Journal of Innovation in Educational Assessment
Vol. 8 No. 1, May (2026)

DOI: https://doi.org/10.66545/s7vd5d22

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

Received: 22nd January 2026
Accepted: 24th April 2026

journal.iaiiea.org

COMPARATIVE ANALYSIS OF PSYCHOMETRIC PROFILING OF AI-GENERATED AND TEACHER-CONSTRUCTED MULTIPLE-CHOICE QUESTION ITEMS

[1]Glory Udoh Evans*, [2]Udoh Felix Evans & [3]John Okon Esin

[1]Department of Educational Foundations Akwa Ibom State College of Education Afaha Nsit
[2]Department of Research and Strategic Development Maritime Academy of Nigeria, Oron
[3]Department of Nautical Science Maritime Academy of Nigeria, Oron

Abstract

The emergence of artificial intelligence (AI) in educational assessment has created new opportunities for automating test item development, yet questions remain regarding the psychometric soundness of AI-generated instruments compared to teacher-constructed tests. This study undertook a comparative analysis of the psychometric properties of multiple-choice questions (MCQs) generated by an AI language model and those designed by experienced teachers. Using a quasi-experimental design, two parallel test forms of 30 items each were administered to 300 senior secondary school students in Akwa Ibom State. Parameters analysed were difficulty index and discrimination index, while test reliability was examined using KR-20, and validity was through content alignment and correlated with external achievement scores. Results revealed that both AI-generated and teacher-made tests achieved acceptable reliability coefficients ($\alpha = 0.746$ and $\alpha = 0.949$, respectively). Teacher-made items demonstrated slightly superior discriminating indices (mean = 0.45) compared to AI-generated items (mean = 0.41), whereas AI-generated items exhibited a more balanced difficulty level, with 74% falling within the optimal difficulty range compared to 69% of teacher-made items. The findings indicate that AI-generated MCQs can produce psychometrically sound items comparable to teacher-made ones, though refinement is needed in discriminative power and distractor plausibility. This study concludes that AI holds promise as a supportive tool for large-scale item generation, but human expertise remains essential for ensuring validity and alignment with pedagogical intent.

Keywords: Multiple-choice questions, AI-generated items, teacher-made items, achievement test, psychometric profiling

To cite this article:

Evans, G., Evans, U., & Esin, J. O. (2026). Comparative Analysis of Psychometric Profiling of AI-Generated and Teacher Constructed Multiple-Choice Question Items. *Journal of Innovation in Educational Assessment, 8*(1), 52-70. https://doi.org/10.66545/s7vd5d22

* Corresponding author:

Akwa Ibom State College of Education Afaha Nsit. Email: evansgloryu@gmail.com

Introduction

Education has significantly been influenced by the digital world; therefore, the integration of artificial intelligence into education has become increasingly famous as technology advancements further reshape various sectors. Artificial Intelligence (AI) is changing every aspect of our lives, including health care, employment, entertainment, and education, sparking debates about personalised learning and the future of assessments (Farazouli et al., 2023). Specifically, integration of AI into education has potential in enhancing personalised learning, automating grading, saving time and increasing students' satisfaction (Dempere et al., 2023; Sallam et al., 2023). AI refers to the field of computer science that involves creating computer programs capable of imitating intelligent behaviour and ideally enhancing human-like abilities (Naqvi, 2020). Artificial intelligence is a technology that allows machines to learn, to reason and to act in a way that typically requires human intelligence. Artificial intelligence has currently become a vital part of the virtual world; it plays an important role in general education (Edtech, 2020).

AI-powered tools and applications are now being used in many industries, including education, to enhance the quality of services provided to students and teachers. To enhance the quality of services provided to students and teachers. AI tools such as Bing and ChatGPT have been referred to as objects individuals can think with, especially in the teaching-learning situation for learners to enhance their ability to think critically and reflectively, foster creativity, acquire problem-solving skills, and grasp concepts effectively (Vasconcelos et al., 2023).

The integration of AI into the education sector has become an inevitable trend for the future of education. The social changes brought about by this new technological revolution are continually remoulding the existing forms and contents of education. The fast-paced technology provides individuals in the area with training and learning with unlimited possibilities. Technology has advanced to the extent that most tasks are no longer performed manually; otherwise, such tasks take longer to accomplish. This explained the widespread use of artificial intelligence (AI) and technology, making it seemingly indispensable in the present-day educational system. They provide personalised learning, efficient evaluation, and creative tools that improve the overall efficacy of teaching and learning. The machine is designed in such a way to think and act like people, which made AI a mini human that acts in different fields. Jain and Jain (2019) say artificial intelligence (AI) is available in different parts of our lives, beginning from smart sensors to individual associates, and that artificial intelligence helps students and teachers to make their educational experience wonderful.

