Asian Journal of Advances in Research

5(1): 682-692, 2022

ITEM PARAMETER ESTIMATES OF UNIFIED TERTIARY MATRICULATION MATHEMATICS EXAMINATION USING ITEM RESPONSE THEORY 3 – PARAMETER LOGISTIC MODEL

CHRISTIANA AMAECHI UGODULUNWA [a], NJIDEKA GERTRUDE MBELEDE [a*], NNEKA CHINYERE EZEUGO [a], LYDIA IJEOMA ELEJE [a] AND IFEOMA CLEMENTINA METU [a]

[a] Evaluation, Research and Statistics Unit, Department of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

AUTHORS' CONTRIBUTIONS

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Received: 02 February 2022
Accepted: 06 April 2022
Published: 12 April 2022

Original Research Article

ABSTRACT

Item response theory has become a unique methodological framework for reviewing response data from assessments in education and other fields of study. This study attempted to assess the Item parameter estimates of Nigeria's unified tertiary matriculation mathematics assessment instrument using item response theory 3 – Parameter logistic model and determined the conformity level of the items of the instrument to the revised Bloom's taxonomy of educational objectives. The descriptive validation design was applied in the study. The population of the study comprised the 3, 320 SS 3 students registered for 2020 Unified Tertiary Matriculation Examination (UTME) who opted for mathematics in the public secondary schools in Abia State out of which 40 students were selected. Simple random sampling and purposive sampling techniques were adopted in selection of the samples using multistage procedure. Mathematics Assessment instrument used in the 2017, 2018 and 2019 UTM examinations were the instruments for data collection. This was made up of 120 (40 items for each year) multiple choice items, each having a single stem with four options including one correct answer and three distractors. *b, a* and *c-parameters* of the individual items of the instruments were established. The findings of the study revealed that a very high percentage (95% in 2017, 95% in 2018 and 85% in 2019) of the items were of moderate difficulty and discriminated adequately among high and low performing students. More so, the pseudo guessing parameter estimate indicated that there were low level of guessing since a high percentage ((90% in 2017, 92.5% in 2018 and 90% in 2019)%) of the items survived. The researcher recommended among others that examination bodies in Nigeria should concretize the quality of their test items by conducting item analysis using IRT model.

Keywords: Item parameter; tertiary matriculation; mathematics examination; 3-parameter logistic model.

**Corresponding author: Email: ng.mbelede@unizik.edu.ng;*

1. INTRODUCTION

The UTME (Unified Tertiary Matriculation Examination) is a screening instrument used to admit young school leavers into Nigerian tertiary institutions. It is a computer-based standardized test for assessing learning, primarily used to examine problem solving, critical thinking, knowledge of scientific concepts, and the importance of each subject taken. The exam had been paper-and-pencil until May 2014, when it was changed to computer-based. It is a test that determines how well someone has learned fundamental skills and competences. As a result, examinees are qualified to continue on to higher education institutions. Exams of high quality are necessary for providing accurate data that may be used to assess student learning, lead to program adjustments, and offer useful information to stakeholders. This poses a substantial burden on item writers who are striving to create proper test items [1]. According to Adeyinka [2], UTME-Computer Based Test (CBT) is a standardized achievement examination that examines individual students' learning performance based on a prescribed syllabus, indicating their preparedness and appropriateness for tertiary education. As a result, Joint Admission and Matriculation Board (JAMB) is paving the way for CBT to become a widespread means of institution examination in Nigeria. By accomplishing this achievement, the age of long months of waiting for results and the consequences associated with it has come to an end, since the majority of applicants now have their examination results disclosed in their entirety. However, it could be said that JAMB adopted these strategies to ensure credibility in its examination with the use of fingerprint detective device in registering and screening candidates prior to the UTME, checking and scanning photographs of candidates, and providing normal mathematical sets and calculator for candidates. As a result, it is necessary to be aware of the format and content of such essential exams. Ojerinde [3] and Ukomadu and Fabian [4] correctly observed that the traditional paper-and-pencil educational assessment techniques in Nigeria may not adequately prepare children to tackle global advancements linked with the desired improvement in technology and education. Furthermore, according to Ubulom and Wokocha [5] and Adebayo [6], using CBT simplifies the entire testing cycle, including test generation, execution, evaluation, and presentation. The authors point out that CBT benefits, such as standardization of test administration conditions, allow test developers to increase their productivity and lead to innovation in their fields, and that CBT enables developers to set the same test conditions for all participants regardless of test population size. The authors of this research believe that these arguments might be recognized as real if the findings obtained through CBT forms are legitimate and accurate.

