

The use of measurement frameworks to explore the qualities of test items

Deborah Oluwafunmilola Tobih¹, Musa Adekunle Ayanwale², Olufemi Abiodun Ajayi¹,
Modupe Victoria Bolaji¹

¹Department of Counselling Psychology and Educational Foundations, College of Specialized and Professional Education, Tai Solarin University of Education, Ijagun, Nigeria

²Department of Science and Technology Education, Faculty of Education, University of Johannesburg, Johannesburg, South Africa

Article Info

Article history:

Received Jan 16, 2022

Revised Feb 18, 2023

Accepted Mar 2, 2023

Keywords:

3-PL

Classical test theory

Item parameters

Item response theory

Uni-dimensionality

ABSTRACT

Psychometricians and researchers have grown increasingly concerned about students' performance. It is noteworthy from observation in the selected university of education in Nigeria that many of these students do not attend lectures for general courses particularly EDU 222 (tests and measurement) due to overcrowded lecture halls, the classroom climate factor, and other latent factors, such as exam conditions, test anxiety among others. Consequently, students' performance in EDU 222 does not significantly correlate with these challenges. The characteristics of the item parameter of the tests and measurement achievement test at the selected University of Education in Ogun State, Nigeria, were therefore examined in this study using measurement theories. The study employed an ex-post facto research of non-experimental design. Purposive sampling was used to select 6,203 second-year undergraduates offering EDU 222 for the 2018/2019 academic year. To analyze students' responses to 30 multiple-choice questions on tests and measurement, factor analysis, item calibration, an independent sample t-test with a significance threshold of 0.05 were utilized with empirical reliability of 0.64. The results showed the test items satisfied the three-parameter logistics model and uni-dimensionality assumption. The findings also showed that the item response theory (IRT) eliminated 7 items due to guessing, while the classical test theory (CTT) eliminated 13 items. The test is prone to guessing, thus, helps pupils do better on the exam. To establish parameter estimates during test development, higher education institutions (HEIs) should employ the use of IRT.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Modupe Victoria Bolaji

Department of Counselling Psychology and Educational Foundations, College of Specialized and Professional Education, Tai Solarin University of Education

Ijagun, Ogun State Nigeria

Email: bolajimodupe101@gmail.com

1. INTRODUCTION

Educational assessment measures academic progress and can be used to identify the different skills and concepts learnt by students throughout the year. Educators, through assessment, can then reflect on the impact of their teachings on students and use obtained information to plan appropriately for their students. As recommended by Rusell and Airasian [1], learning can be advanced and information supplied through the use of assessment. A school-based assessment is an important index of pedagogy that a professional teacher is expected to be skilled in for the attainment of the objectives of teaching and learning. The goals and

objectives of learning must be measured with precision and accuracy by teachers using a precise and accurate instrument that can only be handled by capable teachers [2].

Teacher's ability to provide the type of data required for students' performance is linked to the quality of assessment they provide. A well-designed test allows the teacher to calculate students' mastery of specific content taught in class in an accurate and consistent manner. Such results enable teachers to measure the extent and effectiveness of their classroom instruction and also allow parents know their children's progress. However, when test is poorly designed or poses poor parameter estimates, such a test can advance to in-accurate measurements of learning and produce false information concerning students' performance [3].

In recent times, empirical evidence and personal observation has reveals poor competency of teachers in educational assessments particularly in the region of test development, validation, and explanation of test scores. As stated by Abdullahi and Onasanya [4] in their study established that teachers elicit ignorance and incompetence in the development, validation, and use of test in educational assessments. They went on to say that bad test administration, poor handling of scores and feedback to students, insufficient coverage of instructional content, and incorrect interpretation of test scores are all issues that teachers in Nigeria face. From the foregoing, assessments in Nigeria education is faced with numerous challenges.

The challenges of educational assessments in schools coupled with the relevance of school-based assessment has lured the Nigerian university commission, and other tertiary institution stakeholders, particular education-based institutions to introduce educational courses particularly tests and measurement course into the curricular of teacher trainee. The introduction of tests and measurement course is geared towards fostering the requisite skills in educational measurement in the life of teachers in training bearing in mind that teachers without requisite skills in educational assessments expose students to poorly constructed test which in turn could hinder students' academic growth and cognitive development [5].

According to Hoover and Abrams [6], test and measurement as a course of study in education-based tertiary institutions, developed teacher's competency in formative, summation, and diagnostic evaluation, which are important assessments components in fostering student learning development. They stated further that the test and measurements offer teachers, understanding on how to arrange, control score and use the outcome for quality performance review and knowledge of feedback mechanism to improve learning. Thus, this implies that the potential benefits of tests and measurement course in tertiary institution of education cannot be underestimated.

The aforementioned relevance of tests and measurements as a course in teacher training program justify its inclusion in teacher education training program colleges of education and specialized university of Education in Nigeria. The course is a requisite general course for all undergraduates due to its relevance in equipping teachers in training with requisite educational assessment skills. Test and measurements as a course did not only equipped teachers with requisite skills in educational assessments but also helped to enhance teachers' competency in test construction, validation, and interpretation of tests scores to make meaningful academic or learning decision.