Assessment remains a central component of the educational process, serving as a mechanism for measuring learners' knowledge, diagnosing learning gaps, and guiding instructional improvement. In particular, multiple-choice questions (MCQ items) are

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

widely used in educational assessments because of their objectivity, reliability, and ease of administration and scoring. However, the validity and psychometric quality of MCQs largely depend on how well the items are constructed. Traditionally, teachers design test items based on their subject expertise and curriculum familiarity. Yet, numerous studies have revealed inconsistencies in the quality of teacher-constructed items, often arising from limited training in test construction principles, subjective bias, and time constraints (Kaplan & Haenlein, 2019; Zong & Krishnamachari, 2022; LaFlair and Runge, 2023; Lazarova & Pavlovic, 2024; Adesina, 2024).

The emergence of high-profile artificial intelligence (AI) tools such as advance web search engine like Google, Waymo, generative and creative tools like Yenni, Paperpal, Yomu, ChatGPT, ChatBot, GPT-based test generators, and other natural language processing systems have introduced new possibilities for automating the development of assessment materials. These AI systems can generate large pools of MCQs in seconds, guided by curriculum standards or instructional objectives. Despite this efficiency, there are concerns regarding the psychometric soundness, such as difficulty index, discrimination power, and distractor efficiency of AI-generated items compared to human-constructed ones. Given that psychometric properties determine the reliability and validity of assessments, it is imperative to examine whether AI-generated MCQs meet or exceed the standards of those developed by experienced educators.

Furthermore, the integration of AI into educational assessment raises pedagogical and ethical questions concerning authenticity, contextual relevance, and cognitive alignment. If AI can consistently produce psychometrically robust items, it could support teachers by reducing workload and enhancing test item quality. Conversely, if significant disparities exist, it underscores the irreplaceable value of teacher expertise and contextual judgement in test design.

This study, therefore, seeks to conduct a comparative analysis of the psychometric profiling of AI-generated and teacher-constructed multiple-choice question items. By examining key indices such as item difficulty and discrimination, the research aims to provide empirical evidence on the quality and reliability of AI-generated assessments. The findings will inform educators, curriculum developers, and policymakers about the practical implications of integrating AI tools in education, ensuring that technology complements rather than compromises assessment quality.

A test, whether formal or informal, is an assessment process intending to measure a student's knowledge, skill, and aptitude as well as physical fitness. A test may be administered verbally, on paper, on computer or in a predetermined area that require a test taker to demonstrate or perform a set of skills. Opara (2021) defines a test as an instrument or procedure designed to measure the knowledge, intelligence, ability, traits, skills, aptitude,

interest, and attitude of an individual or group. A test may be developed and administered by an instructor, a clinician, a government body, or test provider. In some instances, the developer of the test may not be directly responsible for its administration. To measure the learning and teaching of any school subject, the measurement instrument must be planned, evaluated and tested to ensure it meets reliability, validity and usability (Ubi & Udemba, 2021). Tests can be standardised, diagnostic, or teacher-made. Teacher-made tests (TMTs) consist of final examinations, unit tests, or weekly tests (Adom et al., 2020). Studies by Tan and Cordova (2019), Espinoza Molina et al. (2021), GDE (2017), and Zatul (2020) contend that educational institutions should rely on TMT data to make reliable judgements on the academic performance of their students.

The teacher-made test, as explained in this study, is the manual and painstaking process of thinking out and writing down possible test items by teachers. This will involve the teacher sitting down with the course content and writing out questions from what he or she has taught in the class. In other words, it could be seen as a manually generated test because the teacher in this process is not aided by any machine. A teacher-made test is one constructed by the classroom teacher to measure the extent of performance of specific objectives within the class (Evans et al., 2016; Ukwuije, 2012). The teacher-made test, which forms the bulk of tests often used in school, especially at the secondary school level, and needs to possess the essential qualities of a good test, including validity, reliability, difficulty index and discrimination power so that the main essence of testing itself, which is to be fair to all testees and to measure exactly what the test purports to measure, is not lost. Studies, such as those by Tan and Cordova (2019), Espinoza Molina et al. (2021), GDE (2017), and Zatul (2020), contend that educational institutions should rely on TMT data to make reliable judgements on the academic performance of their students. A teacher-made test is an alternative to a standardised test, constructed and administered by the teacher in order to measure students' comprehension of the contents being taught. It enables students to demonstrate what they know and are capable of doing (Evans et al., 2022). Teacher-made tests are mostly used for the purpose of formative evaluation in secondary schools. Because of the usefulness of tests in education, it is very important that tests developed present quality test items to achieve educational testing purposes.

Reliability and validity are two of the most essential aspects to consider when assessing the quality of any test instrument. Reliability: the reliability of a measuring instrument is concerned with the accuracy and precision of an evaluation procedure. This is the degree of consistency with which a test measures what it sets out to measure (Kolak, 2014). It refers to the consistency of scores obtained by the same person when re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions (Joshua, 2013).

