Education Science and Management
https://www.acadlore.com/journals/ESM

Psychometric Evaluation of Human-Crafted and AI-Generated Multiple-Choice Questions for Mathematics Instruction

Rachid El Chaal[1*], Rabei Ben Seghir[2], Moulay Othman Aboutafail[1]

[1] Engineering Sciences Laboratory, Data Analysis, Mathematical Modeling, and Optimization Team, Ibn Tofail University National School of Applied Sciences ENSA, 14000 Kenitra, Morocco
[2] Laboratoire Linguistique Générale et Didactique du FLE (LGDFLE), Université Mohammed Premier, 60000 Oujda, Morocco

[*] Correspondence: Rachid El Chaal (rachid.elchaal@uit.ac.ma)

Received: 02-14-2025 **Revised:** 04-01-2025 **Accepted:** 04-07-2025

Citation: Seghir, R. B., El Chaal, R., & Aboutafail, M. O. (2025). Psychometric evaluation of human-crafted and AI-generated multiple-choice questions for mathematics instruction. *Educ Sci. Manag., 3*(2), 78-92. https://doi.org/10.56578/esm030201.

Abstract: The psychometric validity of multiple-choice questions (MCQs) generated by an advanced Artificial Intelligence (AI) language model (ChatGPT) was evaluated in comparison with those developed by experienced human instructors, with a focus on mathematics teacher education. Two parallel 30-item MCQ tests—one human-designed and one AI-generated—were administered to 30 mathematics teacher trainees. A comprehensive psychometric analysis was conducted using six metrics: item difficulty index (P_i), discrimination index (D), point-biserial correlation, item-test correlation (R_{it}), Cronbach's alpha (α) for internal consistency, and score variance. The analysis was facilitated by the Analysis of Didactic Items with Excel (AnDIE) tool. Results indicated that the human-authored MCQs exhibited acceptable difficulty (mean $P_i = 0.55$), moderate discrimination power (mean $D = 0.31$), and strong internal consistency (Cronbach's $\alpha = 0.752$). In contrast, the AI-generated MCQs were found to be substantially more difficult (mean $P_i = 0.22$), demonstrated weak discrimination (mean $D = 0.16$), and yielded negative internal consistency reliability (Cronbach's $\alpha = -0.1$), raising concerns about their psychometric quality. While AI-generated assessments offer advantages in terms of scalability and speed, the findings underscore the necessity of expert human review to ensure content validity, construct alignment, and pedagogical appropriateness. These results suggest that AI, in its current form, is not yet equipped to autonomously generate assessment instruments of sufficient quality for high-stakes educational settings. A hybrid test design model is therefore advocated, wherein AI is leveraged for initial item drafting, followed by rigorous human refinement. This approach may enhance both efficiency and quality in the development of educational assessments. The implications extend to educators, assessment designers, and developers of educational AI systems, highlighting the need for collaborative human-AI frameworks to achieve reliable, valid, and pedagogically sound testing instruments.

Keywords: Educational assessment; Psychometric analysis; Hybrid test design; Difficulty and discrimination index; Cronbach's alpha

1. Introduction

The shift of AI as an integrated part of educational practices disrupts the understanding and execution of teaching and assessment. AI-generated pedagogical tools, especially MCQs (Nasution, 2023), are gaining ground among numerous options because they make evaluation efficient and effective. It is no wonder that MCQs have remained a mainstay of higher education and professional training, providing a scalable and efficient approach to automatically grade exams (Naseer et al., 2024). The standardized nature of such exams enables the assessment of large cohorts simultaneously with minimal lag time, making them quintessential in contemporary educational structures (Yang, 2025). Nonetheless, the effectiveness of MCQs depends on basic psychometric principles. An MCQ has a robust potential for achievement as an assessment tool only if it satisfies the best practice standards of validity, which define the extent to which a test measures what it is intended to measure and the reliability (Cook & Beckman, 2006; Feldt, 1997) or consistency of test results across different populations and settings. These

attributes ensure that assessments are pedagogically appropriate and do not inadvertently disadvantage learners (Lu et al., 2024).