However, as laudable as the importance of tests and measurements course in teacher education program in specialized university of education in Nigeria, observations, and statistical report of students' performance in tests and measurements (EDU 222) in the premier University of Education in South-West Nigeria for four consecutive academic sessions shows some fluctuation in students' performance in the course as shown in Table 1. It was discovered that there was improvement in students' performance from 2015/2016 to 2018/2019 academic session as over 85% of the students pass the course EDU 222 at 45% pass mark. As seen in Table 1, it could be observed that the level of students' performance from each session was progressive except 2017/2018 academic session where there was a slight drop in percentage pass 88.80% while a relatively high performance of 97.95% was recorded in 2018/2019 academic session.

Table 1. Students' performance in tests and measurement in university of education for 2015/2016 to 2018/2019 academic sessions

S/N	Session	Total number of candidate who sat for the exam	Number passed at 45% and above	Percentage pass	Number failed	Percentage failed
1	2015/2016	4035	3520	87.24%	515	12.76%
2	2016/2017	5047	4813	95.36%	234	4.64%
3	2017/2018	5609	4981	88.80%	628	11.20%
4	2018/2019	6203	5995	97.95%	208	2.05%

The performance has become a great concern to psychometricians and researchers with regard to pertinent questions on quality and parameter estimates of the test items. It is worthy to state that from personal observation of students of the selected premier University of Education in Nigeria with regards to their engagement with educational courses particularly EDU 222, it is noted that majority of these students

do not attend educational courses lectures due to over congested lecture room, classroom climate factors and other latent factors such as lack of public address system, poor computer-based test examination conditions among others. From the foregoing, it can be deduced that the performance recorded so far by students in EDU 222 for four consecutive sessions in the selected university of Education do not significantly correlate with learning and examination challenges experienced by students in general courses in the university particularly in EDU222. This implies that the quality and psychometric properties of the test used in assessing students in test and measurement EDU222 need to be assessed.

Ajibade [7] noted that students' performance is excellent in tests that possess poor psychometric properties due to high chance of guessing, poor validation and standardization procedure put in place during test development of such test. Eleje, Onah, and Abanobi [8] stated that students tend to guess correctly and perform very well, when a test is poorly constructed or poses poor psychometric properties and parameter estimates. They mentioned further that a good test must be properly constructed, standardized, validated, and examined with regard to parameter estimates before administration or used in education assessment.

The type of test adopted in educational assessments of learners have relative importance to the educational decision made in this regard about the learner. Thus, it implies that a poorly constructed test items may infer negative learning outcomes and thus may have adverse effect on learning experiences of the learner [9]. Such items when answered correctly or incorrectly due to extraneous factors produced misleading feedback to both examinee and examiner [10]. When test is accurately constructed bearing in mind the principles of test constructions and psychometric properties of a good test, it provides teachers with extensive knowledge of the learner. Thus, teachers' ability to precisely and persistently measures students' proficiency of particular contents taught in class [9]; Such outcomes enable teachers to measure to some extent, the effectiveness of their teaching.

Multiple-choice test is one of the tests used by the teachers to assess the extent of behavioral objectives achieved by the students. According to Onuka and Ogbebor [11], multiple-choice tests are questions with three or more options, one of which is correct; the more options, the more reliable to a point. They went on to say that the option chosen is determined by the person's age and level of ability. They could be designed to investigate objectives in any field. The key element in a multiple-choice test is the stem which is the question, the key (the correct answer) and the distractor (the alternatives wrong answers). The ability of multiple-choice tests to cover myriad of course content without necessarily extending the testing time as well as its impartiality in scoring the examinees' responses has made this type of test to be salient [11].

In spite of the benefits of multiple-choice tests in educational assessments, it is usually faced with problems of written question often referred to as the stem and problem associated with alternative responses called the distractors, due to non-compliance of the principles of item writing [12]. When items are poorly constructed, particular if the distractors are not plausible enough, it may give room for guessing, thus may lead to poor test quality [8]. Therefore, in pursuance of excellence in the field of assessment, experts are needed in Nigeria to develop multiple choice test items that meet the expected psychometric properties using new techniques [12].

In measurement arena, there are two prominent theoretical frameworks used in determining the quality and usability of multiple-choice test: classical test theory (CTT) and item response theory (IRT). Initially, the CTT was the primary framework for creating and analyzing standardized tests. IRT was later developed to address the omission of and then CTT's role is complemented. The CTT assumption is based on the fact that a test-taker has an observed score (X) and a true score (T), which is typically an amalgamation of the test-true taker's scores, plus/minus some unobservable error [13]. The true score takes into account what the test-taker actually knows, but it is constantly contaminated by various sources of error. Measures of item characteristics, item difficulty, and item discrimination are used in the CTT, and their values are based on the distribution of examinee proficiency within a sample [14]. Ordinary people can easily determine, control, and comprehend CTT-based statistical indices, but they vary from sample to sample. CTT's main advantage is its relatively weak theoretical assumptions, which make it simple to use in a variety of testing situations [13]. The item difficulty and item discrimination statistics, both of which are sample dependent, are the cornerstones of CTT. It is unable to include tests that measure an examinee's proficiency level, because the CTT model lacks information about how an examinee is expected to perform on a specific item [15].