Validity is how accurately a test measures what it is supposed to measure, that

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

is, the degree to which an instrument provides information which is relevant to the decision that is to be made, that is, measures what it is constructed to measure. Validity measures the agreement of test results with what the test intended to measure. Validity and reliability are related. An instrument can be reliable but not valid; however, it cannot be valid if it is unreliable. (Joshua, 2013). Educational testing has different kinds of tests, such as essay tests and objective tests. The e-based test predominantly employs multiple-choice items to determine candidate achievement. This has made the multiple-choice test format more popular than other test formats in recent times. Therefore, researches on test formats are tailored to achieving robust, fair, valid and reliable multiple-choice test items. Asim et al. (2020) gave a detailed classification of test formats to be essay-test type and objective-test type. The essay type was further classified into restricted and extended response types, whereas the objective test format was broadly grouped into the selection test type and the completion test type. The completion test type consists of fill-in-the-gap and short-answer types, while the selection test type comprises true/false test items, matching test items and multiple-choice test formats. Multiple-choice questions are regarded as having a high level of reliability since they are scored objectively (Udemba, Jacob and Oluwayemisi, 2024). Designing and making conventional multiple-choice question assessments provides educators with a variety of challenges (Ryan et al, 2020), and poorly constructed questions can lead to misinterpretation, confusion, and erroneous assessment of pupils' knowledge due to double negatives, confusing language, and misleading phrasing (Kaipa, 2021).

Psychometric profiling is a measure of the extent to which a test satisfies some conditions and possesses some qualities which are technically known as 'test psychometric properties'. These include validity, reliability, usability, difficulty index and discrimination power. Omole (2012) defined psychometric properties of a test as certain characteristics inherent in the test upon which an assessment of a candidate is based. These properties include the facility, difficulty and discrimination indices; the power of distractors; and validity and reliability indices. Psychometric properties of tests are the quantifiable aspects of tests that indicate their statistical strengths and weaknesses. Psychometric properties are the intrinsic components of tests that reveal information about the relevance, adequacy and usefulness of tests (Evans et al., 2022). Item difficulty is the proportion of examinees that responded correctly to the item. The item difficulty index equals the number of students who score that item right divided by the number of students who attempted the item (Omirin, 2022). The indices (p-value) vary from zero (0) for a complicated item to 1 for a very easy item. The difficulty coefficient is based on the assumption that a higher difficulty coefficient indicates a more demanding test score. Also, an increase in a test difficulty level results in an increased variability of the test, which is why when a p-value is low, the items are primarily difficult and vice versa.

Item Discrimination

In a study by Obinna (2011), the psychometric analysis of the West African Senior School Certificate Examination and National Examination Council of Nigeria was conducted with the aim of comparing the Standard Error in Measurement (SEM) of biology examinations conducted from 2000 to 2002 using the one-parameter model of item response theory (IRT). SEM is commonly used to produce confidence intervals and estimates of how much error is in a test. Instrumentation research design was used for the study, and a population comprising all year three (SSIII) senior secondary school students who enrolled for May/June/July 2006 Biology SSSCE NECO and WAEC in the three education zones of Benue State was studied. The sample for the study was one thousand eight hundred (1800) students selected using a multistage sampling procedure. NECO and WAEC 2000–2002 objective biology questions were the instruments for the study, and data were analysed using the maximum likelihood estimation techniques of the BILOG-MG computer program and the SPSS. The results showed significant differences in SEM of biology examinations conducted by NECO and WAEC in 2000, 2001 and 2002. It was recommended that IRT analysis should be employed by Nigerian examination bodies.

Day to day, many criticisms on educational tests unfold irrespective of who constructed the test. There is criticism that teacher-made tests are poorly constructed and the results obtained from such poorly constructed tests are not valid and reliable. There is another criticism of standardised tests: that they are biased and complicated and do not discriminate between the dull and bright student (Ubi & Udemba, 2012). This could be one of the reasons for students' poor performance in external examinations. There is a need to develop a quality test to continuously assess student achievement in the subject to improve students' performance. Some reasons that could be responsible for the students' poor performance in school include teacher attitude, student attitude, teacher qualification and poor teacher knowledge in test construction, which may result in poor-quality teacher-made tests.

Based on criticism, the need to develop and present quality test items has been unrelenting in 21st-century educational assessment. This is why algorithms driven by machine-learning technologies, such as ChatGPT, are now being prioritised. ChatGPT is an interactive chatbot created by OpenAI, a California-based artificial intelligence (AI) (Anyawuci, 2018). Artificial intelligence (AI) is the ability of a digital computer-controlled robot to perform tasks commonly associated with intelligent beings (Copeland, 2024). According to Kaplan & Haenlein (2019), some high-profile applications of AI include advanced web search engines like Google , Waymo, generative and creative tools like ChatGPT, ChatBot, etc. Milicevic, Lazarova & Pavlovic (2024) maintain that the use of artificial intelligence has wider applications with limited questionable capabilities. In terms of testing, it is observed that most educational stakeholders, like teachers, now have been involved in the use of AI in generating test items

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

(Bsharat & Khlaif, 2024). They also mention that AI is capable of providing tailored difficulty levels, comprehensive data analysis and objective evaluations and reducing test anxiety.