Multiple Choice Questions (MCQs), on the other hand, have been used by the Joint Admission and Matriculation Board (JAMB) since its inception, among other types of tests. In MCQs, examiners can cover a wide range of topics and learning objectives with properly written questions, and students can choose from a variety of response possibilities. This takes into account Bloom's taxonomy of educational aims, which was amended in the mid-nineties by Lorin Anderson and David Krathwohl as recall, comprehend, apply, analyze, evaluate, and create [7]. Despite that the fact JAMB mathematics assessment tests are uniform; the pass rate of prospective undergraduate students has not been consistent. Since the UTME components are standard tests, this has been attributed to structure and sensitivities. More crucially, Ashikhia [8] identified a number of elements that could affect examinees' UTME Mathematics results - the type of the test items and the characteristics of the examinees. The qualities of a test item can be used to explain an examinee's performance on it. This has become a cause of concern for university education stakeholders. This is because it is still the exclusive method of admission into universities and the predictive potential of this entrance exam must be evaluated on a regular basis. Regardless of significant improvements in teaching, research, and community service, it is noted that studies on the assessment of higher entry examinations are still scarce.

Psychometric qualities are commonly used to describe the quality of an evaluation instrument. The quality of tests, on the other hand, can be determined, at least in part, from the analysis of test items. As a result, it's critical to review and analyze the assessment items once they have been applied. This type of analysis and evaluation is required to improve the items and define the assessment characteristics, such as whether the item is performance-oriented, the thinking sequence it evaluates, and the item's real-life context [1]. Ayanwale, Adeleke, and Mamadelo [9], viewed that developed countries such as the United States, Canada, Ireland, and Germany have highly developed mathematics education programs at the primary, secondary, and post-secondary levels, allowing them to achieve significant success in their respective countries. Mathematics education at the primary, secondary, and post-secondary levels should be carefully handled for any growing country like Nigeria to advance technologically and improve its social and economic standing. Given that numeracy, reasoning, thinking, and problem-solving skills can be exhibited via the learning and application of mathematics, students must develop a greater interest in mathematics and have a good understanding of the basic concepts and fundamental principles [10]. As a result, the researchers used the IRT 3 – parameter logistic model to evaluate the item parameter estimates and level of conformance to the revised Bloom's taxonomy of educational objectives of University Tertiary Matriculation Examination (UTME) assessment items in mathematics.

1.1 Item Response Theory (IRT)

Item response theory is a modeling approach for latent variables, used to reduce preference and ameliorate the measuring power of examinations in educational and psychological as well as other psychometric operations. It uses statistical structural designs that set out to provide an explanation for how test takers react to problems, hence the names "latent trait theory" and "true score theory." It is a testing proposition primarily based on the affiliation among test takers performance ranges on all the capabilities that the test was supposed to measure and the individual test item. The outline is extremely estimable in and of itself; this framework can be used to assess item's overall performance. IRT refers to a group of statistical framework that are targeted at explaining the interconnectedness among hidden characteristics, unnoticeable attributes and their perceptibility, i.e., noticeable outcomes, responses, or overall performance [11]. These frameworks make a connection between the characteristics of items, how people respond to them, and the underlying attributes they measure in a test making use of parameters *b*, *a* and *c*.

1.2 Item Parameter

The psychometric properties of an instrument are referred to as its item parameters. These are statistical indices that determine the level of excellence of a given test item. They include the item fit model, item difficulty, and item discrimination, pseudo-guessing, item information curves, and an item's test information function. These come in handy when choosing items for an instrument. In line with requirements for educational testing, periodic evaluations of the test item stability are necessary to ensure correct accountability [12] because, based on them, results are published. For the sake of this study, item difficulty, item discrimination and guess parameter were applied.

b-parameter:

The *b-parameter* (Item difficulty), exposes the number of students, who correctly respond to an item which depicts the extent of content mastery. The proportion of students who correctly respond to an item corresponds to the level of difficulty. This affects the capacity of an item to distinguish between those who mastered the content being evaluated and those who don't [13,14].

The formula used to calculate the DIF is

$$p = \frac{Ru + Rl}{T} x100$$

Where:

RU = the number in the upper group (quartile Q1) who answered the item correctly.

RL = the number in the lower group (quartile Q3) who answered the item correctly.