With recent improvements in natural language processing like OpenAI's ChatGPT (Naseer et al., 2024), tests can be created differently than in the past. With the proper prompts, these AI models can make thousands of MCQs in seconds! This is excellent news for the advancements in educational technology, but it also opens Pandora's box of questions about the quality of the content created by AI (Ali et al., 2024). A central question is whether these MCQs come close to the nuanced, contextually rich standards of human expert-sourced MCQs, whose test design generally relies on human experts' rich, high-context knowledge in pedagogy, the discipline, psychology, and ethics (Whitehill & LoCasale-Crouch, 2023). A need for systematic and empirical research on AI tool effectiveness in educational assessment is implied by generalized success and failure messages concerning the widespread adoption of this technology (Milić et al., 2024). This research is premised on the following questions: What is AI-generated MCQs' psychometric quality and educational value, and how do they compare against those crafted by human educators? This raises the question: What are the strengths and weaknesses of other tools when it comes to the metrics based on the Classical Test Theory (CTT), such as difficulty indices, discrimination, and reliability coefficients? In what ways might educators and researchers harness and exploit the positive aspects of AI advancements in test generation while minimizing their negative aspects to guarantee that sound and academic testing methods are applied? In alignment with these questions, the precise goals of this research are threefold:

a) Comparative psychometric evaluation: To compare and contrast main metrics, such as difficulty, discrimination, reliability, and internal consistency, between human instructors authored and AI-generated MCQs (Lee & Kim, 2024).

b) Experimental context: To assess the degree to which AI-generated assessments can supplement traditional, human-designed pedagogical tools and when they do not.

c) Recommendations for joint optimization: To suggest strategies that combine human experts' excellence with AI's efficiency, a composite framework of quality evaluation.

Addressing these aims, the study seeks to add to the literature on AI in education as a first step towards an evidence-based AI application for identified educational purposes. Although they are the most popular assessment methods, MCQs have some fundamental drawbacks that require mindful and intentional design. Poorly constructed test items can generate several problems, including excessive guessing, a lack of differentiation between students of different abilities, and inconsistencies that threaten the validity of test scores (Feldt, 1997). These issues highlight the relevance of psychometric analysis (Hussein & Gasmalla, 2023), which applies statistical methods to assess and improve the quality of test items (Hayashi & Eguchi, 2024).

It usually consists of multiple metrics that guarantee tests do what they should (psychometric evaluation). The item difficulty index (P_i) measures the accessibility of an item to learners and the depth of knowledge from the proportion of correct responses (Sharma, 2021). The discrimination index (D) measures how well a question separates high- and low-scoring test-takers (Zakareya et al., 2024). Reliability coefficients like Cronbach's α give an idea of overall consistency and how well the test can be expected to provide similar scores on different test administrations (Wati et al., 2024). Psychometric metrics are intended for sound assessment. For instance, the difficulty index (P_i) assesses an item's ease of use, and the discrimination index (D) assesses its potential to discriminate high-score testers from low-score ones (Lu et al., 2024). These metrics guarantee fairness and adhere to pedagogical outcomes (Hayashi & Eguchi, 2024). These verification processes use dedicated software to calculate these values. An application called AnDIE (Narayanan et al., 2017) was used in this study, which is widely usable and user-friendly and automatically performs psychometric analysis of educational items (Tran et al., 2024). Evaluating MCQs through this tool provides an elaborate analysis of their comparative quality concerning two different approaches to test creation and AI-generated content (Wati et al., 2024).

This study focuses on a specific group of participants (teacher trainees in mathematics education). It thus addresses the increased demand for devising valid assessment measures in language teaching and other education areas. This research, based on an analysis of 30 items from each test type, provides an empirical foundation for identifying relative strengths and weaknesses of AI in test design and a sustainable way forward. This research calls for future assessments to be collaborative between human creativity and AI efficiency and strive to help shape the future of educational assessments.

