Comparative Analysis of Classical Test Theory and Item Response Theory using Chemistry Test Data

Ado Abdu Bichi, Rahimah Embong, Rohaya Talib, Sakinah Salleh, Abdullah bin Ibrahim

Abstract— Assessment of learning involves deciding whether or not the content and objectives of education are down pat by administering quality tests. This study assesses the standard of Chemistry action take a look at and compares the item statistics generated mistreatment CTT and IRT strategies. A descriptive survey was adopted involving a sample of N=530 students. The specialised XCALIBRE 4 and ITEMAN 4 softwares were used to conduct the item analysis. Results indicate that, both the two methods commonly identified 13(32.5%) items as "problematic" and 27(67.5%) were "good". Similarly, a significantly higher correlation exists between item statistics derived from the CTT and IRT models, [(r=-0.985,) and (r=0.801) p<0.05] for item difficulty and discrimination respectively; the study concludes that the Chemistry Achievement test used do not pass through the processes of standardisation. Secondly, CTT and IRT frameworks appeared to be effective and reliable in assessing test items as the two frameworks provide similar and comparable results. The study recommends that the teacher made Chemistry tests used in measuring students' achievement should be made to pass through all the processes of standardisation. Meanwhile, CTT and IRT approaches of item analysis ought to be integrated within the aspects of item development and analysis because of their superiority within the investigation of reliability and minimising measurement errors.

Keywords: Chemistry Tests, Item Response Theory, Classical Test Theory, Item Parameters.

INTRODUCTION

Assessment of students learning is a fundamental aspect of educational process. A significant goal for evaluation in educational settings is to measure learners' achievement so as to make a range of decisions [1].

The persistent students' failure in Chemistry at different examinations in Nigeria continue to draw the attention of stakeholders in education. Teachers of Chemistry in Nigerian schools design, administer, scored and analyse their tests, many of this teachers have no adequate knowledge and training in testing models and principles. They hardly consider about whether their test items are valid enough to quantify learners' abilities and whether the scores from the test are dependable enough to assess learners' achievement in this subject [2]. Conversely, little consideration has been paid, by scholars in Chemistry education, to the investigation of the items contained in the Chemistry achievement tests developed by teachers. A careful investigation of the procedure through which the items were developed with its psychometric properties may propose methods for improving learners' performances.

The development and standardization of test in education and behavioural sciences include various steps that are interrelated and diverse at some particular stage in the development process. These interrelated steps are signified out in the measurements techniques or models. All measurement models are employed to survey the behavior and quantify it by assignment of numerical values.

In principle of evaluation in educational and behavioural sciences there are two contemporary approaches, these are the Classical Test Theory (CTT) and Item Response Theory (IRT). These theories are used to conduct item analysis to ensure that the test items are adequate in quality to assess a sample of behaviour and quantified the behaviour through assignment of numerical values [3]

Several scholars expressed reservation over the estimators provided by CTT, for example [4] opined that the estimates given by CTT are sample dependent (depends on the examinees ability). Consequently, the generalization of its estimators is very difficult more experientially when the population of the examinees are diverse in their abilities. To some researchers, IRT is the answer to the shortcomings of CTT [5]. The IRT used item level measurement to assess the examinees performance. To this ends, various researches have been conducted to assess the test items and evaluate whether the CTT and IRT are comparable in terms of item parameter estimations [6],[7],[8]. However, researchers differ in their perception of the ability of these two popular models in providing a standard measure for assessing learners' abilities.

In the light of these and many concerns, this study therefore is conducted to investigate the comparative nature of the CTT and IRT item parameters using a teacher-made Chemistry achievement test (CAT) in Nigeria. In addressing the above objective, the study provides answers to the following questions, thus;

Revised Version Manuscript Received on April 19, 2019.

