International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.5, Sep-Oct 2026, page no. 410-427

To cite this article: Kamoru O. Usman and John J. Agah (2026). EFFICACY OF ARTIFICIAL INTELLIGENCE-BASED ITEM GENERATION AND TEACHERS' DEVELOPED MATHEMATICS ITEMS FOR ASSESSMENT OF STUDENTS' LEARNING OUTCOMES IN ENUGU STATE, NIGERIA, International Journal of Education and Social Science Research (IJESSR) 9 (5): 410-427 Article No. 1304, Sub Id 1995

EFFICACY OF ARTIFICIAL INTELLIGENCE-BASED ITEM GENERATION AND TEACHERS' DEVELOPED MATHEMATICS ITEMS FOR ASSESSMENT OF STUDENTS' LEARNING OUTCOMES IN ENUGU STATE, NIGERIA

Kamoru O. Usman[1] and John J. Agah[2]

[1]Department of Science Education, University of Nigeria, Nsukka
kamoru.usman@unn.edu.ng

[2]Department of Science Education, University of Nigeria, Nsukka
john.agah@unn.edu.ng
Corresponding Author

DOI : https://doi.org/10.37500/IJESSR.2026.9523

ABSTRACT

The quality of items used for assessing students' achievement in Mathematics and other subjects could cause some inconsistencies and an unsatisfactory level of students' achievement in the subject. The researchers investigated the efficacy of Artificial Intelligence-based item generation and Teachers' developed Mathematics items for assessment of Students' learning outcomes (achievement) in Enugu State, Nigeria. Four specific purposes guided the study. The researchers adopted an instrumentation research design. The sample size for this study was 1056 senior secondary 2 (Grade 11) Students. A multi-stage sampling procedure was used for the study. Artificial Intelligence-Generated Test and Teacher-Developed Test (AIG-Test/TD-Test) were used for data collection. Jmetrik IRT software and Statistical Package for the Social Sciences (SPSS) were used to analyze the data collected. The findings revealed that most items were hard, providing enough challenges, and the discrimination indices were very high, showing proper differentiation between high and low-ability students in Mathematics. It was also found that a higher number of items displayed low guessing indices, reflecting quality distractors and low carelessness indices, suggesting that students responded attentively. Sixty items met the four-parameter criteria and were used to measure students' achievement in Mathematics. There is a significant impact of Artificial Intelligence-generated and Teachers' developed Mathematics test items on students' achievement in Mathematics in favour of Teachers' developed Mathematics tests. Based on the findings, it was recommended that IRT frameworks should be used in refining assessment instruments. Students' achievement should be determined with standardized tests.

KEYWORDS: Artificial Intelligence, Item-generation, Psychometric-indices, Mathematics

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.5, Sep-Oct 2026, p no. 410-427

INTRODUCTION

Teachers' assessment of students' understanding of content and skills is among many of his/her responsibilities, and it is crucial. Assessment of students' achievement is paramount and central at all levels of education. Assessment connotes all tasks teachers and students must accomplish to evaluate performance before, during, and after instruction. According to Delgado and Mark (2021), assessment denotes the wide range of methods or tools that teachers employ to measure, document, and determine students' academic readiness, achievement, learning progress, development of skills, or educational needs. Assessment entails a variety of educational activities that take place during and after instruction to determine students' level of achievement of the intended learning outcomes. One of the fundamental tools for assessment is a test. A test is a structured situation comprising a set of questions that examinees are expected to respond to. Each item in a test has a preferred answer. Test items could be developed by humans or generated through Artificial Intelligence (AI) tools like ChatGPT, Deepseek, among others. Subject teachers and psychometric experts can develop test items, and these are termed human-developed items. Human-developed test items have been criticized for being onerous, discrete, uniform, antiquated, and lacking authenticity (Swiecki et al., 2022). This opens the option for the use of automated item generation with more flexible means like Artificial Intelligence (AI).

Artificial Intelligence (AI) has been adopted in almost all areas of human enterprise due to scientific and technological development, which presents an exciting way of doing things with ease. According to Dohn, Kafai, Mørch, and Ragni (2022), AI is defined as the capacity of a computer system or program to reason and gain knowledge through experience. Chen, Xiang, and Ren (2021) stated that AI is a combination of Machine Learning and Deep Learning with enormous potential in assessment because it is machine-displayed intelligence that mimics human behaviour or thinking and can be trained to solve particular problems. The application of AI in this context offers the chance to improve the assessment process's effectiveness, accuracy, and fairness. AI has the capacity to generate high-quality test items for assessments that measure different cognitive levels in line with Bloom's taxonomy in various subjects when properly instructed (Zhang et al., 2021; Sottilare et al., 2018). However, the negative effects of AI in test item development have been identified to include algorithmic bias, data privacy, ethical issues, and fairness in AI-generated instruments, among others (Ughwubetine & Irighweferhe, 2025). Furthermore, AI technologies like ChatGPT remain underutilized in education due to users' lack of knowledge (González-Calatayud et al., 2021). The integration of AI into assessment is a gradual process, and as such, teacher-developed tests are still essential.

Teacher-developed tests are considered to be test questions that are manually constructed by teachers, subject experts, or examination bodies. They are traditionally developed items and do not use any automation tools. Teaching experience, curriculum, table of specification, teachers' judgment, among others, significantly guide the development of such tests. Many phases are involved in manual test development, and they include: item writing, item review, pilot testing, item analysis, and calibration (Irwing & Hughes, 2018). It is a lengthy process that requires human competence at every turn, and

it can be labour-intensive as well as time-consuming. Its quality relies heavily on the skills, strength, and energy of the item writers.

