measurement errors can occur and reporting true scores and abilities become hampered. Thus, this paper will answer pertinent questions such as the questions of whether or not there exists significant difference between the 2012 and 2013 UTME using the CTT and IRT approaches. It will also ascertain which of the theories favoured or disfavoured the repeaters and finally whether the problem of right application (that is reporting true and error scores) was solved using the IRT approach.

Purpose of the Study

The purpose of this paper is to examine the empirical relationship between CTT and IRT eras in JAMB with the view to ascertaining the impact of both theories on candidates' performances and on JAMB assessment practices in the 2012 and 2013 UTME Use of English (UOE). This paper will explore the extent to which the deployment of the IRT in the 2013 UTME affected on performances of candidates who sat for the 2012 UTME and repeated in 2013. The paper would also establish the parameter estimates and reliability coefficients generated using both theories with a view to ascertaining comparability. Ultimately, the paper hopes to achieve a comparative analysis of performances of candidates using the two measurement frameworks, assessing their implications on the UTME and also highlighting the efficiency of the deployment of innovative technology in assessment.

Related Literature

CTT is a theory about test scores that captures the essence of three concepts. These are: test score (often called the observed score), true score, and error score. The CTT simply postulates a simple linear model linking the observable test score *(X)* to the sum of two unobservable (or often called latent) variables, true score *(T)* and error score *(E)*, that is, $X = T + E$ (Hambleton and Jones, 1993). Nenty (2004) also gave the fundamental formula of CTT to be $\mathbf{X_0 = X + Xe.}$ That is to say that the measurement of any behavioural characteristic is the true score **(X),** while that which results from the measurement is the observed score $\mathbf{(X_0)}$ and $\mathbf{(X_e)}$ is the error inherent in the measurement. The

assumptions in the classical test model are that (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores on parallel tests are uncorrelated. In this formulation, where error scores are defined, true score is the difference between test score and error score. The deployment of IRT enables independent estimation of item and person parameters and local estimation of measurement error. These properties of IRT are also the main theoretical advantages of IRT over CTT. The main advantage of the CTT on the other hand is its simplicity. CTT does not involve truly latent variables despite the fact that the true score is not empirically observable. It can be defined operationally as the average score on the infinite number of equivalent repetitions of the measurement process.

CTT does not have invariance of item and person statistics because items and persons are sample dependent while IRT has invariance of item and person statistics since item and person parameters are sample independent. The item statistics for IRT involves the following: ***a***- the discrimination parameter; ***b*** - the difficulty parameter; and ***c***- the guessing parameter (the parameter ***c*** is the probability of getting the item correct by guessing alone). IRT models are fit for all existing data, so much so that, large sample sizes are used in the estimation of parameters. Items that do not fit the data or model are dropped by IRT. CTT cannot do this because of the total score emphasis. In the CTT, when an outcome measurement is estimated, characterized or selected on the basis of its reliability, tailoring the assessment is not possible. Consequently, validity and reliability measures are not easily achieved. Thus, CTT reliability and results thereof may not be valid and meaningful.

Test theories, therefore, are important to the practice of educational and psychological measurement. These cannot be overemphasized because they provide a better framework for understanding examinees' abilities and what is referred to as their true observed score. Wiberg (2004), Hambleton and Jones (1993), assert that understanding these test theories is critical because they provide better frameworks for handling measurement errors. For example, Hambleton and Jones (1993), posit that a good theory or model helps experts to understand the role that measurement errors play in the following ways: estimating the examinee's ability and how error may be minimized; correlations between variables; and reporting true scores or ability scores.

Research Questions

1. Are the parameter estimates and reliability coefficients of the two test theories using both theories (CTT and IRT) comparable?
2. Is there any significant relationship between the scores of repeaters who sat for UOE in the pre (2012) and post IRT (2013) eras in the UTME?
3. Is there any significant difference between the scores of repeaters in the 2013 as against 2012 UTME across:
   (i) Faculties; and
   (ii) geo-political zones?
4. What is the impact of Gender on the scores obtained by Repeaters in 2012 and 2013 UOE?