While prior studies have explored AI-generated assessments, this study addresses a critical gap by systematically comparing the psychometric properties of human-crafted and ChatGPT-generated MCQs using six key metrics (difficulty, discrimination, reliability, etc.) analyzed via the AnDIE tool. Unlike existing work focused on medical education or vocabulary tests, this research specifically targets mathematics teacher trainees, offering novel insights into AI's limitations in pedagogical contexts requiring higher-order thinking. This is the first study to report negative Cronbach's α (-0.1) for AI-generated tests, highlighting unprecedented reliability concerns.

2. Methodology

2.1 Study Design and Sample

This study was performed with 30 participants, who are teacher trainees following their education at a training institute for mathematics teachers. The participants were convenience-sampled from a mathematics teacher training institute, representing typical end-users of such assessments. While gender and age were balanced, the sample's homogeneity (single institution and subject area) limits generalizability to broader populations. Future work should include diverse disciplines and institutions.

The sample size (N = 30) was determined based on pilot study conventions for psychometric evaluations, balancing feasibility with statistical power for item-level analysis. The participants completed both tests in a controlled classroom setting, with 45 minutes allocated per test. Instructions standardized the test-taking conditions (e.g., no external resources and timed completion). While this sample suffices for initial comparisons, larger cohorts are needed to generalize findings across disciplines. Demographic parity (gender and age) was ensured to minimize confounding effects. The assessment tools include the following two tests:

Test 1 (human-curated): Crafted by an experienced teacher around central topics of the research methodology class. Each question was designed for recall, application, and higher-order thinking access.

Test 2 (AI-generated): ChatGPT with specific commands was used to develop a more aligned and professional style with a balanced difficulty range and identify themes directly from the syllabus.

2.2 Measurement Indices

CTT was chosen over the Item Response Theory (IRT) due to its simplicity and suitability for small samples. While IRT requires larger datasets (N > 200) to estimate latent traits accurately, CTT's aggregate-level analysis (e.g., mean P_i) aligns with the exploratory goals of this study. Descriptive statistics, such as mean and Standard Deviation (SD), complement CTT by contextualizing score distributions. The psychometric parameters were examined at item and test levels; the CTT principles were followed (Hambleton & Jones, 1993; Ohmoto, 2024).

2.2.1 Item-level metrics

Difficulty index (P_i) is a statistical index used to determine the hypothetical difficulty of a single test question or assessment item. It is commonly used in psychometrics, education, and test performance analysis (Sharma, 2021; Taib & Yusoff, 2014). The formula for the difficulty index (P_i) is as follows:

$$P_i = \frac{\text{Number of test-takers who answered correctly}}{\text{Total Number of test-takers who attempted the question}}$$

Difficulty index (P_i) explains the percentage of correct answers. Values between 0.30 and 0.70 (inclusive) indicate optimal difficulty for even-balanced assessments that allow separation. Low values (< 0.30) indicate significant difficulty, whereas high values (> 0.70) represent ease.

The discrimination index (D) is a statistic used in test and measurement analysis to measure how well a question classifies high- versus low-scoring test-takers—a vital ingredient in psychometrics, educational assessment, and AI-generated test evaluations (Hayashi & Eguchi, 2024; Sharma, 2021; Taib & Yusoff, 2014; Zakareya et al., 2024). The formula for the discrimination index (D) is as follows:

$$D = \frac{U - L}{N} \tag{1}$$

where, U denotes the students in the top group for whom the question was answered correctly, L represents the number of students in the bottom group who answered the question correctly, and N denotes the total number of students in the top and bottom groups (i.e., $+L$).

Interpretations for the D values are as follows:

- $D \geq 0.40$: Excellent discrimination
- $0.30 \leq D < 0.40$: Good discrimination
- $0.20 \leq D < 0.30$: Acceptable discrimination
- $0.00 \leq D < 0.20$: Weak discrimination
- $D < 0.00$: Question is flawed (weaker students performed better)

Discrimination index (D) describes an item's ability to distinguish between high and low scorers (i.e., proficient vs. less proficient learners). The scores can be between -1.00 and +1.00. Typical cut-offs can be applied, with items having D values greater than or equal to 0.30 considered "good discriminators." Items with D values less than 0.10 are often discarded.