Many teachers were not trained to utilise AI tools like ChatGPT in developing test items for various subjects. Also, a few teachers who may be aware of AI tools like ChatGPT in developing test items do not have the skills required for assessment. In Nigeria, teachers are only exposed to the traditional mode of test item development, and some of the teachers are weak in developing test items. This could, in turn, hamper the accurate assessment of students' achievement in Nigeria. Exposing teachers to AI tools like ChatGPT and traditional modes of test item development to adequately prepare them for sound academic assessment is still a necessity. Hence, determining the efficiency of Artificial Intelligence-based item generation and Teachers' developed Mathematics items for assessment of Students' learning outcomes in Enugu State, Nigeria, is pertinent.

The introduction of digital technologies in the field of education has brought about changes in the way lessons are planned, delivered, and learning outcomes are assessed. The proliferation and advancement of digital technologies have transformed the roles of the teacher and the entire education system. One of the advanced digital technologies that is paving the way in the field of education is the use of AI like ChatGPT in assessment. ChatGPT, for instance, has great capabilities that can contribute to teaching and assessment in a variety of ways. Lo (2023) reported that ChatGPT can be used for teaching and assessment by teachers, and students can also use it in preparing writing assignments for assessment. Furthermore, GPT-3 can be used to evaluate the quality of the student-generated short-answer questions, while GPT-4, along with human and automated rule-based methods, can be utilized in evaluating the quality of items by identifying item-writing flaws in multiple-choice items (Moore et al. 2022; Moore et al. 2023). Some of the burdens encountered by teachers could be eliminated since AI-based tools can automate traditional assessments by automatically creating tests (Swiecki et al., 2022). Traditional assessment, on the other hand, is fraught with teachers' poorly constructed test items that may not provide precise information about student learning outcomes (Ede et al., 2021). It is based on this backdrop that the research sought to ascertain the efficiency of Artificial Intelligence-based item generation and Teachers' developed Mathematics items for assessment of students' learning outcomes in Enugu State, Nigeria.

The theoretical foundation for this study was the Item Response Theory (IRT). Item Response Theory (IRT) is a statistical model used in educational assessment for scoring, analyzing, and building tests. It is a set of probability functions that connect a person's unobserved ability (θ) to their observed test responses. Thus, IRT assumes that the probability of a correct response is a function of person ability θ and item parameters, and offers a clearer understanding of the quality of assessment instruments. IRT was adopted in this study to evaluate test item quality and provide more robust, interpretable metrics for individual item estimation in Mathematics multiple-choice tests. The selected metrics based on IRT are: difficulty index (b), discrimination index (*a*), guessing index (c) and carelessness (u). When tests are developed with these metrics well determined, they tend to adequately measure the aggregate performance of individual students with a high level of precision.

Educational practices, especially teaching and assessment, now rely heavily on AI tools. It is opined that AI-generated pedagogical tools, particularly multiple-choice questions (MCQs), are gaining ground among numerous options because they make evaluation efficient and effective (Nasution, 2023). The effectiveness of MCQs depends on basic psychometric qualities. The qualities are metrics intended for sound assessment. These include validity, reliability, item indices, among others. The item qualities are the difficulty index (b), discrimination index (*a*), guessing index (c), and carelessness (u) based on item response theory. These qualities ensure that assessments are educationally suitable and do not intentionally disadvantage students (Lu et al., 2024). The validity of test scores from MCQs could be in doubt when the test is poorly constructed. This is because it is susceptible to guessing responses, weak discrimination between more and less able ability students, among other variations (Feldt, 1997). To overcome these limitations, Hussein and Gasmalla (2023) advocated for the psychometric analysis of every item of MCQs.

Psychometric analysis is the statistical method employed to assess and improve the quality of test items (Hayashi & Eguchi, 2024). The item difficulty index (*b*) measures the proportion of learners who provided correct responses to an item (Sharma, 2021). The discrimination index (a) measures how well an item differentiates between high- and low-scoring examinees (Zakareya et al., 2024). Sometimes, examinees earn more marks without being sure of the right answer to an item or set of items in a test, and this strategy is guessing (Royal & Stockdale, 2017). Furthermore, carelessness, or the upper asymptote parameter, accounts for high-ability examinees' responses that don't reflect their true preferences for the items of a test (Jasper, 2024). It is a situation where examinees respond to an item incorrectly when they have the ability to correctly respond to it. Evaluating MCQs using these matrices provides a sound comparative quality concerning human-developed tests and ChatGPT-generated tests (Wati et al., 2024). Thus, the main goal of test development is to construct measurement tools like Mathematics tests that validly and reliably assess students' achievement in the subject.

The use of mathematics tests whose psychometric properties are properly ascertained will accurately measure students' achievement in the subject at the secondary school level. Mathematics achievement is an estimation of students' demonstration of the understanding and skills of what he/she has learnt. It is the ability to apply mathematical knowledge to real-life situations (Kuhfeld et al., 2025), mathematics-related performance outcomes (Breit et al., 2024), and the competency shown by the student in mathematics (Kosoko-Oyedeko et al., 2025). The scores of students in mathematics are most often considered as evidence of their achievement in the subject, and this is in turn taken to represent a person's accomplishment in specific goals, particularly in schools.

The academic achievement of students is often influenced by socio-demographic variables, like gender in most countries and in Nigeria in particular. Some studies suggest that male students frequently outperform their female counterparts in Mathematics. For instance, the findings of Hamza, Saidu and Rabiu (2025) indicated that male students achieve more than female students. It was also found that there were statistically significant differences in the achievement of Mathematics between male and

female distance learners in favour of males (Armah, Akayuure & Armah, 2021). Furthermore, male students excel in Mathematics achievement more than their female counterparts (Allahnana, Akande, Vintseh, Usman & Monica, 2018). However, Oliweh and Oyem (2021) found that there was no significant difference in the performance of boys and girls students in Mathematics. Whereas, it was revealed that girls score better in Mathematics than boys (Muhawenimana & Mutarutinya, 2023). Given the disparity in research outcomes on the achievement of male and female students in Mathematics, it became pertinent to determine the influence of gender on AI-generated test items and teacher-developed items. It is also necessary to determine the effectiveness of Artificial Intelligence-generated and teacher-developed Mathematics items in the assessment of students' learning outcomes (achievement) in Enugu State, Nigeria.