According to Xu et al. (2023), AI can be used in constructing a personalised and accurate feedback system for students. In contrast to traditional assessment. AI gives tailored, real-time feedback. It reveals a student's unique strength and weakness in comprehension. AI scoring can save a lot of time for teachers by providing quick feedback (Ramesh & Sanampudi, 2022). AI assesses objectively, it can provide consistent feedback, and it is unbiased (Schwartz et al, 2022). When comparing AI-based assessments with traditional assessments, it can reduce the workload of teachers and can provide consistent assessment across different classes and schools (Zeeshan et al., 2024). They further noted that difficulty level, discrimination index and reliability of both teachers' tests and AI tests are nearly equal and that AI-based tests can be used along with teacher-made tests at the secondary school level. In this study, the researcher prompted AI apps to generate thirty multiple-choice questions using the SS2 Mathematics scheme of work for the first term. Hence, the objectives of the study are as follows:

1. Determine the difficulty index of multiple-choice test items generated by the teacher manually and by the AI application.
2. Examine the discriminative index of the multiple-choice test item generated by the teacher and the one generated by the AI application.
3. Determine the reliability indices of the multiple-choice test item generated by the teacher manually and the one generated by the AI application.

Research Question

1. What is the difference in terms of difficulty index between a teacher-made test and an AI-generated test in assessing students' knowledge?
2. What is the difference in terms of discrimination index between a teacher-made test and an AI-generated test?
3. Do AI-generated test items significantly differ from teacher-made test items in terms of reliability?

Research Hypotheses

1. There is no significant difference in the difficulty levels of teacher-made tests and AI-generated tests.
2. There is no significant difference in the discrimination index of teacher-made tests and AI-generated tests.
3. There is no significant difference in the reliability of AI-generated and teacher-made tests.

Methodology

Research Design

For effective comparative analysis of AI-generated and teacher-made multiple-choice test items, a quasi-experimental design was deployed. Both types of generated test items were administered to all the students in a control setting. Administering both types of generated test items to the students allows us to directly compare the findings and rule out any potential confounding variables. Furthermore, a quantitative approach enables us to efficiently collect a vast volume of data and evaluate it using statistical approaches in order to reach meaningful conclusions.

The study is delimited to Uyo Educational Zone of Akwa Ibom State, Nigeria, focusing on secondary schools within the zone. Akwa Ibom State is located in the South-South geopolitical zone of Nigeria, divided into six educational zones for administrative and supervisory purposes: Uyo, Ikot Ekpene, Abak, Oron and Ikono. Among these, Uyo Educational Zone holds strategic importance as it accommodates the state capital, Uyo, and serves as a hub for academic, technological and socio-economic activities. Schools within the Uyo educational zone are often at the forefront of adopting modern educational technologies, making it a suitable site for examining the psychometric profiling of test items constructed both by teachers and artificial intelligence tools.

The Uyo educational zone provides a fertile ground for this comparative psychometric study because it embodies a mix of urban sophistication and semi-urban realities. The diversity of its students' population, coupled with varying levels of school resources and teacher expertise, ensures a representation platform for testing the quality of AI-generated and teacher-constructed test items. Selected secondary schools from the zone formed the stratum for which data were obtained.

Participants

The population size of the study was all senior secondary two (SS2) students in Uyo Educational Zone of Akwa Ibom State. The muti-stage random sampling technique was employed for selecting the sample size. This sampling technique ensured systematic and representative selection of samples from five schools selected from the Uyo Educational Zone. At first, five local government areas were selected through simple random sampling from nine LGAs. At the second stage from each of the five schools, simple random sampling was used in selecting one school. Lastly, proportionate sampling was used in selecting 300 students for the study.

Measures

The instruments used in this study were two sets of 30 multiple-choice test items in

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

mathematics; one set was designed by the teacher and the other was generated by AI. Each question offered four options labelled A, B, C, and D from which the single correct option was to be chosen. In or to systematically guide the construction of the test items and guarantee that they adequately capture the learning objectives and content domains prescribed in the curriculum relevant to this study, a comprehensive table of specifications was carefully developed and employed as the blueprint for test item generation.

To ensure the validity of the teacher-made test, constructed test items were reviewed by two subject matter experts in test item construction and two experienced teachers. The review focused on aligning the items drawn from the senior secondary school 2 (SSII) mathematics curriculum using a table of specifications with the curriculum. This was necessary to ensure both content and face validity. The split-half method was used to calculate the reliability. The reliability obtained was 0.83.