T = the total number who tried the item

The p-value which describes the difficulty index ranges from 0-100, with 100 representing the percentage of students that correctly answered the question [13,14]. The higher the index, the simpler the question is. According to Mahjabeen et al. [13], a question is considered good and acceptable if the p-value is between 20 and 90. Furthermore, items with a p-value of less than 20% are too difficult, while those with a p-value of more than 90% are too simple, both of which are unacceptable and should be revised. But Mukherjee & Lahiri [14] had a different description of item difficulty. They reported that items with a p-value of >70% are too simple, 30-49% are average, 50-60% are good, and less than 30% are too difficult. The Table 1 displays the range of index acceptable for this study.

a-parameter:

The item discrimination indicator (DI) can be referred to as the *a-parameter*. Item discrimination can be used to determine the effectiveness of an item in MCQs for discriminating between high-performing and low-performing students [2,13,14]. When taking the DI, test takers were split into quartiles. The highest score students were in the upper quartile. To calculate the DI, first compute the DIF for the upper, lower, and then subtract the DIF from the lower quartiles [14,15] Index of Discrimination = DU - DL.

Conventionally, item discrimination uses a value range of -1 to +1. A higher value item is considered to be discriminating and effective. Mukherjee and Lahiri [14]; Musa Shaheen, Elmardi, Ahmed [16] explain that if all of the test takers from the upper quartile answer correctly, the DI value will be +1.00. If the lower group correctly answers the item and there is no one from the upper group, then the DI value will be -1.00 [17,14]. This may be due to item flaws [18]. Rasiah and Isaiah in Musa et al. [16], wrote that these items should be reviewed carefully for any common causes of poor discrimination such as ambiguous wording or grey areas of opinions, wrong keys and areas of controversy. They further stated that a DI value of 1.00 indicates perfect discrimination among high-performing and low-performing students. If the DI value is less than one, the item should not be allowed to be part of the examining items. Items with a DI greater than 0.40 should be considered exceptional. Items with a DI below 0.25 should be considered marginal. Items with DI values less than 0.25 must be removed. Mahjabeen et al. [13] classified DI items as: items with a value of 0.36 are exceptional, 0.25- 0.35 are good, 0.21- 0.24 are acceptable and 0.20 are poor. In this study, Kelley's formula of 1956 was applied and the range of item discrimination used is shown in Table 2.

c-parameter:

This is also known by pseudo guessing and pseudo chance parameters. The Model predicts the likelihood that a correct reply will be received in the same manner it does the 1 - PL Model. However, the Model is constrained by a third parameter, the guessing parameter. This restricts the probability that a correct response will be given if the respondent has the ability to answer -. The amount of information given by an item drops as respondents guess, and the information item function rises in importance relative to other functions. Items where respondents answer by guessing indicate that their ability is less than its difficulty [11]. Examinees guess because of a lack of knowledge or ability. Mehrens & Lehmann in Meyer [18] identified two types of guessing:

1) Blind guessing - Where the examinee picks an answer at random among all possible options.
2) Informed guessing: Informed guessing is where the examinee applies all his knowledge and abilities in order to find the most correct answer. Normal circumstances are very sparse for blind guessing. Evidence and logic show that more test situations require informed guessing. Students who are motivated to succeed will make use of information available to them to find the right answer. They will eliminate all implausible possibilities and not be restricted to blind guessing. Two main reasons educators try to discourage guessing are:

- First, there is the moral belief that guessing is wrong or sinful since it is gambling.
- Again, the psychometric properties of the test can be affected by guessing. It is important to discourage guessing by providing instructions on the test and scoring the test so that those who make incorrect guesses are penalized by formula scoring (correction-for-guessing). These procedures have been controversial for years. The pseudo chance parameters (S) can be calculated manually using the formula:

$$S = R = \frac{w}{A-1}$$

Where: S = Corrected score, R = number of right answers, W = number of wrong answers and A = number of alternatives per item.

Table 1. Range of difficulty index used in the study

Range of Difficulty Index	P-value (%)	Interpretation	Action
0 – 0.39	< 40	Difficult	Revise or Discard
0.40 – 0.75	40 – 76	Right Difficulty	Retain
0.76 – above	>76	Easy	Revise or Discard

Table 2. Range of item discrimination used in the study

Range of Discrimination index	Quality of an Item	Action
≥0.60	Excellent	Definitely Retain
0.40 – 0.59	Very Good Item	Very Usable
0.25 – 0.39	Good Item	Usable
0.20 – 0.24	Potentially Poor Item	Consider Revising
≤0.20	Very poor item	Possibly Revise Substantially or Discard

1.3 Research Questions

1. What is the difficulty index of the items of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB using 3-parameter model?
2. What is the item discrimination estimate of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB?
3. What is guessing parameter of the items of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB?
4. What is the extent of conformity of mathematics MCQ for UTME from 2016-2018 as constructed by JAMB to the revised Bloom's levels of objectives?