Methodology

(i) The first step was to extract the candidates' responses for analysis for the two years under review. Responses were calibrated to generate the ***a b c*** parameters for IRT, the ***D*** and ***P*** parameters for CTT, reliability coefficients to estimate the ability of the candidates.

(ii) The responses were scored dichotomously to get the aggregate scores.

(iii) Scores of these candidates were equated to bring them to a common scale or platform for comparison (linear equating). A simple linear equating method was used which involves the difference between the target (2012) population's mean score on the reference form and their mean score on the new form (2013). The following linear equating formula was used.

$$Y - \frac{Mean(Y)}{SD(Y)} = X - \frac{\text{Mean (X)}}{SD(X)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots .. (1)$$

$$Y = \frac{SD(Y)}{SD(X)} X + Mean\ Y \left[- \frac{SD\ (Y)}{SD(X)} \right] \text{Mean (X)} = \text{ adjusted X} \ldots\ldots\ldots\ldots ..\ldots . (2)$$

(iv) Difference between scores in the 2012 and 2013 was calculated. The new variable called score difference was tested for significant difference.

(v) Scores in the UOE for 2012 and 2013 were subjected to simple linear regression analysis. The scores in the UOE for 2012 was taken as the dependent variable and 2013 scores as the independent variable.

Design

The study adopted an Ex Post Facto design for data analysis. This is so because researchers had no control over the data manipulation as data was extracted from the UTME Master File of the Information Technology Services (ITS) Department's data base. This extraction contains all details of the UTME candidates' including candidates' responses and scores.

Participants

The population for this study comprised all candidates' that sat for the UTME in the 2012 and repeated in 2013. Population for 2012 is (N =313,202) and 2013 is (N=355,263). Thereafter, a simple random sampling was used in selecting the number of repeaters. A total of 100,000 repeaters only were randomly selected for both years.

Analysis

In the data analysis, the X-Calibre 4.1 was used to generate the parameter estimates of ***a b c*** for the 2013 UTME UOE while Excel software was used to generate the ***D (a) and P*** (***b)*** parameter indices for the 2012 CTT. Chi-square and ANOVA were used to answer questions 2 and 3. Independent T-Test was used to answer question 4 respectively.

Results

Table 1.1
Description of Sample Data according to Faculty and Gender

		Faculty_2013											Total
		Administration	**Agriculture**	**Arts/Humanities**	**Education**	**Engineering/Tech**	**Law/Legal**	**Medicine**	**Sciences**	**Social Sciences**	**Environment**	**Pharmacy**	
Gender	**Male**	5666	204	3408	1119	13065	3235	8487	7735	11414	117	42	54492
	Female	6641	181	4393	1482	1672	3650	11993	5375	9969	26	126	45508
Total		12307	385	7801	2601	14737	6885	20480	13110	21383	143	168	1000000

Table 1.2
Descriptive Statistics of Score Difference

	N	Minimum	Maximum	Mean	Std. Deviation
ScoreDiff	100000	-14.00	11.00	-6.1833	1.49377
Valid N (listwise)	100000				

The data used in the study included data for both male and female repeaters across disciplines. It can be seen that a total number of fifty-four thousand, four hundred and ninety-two (54,492) male repeaters were sampled while female repeaters were forty-five thousand, five hundred and eight (45508) respectively. The scores of repeaters for 2012 was subtracted from that of 2013 to give the score difference. The minimum score is -14.00 and the maximum is 11.00. The mean difference is -6.183 with an std of 1.4947.

Research question 1: Are the parameter estimates and reliability coefficients of the two test theories using both theories (CTT and IRT) comparable?

Research question 1 seeks to find out if the parameter estimates and the reliability coefficients of the items used in 2012 which represents the CTT era is comparable to that used in the IRT era in 2013.