RESEARCH OBJECTIVES:

1. Determine the psychometric indices of Artificial Intelligence-generated and Teachers' developed Mathematics test items
2. To determine the impact of Artificial Intelligence-generated and Teachers' developed Mathematics test items on students' achievement in Mathematics.
3. To determine the influence of gender on students' achievement in Artificial Intelligence-based Mathematics test items.
4. To ascertain the influence of gender on students' achievement in Teachers' developed Mathematics test items.

RESEARCH QUESTIONS:

1. What are the psychometric indices of Artificial Intelligence-generated and Teachers' developed Mathematics test items?
2. What is the impact of Artificial Intelligence-generated and Teachers' developed Mathematics test items on students' achievement in Mathematics?
3. What is the influence of gender on students' achievement in Artificial Intelligence-based Mathematics test items?
4. What is the influence of gender on students' achievement in Teachers' developed Mathematics test items?

HYPOTHESES

H_{01}: There is no significant impact of Artificial Intelligence-generated and Teachers' developed Mathematics test items on students' achievement in Mathematics.

H_{02}: There is no significant influence of gender on students' achievement in Artificial Intelligence-based Mathematics test items.

H_{03}: There is no significant influence of gender on students' achievement in Teachers' developed Mathematics test items.

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.5, Sep-Oct 2026, p no. 410-427

METHODS:

This study adopted an instrumentation research design, which is appropriate for the development of tools, improved lesson plans, or instructional strategies (Emaikwu, 2013). The study was conducted in Enugu State of Nigeria. The sample size for this study was 1056 senior secondary 2 (SS 2 or Grade 11) students. A multi-stage sampling procedure was used for the study. Ten teachers were trained in the use of digital tools like ChatGPT to generate test items and teacher-developed items.

The AI Prompt for generating the items:

Use the following information provided in the test blueprint below and the curriculum content to generate 100 multiple-choice test items for general mathematics.

TABLE OF SPECIFICATIONS FOR GRADE 11 STUDENTS (SS 2)

Content (%)	Cognitive level			
	Knowing (35%)	Apply (40%)	Reasoning (25%)	Total
Theme 1: Number and Numeration (20%)	7	8	5	20
Theme 2: Algebraic Expression (30%)	11	12	7	30
Theme 3: Geometry (20%)	7	8	5	20
Theme 4: Statistics (30%)	11	12	7	30
Total	36	40	24	100

SS 2 General Mathematics:- Theme 1 [Number and Numeration, **and Topics (i)** Logarithm, (ii) Number Approximation, (iii) Error Estimation, (iv) Sequence and Series]; **Theme 2** [Algebraic Expression, and **Topics** (i) Quadratic equation, (ii) Simultaneous linear and Quadratic equation, (iii) Gradient of a curve, (iv) Logical Reasoning, (v) Linear Inequalities, (vi) Algebraic Fraction]; **Theme 3** [Geometry, and **Topics** (i) Chord property, (ii) Circle Theorems, (iii) Trigonometry, (iv) Bearing]; **Theme 4** [Statistics, and **Topics** (i) Measures of Central Tendency for ungrouped data, (ii) Measures of dispersion, (iii) Histogram of group data, (iv) Cumulative frequency graph (Ogive), (v) Measures of Central tendency for grouped data and (vi) Probability]. Ensure that the keys and distractors are evenly distributed across the options for the 100 items; also indicate in a table the keys in one column, the cognitive level the item is measuring, and the topic; suggest possible reasons for each key and distractor.

The teachers were first exposed to basic concepts such as test (Essay and multiple-choice test), stem, options, key, among others. The steps involved in constructing the test blueprint were also thoroughly explained to the teachers, and some of the steps are:

- ❑ Place the objectives along the horizontal part of the table.

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.5, Sep-Oct 2026, p no. 410-427

- Place the content or topics along the vertical part of the table.
- List the objectives guided by modified Bloom's taxonomy (Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating) across the top of the table.
- List the subject matter themes down the table.
- Assign weights (percentage) to each cell in accordance with the relative importance attached to each objective and topic.
- Weights (percentage) should reflect the emphasis placed on each objective and topic
- Determine how many questions will comprise the test/ exam (100).
- Determine the questions per cell (point of intersection between objective and content dimensions). Weight of topic * weight of objective * Number of items (or questions)
- Items were written, reviewed, and validated to ensure that they align with the mathematics curriculum, specific topics, and cognitive levels (modified Bloom's taxonomy) and measure what they ought to measure. Mathematics experts independently reviewed the AI-generated items and teacher-developed items for conceptual accuracy, clarity and found the items appropriate before they were used.

The instruments for data collection were the Artificial Intelligence Generated Test and Teacher-Developed Test (AIG-Test and TD-Test) in Mathematics. Both the AIG-Test and TD-Test were validated by three experts (one in Computer and Robotic Education, and two experts in Mathematics). The reliability of the AIG-Test and TD-Test was determined using Kuder-Richardson 20. Reliability estimates of 0.89 and 0.87 were obtained for AIG-Test and TD-Test, respectively, indicating that the instruments were highly appropriate for examining SS II students. Data collection was done through on-the-spot administration. Item indices for the AIG-Test and TD-Test, as well as students' achievement scores, were determined. Independent sample t-test was also used for data analysis.