2. RESEARCH METHOD

In the study, the Descriptive Validation Design was used. This design allows the researchers describe the study in detail and to examine whether the instrument being used is consistent and does the intended thing [19,20]. The 3,320 senior secondary 3 (SS 3) students who were registered for the 2020 UTME in mathematics were the population of the study and 40 subjects were selected. Out of the 40 students, 18 were males, and 22 were females. These students were selected as prospective undergraduates in JAMB mathematics courses for the academic years 2020-2021. The multistage selection process used simple random sampling and purposive sampling to select the sample for the study. From the three Abia State, education zones, one was chosen. Two public schools were then selected using simple random sampling. The study was limited to the schools that were registered for math-related programs in UMTE. The Mathematics Assessment instrument, which was used in the UTM 2017-2018 and 2019 examinations, was used to collect data. It consisted of 120 multiple choice items (40 each for each year), each with a single stem that had four options. This included one correct answer, three distractors, and incorrect answers. Each item was given one mark. 120 was the highest score, and was converted to percent for easy analysis. There was no minimum score and no item was excluded. The examinations were between 30 and 45 minutes long, just like the UTM examinations. They were completed in three sessions (at two week intervals). The JAMB assessment scheme was used to create a plan of assessment (test blueprint) before the instrument was created. To ensure proper balance and focus on the syllabus, the content areas that would be covered and the cognitive levels that would be reached were included in the scheme of assessment. The Bloom's revised taxonomy of educational goals was used to map the assessment items. Each item covered one level of the Bloom's educational objectives: understanding, remembering, applying, analyzing and evaluating, as well as creating. This was performed by the researchers, 2 measurement and evaluation experts. The instruments were included in the mathematics mock examinations of respondents. They were administered to 40 students as research aid by the math teachers from the sampled schools. The students were explained to why they were subject to this type of assessment. Identification tags were also given to the selected students to make it easy to identify which seats had the samples.

3. RESULTS

1. **What is the difficulty index of the items of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB using 3-parameter model?**

Table 3. Difficulty index of the items of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB using 3-parameter model

Items	2017	Remarks	2018	Remarks	2019	Remarks
Q1	0.727	Good	0.695	Good	0.604	Good
Q2	0.610	Good	0.649	Good	**0.890**	Bad
Q3	0.643	Good	0.623	Good	0.597	Good
Q4	0.714	Good	0.558	Good	0.558	Good
Q5	0.610	Good	**0.312**	Good	0.669	Good
Q6	0.701	Good	0.584	Good	0.584	Good
Q7	0.708	Good	0.623	Good	0.630	Good
Q8	0.662	Good	**0.773**	Good	**0.776**	Bad
Q9	0.461	Good	0.630	Good	0.539	Good
Q10	0.623	Good	0.617	Good	0.565	Good
Q11	0.351	Good	0.429	Good	0.617	Good

Items	2017	Remarks	2018	Remarks	2019	Remarks
Q12	0.127	Good	0.383	Good	0.578	Good
Q13	0.494	Good	0.584	Good	0.494	Good
Q14	0.481	Good	0.636	Good	0.519	Good
Q15	0.519	Good	0.506	Good	0.604	Good
Q16	0.338	Poor	0.649	Good	0.532	Good
Q17	0.584	Good	0.649	Good	0.532	Good
Q18	0.617	Good	0.545	Good	0.604	Good
Q19	0.545	Good	0.481	Good	0.636	Good
Q20	0.662	Good	0.662	Good	0.632	Good
Q21	0.623	Good	0.597	Good	0.552	Good
Q22	0.597	Good	0.368	Good	**0.869**	Bad
Q23	0.617	Good	0.552	Good	0.571	Good
Q24	0.636	Good	**0.291**	bad	**0.832**	Bad
Q25	0.474	Good	0.545	Good	0.526	Good
Q26	0.714	Good	0.578	Good	0.468	Good
Q27	0.545	Good	0.682	Good	0.643	Good
Q28	0.571	Good	0.552	Good	0.636	Good
Q29	0.481	Good	0.708	Good	0.526	Good
Q30	0.623	Good	0.669	Good	**0.799**	Bad
Q31	0.630	Good	0.565	Good	0.526	Good
Q32	0.571	Good	0.643	Good	**0.235**	Bad
Q33	0.519	Good	0.682	Good	0.591	Good
Q34	0.591	Good	0.688	Good	0.558	Good
Q35	**0.383**	Poor	0.610	Good	0.565	Good
Q36	**0.331**	Poor	0.461	Good	0.519	Good
Q37	0.630	Good	0.526	Good	0.584	Good
Q38	0.688	Good	0.558	Good	0.571	Good
Q39	0.662	Good	0.552	Good	0.565	Good
Q40	0.623	Good	0.526	Good	0.578	Good