Ado Abdu Bichi, Graduate Student[a] & Lecturer[b], [a]Universiti Sultan Zainal Abidin, Kuala Terengganu, Malaysia. & [b]Faculty of Education, Yusuf Maitama Sule University, Kano

Rahimah Embong, Assoc. Professor, Department of Education, Dakwah and Islamic Civilization, Faculty of Contemporary Islamic Studies, Universiti Sultan Zainal Abidin, Kuala Terengganu, Malaysia.

Rohaya Talib, Senior Lecturer, School of Education, Universiti Teknologi, Malaysia, Johor Bahru, Malaysia.

Sakinah Salleh, Senior Lecturer, Faculty of Human Sciences, Universiti Pendidikan Sultan Idris, Malaysia.

Abdullah bin Ibrahim, Senior Lecturer, Centre for Fundamental Studies, Universiti Sultan Zainal Abidin, Kuala Terengganu, Malaysia

II. ADVANCED MEASUREMENT THEORIES

The two popular models used in educational, psychological and behavioural measurements are the Classical Test Theory (CTT) and Item Response Theory (IRT). The major difference within the two theories can best be understood from the underlying statistical analyses inherent it [9]. A brief discussion is given to understand the theories as it relates to this work.

1. Classical Test Theory (CTT)

The CTT has been the model used for decades to assess the reliability and validity of measurement instruments. According to [4] CTT is a test score theory that bring with it three concepts (a) Test score also known as observed score, (b) true score and (c) error scores. In CTT a number models were being developed. Example, in what is often referred to as the "classical test model,"

$$X=T+E \quad (1)$$

This model links the test score (X) to the unobservable true score (T) and error score (E). Because the true score is not easily observable, instead, the true score must be estimated from the individual's responses on a set of test items[2]. Thus, the equation cannot solve until some assumptions are made. Some of the major assumptions in CTT are: average error score of the test takers is zero, error scores and true scores are uncorrelated, and error score on the parallel test are not correlated. In CTT the number of correct scores is often taken as ability.

Several researchers expressed reservation over the estimators provided by CTT, for example [4] stated that the estimates given by CTT are sample dependent (depends on the examinees ability). Consequently, the generalization of its estimators is very difficult more especially when the population of the examinees are diverse in their abilities. Similarly [11],[2],[7] has summarized and noted this problem as the estimators coming from CTT are circular dependent i.e. items parameter estimates depend on test taker and abilities test takers are function of the parameter estimates). This circular dependency in the case of easy test can exaggerate the ability estimates of the students, and difficult test can do the invert work by thinking little of the abilities of examinees. Thus, it is hard to sum up the CTT estimators across the population particularly when they are at variation with abilities.

2. Item Response Theory

Item Response Theory (IRT) is a latent technique developed to model the relationship between examinee ability and item stimuli [12]. IRT focus on the form of examinee responses rather than on total score complex variables and regression (linear) theory. In IRT the item responses are considered the outcome (dependent) variables, and the examinee's ability and the items' characteristics are the latent predictor (independent) variables [13].

IRT, item parameters include difficulty (location), discrimination (slope), and pseudo-guessing (lower asymptote). Three most commonly used IRT models are; one-parameter logistic model (1PLM or Rasch model), two-parameter logistic model (2PLM) and three parameter logistics model (3PLM). All these three models have an item difficulty parameter (b-value). In addition to having the b-value, 2PLM and 3PLM possess a discrimination parameter (a-value), this parameter (i.e a-value) allows an item to discriminate differently among the examinees with different abilities. The 3PLM contains a third parameter, referred to as the pseudo-guessing or chance parameter (c-value). The pseudo-guessing parameter corresponds to the lower asymptote of the item characteristic curve (ICC) which represents the probability that low ability examinees will answer the item correctly and provide an estimate of the pseudo-guessing parameter [11].