IRT was developed to conquer the pitfalls detected in CTT through development, interpretation and determination of item parameter estimates of test items. In contrast to the CTT's primary focus on test level information, IRT initially focuses on item-level information. IRT varies greatly from CTT by helping to establish association between examinee's ability and achievement on an item [16]. IRT purposively estimate the dormant attribute premised on people's answers to a series of items and using IRT approach to predict the psychometric properties of the items [17], [18]. These two theories are doubtless the major frameworks employed construct, assess, and analyze test items in educational measurements. Different assumptions and statistical techniques are used in these structures. They are worried about improving the overall quality of test

items rather than just constructing, assessing, or estimating test reliability and validity. Despite the fact that IRT differs significantly from CTT in theory and has some significant theoretical advantages over CTT, many examining institutions in Africa, such as the West African Examinations Council (WAEC) and the National Examinations Council (NECO), continue to use Classical Test Theory [19].

In light of the foregoing, a number of researchers have used a variety of data to assess the similarity of IRT and CTT item parameters. Item parameters estimate for both CTT and IRT were very similar as the findings established high association between item difficulties and discrimination indices of both CTT and IRT [20]–[27]. Both frameworks bring about very close discrimination and difficulties indices in the sampled test items [28], found a very high correlation between the difficulty indices obtained using the CTT and IRT approaches in his study using the MCQ Uganda Certificate of Education (UCE, 2015) examinations. A similar result was found when discrimination indices were compared. The item parameter estimates using the CTT and IRT approaches showed a high degree of correlation. Contrariwise, when comparing and testing for significance differences between item parameter estimates from CTT and those from 2-3-IRT across models for inter and intra model validation, there was no significant inter validation of item parameter estimates between CTT and 2-3-IRT models [29]. In support of previous studies [19], [29] from their study of basic education certificate examination, analysis indicate that CTT and IRT item statistics were just not comparable. It was indicated further that CTT deleted so much items than the IRT measurement framework. This finding suggests that the IRT framework is more advanced than the CTT framework in terms of item analysis, because under the CTT framework, good items were deleted based on the difficulty index. This could be due to the CTT framework's sample dependence of test items.

Based on the aforementioned, it could be observed that contrasting empirical evidence exists in literature as regards comparability and superiority of CTT and IRT in determination of item parameter estimates of multiple-choice test items were established using public examining bodies. It is majorly the examining bodies like WAEC, NECO, and NABTEB that are primarily concerned with finding the validity/parameters of the test elements and how test takers will react to them whereas in teacher-made classroom examinations, some best practices in item and test analysis are applied far too rarely. Moreover, it is necessary to ensure the appropriate validity and reliability of teacher-made tests because a test that is reliable and valid in one culture may not be in another. It is therefore crucial to make sure that careful cross-cultural adaptation has been made when using an instrument that has been validated in one culture with another population or culture.

EDU 222 is a course taken by all students at 200 level to ensure their knowledge of tests and measurements. It is certain that examination items for this course are pooled from different sources to examine the students. Determining the achievement of students therefore necessitate strict adherence to the guidelines for test design, administration, analysis, and reporting which many educators do not take cognizance of when constructing their test items. It is on this premise that the study is based. According to Ghaicha [30], teachers' knowledge of assessment and evaluation is a complex, dynamic, and ongoing activity rather than a static one, test developers are thereby encouraged to be concerned about the quality of test items they developed through the use of item analyses. This study, therefore, sought to compare the two theoretical approaches and also determine which among the two theories is superior in determining item parameter estimates of test and measurement (EDU 222) offered in education-based tertiary institution in Nigeria. The aim of this study therefore is to assess parameter estimates of test and measurement achievement tests (EDU 222) using both test theoretical framework (CTT and IRT).

The following research questions were posed to the participants: i) Does EDU 222 achievement test fulfill the unidimensionality assumption of IRT? ii) Which of IRT models best fits data generated from EDU 222 achievement test items?; iii) What are the CTT and IRT framework item parameters for EDU 222 achievement items? iv) There is no discernible distinction between item discrimination parameter of EDU 222 achievement test items using CTT and IRT frameworks? v) There is no discernible distinction between item difficulty parameter of EDU 222 achievement test items using CTT and IRT frameworks?

2. RESEARCH METHOD

The study adopted ex-post facto research of non-experimental design. The study population comprises of all 2018/2019 academic sessions of 200 level undergraduates' students of Tai Solarin University of Education (TASUED), Ijagun, Ogun State, Nigeria. A simple random sampling was used to select 6,203 second-year undergraduates' students of the 2018/2019 academic session who sat for the EDU 222 examination for that year. A proforma was used to collect the responses of students in the achievement tests from the information and communication technology (ICT) unit of the institution and using the keys to the questions. The data obtained was used for data analysis using statistical tools such as factor analysis and independent sample t-test to establish item parameters of the instrument with reference to CTT and IRT frameworks.

An analysis of principal component analysis (PCA) through orthogonal rotation was conducted to establish the instrument's dimensionality. Meanwhile, a maximum variance can be explained by the first principal component, then this variance can be expressed as a percentage of the total variance been used as an index of unidimensionality. Consequently, the larger the first component's contribution to variance, the closer the set of items is to being one-dimensional. However, there is an obvious question about how high this variance needs to be before one can conclude that a test is one-dimensional. Without any rationale, Hallett *et al.* [31] claimed that a set of items measuring one dimension should account for at least 40% of the total variance. According to Zopluoglu and Davenport [32], the first component should account for at least 20% of the total variance. Retaining only components with eigenvalues exceeding 1.0 [33]–[35] is a common strategy. As an indicator of unidimensionality, eigenvalues greater than 1.0 have been used. Due to the fact that achievement test items inherently satisfy this assumption of unidimensional, it is implicitly assumed that local independence applies to the item in question.