Table 4. Summary of difficulty index of items of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB

Year	Easy items	Moderate Difficulty	Very Difficult items	Total items
2017	0	38	2	40
%	0	95	5	100
2018	1	38	1	40
%	2.5	95	2.5	100
2019	5	34	1	40
%	12.5	85	2.5	100

Table 4 shows that 95% and 85% of the items were moderately difficult in the 2017-2018 examination years. Although no item was considered difficult in 2017, 12.5% and 5% respectively of the 2018 and 2019 items were considered to be very easy. Only 5% of the items in 2017 were considered very difficult, while 2.5% in 2018 was and 2019. These items need to be reviewed for future exams.

2. **What is the item discrimination estimate of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB?**

Table 5. Item discrimination estimate of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB

Items	2017	Remarks	2018	Remarks	2019	Remarks
Q1	0.350	Good	0.350	Good	0.275	Good
Q2	0.325	Good	0.250	Good	0.250	Good
Q3	0.250	Good	0.300	Good	0.250	Good
Q4	0.350	Good	0.250	Good	0.500	Very good
Q5	0.735	Excellent	**0.100***	Poor	0.400	Very good
Q6	**0.175***	Poor	0.320	Good	0.250	Good
Q7	0.325	Good	0.750	Excellent	0.300	Good
Q8	0.325	Good	0.300	Good	**0.200***	Poor
Q9	0.260	Good	**0.200***	Poor	0.250	Good
Q10	0.325	Good	0.375	Good	0.275	Good
Q11	0.250	Good	0.275	Good	0.750	Excellent
Q12	**-0.050***	Very poor	0.275	Good	0.250	Good
Q13	0.375	Good	0.300	Good	0.270	Good
Q14	0.400	Very Good	**0.200***	Poor	0.250	Good
Q15	0.275	Good	0.275	Good	0.250	Good
Q16	0.600	Excellent	0.350	Good	0.300	Good
Q17	0.260	Good	0.375	Good	0.252	Good
Q18	0.620	Excellent	0.275	Good	0.275	Good
Q19	0.255	Good	0.250	Good	0.325	Good
Q20	0.250	Good	0.300	Good	0.250	Good
Q21	0.275	Good	**0.175***	Poor	0.275	Good
Q22	0.500	Very Good	**0.225***	Poor	**0.000***	Poor
Q23	0.255	Good	0.325	Good	0.325	Good
Q24	0.400	Very Good	**-0.125***	Very Bad	**-0.050***	Very poor
Q25	0.275	Good	**0.175***	Poor	0.250	Good
Q26	**0.125***	Poor	0.300	Good	0.500	Very good
Q27	**0.100***	Poor	0.625	Excellent	0.475	Very good
Q28	0.250	Good	**0.200***	Poor	0.250	Good
Q29	0.500	Very Good	**0.000***	Poor	0.750	Excellent
Q30	**0.100***	Poor	0.270	Good	0.250	Good
Q31	0.250	Good	0.300	Good	0.625	Very good
Q32	0.500	Very Good	0.475	Very Good	**0.200***	Poor
Q33	0.750	Excellent	0.400	Very good	0.275	Good
Q34	**0.075**	Very poor	0.250	Good	0.425	Very good
Q35	0.275	Good	0.250	Good	0.265	Good
Q36	**0.100**	Poor	0.250	Good	0.275	Good
Q37	0.475	Very Good	**0.100**	Poor	0.250	Good
Q38	0.275	Good	0.475	Very good	-0.025	Very poor
Q39	0.400	Very Good	0.250	Good	0.300	Good
Q40	0.375	Good	0.270	Good	0.275	Good

Table 6. Summary of item discrimination estimate of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB

Year	Excellent items	Very Good items	Good items	Poor items	Very poor items	Total items
2017	4	7	22	5	2	40
%	10	17.5	55	12.5	5	100
2018	2	3	25	9	1	40
%	5	7.5	62.5	22.5	2.5	100
2019	2	6	27	3	2	40
%	5	15	67.5	7.5	5	100

From Table 6, it was evident that only 17.5%, 23% and 12.5% of the items of the assessment instruments in 2017, 2018 and 2019 respectively were poor items since the items could not discriminate clearly between high and low performers among the test takers. Such items as identified on Table 5 needs to be reviewed or totally expunged from the pool.