Two phases of data analysis were done using Jmetrik IRT software and Statistical Package for the Social Sciences (SPSS). In the first phase, Jmetrik was used to calibrate the parameters of the Mathematics Achievement Test (MAT). Georgiev's (2008) criteria were used to categorize the item difficulty as: very easy ($b < -1.00$); easy ($-1 \leq b < 0.00$), hard ($0.00 \leq b < 1$); and very hard ($b > 1$). Baker and Kim (2017) categorise the item discrimination parameter as: very low ($0 \leq a < 0.34$), low ($0.35 \leq a \leq 0.64$), moderate ($0.65 \leq a \leq 1.34$), high ($1.35 \leq a \leq 1.69$), and very high ($a \geq 1.70$). The criteria for categorizing the guessing parameter were: low ($0 \leq c \leq 0.25$) and high ($c > 0.25$). The carelessness parameter criteria were $0 \leq u < 1$; low carelessness is considered as $u < 0.9$ and high carelessness ≥ 0.90, (Guyer and Thompson, cited in Pardede et al. 2023). In the second phase, the Statistical Package for Social Sciences (SPSS) was used to analyze the means, standard deviation, and independent-samples t-tests on students' achievement in the mathematics multiple-choice test. The research questions were answered using mean and standard deviation, whereas, independent sample t-test was used to test the null hypotheses at 0.05 level of significance.

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.5, Sep-Oct 2026, p no. 410-427

RESULTS

Research Question One: What are the psychometric item qualities of AI-generated and Teachers' developed Mathematics test?

Table 1: Item parameters of AI-generated test items and Teachers' Developed items in Mathematics

Artificial Intelligence (ChatGPT)					Teachers developed			
Items	**b_Para**	**a_para**	**c_para**	**u_para**	**b_para**	**a_para**	**c_para**	**u_para**
1	0.92	2.01	0.11	0.02	0.95	2.20	0.13	0.01
2	0.84	2.45	0.24	0.02	0.95	2.24	0.23	0.02
3	1.09	2.40	0.16	0.02	0.89	2.34	0.14	0.02
4	0.94	2.36	0.16	0.04	0.91	1.82	0.22	0.02
5	0.87	1.70	0.25	0.02	0.86	2.75	0.23	0.02
6	0.87	2.28	0.14	0.02	0.83	1.24	0.13	0.03
7	0.97	1.26	0.13	0.01	0.92	1.38	0.14	0.02
8	0.82	2.18	0.22	0.02	0.88	1.82	0.23	0.02
9	1.96	1.64	0.12	0.03	0.91	2.02	0.11	0.01
10	0.97	2.27	0.22	0.02	0.95	2.10	0.15	0.03
11	0.96	1.73	0.17	0.02	0.97	2.07	0.15	0.02
12	0.88	1.38	0.12	0.06	0.94	1.28	0.13	0.04
13	0.88	1.17	0.24	0.06	0.96	1.35	0.18	0.02
14	0.90	1.67	0.14	0.05	0.93	1.82	0.15	0.03
15	0.87	2.44	0.15	0.04	0.91	2.13	0.13	0.03
16	0.91	2.12	0.08	0.03	0.94	2.18	0.06	0.04
17	0.68	2.23	0.15	0.02	0.71	1.26	0.14	0.03
18	1.75	0.87	0.15	0.05	0.60	1.42	0.13	0.03
19	0.81	1.74	0.18	0.08	0.83	2.07	0.14	0.04
20	0.90	2.13	0.12	0.02	0.94	1.67	0.13	0.03
21	0.82	1.81	0.13	0.03	0.84	1.86	0.13	0.03
22	0.65	1.62	0.17	0.07	0.76	1.73	0.13	0.03
23	1.70	2.10	0.09	0.03	0.70	2.20	0.08	0.02
24	0.77	1.65	0.12	0.03	0.76	1.85	0.10	0.03
25	0.94	2.14	0.24	0.04	0.96	1.47	0.08	0.01
26	0.70	1.33	0.14	0.02	0.89	0.91	0.11	0.02
27	1.80	1.53	0.19	0.04	0.82	1.31	0.11	0.02
28	0.83	1.03	0.13	0.02	0.85	1.38	0.23	0.02
29	0.92	1.08	0.18	0.02	0.94	1.19	0.11	0.03

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.5, Sep-Oct 2026, p no. 410-427

30	0.92	1.82	0.19	0.03	0.98	1.55	0.12	0.02
31	0.91	1.25	0.17	0.04	0.70	1.65	0.14	0.03
32	0.94	2.10	0.17	0.04	0.92	2.33	0.13	0.03
33	0.81	1.56	0.22	0.02	0.91	1.42	0.13	0.03
34	0.87	2.42	0.16	0.06	0.70	2.13	0.17	0.07
35	0.96	1.82	0.23	0.03	0.94	2.19	0.14	0.04
36	1.83	1.29	0.16	0.03	1.79	1.11	0.13	0.02
37	0.89	2.13	0.13	0.03	0.90	2.45	0.15	0.01
38	0.87	1.92	0.12	0.02	0.89	1.71	0.13	0.03

b= difficulty index, a = discrimination index, c = guessing index, and u= carelessness index

Table 1: Item Parameters of Mathematics Achievement Test (continue)