Thresholds for interpretation follow established benchmarks: P_i = 0.30–0.70 ensures balanced difficulty, avoiding floor/ceiling effects; $D \geq 0.30$ indicates effective discrimination, as lower values (< 0.20) fail to distinguish proficiency levels; and Cronbach's $\alpha \geq 0.70$ is considered acceptable for classroom tests. These standards are widely adopted in educational measurement literature.

Item-test correlation (R_{it}) is a statistical measure used in test and assessment analysis to see how well a specific test item (question) correlates with the overall test score. It is commonly used in psychometrics, educational testing, and AI-based evaluation of test exams. It depicts the strength and direction of the relationship of a single question with that of the whole test. A high R_{it} indicates students who performed well on the test also did well on this item (a good question). A low R_{it} indicates that the item does not contribute significantly to discriminating performance. A negative rating means the students who did well on this test were wrong about this item, suggesting it was defective or misleading. The formula for the item-test correlation (R_{it}) is as follows:

$$R_{it} = \frac{\sum(X_i - \overleftarrow{X}_i)\,(T - \overleftarrow{T})}{\sqrt{\sum(X_i - \overleftarrow{X}_i)^2} \cdot \sqrt{\sum(T - \overleftarrow{T})^2}} \tag{2}$$

where, X_i is the score on a specific item (0 for incorrect and 1 for correct in dichotomous scoring), $\overleftarrow{X}_i$ is the mean score for that item, T is the total test score of each student, $\overleftarrow{T}$ is the mean total test score, and $\sum$ is the summation across all test-takers.

Item-test correlation (R_{it}) assesses how closely each body aligns with the overall exam performance. Questionable items have poor correlation or alignment (R_{it} < 0.20). Item-test correlation (R_{it}) is basically the Pearson correlation coefficient between the score of a specific single item and the total test score.

Interpretations for the R_{it} values are as follows:

- ≥ 0.40: Excellent discrimination
- 0.30 – 0.39: Good discrimination
- 0.20 – 0.29: Acceptable discrimination
- 0.10 – 0.19: Weak discrimination
- < 0.10: Poor or no discrimination (should be revised or removed)
- Negative Rit: Problematic item (may be misleading)

2.2.2 Cronbach's alpha: A key metric for assessing test reliability and internal consistency

Cronbach's α is a statistical measure used to determine the reliability of a test or questionnaire. It is commonly used in psychometrics, educational assessments, and AI-generated test evaluations to evaluate whether a set of test items reliably measures a single unidimensional latent trait. Cronbach's α evaluates the degree to which the same test items tend to score together (i.e., reflect the same underlying construct consistently). On a scale of 0 to 1, the higher the value, the better the reliability (Nasution, 2023; Partchev, 2020; Wati et al., 2024). The formula for Cronbach's α is as follows:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_T^2}\right) \tag{3}$$

where, k is the number of items in the test, σ_i^2 is the variance of each item, and σ_T^2 is the variance of the total test score. A simplified explanation is that Cronbach's α is the proportion of variance in test scores due to "true" score variance (consistency) rather than random error.

Interpretations for Cronbach's α are as follows:

- $\alpha \geq 0.90$: Excellent internal consistency
- $0.80 \leq \alpha < 0.90$: Good reliability
- $0.70 \leq \alpha < 0.80$: Acceptable reliability
- $0.60 \leq \alpha < 0.70$: Questionable reliability
- $\alpha < 0.60$: Poor reliability (test needs improvement)

Cronbach's α measures the internal consistency. Educators generally accept tests with $\alpha \geq 0.70$, but scores approaching or above 0.80 are preferable.

2.2.3 Descriptive statistical parameters

Scores regarding average (mean score) and variability (SD) and total variance scores were analyzed. These metrics, in turn, help in understanding the difficulty, stability, and uniformity associated with test questions (Hambleton & Jones, 1993; Hussein & Gasmalla, 2023).