Item response theory (IRT) is, for some researchers, the answer to the limitations of classical test theory [5]. Item response theory (IRT) looks at the examinee performance by using item as the unit of assessment. [14] consider IRT as a modeling technique that tries to describe the relationship between an examinee‘s performance in a test and the latent trait underlying the performance. Similarly, [4] have pointed out the following four characteristics of an item response model. The first is an IRT model must give specification the relationship between the measured score and the underlying unobservable construct. The second is the model must provide a way to estimate scores on the ability. The third is the examinees' scores will be the basis for the estimation of the underlying unobservable construct. The fourth is this model assumes that the performance of an examinee is completely predictable or can be explained from one or more abilities.

3. CTT versus IRT

In comparison of CTT and IRT Models in test development following from the above description, [10] state that, IRT provides a richer set of tools for test development. IRT provides a third parameter (pseudo-guessing) that has no common analog in CTT. IRT also provides a means to assess degree of measurement equivalence at various points on the score scale, based on different sets of items. Therefore, the item analysis results provided by both CTT and IRT are fairly comparable, but IRT provides an additional item statistics and a more sophisticated mechanism for minimising measurement error.

CTT item parameters are specific to a given examinee sample, the item parameters of IRT model hold for the entire population. That is; the item parameters produce by IRT model are said to be invariant across examinee sub-populations (i.e., samples). IRT model is flexibility. For example, different sets of test items could be administered to individual examinees, and yet similar or comparable estimated theta can be obtained from these different set of tests.

IRT models represent the ability of the test-takers and the difficulty of the items as independent parameters however CTT has no way to identify these two constructs separately because in CTT all values are sample-specific. The same items will appear easy to a sample of high-ability test-takers and difficult in low-ability examinees. Therefore, IRT models can separate these empirically intertwined concepts in a way

that no other psychometric models can do.

Estimate of examinee ability depends on individual test taker's responses which usually give complete information. Ability is a continuous variable and IRT gives continuous estimates. CTT gives discrete estimates especially in dichotomously made test and may create discrepancies in assessing students' achievements by total raw scores and IRT students' ability estimates[2].

Table 1: CTT Vs IRT models [2]

SN	Area	CTT	IRT
1	Model	Linear	Nonlinear
2	Level	Test	Item
3	Assumptions	Weak (*easy to meet*)	Strong (*more difficult to meet*)
4	Item-ability relationship	Not specified	Item characteristics functions
5	Ability	Test scores or estimated true scores are reported on the test-score scale	Ability scores are reported on scale – ∞ to + ∞ (or a transformed scale)
5	Invariance of item & person statistics	No-item and person parameters are sample dependent	Yes-item and person parameters are sample independent, if model fits test data.
6	Item statistics	*p, r*	*b, a* and *c* (for the three-parameters model) plus corresponding item information functions
7	Sample size	200 to 500 (in general)	Depends on the IRT model but larger samples i.e over 500, in general are needed

III. MATERIAL AND METHODS

This is a quantitative study with a plan of data collection and analysis using cross sectional survey design was employed to collect the relevant data for the study. Five hundred thirty (530) students were selected from the 17 sciences secondary schools in Kano, using stratified random sampling techniques.

The Chemistry Achievement Test (CAT) developed to assess senior secondary schools student science achievement and to examine their suitability to be sponsored by government to write their final examinations was adopted in this survey. The instrument (CAT) contained 4o items in multiple choice form with five answer options. This test (CAT) was developed using the senior secondary schools Chemistry curriculum designed for senior secondary schools examinations in WASEC and NECO as well as the Chemistry curriculum and assessment procedure prepared by the federal ministry of Education in Nigeria. The content of the developed CAT was validated by the teams of experts in Nigerian universities, teachers institute and college of education in Nigeria

The 40 items CAT were administered to the respondents simultaneously after given them instruction for the test with the help of research assistants in the cooperating schools in July, 2014. The responses of the examinees were coded scored after marking based on designed marking scheme. The final data were used for analysis

The data analysis were conducted using two popular psychometric softwares: XCALBRE 4.2 for IRT analysis and ITEMAN 4 for CTT analysis [19]. The correlation between item parameters of IRT and CTT were obtained using Pearson Product Moment Correlation coefficient (r) generated from SPSS '25. Item were classified based on the item selection standards of ($b = -2.00 \leq 1.00$) and ($a = 0.64 \geq 1.70$) was applied [16].

