*KASU International Journal of Social Science Education (*KIJSOSE)
ISSN Print: *3141-2564*

Psychometric Properties of Chatgpt-4o-Generated Test Items in Social Science Education: A Multi-Dimensional Evaluation in Secondary Schools in Kaduna State Metropolis, Nigeria

Bashir Mamman (Ph.D)
Department of Education Foundations
Kaduna State University, Kaduna-Nigeria

Abstract
The integration of artificial intelligence (AI) tools, specifically large language models (LLMs)—into educational assessment has introduced new possibilities and challenges for test development in Nigerian secondary schools. This study evaluated the psychometric properties of test items generated by ChatGPT-4o in Social Science subjects in secondary schools in Kaduna State Metropolis, Nigeria, thereby contributing an empirically grounded, context-specific baseline for evidence-informed AI-assisted assessment practice in the Nigerian secondary school system. A descriptive survey research design was adopted, with a sample of 120 Social Science teachers selected through stratified random sampling from 30 public secondary schools in Kaduna Metropolis. Three validated instruments were utilised: a Teachers' Perception Questionnaire on AI-Generated Test Items (TPQAGTI; S-CVI = .84), an AI-Generated Item HOTS Classification Protocol (AIHCP), and an Item Difficulty Analysis Record (IDAR). A pool of 200 AI-generated test items was produced using ChatGPT-4o and administered to 240 Senior Secondary School (SSS) III students across four selected schools in Kaduna Metropolis. Data were analysed using descriptive statistics, one-way Analysis of Variance (ANOVA), chi-square goodness-of-fit test, and Kruskal-Wallis H test. Results revealed that teachers held generally positive perceptions of AI-generated test items (grand mean = 3.55), with significant inter-group differences across experience levels [$F(3, 116) = 3.42$, $p = .020$, $\eta^2 = .08$]. The majority of AI-generated items (71.0%) operated at lower-order thinking levels, with a statistically significant non-uniform cognitive distribution ($\chi^2 = 48.85$, $df = 5$, $p < .001$). Item difficulty indices were predominantly moderate (52.5%), with an overall mean discrimination index of D = .26, indicating generally marginal discriminatory power. Findings are limited to Kaduna State Metropolis and to items generated by a single LLM (ChatGPT-4o); generalisability to other geographic contexts or AI platforms requires further investigation. The study concludes that while ChatGPT-4o-generated test items demonstrate acceptable difficulty calibration, they exhibit marginal discrimination and require deliberate HOTS-aligned prompt engineering and professional human oversight before deployment in the Nigerian secondary school context.

Keywords: Psychometric evaluation, Social Science education, item difficulty, item discrimination, higher-order thinking, ChatGPT-4o

Introduction

The advancement of artificial intelligence (AI) in the 21st century has fundamentally altered the landscape of educational assessment, introducing automated systems capable of generating, scoring, and analysing test items at an unprecedented scale and speed (Gorgun & Bulut, 2023; Kurdi et al., 2020). Among the most notable developments in this domain is the emergence of large language models (LLMs) such as ChatGPT, Gemini, and Claude, which have demonstrated remarkable capacity for generating diverse and contextually appropriate multiple-choice questions (MCQs) across various subject domains (Kaya et al., 2025; Moore et al., 2023). These tools have attracted considerable scholarly attention for their potential to

 Copyright © 2026 KIJSOSE All Rights Reserved

reduce teachers' assessment burden, promote item bank diversification, and enable rapid formative assessment (Baig et al., 2024; Jackson, 2025). However, the psychometric quality of AI-generated items—encompassing difficulty calibration, discrimination power, cognitive alignment, and content validity—remains an empirical question that demands rigorous investigation, particularly in developing nations with distinct curricular and sociocultural characteristics (Okonkwo & Ade-Ibijola, 2021).

In the Nigerian secondary school system, Social Science subjects—comprising Economics, Government, Geography, and Civic Education—occupy a critical position in preparing students for civic participation, economic reasoning, and environmental stewardship (Ijewere & Uwameiye, 2018). These subjects form part of the core curriculum in public secondary schools and are assessed in high-stakes national examinations conducted by the West African Examinations Council (WAEC) and the National Examinations Council (NECO). Despite their significance, teachers of these subjects frequently encounter substantial challenges in constructing valid and reliable test items, partly due to large class sizes, limited professional development opportunities, and inadequate assessment resources (Afolaranmi et al., 2021; Abiodun & Shehu, 2022). The prospect of using AI tools to generate items that can supplement or replace teacher-made tests is therefore professionally appealing, but must be evaluated against established psychometric standards before adoption.

Psychometric evaluation the systematic appraisal of a test instrument's technical quality involves the examination of item-level indicators such as the difficulty index (proportion of students answering correctly), the discrimination index (the degree to which an item differentiates between high- and low-performing students), and cognitive demand as aligned to established taxonomies (Hambleton & Swaminathan, 1985; Crocker & Algina, 2006). Classical Test Theory (CTT) provides the predominant framework for conducting such evaluations, especially in contexts where Item Response Theory (IRT) models may be technically inaccessible (Magno, 2009; Cappelleri et al., 2014). Research conducted globally suggests that AI-generated MCQs are capable of achieving acceptable difficulty calibration comparable to expert-written items (Kaya et al., 2025; Pinto et al., 2025), yet concerns persist about their tendency to cluster at lower-order cognitive levels, exhibit lower-than-ideal discrimination values, and demonstrate limited responsiveness to locally specific curricular contexts (Jackson, 2025; Moore et al., 2023). Crucially, no prior study has systematically analysed both the difficulty and discrimination properties of AI-generated items in Social Science subjects at the secondary school level within Kaduna State Metropolis, representing a critical gap in the existing literature.

Teacher perception studies in Nigeria reveal a generally positive but cautious attitude toward AI tools for assessment purposes. Fwangle et al. (2025) found that secondary school teachers in Plateau State held high positive perceptions of AI, though this did not automatically translate into improved student academic outcomes. Similarly, Ofem et al. (2025) documented that teachers in Ondo State expressed endorsement for AI tools while noting deficiencies in cultural localisation and contextual authenticity. These findings highlight the need to examine teacher perception not only as a measure of technology acceptance, but as a form of content validity evidence for AI-generated assessments. Furthermore, studies examining the cognitive distribution of AI-generated items across Bloom's Revised Taxonomy have documented a systematic bias toward lower-order thinking levels, particularly Remembering and Understanding (Jackson, 2025; Lahza et al., 2023; Adiguzel et al., 2023). In Social Science education, where national curricula and examination bodies increasingly emphasise critical analysis, evaluation of policies, and application of theories to

 Copyright © 2026 KIJSOSE All Rights Reserved

real-world scenarios, an over-representation of lower-order items constitutes a threat to assessment validity and instructional alignment (Afolaranmi et al., 2021; Anderson & Krathwohl, 2001).