3. RESULTS AND DISCUSSION

A PCA implemented in SPSS software version 25 was used to determine the number of factors underlying the test items. The PCA technique reduces the dimensionality of a set of correlated features in the high dimension to a set of uncorrelated features in the low dimension. Principal components refer to these uncorrelated features. Also, the PCA is a linear orthogonal transformation, meaning that each principal component is perpendicular to its neighbor. Table 2 shows statistics of the unidimensionality of the 30 test items contained in the EDU 222 through total variance explained. Consequently, the first component of the model tries to explain as much variance as possible from the original data. It explained that the items contained in the test were similar and hung together on one distinct factor. The remaining factors have substantially less than 2.0 [36] suggested that if the first eigenvalue is large compared to the second and the second eigenvalue is not much larger than any of the others, then the test data may be considered approximately unidimensional. Similarly, Figure 1 indicates that a single factor underlies the test data which substantiated the earlier result obtained.

Table 2. Dimensionality of the test items

Component	Initial eigenvalues		
	Total	Of variance (%)	Cumulative (%)
1	2.04	6.81	6.81
2	1.55	5.15	11.96
3	1.47	4.91	16.87
4	1.22	4.07	20.95
5	1.17	3.91	24.85
6	1.10	3.66	28.52
7	1.09	3.64	32.15
8	1.08	3.60	35.75
9	1.06	3.53	39.28
10	1.04	3.48	42.76
11	1.03	3.42	46.18
12	1.00	3.34	49.51
13	0.99	3.30	52.81
14	0.97	3.24	56.05
15	0.97	3.23	59.28
16	0.95	3.17	62.45
17	0.94	3.13	65.59
18	0.92	3.08	68.67
19	0.91	3.05	71.71
20	0.90	2.99	74.71
21	0.89	2.97	77.68
22	0.87	2.88	80.56
23	0.86	2.87	83.43
24	0.84	2.79	86.22
25	0.81	2.71	88.94
26	0.81	2.69	91.62
27	0.74	2.47	94.09
28	0.72	2.41	96.51
29	0.56	1.85	98.36
30	0.49	1.65	100.00

Table 2 reveals that the first to second eigenvalues for the 30 items of the EDU222 achievement test were 2.04 and 1.55, respectively, which are greater than one. This implies that the first eigenvalue accounted for higher percentage of total variance among the observed variables or items, and it is used as an index of unidimensionality [32], [34]. The result is also in relation with the study [33], [37], stated that an eigenvalue greater than one implies that the factor contains same amount of information as a single variable. It is logical to deduce that the IRT models' unidimensional assumption was met for the EDU222 achievement test items used in the study. The result of eigenvalue test produced the scree plot of total variance associated with each factor are shown in Figure 1.

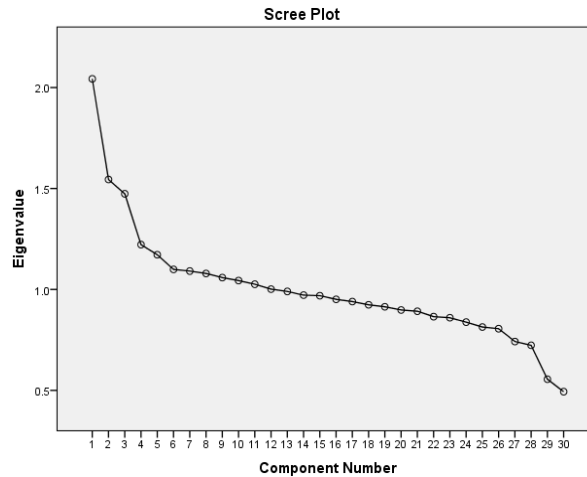


Figure 1. Scree plot for the multiple-choice test items

Furthermore, the model-data fit assessment was assessed using various indicators implemented in BILOG-MG software version 3.0. In determining the most parsimonious model that best fits the data structure, unidimensional models, such as one parameter logistic model (1PL), two parameters logistic model (2PL), and three parameters logistic model (3PL) of item response theory were compared using Akaike information criterion (AIC), Bayesian information criterion (BIC) and -2LogLikelihood (-2LL) ratio values as seen Table 3.

Table 3. Model-data fit assessment for the test items

Model	AIC	BIC	-2LogLikelihood
1PL	14600.5	14776.4	167311.04
2PL	12659.5	13328.9	165543.01
3PL	12250.8	13091.1	156976.83

Table 3 shows the values obtained for -2LogLikelihood (-2LL), AIC, and BIC for each model. To establish which model is most appropriate for the test data, the model with the lowest -2LL, AIC, and BIC values among the models is the most appropriate for model-data fit, and the 3PL model whose values were the lowest among the models given in Table 3, it was decided that 3PL model mostly fit the data structure. Thus, it was used to establish the item statistics of the test data. According to Thorpe and Favia [38], the goal of using the best model fit for the test items is to figure out which model is best for examinee responses.