3. What is the pseudo-guessing parameter estimate of the items of the mathematics MCQ for UTME from 2017-2019 as constructed by JAMB using a 3-parameter model?

Table 7. Pseudo-guessing parameter estimate of the items of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB using IRT 3-parameter model

Items	2017 Asymptote	Remarks	2018 Asymptote	Remarks	2019 Asymptote	Remarks
Q1	0.211	Good	0.000	Good	**0.270***	Bad
Q2	0.200	Good	0.000	Good	0.210	Good
Q3	0.000	Good	0.200	Good	0.000	Good
Q4	0.000	Good	0.000	Good	0.000	Good
Q5	0.000	Good	0.160	Good	0.000	Good
Q6	0.000	Good	0.242	Good	0.210	Good
Q7	0.133	Good	0.000	Good	0.000	Good
Q8	0.000	Good	0.000	Good	0.000	Good
Q9	0.210	Good	0.000	Good	0.172	Good
Q10	0.001	Good	0.172	Good	0.141	Good
Q11	0.002	Good	0.002	Good	0.001	Good
Q12	**0.300***	Bad	0.000	Good	0.210	Good
Q13	0.110	Good	0.000	Good	0.210	Good
Q14	0.000	Good	0.000	Good	0.200	Good
Q15	**0.333***	Bad	0.000	Good	0.183	Good
Q16	0.181	Good	0.211	Good	0.002	Good
Q17	**0.310***	Bad	0.000	Good	0.001	Good
Q18	0.001	Good	0.000	Good	0.240	Good
Q19	0.202	Good	0.000	Good	0.210	Good
Q20	0.000	Good	0.210	Good	0.00	Good
Q21	0.000	Good	0.000	Good	0.000	Good
Q22	0.020	Good	0.000	Good	**3.722***	Bad
Q23	0.000	Good	0.170	Good	**0.325***	Bad
Q24	0.100	Good	**0.416***	Bad	0.050	Good
Q25	0.000	Good	0.000	Good	0.250	Good
Q26	0.000	Good	0.210	Good	0.500	Good
Q27	0.000	Good	0.210	Good	0.000	Good
Q28	0.210	Good	0.202	Good	0.000	Good
Q29	0.200	Good	0.181	Good	0.200	Good
Q30	0.000	Good	0.001	Good	0.000	Good
Q31	0.000	Good	0.000	Good	0.160	Good
Q32	0.000	Good	**0.262***	Bad	**0.275***	Bad
Q33	0.000	Good	0.170	Good	0.000	Good
Q34	0.130	Good	0.000	Good	0.170	Good
Q35	0.000	Good	0.000	Good	0.000	Good
Q36	**0.270***	Bad	0.000	Good	0.000	Good
Q37	0.000	Good	0.000	Good	0.000	Good
Q38	0.000	Good	0.000	Good	0.000	Good
Q39	0.001	Good	0.210	Good	0.000	Good
Q40	0.112	Good	0.000	Good	0.000	Good

Table 8. Level of conformity of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB to the revised Bloom's levels of objectives

Performance Characteristics	Remembering and Understanding (%)	Application (%)	Analyses (%)	Evaluation and Creation (%)
2016	33	16	26	25
2017	33	20	20	27
2018	36	28	23	13

The Table 7 displays the item evaluation using the pseudo-guessing criteria. This ranges from 0 to 1. A good item can only be described as having a value of *ci* that is less than 1/kth of the number of possible choices [18]. There were four options in the UTME math assessment instruments. This premise suggests that the cutoff (ci) value is 0.25. Based on this criterion, 4 items, 3 items, and 4 items were classified as poor in the 2017, 2018 and 2019 assessment tools, respectively. All other items survived. Warm in Obinne [21] stated that items have a wide range of C-values, ranging from 0.00 to 0.40. C-values above.30 are not considered to be very good. It is recommended to have a C value of.20 or lower.

Research Question 4: What is the extent of conformity of mathematics MCQ for UTME from 2017-2019 as constructed by JAMB to the revised Bloom's levels of objectives according to gender?