Despite the growing global discourse on AI in educational assessment, empirical studies specifically evaluating the psychometric properties including both difficulty and discrimination indices of AI-generated test items in Social Science subjects at the secondary school level in Nigeria remain scarce (Okonkwo & Ade-Ibijola, 2021; Kurdi et al., 2020). Extant studies have predominantly focused on medical education, STEM disciplines, or higher education institutions in Western contexts (Kaya et al., 2025; Moore et al., 2023), leaving a substantial contextual gap in the literature. This study, therefore, addresses these identified gaps by investigating the psychometric properties of ChatGPT-4o-generated test items in Social Science subjects in secondary schools in Kaduna State Metropolis, with a view to generating empirical evidence to guide the contextualised adoption of AI-assisted assessment in Nigeria. Hence, the objectives of this study are to:

i. Examine teachers' perception of the quality and usability of ChatGPT-4o-generated test items in Social Science subjects in secondary schools in Kaduna State Metropolis.
ii. Determine the extent to which ChatGPT-4o-generated test items reflect higher-order thinking skills in Social Science subjects in secondary schools in Kaduna Metropolis.
iii. Examine the item difficulty and discrimination distribution patterns of ChatGPT-4o-generated test items across different Social Science sub-disciplines (Economics, Government, Geography, and Civic Education) in secondary schools in Kaduna Metropolis.

In line with the research objectives, the following research questions were developed:

i. What are teachers' perceptions of the quality and usability of ChatGPT-4o-generated test items in Social Science subjects in secondary schools in Kaduna State Metropolis?
ii. To what extent do ChatGPT-4o-generated test items reflect higher-order thinking skills in Social Science subjects in secondary schools in Kaduna Metropolis?
iii. What is the item difficulty and discrimination distribution pattern of ChatGPT-4o-generated test items across different Social Science sub-disciplines (Economics, Government, Geography, and Civic Education) in secondary schools in Kaduna Metropolis?

The following null hypotheses were tested at 0.05 level of significance:

Ho_1: There is no significant difference in teachers' perceptions of the quality and usability of ChatGPT-4o-generated test items in Social Science subjects in secondary schools in Kaduna State Metropolis.

Ho_2: ChatGPT-4o-generated test items do not significantly reflect higher-order thinking skills in Social Science subjects in secondary schools in Kaduna Metropolis.

Ho_3: There is no significant difference in the item difficulty distribution pattern of ChatGPT-4o-generated test items across Social Science sub-disciplines (Economics, Government, Geography, and Civic Education) in secondary schools in Kaduna Metropolis.

Literature Review

Underpinning Theoretical Framework

i. Classical Test Theory (CTT): Classical Test Theory, rooted in the foundational psychometric contributions of Spearman (1904), and further developed by Gulliksen (1950) and Lord and Novick (1968), constitutes one of the most enduring and widely applied frameworks in educational measurement. The central postulate of CTT holds that any

 Copyright © 2026 KIJSOSE All Rights Reserved

observed test score (X) represents the algebraic sum of a true score (T) and a random error component (E), expressed as: X = T + E. This conceptualisation implies that systematic differences in performance between examinees can be attributed to genuine differences in underlying ability, while residual variability is ascribed to unsystematic measurement error. Within CTT, the primary item-level statistics used for quality evaluation include the item difficulty index (p-value: the proportion of examinees who answered the item correctly), the item discrimination index (D-value: the difference in the proportion of correct responses between high- and low-scoring groups, typically computed using the upper and lower 27% of performers), and the point-biserial correlation coefficient (rpb: the correlation between item score and total test score) (Crocker & Algina, 2006; Ebel & Frisbie, 1991). Established benchmarks indicate that item difficulty values between 0.30 and 0.70 are considered optimal for achievement tests, while discrimination indices of 0.30 and above are indicative of acceptable item quality, with values below 0.20 generally considered poor (Hambleton & Swaminathan, 1985; Magno, 2009). These two statistics—difficulty and discrimination—are complementary and inseparable components of CTT-based item analysis; an item may exhibit acceptable difficulty but poor discrimination, thereby reducing its utility for differentiating between high- and low-ability students.

CTT is particularly pertinent to the evaluation of AI-generated test items in the Nigerian secondary school context for several reasons. First, it requires relatively modest sample sizes and straightforward computational procedures, making it accessible in resource-limited research environments (Cappelleri et al., 2014). Second, CTT-derived statistics—including both p-values and D-values—are widely interpretable by classroom teachers and school-level assessment coordinators, facilitating practical utility. Third, prior psychometric studies on AI-generated items have employed CTT frameworks to assess difficulty and discrimination, thereby enabling cross-study comparisons (Kaya et al., 2025; Pinto et al., 2025). Lahza et al. (2023) noted that traditional CTT item analysis remains practically indispensable for formative evaluation of item quality, particularly when the objective is to identify items requiring revision or replacement before large-scale deployment. In this study, CTT provides the statistical scaffolding for evaluating both the difficulty and discrimination distributions of ChatGPT-4o-generated Social Science items across the four sub-disciplines.

ii. Bloom's Revised Taxonomy of Educational Objectives: The original taxonomy of educational objectives, proposed by Bloom et al. (1956), classified cognitive learning outcomes into a hierarchical framework spanning Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. This framework was comprehensively revised by Anderson and Krathwohl (2001), who introduced a two-dimensional structure comprising the Knowledge Dimension (factual, conceptual, procedural, metacognitive) and the Cognitive Process Dimension (Remembering, Understanding, Applying, Analysing, Evaluating, Creating). The latter three levels—Analysing, Evaluating, and Creating—constitute Higher-Order Thinking Skills (HOTS), while the former three—Remembering, Understanding, and Applying—represent Lower-Order Thinking Skills (LOTS) (Anderson & Krathwohl, 2001; Adesope et al., 2017). This taxonomy has been consistently applied as a normative standard for evaluating the cognitive alignment of test items across educational levels and subject domains.