Artificial Intelligence (ChatGPT)					Teachers developed			
Items	**b_Para**	**a_para**	**c_para**	**u_para**	**b_para**	**a_para**	**c_para**	**u_para**
39	-1.81	0.29	0.23	0.07	0.01	1.70	0.03	0.05
40	-1.49	2.10	0.11	0.03	-1.11	2.37	0.11	0.01
41	-0.44	2.85	0.24	0.02	-0.56	2.40	0.17	0.02
42	0.41	2.60	0.19	0.02	1.49	0.82	0.11	0.01
43	0.93	2.50	0.21	0.02	0.65	1.32	0.20	0.03
44	-0.95	2.67	0.19	0.02	-0.93	1.47	0.18	0.01
45	0.54	2.41	0.22	0.03	0.41	1.49	0.21	0.02
46	0.93	2.12	0.02	0.01	0.94	0.91	0.11	0.02
47	1.84	1.74	0.11	0.02	0.87	1.64	0.16	0.01
48	0.97	2.24	0.08	0.03	0.97	2.59	0.07	0.02
49	0.93	1.76	0.12	0.02	0.95	1.55	0.09	0.01
50	-0.44	2.65	0.15	0.02	-0.11	2.24	0.13	0.02
51	0.41	2.60	0.14	0.02	-0.27	2.13	0.12	0.02
52	0.94	2.60	0.17	0.02	0.95	2.35	0.13	0.02
53	0.95	2.47	0.15	0.02	0.95	1.27	0.16	0.03
54	-0.43	2.60	0.19	0.02	-0.25	2.35	0.16	0.02
55	-0.15	2.47	0.17	0.02	0.24	1.27	0.13	0.03
56	0.54	2.21	0.19	0.03	0.43	1.36	0.19	0.04
57	-0.44	2.17	0.12	0.02	-0.24	1.64	0.14	0.03
58	0.98	1.74	0.12	0.02	-0.11	2.17	0.11	0.01
59	-0.84	2.79	0.17	0.02	-0.67	2.20	0.18	0.05
60	-1.98	2.10	0.22	0.03	-0.56	2.20	0.21	0.02

b= difficulty index, a = discrimination index, c = guessing index, and u= carelessness index

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.5, Sep-Oct 2026, p no. 410-427

Table 2: Mean item of item difficulty, discrimination, guessing and carelessness

Parameter	Test	Mean	SEM	Remarks
Difficulty	AI	0.6553	0.1043	Hard
	TD	0.6347	0.0736	Hard
Discrimination	AI	1.9718	0.0681	Very High
	TD	1.7840	0.0601	Very High
Guessing	AI	0.1617	0.0062	Low guessing
	TD	0.1412	0.0055	Low guessing
Carelessness	AI	0.0297	0.0019	Low carelessness
	TD	0.0250	0.0015	Low carelessness

AI = Artificial intelligence (ChatGTP), TD = Teacher-developed

The result in Table 1 shows the difficulty indices (b) of AI-generated test items and teachers developed (TD) test items in Mathematics. The item difficulty values for AIG test ranged from $-1.98 \leq b \leq 1.96$, with item 60 being the easiest and item 9 being the hardest, whereas the item difficulty values for the TD test ranged from $-1.11 \leq b \leq 1.79$, with item 40 being the easiest and item 36 being the hardest. Table 2 indicates that the mean item difficulty index is 0.66 with an SEM of 0.10 for the AI-generated test and 0.64 with an SEM of 0.07 for TD Mathematics test. Table 1 also shows that the discrimination index (a) of AIG test items ranged from $0.29 \leq a \leq 2.85$, with item 39 being the least differentiating and item 41 being the most differentiating, whereas the item discrimination values for TD test ranged from $0.82 \leq a \leq 2.75$, with item 42 being the least differentiating and item 5 being the most differentiating. Table 2 also shows that the mean item discrimination index is 1.97 with an SEM of 0.07 for AIG and 1.78 with SEM of 0.06 for TD Mathematics test.

Furthermore, the results in Table 1 indicate that the guessing index (c) of AIG and TD test items ranged from $0.02 \leq c \leq 0.25$ and $0.03 \leq c \leq 0.23$, respectively. Table 2 also shows that the mean guessing index is 0.16 with SEM of 0.01 for AI-generated and 0.14 with SEM of 0.01 for TD Mathematics test. The results in Table 1 indicate that the carelessness index (u) of AIG and TD test items ranged from $0.01 \leq u \leq 0.08$ and $0.01 \leq u \leq 0.07$, respectively. Table 2 also shows that the mean carelessness index is 0.03 with SEM of 0.002 for AIG and 0.03 with SEM of 0.002 for TD Mathematics test. The mean carelessness indices and standard error of measurement were the same for AI-generated and TD tests.

Research Question two: What is the impact of Artificial Intelligence-generated and Teachers' developed Mathematics test items on students' achievement in Mathematics?

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.5, Sep-Oct 2026, p no. 410-427

Table 3: Independent sample t-test analysis of the mean achievement scores of students in AIGTest and TDTest Mathematics test

variables	N	Mean	SD	t-val	df	Sig	Dec
TDTest	518	25.71	8.15	4.28	1054	0.000	S
AITest	518	23.70	7.11				

Key: t-val = t-test value, df = degree of freedom, Sig. = Exact probability value, Dec. = Decision, S = significant, α = 0.05

Results in Table 3 show the mean responses and standard deviation scores of students in the TDTest and AIGTest Mathematics test**.** The results show that the students in TDTest had a mean of ($\bar{x}$ = 25.271, SD = 8.15) and their counterparts in AIGTest had a mean of ($\bar{x}$ = 23.70, SD = 7.11). The mean score in the AIGTest and TDTest Mathematics test is 2.01. This somewhat implies that students achieved better in TDTest than in AIGTest. Furthermore, the result in Table 3 shows that the t-value of 4.28 with a degree of freedom of 1054 and an exact probability value of 0.000 was obtained. The exact probability value of 0.000 was less than 0.05, and the conclusion drawn was that there is a significant impact of Artificial intelligence-generated and Teachers' developed Mathematics test items on students' achievement in Mathematics in favour of Teachers' developed Mathematics test.

Research Question three: What is the influence of gender on students' achievement in Artificial intelligence-based Mathematics test items?