In addition, based on the item analysis of the test items achieved through BILOG-MG software, Table 4 shows the CTT and IRT framework item parameters. The item difficulty indices are denoted by p , and the discrimination indices are denoted by RPBS in the CTT statistics. On the other hand, the IRT statistics use the letters 'b' for difficulty indices, 'a' for discrimination indices, and 'c' for guessing. More importantly, for the CTT framework, the benchmark for item difficulty and discrimination ranges from $0.20 \leq p \leq 0.80$ and $RPBS \geq 0.30$. Using this criterion, items whose difficulty and discriminating indices fall outside the range of 0.20 to 0.80 and less than 0.3 were considered poor [39]. However, unlike CTT models, which use statistics of discrimination and difficulty indices to determine good items, IRT models are more complicated. In this case, each item was assessed based on its contribution to the assessment model. Figure 2 illustrates the test information function that is produced for the test items in this assessment.

Table 4. Item parameter estimates for the test items

Items	P	RPBS	a	b	c
1	0.49	0.28	1.11	0.70	0.22
2	0.68	0.32	1.66	-0.10	0.34
3	0.64	0.35	2.37	0.23	0.38
4	0.49	0.11	1.14	1.98	0.41
5	0.64	0.44	2.29	-0.43	0.06
6	0.48	0.41	1.72	0.26	0.11
7	0.28	0.12	1.10	2.41	0.21
8	0.46	0.06	1.36	2.51	0.43
9	0.31	0.01	1.91	3.09	0.30
10	0.44	-0.02	0.01	83.31	0.26
11	0.28	0.07	2.08	2.36	0.26
12	0.69	0.36	1.47	-0.69	0.08
13	0.33	0.13	0.73	2.56	0.22
14	0.59	0.19	1.61	1.03	0.50
15	0.40	0.26	1.94	1.34	0.29
16	0.41	0.06	1.42	2.55	0.38
17	0.22	0.02	1.34	4.04	0.21
18	0.65	0.30	1.46	0.05	0.33
19	0.26	-0.05	2.08	3.52	0.26
20	0.49	0.26	1.07	0.80	0.24
21	0.77	0.32	1.22	-0.35	0.50
22	0.63	0.35	2.09	0.19	0.35
23	0.69	0.39	2.00	0.19	0.35
24	0.58	0.13	0.37	-0.00	0.18
25	0.89	0.25	1.45	-1.67	0.31
26	0.65	0.39	1.77	-0.49	0.06
27	0.77	0.31	1.66	-0.32	0.50
28	0.58	0.30	1.07	-0.10	0.14
29	0.33	0.03	0.26	9.36	0.28
30	0.53	0.34	1.91	0.58	0.30

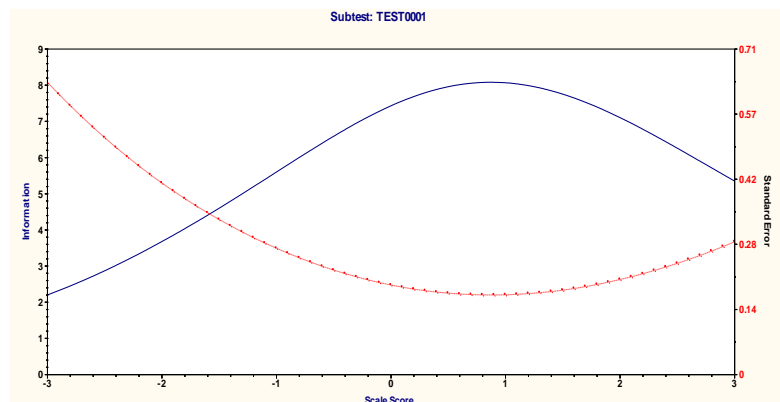


Figure 2. Test information function for the test items

Table 4 shows that the following items 2, 3, 5, 6, 12, 18, 22, 23, 26, 28, and 30 fall within the range and are considered good, while items 1, 4, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 19, 20, 21, 24, 25, 27 and 29 falls outside the criteria set for classical item statistics and are considered poor. Figure 2 also used the test information function to present items. The solid line depicts total data, while the dotted line depicts the standard error for a particular ability, implying that 7.8 was the maximum information provided at a competence level of 0.9 by the test items. According to Thompson [40] (that is the peak of the curve), Items with a difficulty level between -1.6 and 3.0 were considered good items based on the test information function. As a result, 2PLM kept 27 good items and eliminated three others (9, 10, and 25). The CTT removed more items than the 2PLM, whereas 3PLM accepted items 1, 2, 5, 6, 7, 9, 10, 11, 12, 13, 15, 17, 18, 19, 20, 22, 23, 24, 25, 26, 28, 29, and 30. that have low guessing parameters and rejected items 3, 4, 8, 14, 16, 21, and 27. that has a high level of guessing parameters [41]. Next, the researchers examined whether there is statistical evidence that the means of the two independent groups in terms of the item discrimination and difficulty parameters differ significantly or not. This feat is achieved using the independent samples t-test as seen in Tables 5 and 6.

Table 5 shows the mean value for CTT and IRT discrimination parameters with ($M=0.21$, $SD=0.14$; $M=1.45$, $SD=0.58$). The result implies that the IRT framework distinguishes better between those that know the materials and those that did not compare with CTT parameters. Also, the observed difference was further assessed using an independent sample t-test. The results revealed statistically significant mean differences in the discrimination of both frameworks with ($t=-12.67$, $df=29$, $p<0.05$). This implies that discrimination parameters coming from CTT and IRT differs across the test items. In essence, the result showed that IRT discriminates between high and low-ability examinees very well compared to the CTT framework.