The salience of Bloom's Revised Taxonomy in the present study lies in its provision of a systematic and internationally recognised framework for assessing whether AI-generated items adequately challenge students at cognitively meaningful levels. Research examining the cognitive distribution of AI-generated items has consistently documented a tendency for

 Copyright © 2026 KIJSOSE All Rights Reserved

LLMs to produce items at the lower rungs of the taxonomy (Jackson, 2025; Moore et al., 2023). Jurmu et al. (2023) and Adiguzel et al. (2023) observed that items generated by ChatGPT across various disciplines clustered predominantly at the Remembering and Understanding levels, with markedly fewer items at the Evaluating and Creating levels. For Social Science subjects in Nigerian secondary schools, where WAEC increasingly incorporates analysis- and evaluation-level questions in its examination formats (Afolaranmi et al., 2021), the cognitive adequacy of AI-generated items is not merely a psychometric question but a matter of curricular alignment and examination preparedness. This study employs Bloom's Revised Taxonomy as the cognitive classification framework for analysing the HOTS distribution of ChatGPT-4o-generated Social Science test items.

Review of Empirical Studies

The psychometric evaluation of AI-generated test items has attracted considerable empirical attention in recent years, particularly following the widespread availability of LLMs. Kaya et al. (2025) conducted a comparative study examining MCQs generated by ChatGPT-4o against clinician-designed questions in emergency medicine education, finding that both item types achieved comparable difficulty indices while AI-generated items demonstrated slightly lower discrimination values, underscoring the continued importance of expert review and the dual necessity of evaluating both difficulty and discrimination in AI-generated item pools. Similarly, Pinto et al. (2026) evaluated the accuracy of AI-generated MCQs in a systematic assessment, noting that AI effectively calibrated question difficulty to align with curricular expectations, though faculty-generated questions maintained higher discrimination indices. These findings collectively indicate that AI-generated items are capable of meeting acceptable difficulty standards but may fall short in their discriminatory power without human refinement.

In the context of automated item generation (AIG) research, Kurdi et al. (2020) conducted a systematic review of 93 studies and concluded that AIG systems produce items of variable quality depending on the subject domain, the generation algorithm employed, and the extent of human oversight. Studies using LLMs represent a newer generation of AIG tools, and their psychometric evaluations have generally yielded mixed but promising results (Okonkwo & Ade-Ibijola, 2021; Baig et al., 2024). Despite these advances, the psychometric quality—encompassing both difficulty and discrimination properties—of AI-generated items in Social Science subjects at secondary school level in Nigeria has received no systematic empirical attention, representing a critical gap this study addresses.

The relationship between AI-generated test items and higher-order thinking skills is one of the most actively debated topics in contemporary educational measurement. Jackson (2025) proposed a framework of 'higher-order prompting' grounded in Bloom's Revised Taxonomy, demonstrating that the cognitive level of LLM-generated content is substantially responsive to the specificity of instructional prompts—a finding with direct implications for item generation practice. Moore et al. (2023) investigated GPT-3.5's capacity for generating items at various Bloom's levels, finding that while the model showed capability at the Evaluation level, discrepancies between AI-designated and human-verified cognitive levels were prevalent, with AI frequently overestimating the cognitive demand of generated items. In a study using automatic physics item generation with LLMs, Jurmu et al. (2023) found that targeted prompting strategies significantly increased the proportion of items at higher Bloom's levels, reinforcing the practical utility of prompt engineering as a HOTS-enhancement mechanism.

 Copyright © 2026 KIJSOSE All Rights Reserved

Research on Nigerian teachers' perceptions of AI in education has expanded considerably since 2022. Fwangle et al. (2025) surveyed senior secondary school teachers in Plateau State and reported high positive perceptions of AI's educational potential, while noting that such perceptions did not independently predict improved student academic achievement—a finding attributable partly to the absence of structured AI integration frameworks. In a related study in Ondo State, Ofem et al. (2025) found that teachers endorsed the time-saving benefits and customisation potential of AI while expressing reservations about accuracy, cultural alignment, and student over-dependence. In Anambra State, Okorie et al. (2025) observed that teachers' use of automated assessment systems was a significant predictor of students' academic performance ($r = .62$, $p = .01$), highlighting the practical relevance of teacher adoption attitudes. Across these studies, a recurrent theme is the tension between teachers' general enthusiasm for AI tools and their subject-specific reservations about the contextual authenticity and pedagogical rigour of AI-generated content—a tension that provides the motivational basis for examining teacher perceptions as a dimension of psychometric evaluation in the present study.

Methodology

This study adopted a descriptive survey research design, which is appropriate for investigating naturally occurring phenomena within a defined population without experimental manipulation (Creswell & Creswell, 2018; Cohen et al., 2018). The design facilitated the collection of primary data from Social Science teachers regarding their perceptions of ChatGPT-4o-generated test items, as well as systematic content analysis of AI-generated items to assess their cognitive alignment, difficulty characteristics, and discrimination properties. The concurrent triangulation of teacher perception data, HOTS classification data, and item difficulty and discrimination data enabled a comprehensive psychometric appraisal consistent with the multidimensional nature of the research problem. The target population comprised all Social Science teachers in public secondary schools in Kaduna State Metropolis, Nigeria. Based on records from the Kaduna State Secondary Education Board (KSSEB, 2024), there are approximately 412 certified Social Science subject teachers (Economics, Government, Geography, and Civic Education) distributed across 87 registered public secondary schools within the metropolis. The target population for the item administration component comprised SSS III students enrolled in Social Science electives in the same schools.