Table 4: Independent sample t-test analysis of the mean achievement scores of male and female students in Artificial intelligence-generated Mathematics test

variables	N	Mean	SD	t-val	df	Sig	Dec
Male	253	24.58	6.92	2.73	526	0.006	S
Female	275	22.89	7.19				

Key: t-val = t-test value, df = degree of freedom, Sig. = Exact probability value, Dec. = Decision, S = significant, α = 0.05

Results in Table 4 show the mean responses and standard deviation scores of male and female students in the Artificial Intelligence Generated Mathematics test**.** The result shows that the male students in AIGTest had a mean of ($\bar{x}$ = 24.58, SD = 6.91) and their female counterparts in AIGTest had a mean of ($\bar{x}$ = 22.89, SD = 7.19). The difference in the mean score of male and female students in AIGTest

Mathematics test is 1.69. This somewhat implies that male students achieved better results in AIGTest than female. Furthermore, the result in Table 4 shows that a t-value of 2.73 with a degree of freedom of 526 and an exact probability value of 0.006 was obtained. The exact probability value of 0.006 was less than 0.05, and the conclusion drawn was that there is a significant influence of gender on students' achievement in the Artificial Intelligence-generated Mathematics test in favour of male students.

Research Question four: What is the influence of gender on mean achievement scores of students in the teachers' developed Mathematics test?

Table 5: Independent sample t-test analysis of the mean achievement scores of male and female students in the teachers' developed Mathematics test

variables	N	Mean	SD	t-val	df	Sig	Dec
Male	253	26.33	7.95	1.68	526	0.093	NS
Female	275	25.13	8.31				

Key: t-val = t-test value, df = degree of freedom, Sig. = Exact probability value, Dec. = Decision, S = significant, α = 0.05

Results in Table 5 show the mean responses and standard deviation scores of male and female students in the teachers' developed Mathematics test**.** The results show that the male students in TDTest had ($\bar{x}$ = 24.58, SD = 6.91) and their female counterparts in TDTest had ($\bar{x}$ = 22.89, SD = 7.19). The difference in the mean score of male and female students in the TDTest Mathematics test is 1.20. This somewhat implies that male students achieved slightly better in TDTest than female. Furthermore, the result in Table 5 shows that the t-value of 1.68 with a degree of freedom of 526 and an exact probability value of 0.093 was obtained. The exact probability value of 0.093 was greater than 0.05, and the conclusion drawn was that there is a significant influence of gender on students' achievement in the teacher-developed Mathematics test.

DISCUSSION OF THE FINDINGS

Sixty items of the mathematics test for AI-generated and teacher-developed tests, respectively, had difficulty values that fell within the acceptable range as shown in Table 1. The 60 test items for both AI-generated and teacher-developed tests were spread across easy, moderate, and difficult levels. The mean item difficulty index of 0.66 with SEM of 0.10 for AI-generated and 0.64 with SEM of 0.07 for TD indicates that the mathematics test items were somewhat difficult, tending slightly toward the hard end of the spectrum. The variation in difficulty between the AIG test and TD test suggests that the AIG test was slightly harder than the TD test. The findings indicate that most of the items of both tests provided the students with challenging experiences and were appropriately spread across the ability continuum. The finding aligns with the assertion of some authors who posit that item difficulty

enhances the high quality of test items when they fall within the acceptable range, and is a crucial parameter in various models used to estimate examinees' ability and item characteristics (Rudolph et al., 2019; Okoye & Ndubuzor, 2016). Although the difficulty values of the AIG test and TD test suggest that the items were somewhat hard, most of them may be slightly below or above the average ability estimate of the SS2 students who constituted the population.

The findings from this study revealed that most items (over 60) had very high discrimination values. The mean item discrimination index of 1.97 with SEM of 0.07 for AI-generated and 1.78 with SEM of 0.06 for TD Mathematics tests indicates that both tests differentiated maximally between low and high-ability groups. The low standard error of measurement (SEM) reinforces the test's strength in measuring students' true ability levels in a consistent pattern across AIG and TD tests. This result implies that the items of both tests were outstandingly effective at differentiating between high- and low-ability students in the AIG test and the TD test. This result is in agreement with other authors (Baker & Kim, 2017; DeMars, 2021; Adodo & Ojerinde, 2022) who assert that high discrimination parameters help to detect or distinguish variation among learners across a continuum of behaviour measured. This is a desirable result because it ensures that item responses accurately reflect the underlying latent trait of students' achievement in mathematics. AIG and TD tests were very sensitive in differentiating varying levels of students' ability.

It was further found that most items (over 60) had very good guessing parameters in all the items in Mathematics for the AI and TD. The mean guessing index of 0.16 with SEM of 0.01 for AI-generated and 0.14 with SEM of 0.01 for TD Mathematics test implies that students cannot easily use guessing to provide correct answers to the items. In both AI and TD, examinees cannot randomly provide responses that do not reflect their true abilities. With AI and TD guessing indices being less than 0.17, which is quite far from the criteria of low guessing, the worries of authors (Soland, Jensen, Keys, Bi & Wolk, 2019; Royal & Stockdale, 2017) that guessing raises measurement error by increasing the variance error of test scores and reduces the reliability of the test, as well as makes examinees earn more marks than their true abilities could produce, are solved. Since all the items of AI and TD tests have guessing values less than the ideal threshold of 0.25, it implies distractors were of high quality. This, in turn, means that all the incorrect response options (distractors) may have been sufficiently plausible, not allowing students to guess the answer correctly.

The mean carelessness (or upper asymptote parameter) index for AI-generated and TD Mathematics tests was 0.03 (SEM = 0.002) and 0.03 (SEM = 0.002), respectively. The carelessness values of all the items of AI and TD were good and the same for both tests. These indicate that most students who had the requisite ability answered the items correctly. This low average and low standard error of measurement confirm that students were generally sure of their responses, and that errors were more likely due to an actual lack of knowledge rather than guessing. This implies students gave answers that reflected their true preferences and abilities. The study's findings indicate that AI and TD could be effectively deployed in the development of Mathematics achievement test items. All the psychometric

qualities of the AI-generated and TD test items were similar and did not show substantial variations in all the parameters determined. Both formats of test item development for Mathematics should be encouraged despite the possible lack of time spent on teacher-developed tests.