Table 6 presents the mean value for CTT and IRT difficulty parameters with ($M=0.52$, $SD=0.17$; $M=1.94$, $SD=0.36$). The results imply that the test is difficult. Also, the observed difference was further assessed using an independent sample t-test. The findings revealed that there are statistically significant mean differences in the difficulty indices of both frameworks with ($t=-14.62$, $df=29$, $p<0.05$). This implies that difficulty parameters from CTT and IRT differ across the test items.

Table 5. Sample t-test of item discrimination parameter using CTT and IRT

	Group	Mean	SD	Mean diff.	t-value	p-value
Item discrimination	CTT	0.21	0.14	-1.24	-12.67	0.00
	IRT	1.45	0.58			

Table 6. Sample t-test of item difficulty parameter using CTT and IRT

	Group	Mean	SD	Mean diff.	t-value	p-value
Item difficulty	CTT	0.52	0.17	-1.42	-14.62	0.02
	IRT	1.94	0.36			

This study assess quality of parameter estimates of test and measurement achievement using measurement theories. Uni-dimensionality is a key feature of measurement theory that can be used to evaluate variable interrelationships. When the eigenvalue of the first value has higher percentage over the next one, uni-dimensionality is usually supported [37], [42]. The finding revealed that EDU222 achievement test sufficiently meets the assumption of unidimensionality, as determined by [33] using factor analysis with the eigenvalues-greater-than-one rule. Furthermore, as Zóltak and Golonka [41] pointed out that determining the IRT model fit is critical when drawing conclusions from any data. Assessing the model data fit of EDU222 considering the submission of previous study [38], [43], the best fit is provided by the model with the lowest AIC, BIC, and -2 likelihood ratio value. Thus, the finding shows that the 3-parameter logistic model provides the best fit to the data, enabling tendency for guessing.

In other words, result revealed that students' performance in EDU 222 test are determined by three factors: the intellectual ability of the students, difficulty level of the test items, and ability to guess correctly. Items with 3 parameter model greater than 0.35 are not good items [39], as a results of this, (23%) tests items were prone to guessing which makes students to perform better in the examination. This implies that when learner do not know for sure about the key to a certain item, they are attracted to the distracters and guessed randomly [44]. Also, students guess when they do not have high adequate knowledge or ability to provide correct answer. Furthermore, when comparing the quality of the test items CTT accepted 11 (36%) items as good items whereas IRT (3PLM) accepted 23 (77%) items and in terms of difficulty and discriminating indices, the CTT and IRT parameter estimates differed significantly. This is in line with the findings of [19], who discovered that item parameter estimates for CTT and IRT basic education certificates examinations were not comparable. This finding support the research [9] which found a disparity between CTT and IRT and explained further that in terms of very difficult items and the most appropriate model for data assessment is IRT. On the other hand, there is a substantial difference in between 3-parameter model and CTT.

4. CONCLUSION

The result concluded that EDU222 test items has tendency to guessing which boost students' performance and makes students pass successfully without attending lectures, having over congested learning environment and poor computer-based test examination conditions among others. The study recommends that item response theory method should be embraced by higher education institutions during their test development to establish parameter estimates and provision of quality assurance in measurement. Lecturers of EDU222 should introduce measures that could discourage guessing (like negative marking) in examination. Also, undergraduates from university of education should be exposed to psychometrics theories used in assessing test items through workshops and seminars. Finally, teachers in training should be allowed to learn different software packages that deal with item analysis.