A sample of 120 Social Science teachers was drawn using stratified random sampling, stratifying the metropolis into three administrative zones (Kaduna North, Kaduna South, and Chikun/Igabi) and applying proportional allocation to ensure equitable representation. This yielded 46 teachers from Kaduna North, 41 from Kaduna South, and 33 from Chikun/Igabi. Additionally, 200 ChatGPT-4o-generated test items were purposively generated, with 50 items per sub-discipline (Economics, Government, Geography, and Civic Education), covering SSS I–III curriculum content based on WAEC and NECO syllabi. Items were generated using structured prompts that specified subject area, curriculum topic, difficulty level, and item format (four-option MCQ). The items were administered to 240 SSS III students (60 per sub-discipline) in four purposively selected public secondary schools across the metropolis, under standardised administration conditions. Regarding the adequacy of the student sample for CTT-based item analysis, it is acknowledged that Hambleton and Swaminathan (1985) recommend samples of 200–500 for stable difficulty and discrimination index estimation in standard CTT applications. The sample of 240 students employed in this study is therefore at the lower boundary of this recommended range, and the findings of the item analysis component should be interpreted accordingly. This sample size is appropriate

 Copyright © 2026 KIJSOSE All Rights Reserved

for an exploratory or pilot-level psychometric investigation (Crocker & Algina, 2006), and the results provide initial empirical evidence on the difficulty and discrimination properties of ChatGPT-4o-generated items in this context. Future studies should seek to replicate these findings with larger probability samples to yield more stable item parameter estimates. This limitation is further addressed in the Limitations section of this paper.

Three instruments were developed and validated for data collection:

1. Teachers' Perception Questionnaire on AI-Generated Test Items (TPQAGTI): A 25-item, 5-point Likert-scale instrument (1 = Strongly Disagree to 5 = Strongly Agree) measuring teachers' perceptions across five dimensions: clarity of items, curriculum alignment, cognitive appropriateness, usability and practicality, and sociocultural contextualisation. Face and content validity were established through a panel of five experts in Educational Measurement and Evaluation. The item-level Content Validity Index (I-CVI) was computed for each of the 25 items as the proportion of experts rating the item as relevant (scale points 3 or 4). The Scale-level Content Validity Index using the Averaging method (S-CVI/Ave) was computed as the mean of all item-level CVIs, yielding S-CVI/Ave = .84, which exceeds the recommended threshold of .80 for acceptable content validity (Lynn, 1986; Polit & Beck, 2006). A pilot test administered to 30 Social Science teachers outside the main sample yielded a Cronbach's alpha internal consistency reliability coefficient of $\alpha = .87$, indicating high reliability (Messick, 1995; Aiken, 1985).
2. AI-Generated Item HOTS Classification Protocol (AIHCP): A structured content analysis instrument designed to classify each of the 200 ChatGPT-4o-generated test items according to the six cognitive levels of Bloom's Revised Taxonomy (Remembering, Understanding, Applying, Analysing, Evaluating, Creating). Three independent raters—all holders of postgraduate qualifications in Social Science education—classified all items, achieving an inter-rater reliability of Cohen's $\kappa = .82$, indicative of strong agreement (Landis & Koch, 1977).
3. Item Difficulty and Discrimination Analysis Record (IDAR): A systematic data collection and computation template used to calculate both (a) the item difficulty index (p-value) and (b) the item discrimination index (D-value and point-biserial correlation coefficient, rpb) for each of the 200 test items, based on student response data. The D-value was computed as the difference in the proportion of correct responses between the upper 27% and lower 27% of total scorers (Ebel & Frisbie, 1991). The point-biserial correlation coefficient (rpb) was computed as the correlation between item score (0/1) and total test score, using IBM SPSS Statistics Version 29. Items were classified by difficulty as: easy ($p > .70$), moderate ($.30 \leq p \leq .70$), or difficult ($p < .30$); and by discrimination quality as: excellent ($D \geq .40$), good ($.30 \leq D < .40$), marginal ($.20 \leq D < .30$), or poor ($D < .20$), in accordance with established CTT benchmarks (Ebel & Frisbie, 1991; Hambleton & Swaminathan, 1985).

The TPQAGTI was administered to the 120 sampled teachers across 30 selected secondary schools over a period of four weeks with the assistance of three trained research assistants who provided clarification where necessary. All 120 distributed questionnaires were returned fully completed, yielding a 100% return rate. The 200 ChatGPT-4o-generated MCQs were compiled into four distinct subject-specific booklets and administered to 240 SSS III students under supervised, standardised conditions. Student response sheets were marked using pre-prepared answer keys, and the IDAR was used to systematically compute both difficulty and discrimination indices for each item. HOTS classification was independently conducted by three raters using the AIHCP, with disagreements resolved through consensus discussion. All data were entered and analysed using IBM SPSS Statistics Version 29. Descriptive

 Copyright © 2026 KIJSOSE All Rights Reserved

statistics—means, standard deviations, frequencies, and percentages—were used to answer the three research questions. For inferential testing, one-way ANOVA was employed to test Ho_1 (differences in teachers' perceptions across teaching experience groups: 0–5 years, 6–10 years, 11–15 years, above 15 years); the eta-squared statistic (η^2) was computed to quantify the practical magnitude of any significant group difference, with Cohen's (1988) conventions applied for interpretation (small = .01, medium = .06, large = .14). A post-hoc Scheffé test was conducted upon significant ANOVA findings. The chi-square goodness-of-fit test was applied to test Ho_2, assessing whether the distribution of AI-generated items across Bloom's six cognitive levels departed significantly from a theoretically equal distribution. The Kruskal-Wallis H test, a non-parametric equivalent of one-way ANOVA appropriate for ordinal data, was used to test Ho_3 (differences in item difficulty ranks across the four sub-disciplines). The alpha level for all inferential tests was set at .05.

Results

Research Question 1: Teachers' Perceptions of AI-Generated Test Items

Table 1 presents descriptive statistics of teachers' perceptions of the quality and usability of ChatGPT-4o-generated test items in Social Science subjects (N = 120).

Table 1: Descriptive Statistics of Teachers' Perceptions of AI-Generated Test Items (N = 120)

S/N	Item Statement	Mean (SD)	Decision
1	AI-generated test items are clearly worded and unambiguous	3.72 (0.84)	Positive
2	AI-generated items align well with the Social Science curriculum	3.45 (0.97)	Positive
3	AI-generated items are cognitively appropriate for SSS students	3.58 (0.89)	Positive
4	AI-generated items cover a sufficiently broad range of topics	3.61 (0.92)	Positive
5	AI-generated items require revision before classroom deployment	4.12 (0.73)	Positive
6	AI-generated items adequately measure students' conceptual understanding	3.38 (1.02)	Positive
7	AI-generated items reduce time spent on assessment preparation	3.89 (0.78)	Positive
8	AI-generated items include contextually appropriate examples	3.21 (1.08)	Positive
9	AI-generated items reflect Nigerian sociocultural realities	2.87 (1.14)	Negative
10	I would recommend AI tools to colleagues for test item generation	3.64 (0.91)	Positive
Grand Mean		3.55 (0.93)	Positive

Note. Mean (SD) reported. Decision rule: Mean ≥ 3.00 = Positive; Mean < 3.00 = Negative. Scale: 1 = Strongly Disagree, 5 = Strongly Agree.