The formula for the mean $\bar{X}$ is as follows:

$$\bar{X} = \frac{\sum X_i}{N} \tag{4}$$

where, $\bar{X}$ is the mean (average) score, $\sum X_i$ is the sum of all test scores, and N is the number of test-takers. A high mean indicates that the test is primarily simple. A low mean indicates that the test is about challenging you in general. A mean close to 50% of the total obtainable score suggests a relative balance in question difficulty. The formula for SD is as follows:

$$\sigma = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N}} \tag{5}$$

where, σ is the SD, X_i is each test score, $\bar{X}$ is the mean score, and N is the number of test-takers. A low SD shows that scores fell within a tight range, indicating that the test may have been of similar difficulty for all students. A high SD shows a wide performance spread, indicating that some questions were too hard or too easy for students.

Variance (σ^2) indicates the general spread of scores. This is just the variance, which is simply the SD squared. The formula for variance (σ^2) is as follows:

$$\sigma^2 = \frac{\sum(X_i - \bar{X})^2}{N} \tag{6}$$

A high variance shows a wide distribution of scores, indicating that the test may be too hard/easy in general. A low variance shows that scores are near the average, indicating that the test is more reliable.

2.2.4 Analytical procedure

All participants performed the same battery of tests administered using a standardized protocol. The test results were documented, and an analysis was performed using the AnDIE software for those test scores. The data underwent initial processing with AnDIE, which produced visual summaries, histograms and scatterplots, to identify trends and outliers that may influence each other. AnDIE was selected for its accessibility and automated psychometric calculations. The tool computes P_i, D, and R_{it} via Excel macros, reducing manual errors. Its outputs include histograms and scatterplots for visual validation. While AnDIE assumes dichotomous scoring (right/wrong), this aligns with the MCQ design of this study. Limitations include the inability to handle polytomous items or IRT parameters.

Along with AnDIE, Python was used for further data analysis and plotting. To develop extra statistical summaries and in-house designated illustrations, Python's information examination libraries were utilized for information control (Pandas), numerical tasks (NumPy), and visualization (Matplotlib). Together, these tools allowed for a thorough examination of the data to really identify trends and linkages.

2.3 Theoretical Framework

CTT was adopted in this study to evaluate item and test quality, as it provides robust, interpretable metrics for small-scale educational assessments. CTT's assumptions, unidimensionality and linear relationships between items align with the research objectives, as this study focuses on aggregate test performance rather than individual skill estimation (which would require IRT). The selected metrics, difficulty index (P_i), discrimination index (D), and Cronbach's α, are CTT standards validated for MCQ analysis. P_i assesses accessibility, D measures differentiation between high/low performers, and Cronbach's α quantifies internal consistency, collectively addressing the research questions about validity and reliability.

3. Results

3.1 Statistical Overview

The results are the psychometric performance of the human-designed test and an AI-generated test. Below are summaries of key statistical parameters at the item and test levels, as shown in Figure 1.

The graph displays participants' human test (blue) and AI test (green) scores in order. The human test always scores higher, with fewer bumps in the road, indicating better-structured questions. The AI test scores are lower and more dispersed, implying that there are questions that are easy to perform and test the limit of AI.

As shown in Figure 2, the scores obtained by participants for both tests are plotted in this histogram. This suggests that the test is very well-balanced in terms of difficulty, whereas the human test has a higher scoring skew. The distribution is more widespread, showing some variability in the difficulty of the test items in the AI test. The

comparison of both tests is as follows:

Test 1 (human-crafted)

- Mean score: 15.2/30
- SD: 4.8 points
- Variance: 23.04
- Distribution: The symmetric curve indicates the well-placed item challenge. The Coefficient of Variation (CV) is 31.6%, which means a moderate dispersion of scores, thus allowing a reliable discriminator between test-takers.

Test 2 (AI-generated)

- Mean score: 8.7/30
- SD: 6.2 points
- Variance: 38.44
- Distribution: The steeply negative curve is associated with lower scores, indicating multiple difficult items. CV is 71.3%, which indicates very poor homogeneity, and thus, items are unable to furnish the same information on participants.