REFERENCE

- [1] M. Rusell and G. Airasian, "Measurement and evaluation literacy levels of science teachers," Paper presented at the *Sixth Congress on National Science and Mathematics Education*, İstanbul: Marmara University, 2012.
- [2] S. Yi, R. Shadiev, R. Yun, and Y. Lu, "Developing and validating an instrument for measuring teachers' informatization teaching ability in primary and secondary schools in China for the sustainable development of education informatization," *Sustainability (Switzerland)*, vol. 14, no. 11, p. 6474, 2022, doi: 10.3390/su14116474.
- [3] L. Tay, A. W. Meade, and M. Cao, "An overview and practical guide to IRT measurement equivalence analysis," *Organizational Research Methods*, vol. 18, no. 1, pp. 3–46, 2015, doi: 10.1177/1094428114553062.
- [4] O. E. Abdullahi and S. A. Onasanya, "Challenges facing the administration of educational assessment measures at the secondary school level in Nigeria," *Journal of Applied Sciences*, vol. 10, no. 19, pp. 2198–2204, 2010, doi: 10.3923/jas.2010.2198.2204.
- [5] O. Y. Saka, "Development and validation of pool of 100 multiple choice test in Senior secondary school Biology," Unpublished M.Ed. Mini Project on Principle of Test Construction, Tai Solarin University of Education, 2016.
- [6] N. R. Hoover and L. M. Abrams, "Teachers' instructional use of summative student assessment data," *Applied Measurement in Education*, vol. 26, no. 3, pp. 219–231, 2013, doi: 10.1080/08957347.2013.793187.
- [7] O. E. Ajibade, "Opinions of prospective teachers on alternative assessment-evaluation methods," *Journal of Education and Research*, vol. 22, no. 1, pp. 123–142, 2014.
- [8] L. I. Eleje, F. E. Onah, and C. C. Abanobi, "Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results," *European Journal of Educational & Social Sciences*, vol. 3, no. 1, 2018.
- [9] O. B. K and A. A. Benson, "Using test theories models to assess senior secondary students ability in constructed-response mathematics tests," *Journal of Education and Practice*, vol. 11, no. 7, 2020, doi: 10.7176/jep/11-7-05.
- [10] M. A. Metibemu and T. Omole, "Achievement test in 21st century," In B.A. Adegoke, O. Popoola, and O. E. Babatunde, Eds., *Public Examining in Sub-Saharan Africa: Issues, Challenges and Prospects. A book of readings in honour of Emeritus Prof. 'Dibu Ojerinde*. Abuja, Marvelous Mike Press Limited, 2016, pp. 215–227.
- [11] A. O. U. Onuka and U. Ogbor, "An Introduction to assessment," In B.A. Adegoke, O. Popoola, and O. E. Babatunde, Eds., *Public Examining in Sub-Saharan Africa: Issues, Challenges and Prospects. A book of readings in honour of Emeritus Prof. 'Dibu Ojerinde*. Abuja, Marvelous Mike Press Limited, 2016, pp. 145–245.
- [12] M. A. Ayanwale, "Efficacy of item response theory in score ranking and concurrent validity of dichotomous and polytomous response mathematics achievement test in Osun State, Nigeria," Unpublished Ph.D. Thesis, University of Ibadan, 2017.
- [13] R. Jabrayilov, W. H. M. Emons, and K. Sijtsma, "Comparison of classical test theory and item response theory in individual change assessment," *Applied Psychological Measurement*, vol. 40, no. 8, pp. 559–572, 2016, doi: 10.1177/0146621616664046.
- [14] J. C. Cappelleri, J. Jason Lundy, and R. D. Hays, "Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures," *Clinical Therapeutics*, vol. 36, no. 5, pp. 648–662, 2014, doi: 10.1016/j.clinthera.2014.04.006.
- [15] O. O. Adedoyin, H. J. Nenty, and B. Chilisa, "Investigating the invariance of item difficulty parameter estimates based on CTT and IRT," *Educational Research and Review*, vol. 3, no. 2, pp. 83–93, 2008.
- [16] Z. Hu, L. Lin, Y. Wang, and J. Li, "The integration of classical testing theory and item response theory," *Psychology*, vol. 12, pp. 1397–1409, 1950, doi: 10.4236/psych.2021.129088.
- [17] F. M. Yang and S. T. Kao, "Item response theory for measurement validity," *Shanghai Archives of Psychiatry*, vol. 26, no. 3, pp. 171–177, 2014, doi: 10.3969/j.issn.1002-0829.2014.03.
- [18] J. O. Amusa, M. A. Ayanwale, L. A. Oladejo, and F. Ayedun, "Undergraduate Physics Test Dimensionality and Conditional Independence: Perspective from Latent Traits Model Package of R Language," *International Journal of Assessment and Evaluation*, vol. 29, no. 2, pp. 47–61, 2022, doi: 10.18848/2327-7920/CGP/v29i02/47-61.
- [19] M. A. Ayanwale, J. O. Adeleke, and T. I. Mamadelo, "An assessment of item statistics estimates of basic education certificate examination through classical test theory and item response theory approach," *International Journal of Educational Research Review*, vol. 3, no. 4, pp. 55–67, 2018, doi: 10.24331/ijere.452555.
- [20] M. A. Metibemu, "Comparison of classical test theory and item response theory in the development and scoring of senior secondary school physics tests in Ondo State," Unpublished Ph.D. Thesis, University of Ibadan, 2016.
- [21] O. L. Ojerinde, "Classical test theory (CTT) vs item response theory (IRT): an evaluation of comparability of item analysis results," Lecture Presentation at the Institute of Education, University of Ibadan, 2013.
- [22] N. Kohli, J. Koran, and L. Henn, "Relationships among classical test theory and item response theory frameworks via factor analytic models," *Educational and Psychological Measurement*, vol. 75, no. 3, 2015, doi: 10.1177/0013164414559071.
- [23] A. A. Bichi, R. Embong, R. Talib, S. Salleh, and A. Bin Ibrahim, "Comparative analysis of classical test theory and item response theory using chemistry test data," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5 C, pp. 1260–1266, 2019, doi: 10.35940/ijeat.E1179.0585C19.
- [24] B. A. Adegoke, "Comparison of item statistics of physics achievement test using classical test and Item Response Theory Frameworks," *Journal of Education and Practice*, vol. 4, no. 22, pp. 87–96, 2013.
- [25] M. Erguven, "Two approaches in psychometric process: Classical test theory & item response theory," *Journal of Education*, vol. 2, no. 2, pp. 23–30, 2013.
- [26] O. A. Awopetu and E. R. I. Afolabi, "Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination," *European Scientific Journal, ESJ*, vol. 12, no. 28, p. 263, 2016, doi: 10.19044/esj.2016.v12n28p263.
- [27] C. Brown, J. Templin, and A. Cohen, "Comparing the two- and three-parameter logistic models via likelihood ratio tests: A commonly misunderstood problem," *Applied Psychological Measurement*, vol. 39, no. 5, 2015, doi: 10.1177/0146621614563326.
- [28] S. Pido, "Comparison of item analysis results obtained using item response theory and classical test theory approaches," *Journal of Educational Assessment in Africa*, vol. 7, pp. 192–207, 2012.
- [29] O. O. Adedoyin, H. J. Nenty, and B. Chilisa, "Investigating the invariance of item difficulty parameter estimates based on CTT and IRT," *Educational Research and Review*, vol. 3, no. 2, pp. 83–93, 2008.
- [30] A. Ghaicha, "Theoretical framework for educational assessment," *A Synoptic Review Journal for Educational Practice*, vol. 7, no. 24, 2016.
- [31] N. Hallett, J. Huber, J. Sixsmith, and G. L. Dickens, "Measuring the violence prevention climate: Development and evaluation of the VPC-14," *International Journal of Nursing Studies*, vol. 88, pp. 97–103, 2018, doi: 10.1016/j.ijnurstu.2018.09.002.
- [32] C. Zopluoglu and E. C. Davenport, "A note on using eigenvalues in dimensionality assessment," *Practical Assessment, Research and Evaluation*, vol. 22, no. 7, pp. 1–10, 2017, doi: 10.7275/zh1k-zk32.