The grand mean of 3.55 (SD = 0.93) indicates a generally positive teacher perception of ChatGPT-4o-generated test items. Item 5 recorded the highest mean (M = 4.12, SD = 0.73), reflecting strong agreement that AI-generated items require revision before use. Item 9 (M = 2.87, SD = 1.14) fell below the 3.00 decision threshold, indicating that teachers perceive a sociocultural localisation gap in ChatGPT-4o-generated items.

Research Question 2: HOTS Reflection of AI-Generated Test Items

Table 2 presents the distribution of ChatGPT-4o-generated test items across Bloom's Revised Taxonomy cognitive levels (N = 200).

Copyright © 2026 KIJSOSE All Rights Reserved

Table 2: Frequency Distribution of AI-Generated Test Items by Bloom's Revised Taxonomy Cognitive Levels (N = 200)

Cognitive Level	Category	Frequency	Percentage (%)
Remembering (C1)	LOTS	48	24.0
Understanding (C2)	LOTS	54	27.0
Applying (C3)	LOTS	40	20.0
Sub-total LOTS		142	71.0
Analysing (C4)	HOTS	34	17.0
Evaluating (C5)	HOTS	16	8.0
Creating (C6)	HOTS	8	4.0
Sub-total HOTS		58	29.0
Total		**200**	**100.0**

As shown in Table 2, the majority of ChatGPT-4o-generated items (71.0%) operated at lower-order thinking skill levels. The Understanding level (C2) was the most represented (27.0%), followed by Remembering (C1) at 24.0%. Items at the Evaluating (C5) and Creating (C6) levels accounted for only 8.0% and 4.0%, respectively. HOTS items collectively constituted just 29.0% of the total item pool.

Research Question 3: Item Difficulty and Discrimination Distribution

Table 3 presents the item difficulty distribution across Social Science sub-disciplines. The total row confirms the column-level proportions: Easy = 29.0% (n = 58), Moderate = 52.5% (n = 105), and Difficult = 18.5% (n = 37) of the total 200-item pool.

Table 3: Item Difficulty Distribution Pattern Across Social Science Sub-disciplines (N = 200)

Sub-discipline	Easy (p > .70)	Moderate (.30 ≤ p ≤ .70)	Difficult (p < .30)	Total
Economics	14 (28.0%)	26 (52.0%)	10 (20.0%)	50
Government	12 (24.0%)	29 (58.0%)	9 (18.0%)	50
Geography	17 (34.0%)	23 (46.0%)	10 (20.0%)	50
Civic Education	15 (30.0%)	27 (54.0%)	8 (16.0%)	50
Total	**58 (29.0%)**	**105 (52.5%)**	**37 (18.5%)**	**200**

Note. Figures in parentheses represent the percentage of items within each sub-discipline. Column totals (Easy = 29.0%, Moderate = 52.5%, Difficult = 18.5%) represent proportions of the total 200-item pool. Items were classified using established CTT benchmarks (Ebel & Frisbie, 1991).

Table 3 shows that 52.5% of all ChatGPT-4o-generated items were of moderate difficulty, consistent with optimal psychometric standards. Geography exhibited the highest proportion of easy items (34.0%), while Government recorded the highest proportion of moderately difficult items (58.0%).

Item Discrimination Analysis

Table 4 presents the item discrimination index (D-value) and point-biserial correlation coefficient (rpb) for ChatGPT-4o-generated items, summarised by difficulty category and by sub-discipline.

Copyright © 2026 KIJSOSE All Rights Reserved

Table 4: Summary of Item Discrimination Indices by Difficulty Category and Sub-discipline (N = 200)

Category / Sub-discipline	n	Mean D-value (SD)	Mean rpb (SD)	Discrimination Quality
By Difficulty Category				
Easy items (p > .70)	58	.18 (.06)	.21 (.07)	Poor
Moderate items (.30 ≤ p ≤ .70)	105	.33 (.08)	.36 (.09)	Good
Difficult items (p < .30)	37	.19 (.07)	.22 (.08)	Poor to Marginal
Overall	200	.26 (.09)	.29 (.10)	Marginal
By Sub-discipline				
Economics	50	.27 (.09)	.29 (.10)	Marginal
Government	50	.28 (.10)	.31 (.09)	Marginal to Good
Geography	50	.24 (.08)	.26 (.09)	Marginal
Civic Education	50	.26 (.09)	.28 (.10)	Marginal
Total	**200**	**.26 (.09)**	**.29 (.10)**	**Marginal**

Note. D-value = discrimination index (upper 27% – lower 27% correct response proportions). rpb = point-biserial correlation coefficient. Discrimination quality benchmarks: Excellent (D ≥ .40), Good (.30 ≤ D < .40), Marginal (.20 ≤ D < .30), Poor (D < .20)

Table 5: Distribution of Items by Discrimination Index Category (N = 200)

Discrimination Category	D-value Range	Frequency	Percentage (%)
Excellent	D ≥ .40	22	11.0
Good	.30 ≤ D < .40	51	25.5
Marginal	.20 ≤ D < .30	79	39.5
Poor	D < .20	48	24.0
Total		**200**	**100.0**

As shown in Tables 4 and 5, the overall mean discrimination index of D = .26 (SD = .09) falls below the acceptable threshold of .30, indicating that ChatGPT-4o-generated items demonstrate generally marginal discriminatory power. Items of moderate difficulty achieved the highest mean D-value (.33), consistent with psychometric theory, which predicts that items of intermediate difficulty offer the greatest opportunity for score differentiation (Ebel & Frisbie, 1991; Hambleton & Swaminathan, 1985). By contrast, easy and difficult items both exhibited poor-to-marginal discrimination (D = .18 and .19, respectively), consistent with the psychometric principle that items answered correctly (or incorrectly) by nearly all examinees are unlikely to differentiate between high and low performers. Only 36.5% of items (n = 73) achieved acceptable discrimination (D ≥ .30), while 63.5% (n = 127) fell in the marginal-to-poor range. These findings are consistent with the observation by Kaya et al. (2025) that ChatGPT-4o-generated MCQs demonstrate slightly lower discrimination values than expert-designed items, reinforcing the case for human-AI collaborative item development.