IV. RESULTS

Summary statistics

The descriptive statistics obtained showed that, the 40 items CAT administered to 530 students has a mean score of 16.6 with standard deviation of 7.22. The Kuder Richardson 20 (KR-20) as measure of internal consistency reliability showed a coefficient value of 0.85 for and 0.86 for IRT. This coefficient of test reliability indicated that, the test items are reliable enough to measure the Chemistry objective as contained in the curriculum. By this coefficient the test have proved to have substantial reliability as the values exceeded the recommended 0.70 [15].

5.1 Unidimensionality

This assumption of unidimensionality was exactly measured by exploring whether overwhelming factor exists among all the test items. Accordingly, principal factor analysis was completed, and the eigenvalues were checked as prescribed by[16],[17]. The outcome of the factor analysis produces fourteen items with eigenvalues higher than one. These fourteen factors explain 64.67% of the variation. The main eigenvalue was 6.657 higher than the following eigenvalue. The first factor explained 16.64% of the change; the following factor explained 6.31% of the rest of the difference. The remainder of the change was by other 26 factor. Henceforth, there is one dominant factor in the factor structure of the items set. Since there is strong factor that explained 16.64% of the variation, the assumption of unidimensionality is established[18].

Model-data fit was evaluated by checking if the individual test items fitted the given IRT model, using a likelihood-ratio standardized residual (z Resid) test, This is done for every item [19]. The p-values related with statistical tests for distinguishing item fit, standardized residual (z) with probability of under 0.05 ($p < .05$) is denoting items misfit [20]. Table 2 demonstrates the outline of standardized residual (z Resid) result fit test.

Table 2: Items fitting the models

IRT Model	1PL	2PL	3PL
Items fitting the model	39	40	38
Misfitting Items	30	0	11 & 39
% of fitting items	97.5%	100%	95%

The z Resid values associated with the item in the test, it is evident that one item (Item30) representing 2.5% of the total items in the test was statistically significant and do not fit the 1PL model. Two items (Item11 and Item39) representing 5% of the total test were also statistically significant and do not fit the 3PL model. However, all the 40 items were not statistically significant and fitted the 2PL model determined at 0.05 level of significance. Thus 2PL as the model with the most compatibility to the test data, where the entire 40 test item fitted in was found suitable and therefore used to estimate the item statistics based on the IRT model in this study.

Research Question 1: What are the item parameters of the CAT using CTT and IRT models?

The data was analysed to produce the CAT item parameters for CTT and IRT respectively. Table 3 presents the item statistics based on CTT and IRT models; item statistics [difficulty (p) and discrimination indices (rpbi) for CTT and difficulty (b) and discrimination indices (a) for IRT model."

Research Question 2: Which of the item (s) are considered 'faulty' on the basis of item parameters generated using CTT and IRT?

On the basis of standards for interpreting the item parameters in botth CTT and IRT, using CTT 27 (67.5%) of the Items were of moderate difficulty, 1(2.5%) was easy, and 12(30%) were found to be difficult. Similarly, using IRT 26 (65%) of the items were found to be of moderate difficulty, 2(5%) were easy and 12(30%) were difficult.

On the basis of discriminating index criteria set, the results using CTT indicates that 9 (22.5%) of the items are poor, 5(12%) items were marginal need to be reviewed, 8(20%) of the items are reasonable good and 18(45%) of the items functions very well. Similarly, using IRT none of the items were found to be very poor, 3(7.5%) of the items were of marginal discriminating ability, and the remaining 37(92.5%) were reasonably good or satisfactory.