AI-Generated Test

The researcher generated a test consisting of 30 items using ChatGPT, according to the test blueprint. To ensure the test accurately reflects the curriculum and is appropriate for a multiple-choice test format. Each question offered four answer choices labelled A, B, C, and D. The AI-generated test items were reviewed by two subject matter experts and two experienced teachers. The review focused on aligning the item (drawn from the senior secondary school 2 (SSII) mathematics curriculum using a table of specifications with the curriculum. This was necessary to ensure both content and face validity. Reliability was carried out using the Kuder-Richardson 20 (KR-20) reliability method. The rules of thumb used for the judgement on item analysis were P-value indices from 0.40 to 0.70 were moderate, and then 0.00 to 0.39 and > 0.70 were considered inappropriate (too difficult and too easy). In terms of item discrimination, a cut-off score of >0.40 was adopted as proposed by Aljehani et al. (2020). The K-R20 and Cronbach's Alpha were used in determining the reliability estimate of teacher-made items (TMI) and AI-generated items (AIGI). This enables the study to compare the two approaches across key psychometric indices such as difficulty index, discrimination index, and reliability (KR-20 and Cronbach's alpha) estimates.

Results

HO1: There is no significant difference in the difficulty levels of teacher-made test and AI-generated tests.

Table 1

P-value of teacher-made test and AI-generated test

Item	P-value (TMI)	Description	Item	P-value (AIGI)	Description
1	0.52	Moderate Item	1	0.58	Moderate Item
2	0.34	Too Difficult item	2	0.59	Moderate Item
3	0.61	Moderate Item	3	0.52	Moderate Item
4	0.51	Moderate Item	4	0.45	Moderate Item
5	0.52	Moderate item	5	0.42	Moderate Item
6	0.54	Moderate Item	6	0.54	Moderate Item
7	0.61	Moderate item	7	0.57	Moderate Item
8	0.49	Moderate Item	8	0.60	Moderate Item
9	0.33	Too Difficult item	9	0.65	Moderate Item
10	0.62	Moderate item	10	0.53	Moderate Item
11	0.70	Moderate Item	11	0.57	Moderate Item
12	0.44	Moderate Item	12	0.10	Too Difficult item
13	0.79	Too Easy item	13	0.41	Moderate Item
14	0.56	Moderate item	14	0.61	Moderate Item
15	0.53	Moderate item	15	0.68	Moderate Item
16	0.56	Moderate Item	16	0.64	Moderate Item
17	0.61	Moderate item	17	0.15	Too Difficult item
18	0.59	Moderate Item	18	0.49	Moderate Item
19	0.63	Moderate Item	19	0.32	Too Difficult item
20	0.50	Moderate Item	20	0.50	Moderate Item
21	0.91	Too Easy item	21	0.25	Too Difficult item
22	0.29	Too Difficult item	22	0.33	Too Difficult item
23	0.54	Moderate Item	3	0.64	Moderate Item
24	0.65	Moderate Item	24	0.49	Moderate Item
25	0.50	Moderate Item	25	0.57	Moderate Item
26	0.52	Moderate Item	26	0.55	Moderate Item
27	0.63	Moderate Item	27	0.81	Too Easy Item
28	0.51	Moderate Item	28	0.51	Moderate Item
29	0.57	Moderate Item	29	0.19	Too Difficult item
30	0.65	Moderate Item	30	0.55	Moderate Item

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

Table 2

Item Difficulty Result for Teacher-Made Test

S/N	Category	Items	Frequency	Percentage
1	Difficult	2,9,22	3	10%
				83.3%
2	Moderate	1,3,4,5, 6,7,8,10,11,12,14,15, 16.17, 18,19,20 ,23,24,25,26,27,28,29,30	25	
3	Easy	13,,21	2	6.7%

Based on the data above, it can be seen in the teacher-made test that the highest number of items are moderate 25 (83.3%). Whereas only two items (6.7%) can be categorised as easy, while three items (10%) can be categorised as difficult. In the AI-generated test, the highest number of items are moderate, i.e., 24 (80%), whereas 5 (16.7%) items can be categorised as difficult and 1 (3.3%) can be categorised as easy. In all, it indicates that the teacher-made test (TMT) had 25 items with acceptable difficulty, while the AI-generated test had 24 items. It could be seen that the teacher-made test was better in producing items with moderate difficulty than the AI-generated test.

Table 2 shows that the highest number of items are moderate, i.e., 25 (83.3%), whereas 3 (10%) items can be categorised as difficult and 2 (6.6%) can be categorised as difficult in teacher-made tests.