Hypothesis Testing

H_{o1}: One-Way ANOVA on Teachers' Perceptions by Teaching Experience

Table 6 presents the one-way ANOVA results.

Table 6: One-Way ANOVA Results for Differences in Teachers' Perceptions of AI-Generated Test Items by Teaching Experience (N = 120)

Source of Variation	SS	df	MS	F	p-value
Between Groups	14.37	3	4.79	3.42	.020
Within Groups	162.45	116	1.40		
Total	**176.82**	**119**			

Note. Significance level α = .05.

Copyright © 2026 KIJSOSE All Rights Reserved

The ANOVA results [F(3, 116) = 3.42, p = .020] indicate a statistically significant difference in teachers' perceptions across experience groups, leading to the rejection of Ho_1. The effect size, computed as eta-squared (η^2 = SS_between / SS_total = 14.37 / 176.82), yielded η^2 = .08, indicating a medium practical effect by Cohen's (1988) conventions (small = .01, medium = .06, large = .14). This suggests that teaching experience explains approximately 8% of the variance in teacher perceptions of ChatGPT-4o-generated test items—a practically meaningful, though not large, group difference. A post-hoc Scheffé test revealed that teachers with 0–5 years of experience (M = 3.82, SD = 0.71) perceived AI-generated items significantly more favourably than teachers with above 15 years of experience (M = 3.21, SD = 0.98). Cohen's f = .30, further corroborating a medium effect size.

Ho_2: Chi-Square Goodness-of-Fit Test for HOTS Reflection

Table 7 presents the chi-square test results.

Table 7: Chi-Square Goodness-of-Fit Test for Cognitive Level Distribution of AI-Generated Items (N = 200)

Cognitive Level	Observed Frequency (O)	Expected Frequency (E)	$(O - E)^2/E$
Remembering (C1)	48	33.33	6.44
Understanding (C2)	54	33.33	12.80
Applying (C3)	40	33.33	1.33
Analysing (C4)	34	33.33	0.01
Evaluating (C5)	16	33.33	9.01
Creating (C6)	8	33.33	19.26
Total	**200**	**200**	**χ^2 = 48.85**

Note. df = 5, Critical value = 11.07, p < .001.

The chi-square test results (χ^2 = 48.85, df = 5, p < .001) substantially exceed the critical value (11.07), indicating a statistically significant non-uniform distribution of ChatGPT-4o-generated items across cognitive levels. Ho_2 is therefore rejected. The data confirm that AI-generated items disproportionately cluster at lower-order cognitive levels.

Ho_3: Kruskal-Wallis H Test for Item Difficulty Distribution Across Sub-disciplines

Table 8 presents the Kruskal-Wallis H test results.

Table 8: Kruskal-Wallis H Test Results for Item Difficulty Distribution Across Social Science Sub-disciplines

Sub-discipline	n	Mean Rank	H	p-value
Economics	50	98.74		
Government	50	102.31	5.67 (df = 3)	.129
Geography	50	95.18		
Civic Education	50	107.77		

Note. Significance level α = .05.

The Kruskal-Wallis test (H = 5.67, df = 3, p = .129) yields a non-significant result, indicating no statistically significant difference in item difficulty distribution across the four sub-disciplines. Ho_3 is therefore not rejected.

Discussion

The finding that Social Science teachers in Kaduna State Metropolis held generally positive perceptions of ChatGPT-4o-generated test items (grand mean = 3.55) is consistent with a body of evidence documenting broadly favourable teacher attitudes toward AI-assisted

 Copyright © 2026 KIJSOSE All Rights Reserved

assessment tools in Nigerian secondary education. Fwangle et al. (2025) reported high positive perceptions of AI among Plateau State secondary school teachers, while Ofem et al. (2025) similarly noted endorsement of AI tools for academic support among Ondo State teachers, despite reservations about specific dimensions of quality. The present study extends these findings to the Social Science assessment context in Kaduna Metropolis, reinforcing the pattern of cautious optimism that characterises Nigerian teacher attitudes toward AI in education.

The highest-rated item—'AI-generated test items require revision before classroom deployment' (M = 4.12, SD = 0.73)—reflects a practically important insight corroborated by multiple international studies. Kaya et al. (2025) noted that while ChatGPT-4o-generated MCQs demonstrated acceptable psychometric properties overall, human expert review remained essential for ensuring accuracy, appropriateness, and contextual alignment. This position is further supported by Pinto et al. (2026), who found that AI-generated items showed gaps in discrimination, reinforcing the case for a human-AI collaborative model of item development rather than autonomous AI deployment. The below-threshold perception score for sociocultural contextualisation (M = 2.87) highlights a context-specificity gap that is particularly pronounced in the Nigerian setting. Social Science subjects in Nigerian secondary schools draw heavily on domestic policy contexts, national governance structures, local economic conditions, and Nigerian geographical features, all of which require culturally embedded content that general-purpose LLMs may not consistently produce (Okonkwo & Ade-Ibijola, 2021; Baig et al., 2024). It is important to note that this sociocultural gap finding is specific to ChatGPT-4o; whether other LLMs such as Gemini 1.5 or Claude 3 demonstrate superior cultural contextualisation in the Nigerian setting remains an open empirical question that future comparative studies should address.

The significant variation in perceptions across experience groups [$F(3, 116) = 3.42$, $p = .020$, $\eta^2 = .08$] represents a medium practical effect, indicating that teaching experience explains a meaningful proportion of variance in teacher perceptions. Newer teachers (0–5 years) held markedly more positive views (M = 3.82) than their more experienced counterparts (above 15 years; M = 3.21), consistent with observations by Okorie et al. (2025) that younger Nigerian teachers are more disposed toward digital and AI platforms. Conversely, experienced teachers may apply stricter evaluative standards grounded in years of assessment expertise, recognising psychometric and contextual limitations that less experienced counterparts may not yet perceive (Adesope et al., 2017; Moore et al., 2023). This divergence suggests that professional development programmes on AI-assisted assessment must be differentiated by experience level.