Table 3: IRT and CTT based Item Statistics (Parameters)

	IRT			**CTT**		
Item	**(b)**	**(a)**	**Flag**	**(P)**	**(r_{pbi})**	**Flag**
1	0.97	**0.44***	**F**	0.40	**0.22***	**F**
2	**-1.90***	**0.61***	**F**	**0.78***	0.30	**F**
3	**-1.15***	0.66	**F**	0.68	**0.22***	**F**
4	-0.51	0.91		0.60	0.46	
5	-0.45	1.09		0.60	0.53	
6	0.83	0.88		0.34	0.32	
7	0.88	0.98		0.32	0.44	
8	0.78	0.96		0.34	0.39	
9	**2.56***	0.83	**F**	**0.10***	**-0.09***	**F**
10	0.02	0.80		0.50	**0.20***	**F**
11	-0.57	1.29		0.64	0.64	
12	0.10	1.00		0.48	0.48	
13	0.28	1.02		0.44	0.48	
14	**2.07***	0.77	**F**	**0.16***	**-0.07***	**F**
15	0.74	0.83		0.36	**0.24***	**F**
16	0.27	1.05		0.44	0.52	
17	0.46	0.72		0.42	**0.14***	**F**
18	**1.37***	0.78	**F**	**0.26***	**0.13***	**F**
19	-0.64	1.03		0.64	0.47	
20	**1.06***	1.02	**F**	**0.28***	0.45	**F**
21	0.13	0.72		0.48	**0.12***	**F**
22	0.77	1.20		0.32	0.61	
23	0.69	0.73		0.38	**0.12***	**F**
24	0.87	0.80		0.34	**0.21***	**F**
25	**1.45***	1.11	**F**	**0.20***	0.50	**F**
26	**1.10***	0.95	**F**	**0.28***	0.37	**F**
27	0.56	1.01		0.38	0.46	
28	0.01	0.90		0.50	0.39	
29	**1.23***	1.29	**F**	**0.22***	0.64	**F**
30	**1.72***	0.60	**F**	**0.24***	**-0.41***	**F**
31	0.18	1.08		0.46	0.53	
32	**1.76***	0.77	**F**	**0.20***	**0.02***	**F**
33	-0.83	1.03		0.68	0.46	
34	**1.31***	0.97	**F**	**0.24***	0.39	**F**
35	**1.79***	0.75	**F**	**0.20***	**-0.02***	**F**
36	-0.38	0.93		0.58	0.43	

37	-0.58	0.89		0.62	0.34	
38	-0.09	0.96		0.52	0.43	
39	-0.59	1.21		0.64	0.60	
40	**1.14***	0.90	**F**	**0.28***	0.30	**F**

*Unacceptable item parameter estimates, to be modified or completely eliminated

In IRT a: discrimination, b: difficulty. In CTT P: difficulty, rpbi: discrimination.

Table 4: Items deleted using CTT and IRT frameworks

Model	Number deleted	Items deleted
CTT	16	2, 9, 14, 17, 18, 20, 21, 23, 25, 26, 29, 30, 32, 34, 35, 40
IRT	14	2, 3, 9, 14, 18, 20, 25, 26, 29, 30, 32, 34, 35, 40
Common items deleted	13	2, 9, 14, 18, 20, 25, 26, 29, 30, 32, 34, 35, 40
Total items deleted using both	17	2, 3, 9, 14, 17, 18, 20, 21, 23, 25, 26, 29, 30, 32, 34, 35, 40

Table 5: Correlations between CTT and IRT-based item parameters

Item Parameters	CTT and IRT	N	Mean	S.D	r-cal	df	P-value
Item difficulty	CTT Item Difficulty	40	0.414	0.169	-0.985*	38	0.00
	IRT Item Difficulty	40	0.485	0.951			
Item Discrimination	CTT Discrimination	40	0.324	0.226	0.801*	38	0.00
	IRT Discrimination	40	0.912	0.187			

Additionally, the number of items deleted by each model, the common items deleted by both as well as the total number of items deleted using both the CTT and IRT model are presented in Table 4 (i.e the poor items were identified and deleted based on their difficulty and discrimination estimates)

Research Question 3: How comparable are the CTT-based and IRT-based item Parameters (Item difficulty and Discrimination) estimates in chemistry Achievement tests?