Table 3:

Item Difficulty Result of AI-Generated Test

S/N	Category	Items	Frequency	Percentage
1	Difficult	12,17,19,21,22	5	16.7%
2	Moderate	1,2,3,4,5,6,7,8,9.10,11,13,14,15,16, 18,20,21,23,24,25,26,28,29,30	24	80%
3	Easy	27	1	3.3 %

From Table 2 and Table 3, it can be seen that 83.3% of items are moderate in teacher-made tests and 80% are moderate in AI-generated tests. In teacher-made tests, only (10%) of the items are difficult. Whereas 16.7% of items are difficult in AI-generated tests. Therefore, it can be concluded that the teacher-made test has performed well in terms of difficulty indices of the items.

HO2: There is no significant difference in the discrimination index of teacher-made tests and AI-generated tests.

Table 4:

Item Discrimination Result of Teacher-Made Test and AI-Generated Test

Items	Discrimination TMT	Remarks	Items	Discrimination of AI Test	Remarks
1	0.50	Good item	1	0.47	Good item
2	0.12	Poor item	2	0.51	Good item
3	0.40	Good item	3	0.45	Good item
4	0.52	Good item	4	0.55	Good item
5	0.30	Poor item	5	0.57	Good item
6	0.51	Good item	6	-0.09	Poor item
7	0.49	Good item	7	0.41	Good item
8	0.33	Poor item	8	0.40	Good item
9	0.69	Good item	9	0.05	Poor item
10	0.57	Good item	10	0.51	Good item
11	0.60	Good item	11	0.50	Good item
12	0.45	Good item	12	-0.04	Poor item
13	0.49	Good item	13	0.54	Good item
14	0.62	Good item	14	0.61	Good item
15	0.02	Poor item	15	0.40	Good item
16	0.53	Good item	16	0.50	Good item
17	-0.12	Poor item	17	0.10	Poor item
18	-0.19	Poor item	18	0.42	Good item
19	0.61	Good item	19	0.54	Good item
20	0.48	Good item	20	0.41	Good item
21	0.08	Poor item	21	0.44	Good item
22	0.14	Poor item	22	0.53	Good item
23	0.56	Good item	23	0.02	Poor item
24	0.61	Good item	24	0.15	Poor item
25	0.53	Good item	25	0.57	Good item
26	0.55	Good item	26	0.49	Good item
27	0.47	Good item	27	0.55	Good item
28	0.43	Good item	28	0.18	Poor item
29	0.51	Good item	29	0.52	Good item
30	0.38	Poor item	30	0.42	Good item

From Table 4, 21 items in the teacher-made test (TMT) had good discrimination. They include 1, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 16, 19, 20, 23, 24, 25, 26, 27, 28 and 29. On the

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

other hand, items 2, 5, 8, 15, 17, 18, 21, 22 and 30, respectively, discriminated poorly. For AI-generated items (AIGT), 23 items had good discrimination indices. These include items 1, 2, 3, 4, 5, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 20, 21, 22, 25, 26, 27, 29 and 30. On the contrary, items 6, 9, 12, 17, 23, 24 and 28 had poor discrimination indices. Comparatively, this implies that AI-generated tests (AIGT) had better discrimination (23>21) than teacher-made tests (TMT).

HO3: There is no significant difference in reliability of AI-generated and teacher-made tests.

Table 5:

Internal Consistency of Teacher-Made Test and AI-Generated Test

K: Teacher-Made Items	**∑pq**	**δ 2**	**K: AI-Generated Items**	**∑pq**	**δ 2**
30	9.23	112.024	30	8.0125	32.620
Reliability= $\frac{k}{(k-1)}\left(\frac{1-(\Sigma pq}{\delta^2}\right)$			Reliability= $\frac{k}{(k-1)}\left(\frac{1-(\Sigma pq}{\delta^2}\right)$		
Where k = number of items & δ^2= variance			Where k = number of items & δ2= variance		
$K_{r20}= \frac{30}{30-1}\left(1-\frac{9.23}{112.024}\right)$			$Kr_{20}= \frac{30}{30-1}\left(1-\frac{8.0125}{32.620}\right)$		
$Kr_{20}= \frac{30}{29}\left(1-0.08239\right)$			$K_{r20} = \frac{30}{29}\left(1-0.24563\right.$		
Kr_{20}= 1.03448(0.91761			Kr_{20}= 1.03448(0.75437		
Kr_{20}= 0.94924			K_{r20} = 0.74590		

From Table 5 above, the reliability estimate using Kr-20 shows that the teacher-made test has reliability coefficient of 0.94924, and the AI-generated test has coefficient of 0.74590, indicating that both tests are highly reliable, but the teacher-made test has a better reliability coefficient compared to the AI-generated test.