Table 2.1
Summary of Parameter Statistics for all Caliberated Items Total Scores for 2012 and 2013

	Parameter	**Items**	**Mean**	**SD**	**Min**	**Max**
IRT (2013)	***a***	95	1.2276	0.7688	0.0942	3.6857
	b	95	0.2311	0.8724	-1.6428	2.6349
	c	95	0.1778	0.1045	0.0133	0.5808
CTT (2012)	***D (a)***	100	0.2483	Nil	-0.00372	0.40662
	P (b)	100	0.5723	Nil	0.0357	0.82129

Table 2.2
Summary Statistics for Total Scores

Model	Items	Alpha	Mean	SD	Median	Min	Max
IRT (2013)	95	0.9473	52.5983	18.4575	52	0	89
CTT (2012)	100	0.8202	52.8201	18.1902	52	0	89

Table 2.3
Theta Estimates for the Post (2013) IRT Era

Model	**Test**	**Examinees**	**Mean**	**SD**	**Skew**	**Min**	**Q1**	**Items**	**Chi-square**	**df**
IRT (2013)	Full Test	100000	-0.0237	1.0514	-1.1402	-7	-0.534	95	152207	1140

The analysis revealed that the items used in the 2013 were more reliable than that used in 2012. Test items used during the 2012 UTME were analyzed using the CTT model involving the determination of the ***D*** (item discrimination index) and ***P*** (item difficulty index) parameters while the items used during the IRT era were analyzed using the 3-parameter IRT model of ***a b c***. The reliability coefficients of the items used in the 2012 and 2013 UTME were **0.8202 and 0.9473** respectively. The mean and standard deviation of the items used in the 2012 UTME and that of 2013 were 52.8201, 18.1902 and 52.5983, 18.4575 respectively. Although the mean score was slightly greater than that of 2013, the standard deviation of the 2013 test items was higher. Also, the summary statistics for all calibrated items showed that the maximum value for the item parameters were higher for the 2013 items than that of the 2012. The maximum values for the items used for the 2013 were 3.6857 and 2.6349 respectively while that of 2012 were 0.4066 and .8213 respectively for the ***D*** (***a)*** and ***P*** (***b)*** parameters.

The summary statistics for theta estimates, applicable to IRT alone, revealed that 95 out of 100 items set for calibration were accepted for final analysis while all the 100 items used for analysis through the CTT were accepted and used. As stated earlier, this is because the CTT does not reject any item even if the items have low point biserial correlations. In spite the fact that the item statistics obtained using the two approaches were comparable, the reliability coefficient obtained using the IRT showed greater chances in the internal consistency of the items than that obtained using the CTT approach.

Research Question 2: Is there any significant relationship between the scores of repeaters who sat for UOE in the pre (2012) and post IRT (2013) eras in the UTME?

Table 3.1
Chi-Square Test of Relationship between the Pre and Post IRT Eras in the performances of Repeaters in the UTME

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.88E+06	3220	0
Likelihood Ratio	590576.684	3220	0
Linear-by-Linear Association	97293.102	1	0
Number of Valid Cases	100000		

a. 1730 cells (51.8%) have expected count less than 5. The minimum expected count is .00.

Result of the analysis of research question 2 showed that there is significant relationship between the scores obtained in 2012 and 2013 by the repeaters. The Pearson's Chisquare value is 2.882+E6 at 3220. d.f is significant. Also, the linear-to-linear association of the repeaters score in 2012 and that of 2013 is significant at $p<0.05$ or $p = .000<$ p= .05 level of significance. This result shows that a high level of relationship exists between the performances of the 2012 UTME candidates' who repeated in 2013.

Research question 3: Is there any significant difference between the scores of repeaters in the 2013 as against the 2012 UTME across:

(i) Faculties; and
(ii) geo-political zones?