To answer this question, the difficulty (b-values) of IRT and difficulty (p-values) of CTT were correlated. The correlation coefficient of the relationship was determined using Pearson Product Moment Correlation coefficient (r); the result is presented in Table 5

To answer this question, the discrimination (a-values) of IRT and discrimination (rpbi) of CTT were correlated. The correlation coefficient of the relationship was determined using Pearson Product Moment Correlation coefficient (r); as presented in Table 5 below;

Information from table 5 indicates a significant higher and negative correlation r (38) =-0.985, P=0.00 (P<0.05) between CTT item difficulty values (p-values) and IRT item difficulty values (b-values). This shows that the test items on average have 0.414 level of difficulty in CTT and 0.485 difficulty level in IRT. Similarly, this shows that high correlation exists between the item difficulties under the two different frameworks.

V.DISCUSSION OF FINDINGS

Assessment of test items quality and comparison of item statistics under different frameworks especially CTT and IRT has been a major topic of researches related to item analysis in the field of educational and psychological measurement. The findings of this research were similar to many previous studies conducted on item analysis to validate assessment instruments.

On the basis of difficulty and discrimination indices 16 items were classified as 'poor' or 'faulty', the remaining items are the 'good' ones. Example the most difficult item in the entire test is item9 with the p-value of 0.10; this item is difficult because only 10% of the total examinees got the item correct.

As indicated also by the item discrimination indices, items 9, 14, 30 and 35 are very poor. For example, Item 30 having the largest negative discriminating value of -0.41 is the very poor item in terms of discriminating power this tells us that higher ability students got the item incorrect, while low ability students are the ones that performed better on this item. In item analysis using CTT negative discrimination value is suggesting that we should look carefully at the item to see why the higher ability students should have more trouble with it than the weaker students. This reveals that the items could not discriminate between high achievers and low achievers. According to [21],[22] those who got the items with negative discriminating values might have probably guessed before they got it right.

Using IRT on the basis of discriminating index criteria set, the results indicates that 3 items were marginal need to be reviewed and 37 of the items are reasonably good. However, none of the items were classified as very poor item. Therefore, on the basis of difficulty and discrimination indices 14 items were classified as 'poor' or 'faulty', the remaining items are the 'good' can be used without modification. Example the most difficult item in the entire test also identified by the use of IRT model is item9 with the b-value of 2.51; this item is difficult because the proportion correct was 0.10.

As can be seen, for example, item 29 is the most discriminating item, with the highest value of its discrimination index (a = 1.291) but at the same time it is a difficult item (b = 1.227). That is, Item 29 discriminates well between examinees with different abilities. Item 1 is least discriminating item (a = 0.437).

Going by the item parameters estimates by the two measurement models. Using CTT-based estimates more items (16) were considered 'faulty' or 'problematic' and the remaining 24 items classified as "good" or of "moderate difficulty" on the 40-Items CAT than when IRT-based items Statistics estimates were used. Using IRT-based statistics 14 items were considered "faulty" or "problematic" and the remaining 26 items classified as "good" items. Similarly, both the two frameworks commonly identified 13 items as "problematic" or "faulty" these were items 2, 3, 9, 14, 18, 20, 25, 26, 29, 30, 32, 34, 35 and 40. Items 9, 14, 30 and 35 have negative discrimination indices. In test development and item evaluation as opined by [17] such problematic items having failed to satisfy the standards should be modified, dropped, replaced or eliminated completely from the test. This finding is consistent with that of [17],[6],[7] whose studies revealed deleting more items by using CTT than IRT-based statistics.