Discussions

There was no significant difference in difficulty score between the teacher-made test and the AI-generated test. Therefore, it can be concluded that the difficulty of AI-generated tests is nearly equal to the difficulty of teacher-made tests. The current finding is in agreement with findings of Zong and Krishnamachairari (2022), who reported that AI-generated tests lack significant clarity, which serves as a disadvantage to students when compared with manually generated tests. These results are contradictory to the findings of Agarwal et al. (2023) in the study 'Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology', which found that ChatGPT generated the

most valid MCQs but the least difficult ones. The possible reason for the contradiction in the findings might be that they were simply focused within the time limit of responding to the items; it could also be that the students did not understand these questions or they may have been too confused to know the exact answer and missed out on attempting them due to time.

From the second finding, based on Aljehani et al. (2020). Item discrimination is "good" if the index is above 40. From this premise, it is clear that the teacher-made test's 21 items had good discrimination, while 6 items showed poor discrimination. This means that 9 of the items of the teacher-made test were not able to distinguish between the high-performing student and the low-performing one, From AI generated test, 23 item had good discrimination, while 7 items showed poor discrimination. Comparatively it indicates that AI generated test had better discriminating power than teacher made test. The result of the findings here could be that AI have better algorithms in delineating specific constructs more compared to individual teachers' ability. This has been verified by Attali, LaFlair and Rung (2023), who noted that leveraging AI to write unit tests empowers developers to improve test coverage, enhance accuracy and efficiency, and optimise resource utilisation.

Both AI-generated tests and teacher-made tests show very good reliability in assessing knowledge. The teacher-made test, with a KR20 value of 0.94924. It demonstrates a marginally higher level of reliability compared to the AI-generated test. Which has a KR20 value of 0.74590. This suggests that teacher-made tests can be highly reliable and potentially even more consistent than the AI-generated test in evaluating student knowledge. The findings mean that it may be attributed to the fact that AI algorithms can analyse a vast amount of data and generate test items that are more consistent and accurate. It can also mean that AI-generated test items can be designed to meet specific learning objectives and can be tailored to individual students' needs. The findings are supported by Dodan, Goru Dogan and Bozkurt (2023), who reported significant teacher-made tests have higher reliability than that of AI-generated tests.

Conclusion

A comparative analysis of teacher-made and AI-generated mathematics multiple-choice test terms using senior secondary school II students was conducted. The study reveals that teacher-made multiple-choice mathematics questions were of slightly superior discriminating indices to AI-generated items. A similar observation holds in the reliability coefficients between teacher-made tests and artificial intelligence-generated test items. However, the reliability of both test items proved to be very good in assessing students' achievement by meeting acceptable standards for test reliability. Interestingly, difficulty level indicates that AI-generated test items were more difficult than teacher-made test items. Therefore, AI-generated test items should not be used in isolation to test students.

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

Recommendations

Based on the findings of the study, the following recommendations were made:

1. Difficulty level, discriminating index and reliability of both teacher-made tests and AI tests are almost equal; therefore, it is recommended that AI-based tests can be used along with teacher tests at the secondary school level.
2. AI developers should work closely with teachers and educational experts to get technical details and improve assessments. Furthermore, AI developers can get insight from subject experts and psychometricians to make assessments technically sound at different standards.

References

Adesina, I.O (2024). The Role of Artificial Intelligence in Teaching of Science Education in Secondary Schools in Nigeria. *European Journal of computer Science and Information Technology, 12*(1), 57-67.

Adom, D., Adu Mensah, J., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education*, *9*(1), 109–119. https://doi.org/10.11591/ijere.v9i1.20457.

Attali, Y., LaFlair, G. ,&Runge, A. (2023). A new paradigm for test development (Duolingo webinar series). https://www.youtube.com/watch?v=rRc960e9bzk&t2s

Asim, A.E., Evans, G.U., Idaka, I.E. (2020). Analysis of multiple-choice item format and Secondary School Student Achievement in Mathematics in Akwa Ibom State, Nigeria, *African Journal of Theory and Practice of Educational Research* (AJTPER) 8, 58-72.

Bsharat, T. & Khlaif, Z. (2024). Generative AI-Powered Adaptive Assessment. In 430 https//doi.org/10.4018/979-8-3693-6397-3.

Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. Frontiers in Education.

Edtech, (2020). Successful AI Example in Higer Education That Can Inspire Our Future EdTech Magazine

Espinoza M, F. E., Arenas R., B. D. V., Aparicio, F. & Zúñiga O, D. C. (2021). Road safety perception questionnaire (RSPQ) in Latin America: a development and validation study. *International Journal of Environmental Research and Public Health*, *18*(5), 2433. https://doi.org/10.3390/ijerph18052433.

Evans, G. U. (2016). Students' perception of multiple-choice item format and Mathematics achievement test in junior secondary schools in Uyo Educational Zone of Akwa Ibom State, Nigeria. *Academic Journal of Educational Research*, *4*(7), 111-116.