Table 4.1
Analysis of Variance of Score Difference in Performances of Repeaters in 2012 and 2013 in the UOE across Faculties

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8.858	8	1.107	0.497	0.859
Within Groups	220913	99245	2.226		
Total	220921	99253			

Table 4.2
Analysis of Variance of Score Difference in Performances of Repeaters in 2012 and 2013 in the UOE across Geo-political Zones

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2984.196	5	596.839	271.093	.000
Within Groups	220147.182	99994	2.202		
Total	223131.378	99999			

Table 4.3
Multiple Comparison of Score Difference of Repeaters in the 2012 and 2013 UTME across Geo-political Zones

(I)	(J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
North West	North East	-.21473*	.02811	.000	-.2698	-.1596
	North Central	.12947*	.01600	.000	.0981	.1608
	South West	.44878*	.01367	.000	.4220	.4756
	South East	.21128*	.01345	.000	.1849	.2376
	South South	.25359*	.01482	.000	.2245	.2826
North East	North West	.21473*	.02811	.000	.1596	.2698
	North Central	.34420*	.03002	.000	.2854	.4030
	South West	.66351*	.02884	.000	.6070	.7200
	South East	.42601*	.02874	.000	.3697	.4823
	South South	.46832*	.02941	.000	.4107	.5260
North Central	North West	-.12947*	.01600	.000	-.1608	-.0981
	North East	-.34420*	.03002	.000	-.4030	-.2854
	South West	.31931*	.01726	.000	.2855	.3531
	South East	.08181*	.01708	.000	.0483	.1153
	South South	.12412*	.01818	.000	.0885	.1598
South West	North West	-.44878*	.01367	.000	-.4756	-.4220
	North East	-.66351*	.02884	.000	-.7200	-.6070

	North Central	-.31931*	.01726	.000	-.3531	-.2855
	South East	-.23750*	.01492	.000	-.2667	-.2082
	South South	-.19519*	.01617	.000	-.2269	-.1635
South East	North West	-.21128*	.01345	.000	-.2376	-.1849
	North East	-.42601*	.02874	.000	-.4823	-.3697
	North Central	-.08181*	.01708	.000	-.1153	-.0483
	South West	.23750*	.01492	.000	.2082	.2667
	South South	.04231*	.01599	.008	.0110	.0736
South South	North West	-.25359*	.01482	.000	-.2826	-.2245
	North East	-.46832*	.02941	.000	-.5260	-.4107
	North Central	-.12412*	.01818	.000	-.1598	-.0885
	South West	.19519*	.01617	.000	.1635	.2269
	South East	-.04231*	.01599	.008	-.0736	-.0110

*. The mean difference is significant at the 0.05 level.

Research question 3 sought to find out if a significant relationship exists across faculties of study and across geo-political zones. Table 4.1 therefore shows the ANOVA result of performances of repeaters across faculties. The *F*-ratio obtained from the analysis of variance showed an F (8, 99245) = .497 which is low and not significant at $p<0.05$ level. Table 4.2 shows ANOVA result of score difference in performances across geo-political zones. The *F*-ratio obtained showed an F (5, 99994) = 271.093 at $p < 0.05$. This result is significant. This depicts that significant difference exist across geo-political zones but not across faculties. However, a Post Hoc of analysis results according to faculties showed no significant differences between Pharmacy and Administration, Agriculture, Arts/Humanities, Education, Engineering/Tech, Law/Legal, Medicine, Sciences, Social Sciences. Law/Legal had significant differences with Sciences and Arts/Humanities **(the Post Hoc Table is not included in this report).**

Research question 4: What is the impact of Gender on the scores obtained by Repeaters in 2012 and 2013 UOE?