- [33] M. Björklund, "Be careful with your principal components," *Evolution*, vol. 73, no. 10, pp. 2151–2158, 2019, doi: 10.1111/evo.13835.
- [34] T. O. Ajeigbe and E. R. I. Afolabi, "Assessing Unidimensionality and Differential Item Functioning in Qualifying Examination for Senior Secondary School Students, Osun State, Nigeria," *World Journal of Education*, vol. 4, no. 4, pp. 30–37, 2014, doi: 10.5430/wje.v4n4p30.
- [35] D. Ojerindeand and B. Ifewulu, "Item unidimensionality using 2010 unified tertiary matriculation examination mathematics pre-test," A paper presented at the 2012 International Conference of IAEA, 2012, pp. 5–18.
- [36] F. M. Lord, *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum, 1980.
- [37] B. Williams, A. Onsmann, and T. Brown, "Exploratory factor analysis: A five-step guide for novices," *Journal of Emergency Primary Health Care*, vol. 8, no. 3, pp. 1–13, 2010, doi: 10.33151/ajp.8.3.93.
- [38] G. L. Thorpe and A. Favia, "Data analysis using item response theory methodology: An introduction to selected programs and applications," Psychology Faculty Scholarship, 2012.
- [39] F. I. Ukozor, F. E. Onah, and F. O. Ukah, "Application of three parameter latent trait model in development of secondary school economic achievement test," *Nigerian Journal of Educational Research and Evaluation*, vol. 16, no. 1, 2017.
- [40] N. Thompson, *Test information functions from item response theory*. ASC Smarter Assessment, 2021.
- [41] T. Zóltak and G. Golonka, "Does guessing matter? Differences between ability estimates from 2PL and 3PL IRT models in case of guessing," *EDUKACJA Quarterly*, vol. 134, no. 3, pp. 63–76, 2015.
- [42] M. A. Ayanwale, F. O. Isaac-Oloniyo, and F. R. Abayomi, "Dimensionality Assessment of Binary Response Test Items: A Non-parametric Approach of Bayesian Item Response Theory Measurement," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 9, no. 2, pp. 412–420, 2020, doi: 10.11591/ijere.v9i2.20454.
- [43] M. A. Ayanwale, J. O. Adeleke, and T. I. Mamadelo, "Invariance person estimate of Basic Education Certificate Examination: Classical test theory and item response theory scoring perspective," *Journal of the International Society for Teacher Education*, vol. 23, no. 1, pp. 18–26, 2019.
- [44] A. D. E. Obinne, "Using IRT in determining test item prone to guessing," *World Journal of Education*, vol. 2, no. 1, 2012, doi: 10.5430/wje.v2n1p91.

BIOGRAPHIES OF AUTHORS






Deborah Oluwafunmilola Tobih    holds her PhD in Tests and Measurements from Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria. She is currently an Associate Professor of Tests and Measurements in the Department of Counselling Psychology and Educational Foundations, College of Specialized and Professional Education in Tai-Solarin University of Education. Her research interests include, tests, measurements, mathematics education and gender Evaluation. She can be contacted at drtobih002@gmail.com.






Musa Adekunle Ayanwale    is currently a postdoctoral research fellow at the Department of Science and Technology Education, Faculty of Education at the University of Johannesburg, South Africa. He holds a PhD in educational research, measurement, and evaluation credentials him as an expert in the field of assessment and psychometrics. His research interests include testing theories, instrument development and validation, psychometrics, generalizability theory-based reliability analyses and evaluations, Q-Methodology, structural modelling, and computerized adaptive testing. His current research involves advancing technological solutions assessment for the purpose of developing a more effective and efficient method of educational assessment in Africa. He can be contacted at ayanwalea@uj.ac.za.



Olufemi Abiodun Ajayi    is an Associate Professor of science Education and educational evaluation in the Department of Counselling Psychology and Educational Foundations, Tai Solarin University of Education, Ijagun, Nigeria. Dr. Ajayi was a first-class graduate of Education/Biology from University of Nigeria, Nsukka. He bagged M.Ed and PhD in Educational Evaluation from University of Ibadan, Ibadan. He is currently the Postgraduate coordinator of his department. He can be contacted at ajayioa@tasued.edu.ng.



Modupe Victoria Bolaji    holds her Masters in Educational Evaluation from Tai Solarin University of Education, Ijagun, Ogun State, Nigeria. She had her first degree in the same University. Research interest includes educational management and developing test instrument. She can be contacted at bolajimodupe101@gmail.com.