Table 8 lists the percentages for each year of study of the instruments and the objective levels to which they are conformed. An increased proportion of items correspond to the understanding and remembering level. Test experts classified the items according to Bloom's educational goals levels. From Table 8, it was clear that each item covered one level of Bloom's educational learning objectives levels.

4. DISCUSSION

The current study focused on the quality assessment items for the mathematics MCQ for UTME (2017-2019) as constructed and conducted by JAMB. The study yielded three important and valuable results that may be of value to item developers as well as examination boards. (1) Very high percentages of the assessment instruments were of moderate difficulty. They also discriminated well between students who perform well and those who do not. The 2018 assessment instrument showed that 25% of the items did not discriminate among test takers. However, this is still a small percentage when compared to the other 75%. These findings exonerate the developers of the UTME assessments instruments. (2). It also shows that there was no reason for the low performance across the years due to extremely difficult items. (3) The 2018 and 2019 UTME mathematics assessments items were closely aligned with the revised Bloom's taxonomy, which had an impact on their psychometric properties.

These results show that the C-values of 2017 items ranged between 0.0001 and 0.333. C-values of all items fell within the recommended range. About 36 (90%) of 40 test items had C values below 0.250. That is the most desirable C value. The C-values for four items (items 12, 15, 17 and 36) were higher than the recommended range. This indicates that these items are not very good as they are easy to guess. C-values for 2018 mathematics UTME items ranged from 0.000 up to 0.416. Only 38 items (95%) had the desirable C value of 0.250 or less. This demonstrated that most items were of good quality. A few items were not good enough. C-values of 0.250 and less were the norm for the majority of items. This indicates that the items are good with low chances of being correctly answered by low-ability examinees. This finding is consistent with those of Obinne [21] and Tjabolo and Otaya [22]. If low-ability examinees do not know the answer, they are more attracted to distractors that the correct answer. This means they get the item wrong much more often than if it was guessed randomly. Evidently, JAMB, other examining bodies in Nigeria and elsewhere have been concerned about the quality and response of examinees to test items. Items writers should be aware of their ability to guess and avoid creating too complicated items that are easy to guess. Setiawati, Setiawati and EkaIzzaty [23] highlighted that educational practice has one of its main tasks, which is the development of test items that measure the learning aspects with the greatest precision. IRT is gaining popularity due to the advancements in psychometrics, computer adaptive testing, and other areas. Adedoyin and Mokobi [24] state that IRT offers a number advantages over CTT, including the ability to evaluate learning, develop better measures and assess change over time. Its models give invariant trait and latent item estimates. IRT psychometric methodology has been used to resolve assessment challenges. IRT is used to determine differential item functioning (DIF), or distinguish between biases and real differences in traits within groups. This tends to be the case.

5. CONCLUSIONS

According to the study, 118 of the one hundred twenty (120) items that were studied over the three years 2017-2018 and 2019 were classified as good test items. The revised Bloom's Taxonomy of Educational Objectives was used to classify the items. When large-scale testing is required, such as the JAMB UTME, it is important to analyze test data in order to determine the quality of the test. Education assessment efforts are judged by the quality of the instruments used, which is the tools and techniques used. Badly designed instruments can result in a wasteful use of time and money. The main source of information regarding student achievement at school is the educational test.

6. RECOMMENDATIONS

The researchers recommended that:

To improve the quality and reliability of the test items, especially in Nigeria, examining bodies should look into using the IRT model. This will undoubtedly strengthen the instrument's validity. I wish that Nigerian psychologists would embrace the challenges of the measurement community and recognize the importance of IRT when developing high-quality exam items. It is essential to shift from CTT to IRT in order to construct and analyze items in tests. This is especially true for Nigerian public examinations.

It is important to recognize the possibility of guessing when writing an item. Use the IRT method for item analysis during construction to eliminate items that are prone or likely to guess. This will ensure that the item is not blamed for guessing.