Evans, G. U., Uko, M. P. & Ekim, R. E. D. (2022). Investigation of differential item functioning

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

of basic education certificate examination (BECE) Mathematics items in Akwa Ibom State. The *African Journal of Behavioural and Scale Development Research*, *4*(1), 62-75.

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2023). Hello GPT!Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, 1-11.

Georgia Department of Education. (2017). An assessment & accountability brief: 2016-2017 Georgia milestones validity and reliability.

Joshua, M. T. (2013). Fundamentals of Test and Measurement in Education. ANITA Press, Eyo-ita Street, Calabar, Nigeria.

Kaipa, R. M. (2021). Multiple choice questions and essay questions in curriculum. *Journal of Applied Research in Higher Education, 13*(1), 16-32.

Kaplan, J.D. & Haenlein, G. (2019). Language model are few-shot learners. Advance in Neural Information Processing Systems 33 (NeuraIPS2020). https://proceedings Neurips.cc/paper./2020hash/1457c0d6bfeb4967418bfb8ac142f64a-Abstract.html.

Kolak, A. (2014), Teachers' attitude towards evaluation process. Retrieved on 20/07/2017 from www.google/hrcak.srce.hr.

Milicevice, V., Lazarova, L. k & Pavlovic, M. J. (2024). The Application of Artificial Intelligence in Education-The current State and Trends. *International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE), 12*(12), 259-272.

Naqvi, (2020). Artificial intelligent for adult, forensic accounting and valuation: *A strategic perspective*. John Wiley & Sons. https://doi.org/10.1002/97811119601906.

Obinne, A.D.E. (2011). A Psychometric Analysis of Two Major Examinations in Nigeria: Standard Error of Measurement. Retrieved 7-7-2015 from https//sites.google.com/site/hopecep900/rdp-1/annotated-referemces.

Omole, D. O. K. (2012). A comparative Analysis of the Psychometric Characteristics of JAMB,NABTEB, NECO AND WAEC Conducted Biology Examinations in Nigeria. Keffi *Journal of Educational Studies (KEJES). A Publication of the Faculty of Education Nasarawa State University Keffi Nigeria 3,* (1), June 2012.

. Ryan, A., Judd, T., Swanson, D., Larsen, D. P., Elliott, S., Tzanetos, K., & Kulasegaram, K. (2020).Beyond right or wrong: More effective feedback for formative multiple-choice tests.Perspectives on Medical Education, 9, 307-313.

Sallam, M., Salim, N. A., Barakat, M., & Ala'a, B. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: *A descriptive study highlighting the advantages and limitations. Narra J, 3*(1).

Schwartz, R. Vassiley, A., Green, K., Perine, L. Burt, A. & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. NSIT special publication, 1270(10.6028).

Tan, D. A., & Cordova, C. C. (2019). Development of Valid and Reliable Teacher-Made Tests for Grade 10 Mathematics. *International Journal of English and Education*, *8*(1)

Ubi, I.O. &Udemba, E.C. (2021). Age Differentials in Calibrated items of WAEC English Language Objective Test Taken by Students in Nigeria. *Global Journal of Educational Research, 20*(1), 45-54. 6th August. Publisher AJOL www.global journal series.com; globaljournalseries@gmail.com.

Udemba, E.C. Jacob, E.O.& Oluwayemisi, D. A. (2024). Psychometrics Properties of Artificial Intelligence (CHATGPT Bread Economics Multiple- Choice items. *African Journal of Theory and Practice of Educational Assessment, 13*(2) 26-35.

Ukwuije, R P.I (2012). Educational Assessment: A Sine Qua Non for quality Education 83rd Inaugural Lecture, University of Port Harcourt, Choba.

Vasconcelos, M.A.R. & Dos Santos, R. P. (2023). Enhancing STEM learning with Chart GPT and Bing Chat as objects to think with: A case study. EURASIA *Journal of Mathematics, Science and Technology Education. 19*(7), https://doi.org/10.29333/ejmste/13313.

ISSN: 2705-3830 (Print)
ISSN: 2705-3857 (Online)

Xu, W. Meng, J.,Raja, S.K.S., Priya, M.P,& Kirurhiga Devi, M. (2023).Artificial intelligence inconstructing personalized and accurate feedback system for students. *International Journal of Modeling, Simulation, and Scientific Computing, 14*(01), 2341001.

Zatul, T. (2020). Investigating reliability and validity of student performance assessment in Higher Education using Rasch Model. Journal of Physics: Conference Series, 1529, 042088.

Zeeshn, .M.,Iqbal, M.,Sahibzada, S. Malik, G. M. (2023).*A Comparative Analysis of Psychometric Properties in AI-Generated and Teacher-Made MCQs Kurdish Studies 12*(4).188-18200 www.kurdishstudies.net DOI:10.53555/ks.v12j4.3653

Zong, M. & Krishnamachari,B. (2022). A survey on GPT-3. Preprint. https://dio.org/10.48550/arXiv.2212.00857.