Table 5.1

Group Statistics of Score Difference according to Gender

	Gender	N	Mean	Std. Deviation	Std. Error Mean
ScoreDiff	Male	3485	-6.1435	1.21631	0.0206
	Female	577	-5.8128	1.33839	0.05572

Table 5.2
Independent Samples Test for Score Difference in Performances by Gender

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ScoreDiff	Equal variances assumed	13.486	.000	-6	4060	0	-0.3306	0.05548	-0.4394	-0.2218
	Equal variances not assumed			-5.6	742	0	-0.3306	0.05941	-0.4472	-0.2140

Result of the independent samples t-test carried out showed that the calculated t-value of -5.6 is greater than the critical t-value of 1.645 in absolute terms. That is t-calculated > t-critical at $p < .05$ is significant. This means that there is no gender difference in the performances of the repeaters.. This is to infer that the items used for the 2013 UTME was not biased against any group in both years.

Conclusion

The paper thus concludes by reiterating the fact that although both measurement frameworks are useful to test development experts in understanding and measuring psychological phenomena and construct, experts should explore the benefits inherent in the deployment of IRT which gives better measurement results than the CTT. The study found that IRT provides a more reliable measurement framework as can be seen from the application of the theory on the data used. Statistics generated from its use, provides greater information about behaviour of items. IRT has the ability to accept good items and at the same time, redeem or reject bad ones. The framework tests the validity of the model used while CTT does not have this ability. This was observed in the study where out of a 100 UOE items, 5 were rejected by the theory. CTT does not take into consideration the ability of the candidate while IRT estimates ability of the candidates. This is a greater advantage over the CTT which accepts all items whether or not they are good. It was also deduced from this study that the items for both years had no gender bias. As earlier stated, the IRT impacted greatly on the quality of items used by the repeaters in the 2013 UTME. This invariably contributed to a better performance by repeaters in the 2013 UTME UOE as compared to the 2012 UTME. The IRT measurement framework also makes possible the use of parallel forms in test which is not possible in the CTT. As the Board looks at the prospect of delving into Computer Adaptive Test (CAT), it is hard to imagine CAT without IRT.

Recommendation

- Test development experts are advised to ensure that measurement is aimed at understanding or ascertaining underlying ability or trait which produces quality test performances in their quest for valid measurement results.
- It is recommended that they also use the IRT because of its reliability and validity in measuring traits over the CTT. Experts should not just deploy this statistical package because some authority suggests its superiority over the CTT but should seek to find out themselves to be able to make informed choice and decision like what the Joint Admissions and Matriculation Board (JAMB) has done.

- For measurement experts who value invariant item and person statistics, the solution lies in the deployment of the IRT in its test development process.
- Finally, IRT should be adopted because of its ability to reject bad items for low point biserial correlations and its ability to redeem bad items for possible revalidation and re-trial testing. The CTT does not have this ability for it accepts all items irrespective of whether they are good or bad.

References

Hambleton, R. K. and Jones, R.W. (1993). Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development. An NCME Instructional Module 16, Fall 1993. Retrieved on Saturday, 9th March, 2013 from http://www.ncme.org/pubs/items/24.pdf

Nenty, H. J. (1998). Introduction to Item Response Theory cited in Scientific Notes, Global Journal of Pure and Applied Science, Vol 4, N0. 1, pp 93 – 100, ISSN 1118 – 0579.

Nenty, H. J. (2004). From Classical Test Theory (CTT) to Item Response Theory (IRT): An Introduction to a Desirable Transition. In O. A. Afemikhe and J. G. Adewale (Eds.), Issues in Educational *Measurement and Evaluation* in *Nigeria (in honour of* 'Wole Falayajo) (Chapter 33, pp.371 – 384). Ibadan, Nigeria: Institute of Education, University of Ibadan.

Ojerinde, D, Popoola, K., Ojo, F., Onyeneho, P. (2012): Introduction to Item Response Theory: Parameter Models, Estimation and Application, Lagos, Nigeria, Goshen Print Media Limited.

Wiberg, M. (2004). Classical Test Theory vs Item Response Theory – An Evaluation of the Theory Test in the Swedish driving-license test. EM No. 50, 2004, Umea University, Sweden retrieved Saturday, 9th March, 2013: http://www.umu.se/edmeas/publikationer/pdf/Em%20no%2050.pdf