Teachers in Nigerian schools who do not have adequate knowledge in the area of measurement and tests should be permitted to take on-the job training.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. AlKhatib HS, Brazeau G, Akour A, Almuhaissen SA. Evaluation of the effect of items' format and type on psychometric properties of sixth year pharmacy students clinical clerkship assessment items. BMC Medical Education. 2020;20.
2. Adeyinka T. Variables that may determine secondary school students' preparedness for Utme-Cbt. Global Scientific Journal; 2019.
3. Ojerinde D. Innovations in assessment: JAMB experience, Nigerian Journal of Educational Research and Evaluation. 2015; 14(3).
4. Ukomadu C, Fabian B. Effect of public service reform on service delivery of Joint Admissions and Matriculation Board (JAMB) in Nigeria. Quest Journals Journal of Research in Humanities and Social Science. 2018;6(8):41-49.
Available:www.questjournals.org
5. Ubulom WJ, Wokocha KD. Readiness and acceptability of computer-based test (CBT) for post-university matriculation examinations (PUME) among urban and rural senior secondary school students in Rivers State. International Journal of Innovative Social & Science Education Research. 2017;5(3):51-60.
Available:www.seahipaj.org
6. Adebayo FO. Using computer based test method for the conduct of examination in Nigeria: prospects, challenges, and strategies. Mediterranean Journal of Social Sciences MCSER Publishing, Rome-Italy. 2014;5(2). DOI: 10.5901/mjss.2014.v5n2p47
7. Wilson CO. Understanding the New Version of Bloom's Taxonomy; 2016.
Available:https://quincycollege.edu/wp-content/uploads/Anderson-and-Krathwohl_Revised Blooms Taxonomy.pdf.
8. Asikhia OA. Students and teachers' perception of the causes of poor academic performance in Ogun state secondary schools: Implications for counseling for National development. European Journal of Social Sciences. 2010; 13(2):28-36.
9. Ayanwale MA, Adeleke JO, Mamad A. Relational analysis of personal variables and marking skills of national examinations council's examiners. African Journal of Pedagogy, Kampala International University College, Tanzania. 2018;8:25-38.
10. Adegoke BA. Comparison of item statistics of physics achievement test using classical Test theory and item response theory frameworks. Journal of Education and Practice. 2013;4(22): 87–96.
11. Columbian Public Health; 2021.
Available:https://www.publichealth.columbia.edu/research/population-health-methods/itemresponse theory
12. Mehta G, Mokhasi V. Item analysis of multiple choice questions- An assessment of the

assessment tool. International Journal of Health Sciences & Research. 2014;4(7):197-202.

13. Mahjabeen W, Alam S, Hussan U, Zafar T, Butt R, Konain S, Rizvi M. Difficulty index, discrimination index, and distractor efficiency in multiple choice questions. Annuals of Pakistan Institute of Medical Sciences. 2018;4: 310-315.
    Available:https://www.researchgate.net/publication/323705126
14. Mukherjee P, Lahiri SK. Analysis of multiple-choice questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal. Journal of Dental and Medical Sciences. 2015;14(12):47-52.
    Available:www.iosrjournals.org
15. Pathak CK, Patro KC, Pathak JA, Valenha CL. An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. Journal of Applied Psychology. 2013;67:826-834.
16. Musa A, Shaheen S, Elmardi A, Ahmed A. Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine, Khartoum University. Khartoum Medical Journal. 2018;11(02):1477-1486.
    Available:https://www.researchgate.net/publication/328583573
17. D'Sa JL, Visbal-Dionaldo ML. Analysis of multiple choice questions: Item difficulty, discrimination index and distractor efficiency. International Journal of Nursing Education. 2017;9(3):109-114.
    DOI: 10.5958/0974-9357.2017.00079.4.
18. Meyer JP. Applied Measurement with jMetrik. Published by Routledge 55 B/W Illustrations; 2014.
    ISBN: 9780415531979.
19. Houser J. Nursing research: Reading, using, and creating (4th ed.). Burlington, MA: Jones & Bartlett Learning; 2018.
20. Polit DF, Yang FM. Measurement and the measurement of change. Philadelphia: Wolters Kluwer. Obon & Rey, Analysis of Multiple-Choice; 2015.
21. Obinne ADE. Using IRT in Determining Test Item Prone to Guessing. World Journal of Education. 2012;2(1).
    Retrieved 22/12/2021
    Available:https://files.eric.ed.gov/fulltext/EJ1158955.pdf,
    http://dx.doi.org/10.5430/wje.v2n1p91
22. Tjabolo SA, Otaya LG. Quality of school exam tests based on item response theory. Universal Journal of Educational Research. 2019;7(10): 2156-2164.
    DOI: 10.13189/ujer.2019.071013
23. Setiawati FA, Setiawati R, EkaIzzaty VH. Items parameters of the space-relati subtest using item response theory. Data in Brief. 2018;19:1785-1793.
    Available:https://doi.org/10.1016/j.dib.2018.06.061i
24. Adedoyin OO, Mokobi T. Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. International Journal of Asian Social Science. 2013;3(4): 992-1011.