Other findings of the study show that students achieved better in TDTest than AITest, and there is a significant impact of Artificial Intelligence-generated and Teachers' developed Mathematics test items on students' achievement in Mathematics in favour of Teachers' developed Mathematics test. Furthermore, male students achieved better in AITest than their female counterparts, and the influence of gender on students' achievement in the Artificial Intelligence-generated Mathematics test was significant in favour of male students. This finding supported other researchers who revealed that there were statistically significant differences in the mathematics achievement of male and female learners in favour of males (Hamza, Saidu & Rabiu, 2025; Armah, Akayuure, & Armah, 2021). However, it deviates from that of Muhawenimana and Mutarutinya (2023), who revealed that girls score better in mathematics than boys.

Although the male students achieved better than their female counterparts in the TD test, the difference was significant. This result corroborated the work of Oliweh and Oyem (2021), who found that there was no significant difference in the performance of boys and girls students in Mathematics. These findings are highly informative. First, the students appear more familiar with teacher-generated items than AI-generated items; this may have impacted their achievement. The AI may have generated items that validators thought were appropriate for the class but are really not. Secondly, male students who participated in the studies may have a stronger affinity for AIG items or be more attuned to digital technology-based instruments than their female counterparts. Finally, the TD test item contained information that both male and female students are conversant with, and probably the reason why the results of both groups were not significant.

CONCLUSIONS

It was concluded that both artificial intelligence (automated) and Teachers developed (manual) ways are efficacious in generating Mathematics items for assessment of students' learning outcomes (achievement) in Enugu State, Nigeria. The item indices for both AI-generated and TD tests were suitable for two versions of the test. The AIG test and TD test contained 60 items each. The test items in both versions provided students with similar challenges and sufficiently distinguished between low- and high-ability students. The probability of guessing an item correctly in the AI and TD tests was negligible or very low. Students' carelessness was minimal, and they were generally sure of their responses. The students' answers reflected their true abilities, and any incorrect responses were likely due to a genuine lack of knowledge rather than guessing. The AI-generated (AIG) test items and teachers' developed (TD) test items have strong psychometric properties for assessing students' achievement in Mathematics.

Students' achievement in the teachers' developed Mathematics test was significantly higher than that of the Artificial Intelligence-generated test. Male students achieved significantly better than their female counterparts in the AIGTest it implies gender influences students' achievement in Artificial Intelligence-generated mathematics. Although the male students achieved significantly better than their female counterparts in the AIGTest, it implies gender influences students' achievement in Artificial Intelligence-generated mathematics.

LIMITATIONS

Some limitations of the study included: (i) the researchers did not examine whether items function differently across gender and other demographic groups. (ii) Student perceptions of AI-generated versus teacher-developed items were not determined. (iii) The time, effort, and resources required for AI-generated items were less than teacher-developed items.

RECOMMENDATIONS

1. Teachers should be trained on item development.
2. Assessment policy should emphasize the use of standardized tests for formative and summative assessment of learning outcomes in Mathematics.
3. Further research should be conducted on the use of AI-generated tests in Nigerian schools.

ACKNOWLEDGEMENTS: The researchers sincerely appreciate the financial support received from TETFUND to conduct and complete this work.

REFERENCES

Adodo, S. O., & Ojerinde, D. (2022). Students' engagement and academic achievement in mathematics: The mediating role of emotional and cognitive participation. *Journal of Educational Research and Development*, *13*(2), 84–98.

Allahnana, K. M., Akande, M. T., Vintseh, I. M., Usman, A. E. A., & Monica, E. A. (2018). Assessment of Gender and Interest in Mathematics Achievement in Keffi Local Government Area of Nasarawa State, Nigeria. *International Journal of* Operational *Research in Management, Social Sciences & Education, 4*(1), 127-140

Armah, S. E., Akayuure, P., & Armah, R. B. (2021). A Comparative Study of Male and Female Distance Learners' Mathematics Achievement. *Contemporary Mathematics and Science Education, 2*(1), ep21001. https://doi.org/10.30935/conmaths/9288

Baker, F. B., & Kim, S.-H. (2017). The basics of item response theory using R. Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8

Breit, M, Schneider, M. & Preckel, F. (2024). Mathematics achievement and learner characteristics: A systematic review of meta-analyses, *Learning and Individual Differences,* 118, www.elsevier.com/locate/lindif

Chen, S., Xiang, J., & Ren, Z. (2021). To Leverage the Innovation Capability by Co-creation of Human-Machine. *Forest Chemicals Review*, 345-360

Delgado, S. M. & Mark, S. (2021). Assessment and evaluation. *Introduction to Education*.

DeMars, C. E. (2021). *Item response theory*. Oxford University Press.

Dohn, N. B., Kafai, Y., Mørch, A. & Ragni, M. (2022) Survey: artificial intelligence, computational thinking and learning. *KI-KünstlicheIntelligenz*, *36*(1), 5-16.

Ede M. O., Agah J. J., Okeke C. I., Chuks Z. O., Oguguo B. C., Agu P. U., Ene C. U., Ekesionye N. E., Nji I. A., Eze C., Manafa I. F. (2021) Effect of cognitive behavioral active engagement training on test item construction skills among primary school teachers in Nigeria: implication for educational policy makers. *Medicine* (e26876).

Feldt, L. S. (1997). Can validity rise when reliability declines? *Applied Measurement Education*, *10*(4), 377-387. https://doi.org/10.1207/s15324818ame1004_5

Georgiev, N. (2008). Item analysis of C, D and E series from Raven's standard progressive matrices with item response theory two-parameter logistic model. *Europe's Journal of Psychology*, 4(3). https://doi.org/10.5964/ejop.v4i3.431

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, *11*(12), 5467.