The finding on the comparability of the CTT and IRT-based item difficulty estimates, the difficulty estimates from the two frameworks were correlated using Pearson Product Moment Correlation coefficient(r) and the result was presented. The CTT-based item difficulty (p-values) had a very higher correlation with the IRT-based item difficulty estimates (b-values). As observed, the item parameters obtained by using the two frameworks, produced a strong correlation coefficient of 0.99. This strong correlation coefficient is an indication that the item difficulty estimates (p and b-values) were almost perfectly related. This means that the CTT and IRT produce similar item difficulty estimates and can be used interchangeably in test development and evaluation. It is also clear that the correlation obtained is negative; this is because the CTT p-values were not reversed. The finding of this study agrees with that of the earlier studies e.g., [23],[5],[17],[24],[6],[7]. This finding is consistent with these previous studies because their studies revealed a very strong and negative correlation between the item difficulty produced by the two competing model. Similarly, this finding led credence to the [4] that, the correlation between CTT p-values and IRT b-values should be higher and negative.

Similarly, the CTT-based and IRT-based item discrimination estimates demonstrated a strong positive correlation. As observed the correlation coefficient obtained by correlating the item discrimination values was 0.80, a significant higher and positive correlation between CTT item discrimination values (rpbi) and IRT item discrimination values (a-values). This shows that a high correlation exists between the item discrimination values (rpbi and a-values) under the two different frameworks. Thus, an indication that, the CTT and IRT produce similar item discrimination estimates and can be used interchangeably in test development and evaluation. This finding is consistent with the results of similar studies i.e., [11], [5], [23], [24], [6], [17], [25], [26], [7], whose findings revealed a strong positive correlation between CTT-based and IRT-based item discrimination indices. This is supported by the claim that a correlation coefficient of the relationship between IRT a-values and CTT Point Biserial Correlation should be high and positive[4].

VI. CONCLUSION

This study evaluated the teacher made chemistry achievement test (CAT) and compare the item parameters generated from the CTT and IRT approaches. In line with the findings of this comparative analysis, it can be concluded that, the CAT used in this study was an organised and reliable measure of students' abilities. However, this is in spite of the fact that, using the two frameworks many items were identified as problematics, which is an indication that, the CAT used in to examine students Chemistry achievement do not pass through the needed formal process of validation and standardisation. In addition, the CTT and IRT frameworks appeared to be effective in the item analysis of the CAT, as the result obtained from the two frameworks provide similar and comparable results. Similarly, both the two frameworks were found to be reliable in assessing test items parameters and the result from any of the frameworks can be used to judge the quality of tools used in assessing and evaluating learning outcomes in education and psychology. This notwithstanding, the shortcomings attributed to the CTT as well as the theoretical superiority of IRT over the CTT framework. It is recommended that; The teacher made CAT should be made to pass through all validation process to improve it utility, the poor items detected in this should be modified, or completely eliminated from the Test and lastly,

The two approaches investigated in this study should be integrated by teachers into the construction and anlysis of CAT in Nigeria morew essentially, because the approaches confirmed their superiority in investigating reliability and minimizing measurement errors.