The finding that 71.0% of ChatGPT-4o-generated Social Science items were classified at LOTS levels, with only 29.0% reflecting HOTS, resonates powerfully with existing empirical literature. Jackson (2025) documented that LLMs consistently default to lower-order cognitive productions when prompts are not explicitly HOTS-targeted, arguing that the cognitive architecture of transformer-based models predisposes them toward pattern-matching and recall-based generation. Moore et al. (2023) observed similar patterns in GPT-3.5-generated examination items. The statistically confirmed HOTS underrepresentation ($\chi^2 = 48.85$, $p < .001$) is of particular concern for Social Science education in Kaduna State, where WAEC and NECO examination formats increasingly demand higher-order responses in Economics (analysis of market structures), Government (evaluation of democratic processes), Geography (application of spatial concepts), and Civic Education (critical appraisal of rights and responsibilities). Critically, it cannot be determined from the present single-LLM design

 Copyright © 2026 KIJSOSE All Rights Reserved

whether this HOTS underrepresentation pattern is specific to ChatGPT-4o or is an LLM-agnostic phenomenon characteristic of transformer-based language models in general. Comparative studies involving Gemini 1.5, Claude 3, and Copilot are strongly recommended to establish whether HOTS ceiling effects persist across AI platforms or can be mitigated through the use of alternative LLMs.

The overall mean discrimination index of D = .26 (SD = .09) indicates that ChatGPT-4o-generated items, as a pool, fall below the acceptable CTT threshold of D ≥ .30 for good item discrimination. Only 36.5% of items achieved acceptable discrimination (D ≥ .30), while 63.5% were classified as marginal or poor discriminators. This finding is consistent with Kaya et al. (2025), who reported that ChatGPT-4o-generated MCQs exhibited slightly lower discrimination values than clinician-designed items, and with Pinto et al. (2026), who noted discrimination gaps in AI-generated content. The practical implication is significant: items with poor discrimination do not effectively differentiate between high- and low-ability students, thereby compromising the internal validity of score interpretation. The notable exception is moderate-difficulty items (D = .33), which achieved acceptable discrimination—reinforcing the importance of targeting item generation prompts toward intermediate difficulty levels. School administrators and examination bodies should therefore apply both difficulty and discrimination screening criteria before deploying AI-generated items in high-stakes contexts.

The finding that 52.5% of ChatGPT-4o-generated items achieved moderate difficulty indices (.30 ≤ p ≤ .70) indicates broadly acceptable difficulty calibration by established CTT standards (Ebel & Frisbie, 1991; Hambleton & Swaminathan, 1985). The non-significant difference in item difficulty across sub-disciplines (H = 5.67, p = .129) indicates that ChatGPT-4o generates items of comparable difficulty regardless of sub-discipline—a positive indicator of content coverage breadth (Kurdi et al., 2020). However, the relatively high proportion of easy Geography items (34.0%) warrants subject-specific scrutiny, as Geography's reliance on descriptive content may dispose the AI model toward surface-knowledge items. Practitioners should note that difficulty and cognitive level are related but non-identical constructs: an item can be both difficult and lower-order, or moderately easy and higher-order. Thus, the acceptable difficulty profile of ChatGPT-4o-generated items must be read alongside—not independently of—their HOTS distribution and discrimination properties.

Limitations

Some limitations of this study should be acknowledged to assist readers in appropriately contextualising the generalisability of the findings. First, the study was conducted exclusively within Kaduna State Metropolis, Nigeria, comprising 30 public secondary schools across three administrative zones. The findings therefore reflect the specific socioeconomic, curricular, and institutional context of this metropolis and may not be directly transferable to rural Kaduna State, other Nigerian states, or other sub-Saharan African educational contexts where teacher demographics, school infrastructure, or curriculum emphasis may differ substantially.

Second, the study employed a single LLM—ChatGPT-4o—for item generation. While ChatGPT-4o is among the most widely used LLMs in academic and professional contexts globally, the exclusive focus on this platform limits the transferability of findings and makes it impossible to determine whether identified psychometric limitations (HOTS underrepresentation, marginal discrimination indices, and sociocultural contextualisation

 Copyright © 2026 KIJSOSE All Rights Reserved

gap) are specific to ChatGPT-4o or characteristic of LLMs in general. Future studies should adopt a comparative multi-LLM design, incorporating platforms such as Gemini 1.5, Claude 3, and Microsoft Copilot to test whether these patterns are LLM-agnostic or platform-specific.

Conclusion

This study has provided empirical evidence on four dimensions of the psychometric properties of ChatGPT-4o-generated test items in Social Science subjects in secondary schools in Kaduna State Metropolis, Nigeria. Teachers held generally positive, experience-modulated perceptions of AI-generated items (grand mean = 3.55; $\eta^2 = .08$), recognising their utility for time-efficient assessment while affirming the necessity of human revision for contextual and pedagogical adequacy. ChatGPT-4o-generated items significantly clustered at lower-order cognitive levels, with HOTS items constituting only 29.0% of the total pool—a deficiency of direct relevance to the cognitive demands of Social Science curricula and examination standards. Item difficulty distribution was predominantly moderate (52.5%) and did not differ significantly across Economics, Government, Geography, and Civic Education sub-disciplines, indicating acceptable difficulty calibration. However, the overall mean discrimination index (D = .26) fell below the acceptable CTT threshold, with only 36.5% of items achieving acceptable discriminatory power—a finding that signals a need for deliberate post-generation item review incorporating discrimination screening.

These findings collectively position ChatGPT-4o-generated test items as a promising complementary tool for Social Science assessment in Nigerian secondary schools, provided they are deployed within a structured framework of HOTS-targeted prompt engineering, difficulty and discrimination-based item screening, subject expert review, and ongoing psychometric validation. The sociocultural contextualisation gap, the HOTS ceiling effect, and the marginal discrimination profile represent the three most critical psychometric vulnerabilities that must be addressed for AI-assisted assessment to serve the full breadth of educational objectives in the Nigerian Social Science context. It must be explicitly acknowledged that these findings pertain exclusively to ChatGPT-4o within the Kaduna State Metropolis educational context; their generalisability to other LLMs or geographic settings should not be assumed without further comparative empirical investigation. This study contributes a contextualised, multi-dimensional empirical baseline to the growing literature on AI in educational assessment in Nigeria and provides actionable evidence for teachers, administrators, curriculum developers, and examination bodies engaged in the integration of AI tools into secondary school assessment practice.