Hayashi, K. & Eguchi, S. (2024). A new integrated discrimination improvement index via odds. *Stat. Pap.*, *65*, 4971-4990. https://doi.org/10.1007/s00362-024-01585-7

Hussein, A. & Gasmalla, H. E. E. (2023). Introduction to the psychometric analysis. In *Written Assessment in Medical Education* (pp. 111-135). Springer International Publishing. https://doi.org/10.1007/978-3-031-11752-7_9

Hamza, B. K., Saidu, S., & Rabiu, A. T. (2025). Effects of gender and school location on students' achievement in mathematics olympiad among junior secondary school students in Kaduna State. *FNAS Journal of Mathematics and Science Education, 7*(1), 9-14. https://doi.org/10.63561/fnas-jmse.v7i1.1057

Irwing, P. & Hughes, D. J. (2018) *Test development*. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development, 1-47

Jasper, V. (2024). *Choice model using a Four-Parameter Logistic Function*. Unpublished master's thesis, Erasmus University, Rotterdam, Netherlands.

Kosoko-Oyedeko, G. A., Odupe T. A. A., Ola-Michael O., & Salami, I. A. (2025). Effectiveness of AI-Designed Instructional Materials on Mathematics Achievement of Secondary School Students in Lagos State. *International Journal of Scientific Research in Social Sciences & Management Studies, 8(2), 227-239.* https://doi.org/10.48028/iiprds/ijsrssms.v8.i2.16

Kuhfeld, M., Soland, J., Lewis, K., & Morton, E. (2025*).* Academic achievement trends postpandemic: Evidence from large-scale assessments. *Educational Researcher, 54*(2), 123–135. https://doi.org/10.3102/0013189X24123456

Lu, W., Laffey, J., Sadler, T. D., Griffin, J., & Goggins, S. P. (2024). A scalable, flexible, and interpretable analytic pipeline for stealth assessment in complex Digital Game-Based Learning environments: Towards eneralizability. *J. Educ. Data Min.*, *16*(2), 214-303. https://doi.org/10.5281/zenodo.14503598

Lo, C.K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences. 13*(4), 410, 1-15. https://doi.org/10.3390/educsci13040410

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. In I. Hilliger, P. J. Munoz-Merino, T. D. Laet, A. Ortega-Arranz & T. Farrell (Eds.), *Educating for a new future: Making sense of technology-enhanced learning adoption* (pp. 243-257). Springer.

Moore, S., Nguyen, H.A., Chen, T., & Stamper, J. (2023). Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods. In O. Viberg, I. Jivet, P. K. Munoz-Merino, M. Perifanou & T. Papathoma (Eds.), *Responsive and Sustainable Educational Features* (pp. 229-245). Springer.

Muhawenimana, P. & Mutarutinya, V. (2023). Impact of gender differences on students' performance in mathematics within selected secondary schools of Gicumbi District, Rwanda. *Journal of Research Innovation and Implications in Education, 7*(1), 188 – 199.

Nasution, N. E. A. (2023). Using artificial intelligence to create biology multiple choice questions for higher education. *Agric. Environ. Educ.*, *2*(1), em002. https://doi.org/10.29333/agrenvedu/13071

Okoye, R. O., &Ndubuzor, J. N. (2016). Empirical Analysis of item difficulty and discrimination indices of national business and technical examination board (NABTEB) economics essay test items from 2013-2015.

Oliweh, I. S., & Oyem, I. M. (2021). Gender differential mathematics achievement of students in selected senior secondary schools in delta state, Nigeria. *Abacus (Mathematics Education Series) 46*(1), 58-64

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL I RT model and strategies for estimating it. *REID* (*Research and Evaluation in Education*), *9*(1), 86-117. https://doi.org/10.21831/reid.v9i1.63230

Royal, K. D. & Stockdale, M. R. (2017) The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. *Journal of Advances in Medical Education & Professionalism*, *5*(2), 84

Rudolph, J., Tan, S. & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6* (1)

Sharma, L. R. (2021). Analysis of difficulty index, discrimination index and distractor efficiency of multiple-choice questions of speech sounds of English. *Int. Res. J. MMC*, *2*(1), 15-28. https://doi.org/10.3126/irjmmc.v2i1.35126

Soland, J., Jensen, N., Keys, T. D., Bi, S. Z. & Wolk, E. (2019) Are test and academic disengagement related? Implications for measurement and practice. *Educational Assessment*, *24*(2), 119-134.

Sottilare, R. A., Brawner, K. W., & Goldberg, B. (2018). Adaptive intelligent tutoring systems: AI in education. *International Journal of Artificial Intelligence in Education, 28*(2), 123 145

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J.M., Milligan, S., ... & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075

Ughwubetine, S. & Irighweferhe, S. (2025). Exploring the Implication of Artificial Intelligence in Mathematics Test Items. *Journal of Computing, Science &Technology,* 3, 2025

Wati, E. R. K., Nengsih, Y. K., Handrianto, C., & Rahman, M. A. (2024). The quality of teacher-made summative tests for Islamic education subject teachers in Palembang, Indonesia. *Journal of Cakrawala Pendidik.*, *43*(1), 186–197. https://doi.org/10.21831/cp.v43i1.53558

Zakareya, S., Alsaleem, N., Alnaghmaish, A., Alnaim, N., & Alojail, F. (2024). Evaluating the discrimination index of AI-generated vs. human-generated multiple-choice questions: action research. *In 17th Annual International Conference of Education, Research and Innovation, Seville, Spain*, 221-226. https://doi.org/10.21125/iceri.2024.0137

Zhang, Y., Wang, X., & Liu, P. (2021). Automated item generation using deep learning techniques. *Educational Measurement: Issues and Practice, 40*(2), 45-58