REFERENCES

1. S. P. Klein and L. Hamilton, Large-Scale Testing: Current Practices and New Directions. Santa Monica: CA: RAND, 1999.
2. A. A. Bichi, R. B. Embong, M. Mamat, and D. A. Maiwada, "Australian Journal of Basic and Applied Sciences Comparison of Classical Test Theory and Item Response Theory: A Review of Empirical Studies," Aust. J. Basic Appl. Sci., vol. 9, no. April, pp. 549–556, 2015.
3. A. A. Bichi, R. Talib, H. Mohamed, J. Ahamad, and N. A. Khairuddin, "Exploratory Sequential Design to Develop and Validate Economics Placement Test for Nigerian Universities," Int. J. Recent Technol. Eng., vol. 7, no. 6, pp. 769–772, 2019.
4. R. K. Hambleton and R. W. Jones, "An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development," Educ. Meas. issues Pract., vol. 12, no. 3, pp. 38–47, 1993.
5. T. G. Courville and B. Thompson, "etd-tamu-2004B-EPSY-Courville-2.pdf," no. August, 2004.
6. B. A. Adegoke, "Comparison of item statistics of Physics achievement test using classical test and item response theory frameworks," J. Educ. Pract., vol. 4, no. 22, pp. 87–96, 2013.
7. N. Guler, G. K. Uyanik, and G. T. Teker, "Comparison of classical test theory and item response theory in terms of item parameters," Eur. J. Res. Educ., vol. 2, no. 1, pp. 1–6, 2014.

8. H. Nenty and O. O. Adedoyin, "Test for invariance: inter and intra model validation of classical test and item response theories," Asia Pacific Journal Res. I, 2013.
9. A. A. Bichi and R. Talib, "Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development," Int. J. Eval. Res. Educ., 2018.
10. A. D. Mead and A. W. Meade, "Item selection using CTT and IRT with unrepresentative samples," in twenty-fifth annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA. Retrieved December, 2010, vol. 22, p. 2018.
11. X. Fan, "Item response theory and classical test theory: An empirical comparison of their item/person statistics," Educ. Psychol. Meas., vol. 58, no. 3, pp. 357–381, 1998.
12. L. Crocker and J. Algina, Introduction to classical and modern test theory. ERIC, 1986.
13. D.-T. Le, "Applying item response theory modeling in educational research.," Diss. Abstr. Int. Sect. B Sci. Eng., vol. 75, no. 1, 2014.
14. X. An and Y. Yung, "Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It," SAS Inst. Inc., pp. 1–14, 2014.
15. J. C. Nunnally, "Psychometric theory (2nd edit.) mcgraw-hill," Hillsdale, NJ, vol. 416, 1978.
16. N. Georgiev, "Item analysis of c, d and e series from raven's standard progressive matrices with item response theory two-parameter logistic model," Eur. J. Psychol., vol. 4, no. 3, 2008.
17. D. Ojerinde, K. Popoola, F. Ojo, and P. Onyeneho, "Introduction to item response theory: Parameter models, estimation and application," Abuja Marvelouse Mike Press Ltd, 2012.
18. M. D. Reckase, "Unifactor latent trait models applied to multifactor tests: Results and implications," J. Educ. Stat., vol. 4, no. 3, pp. 207–230, 1979.
19. R, Guyer. & N. A, Thompson, User's Manual for Xcalibre item response theory calibration software, version 4.2, vol. 10, no. 2. Woodbury MN: Assessment Systems Corporation., 2014.
20. D. M. Dimitrov, "An Approach to Scoring and Equating Tests with Binary Items: Piloting with Large-Scale Assessments," Educ. Psychol. Meas., vol. 76, no. 6, pp. 954–975, 2016.
21. A. Field, Discovering statistics using SPSS:(and sex, drugs and rock'n'roll), vol. 497. Sage, 2000.
22. S. Varma, "Preliminary item statistics using point-biserial correlation and p-values," Educ. Data Syst., vol. 16, no. 7, pp. 1–7, 2006.
23. C. Stage, Classical test theory or item response theory: The Swedish experience, vol. 42. Univ., 2003.
24. S. Pido, "Comparison of item analysis results obtained using item response theory and classical test theory approaches," J. Educ. Assess. Africa, vol. 7, pp. 192–207, 2012.
25. M. Erguven, "Two approaches in psychometric process: Classical test theory & item response theory," J. Educ., vol. 2, no. 2, pp. 23–30, 2013.
26. O. O. Adedoyin, "Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories," Int. J. Educ. Sci., vol. 2, no. 2, pp. 107–113, 2017.