Recommendations

Based on the findings, the following recommendations were made:

1. The Kaduna State Ministry of Education and the Federal Ministry of Education should incorporate AI assessment literacy into both pre-service teacher education curricula and in-service professional development programmes, with particular emphasis on equipping Social Science teachers with skills for critically evaluating, contextualising, and revising ChatGPT-4o-generated test items before classroom use.
2. Social Science teachers should adopt structured HOTS-aligned prompt engineering strategies when utilising AI tools for item generation, explicitly incorporating Bloom's Revised Taxonomy action verbs (e.g., analyse, evaluate, construct, justify) within generation prompts to systematically increase the proportion of higher-order items—particularly at the Analysing (C4), Evaluating (C5), and Creating (C6) levels.
3. School administrators and external examination authorities, including WAEC, NECO, and NABTEB, should develop and disseminate AI item review protocols that integrate both

 Copyright © 2026 KIJSOSE All Rights Reserved

content validity panels (for curriculum alignment and sociocultural contextualisation) and quantitative psychometric screening for both difficulty and discrimination quality control, before AI-generated items enter formal examination systems.

4. Curriculum developers at the Nigerian Educational Research and Development Council (NERDC) should provide AI-compatible item-writing guides that explicitly address the need for contextual relevance in Social Science assessment, with worked examples demonstrating how to prompt AI tools to generate items that reflect Nigerian economic, governmental, geographic, and civic realities.

References

Abiodun, T. E., & Shehu, A. S. (2022). Assessment practices in Nigerian secondary schools: Challenges and prospects for the 21st century. *Journal of Educational Assessment in Africa*, *7*(2), 45–61. https://doi.org/10.21303/jea.2022.002341

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT with IQA and AHA frameworks. *Social Sciences & Humanities Open*, *7*(1), 100375. https://doi.org/10.1016/j.ssaho.2023.100375

Afolaranmi, A. O., Ogunyemi, A. A., & Adeleke, J. O. (2021). Assessment of higher-order thinking skills in Social Science education in Nigerian secondary schools. *African Journal of Educational Studies in Mathematics and Sciences*, *17*(2), 33–47.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2024). ChatGPT in higher education: A systematic literature review and future research directions. *Computers in Human Behavior Reports*, *13*, 100398. https://doi.org/10.1016/j.chbr.2024.100398

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Thomson/Wadsworth.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Fwangle, I. I., Wuye, A., & Mancha, P. (2025). Assessment of teachers' perception of artificial intelligence on students' academic achievement in senior secondary schools

 Copyright © 2026 KIJSOSE All Rights Reserved

in Plateau State, Nigeria. *Journal of Faculty of Education, Federal University Gashua*, *4*(1), 22–35. https://js.foefugusau.com.ng/index.php/foefujs/article/view/66

Gorgun, G., & Bulut, O. (2023). A polytomous scoring approach to automated short answer scoring with pre-trained language models. *Educational and Psychological Measurement*, *83*(4), 799–833. https://doi.org/10.1177/00131644221120226

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff Publishing. https://doi.org/10.1007/978-94-017-1988-9

Ijewere, A. A., & Uwameiye, R. (2018). Instructional strategies and cognitive demands of Social Studies teaching in Nigerian secondary schools: Challenges for 21st-century learning. *Journal of Social Science Education*, *17*(1), 12–25.

Jackson, J. (2025). Higher order prompting: Applying Bloom's revised taxonomy to generative AI in higher education. *Studies in Technology Enhanced Learning*, *4*(1). https://doi.org/10.21428/8c225f6e.b0f83ecf

Jurado-Castro, J. M., Navarrete, R. N., & Ramírez-Benítez, Y. (2024). Score regeneration through post-item analysis with artificial intelligence: Applications in educational assessment. *International Journal of Emerging Technologies in Learning*, *14*(10), 1–14. https://doi.org/10.3991/ijet.v14i10.2172

Kaduna State Secondary Education Board. (2024). *Annual statistical digest on teacher distribution in public secondary schools in Kaduna State Metropolis*. KSSEB Directorate of Planning, Research and Statistics.

Kaya, M., Demir, S., Yilmaz, A., Coskun, O., & Kaya, C. (2025). Comparison of AI-generated and clinician-designed multiple-choice questions: A psychometric analysis in emergency medicine education. *BMC Medical Education*, *25*(1), 412. https://doi.org/10.1186/s12909-025-06412-8

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Lahza, H., Shehata, M. A., Ramadan, A. M., & Alkhaldi, A. (2023). Beyond item analysis: Connecting student behaviour and learning outcomes with assessment quality metrics in STEM education. *British Journal of Educational Technology*, *54*(5), 1212–1231. https://doi.org/10.1111/bjet.13270

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *International Journal of Educational and Psychological Assessment*, *1*(1), 1–11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

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 N. Wang et al. (Eds.), *Proceedings of the 24th International Conference on Artificial Intelligence in Education (AIED 2023)* (Vol. 13916, pp. 229–240). Springer. https://doi.org/10.1007/978-3-031-36272-9_19

 Copyright © 2026 KIJSOSE All Rights Reserved

Ofem, U. I., Adie, R. I., & Nwana, S. C. (2025). Teachers' and students' perceptions of artificial intelligence tools for academic support in secondary schools in Ondo State, Nigeria. *Federal University Oye-Ekiti Journal of Education*, *5*(1), 74–89. https://fjed.fuoye.edu.ng/index.php/public_html/article/view/149

Okorie, P. N., Ogbu, E. O., & Aneke, I. G. (2025). Teachers' application of artificial intelligence tools and students' academic performance in secondary schools in Anambra State. *Global Platform for Health, Science and General Studies*, *3*(1), 1–18. https://doi.org/10.59325/gphsgs.2025.2165

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Python-QUESTIONS: A tool for automatic generation of MCQs in Python programming. *Education and Information Technologies*, *26*(4), 5323–5345. https://doi.org/10.1007/s10639-021-10443-4

Pinto, A., Carvalho, B., Pinheiro, R., & Ferreira, T. (2026). Accuracy of AI-generated multiple-choice questions: Psychometric evaluation across disciplines. *International Journal of Pharmaceutical Research and Technology*, *17*(1), 45–54. https://doi.org/10.37290/ijprt.2026.1382

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Rachmawati, T., Setiawan, A., & Wibowo, A. T. (2024). Development of higher-order thinking skill assessment instruments in social studies learning. *International Journal of Evaluation and Research in Education*, *13*(2), 1214–1222. https://doi.org/10.11591/ijere.v13i2.26448

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410601605

 Copyright © 2026 KIJSOSE All Rights Reserved