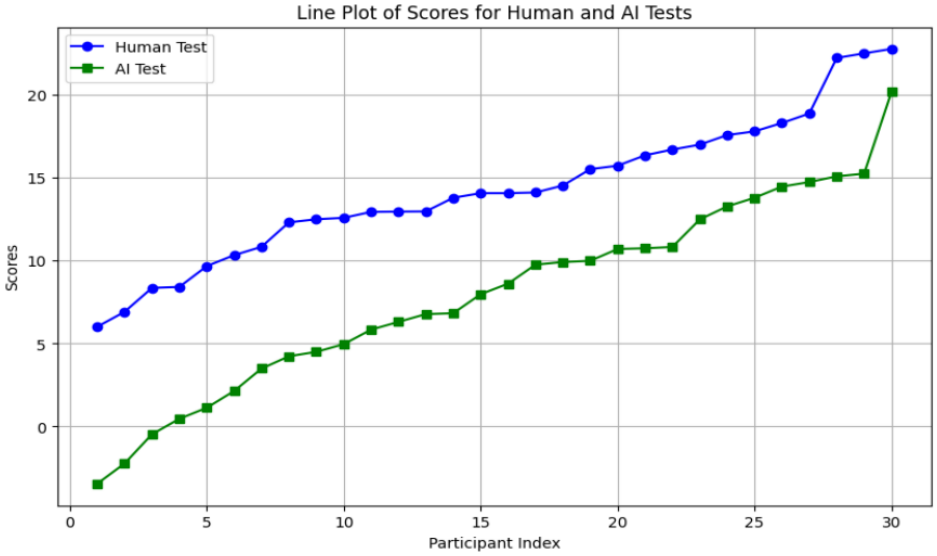

Figure 1. Distribution of student scores in human vs. AI tests

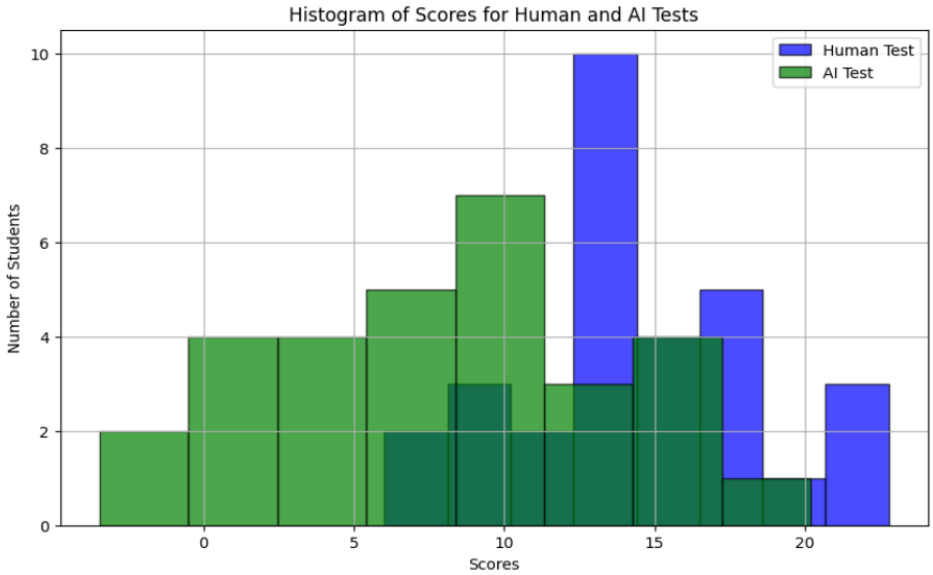

Figure 2. Frequency distribution of scores on both tests

Differences in statistical variability, score distributions, and central tendencies are so striking that they indicate core psychometric weaknesses of an AI-generated test and identify serious problems with its use in an educational

setting. The average score of 8.7 out of 10 on the AI test suggests it was too hard, with most entrants failing to correct a single question. A high variance (38.44) indicates that the proposed model varied questions significantly more straightforward or complicated than the human test (variance = 23.04), which is a problem for a consistent difficulty level.

3.2 Item-Level Psychometric Analysis

Comparing item-level metrics highlights some significant differences between the human- and AI-generated tests.

3.2.1 Difficulty index (P_i)

The difficulty index, calculated to determine the proportion of correct responses per item, revealed significant contrasts in item calibration.

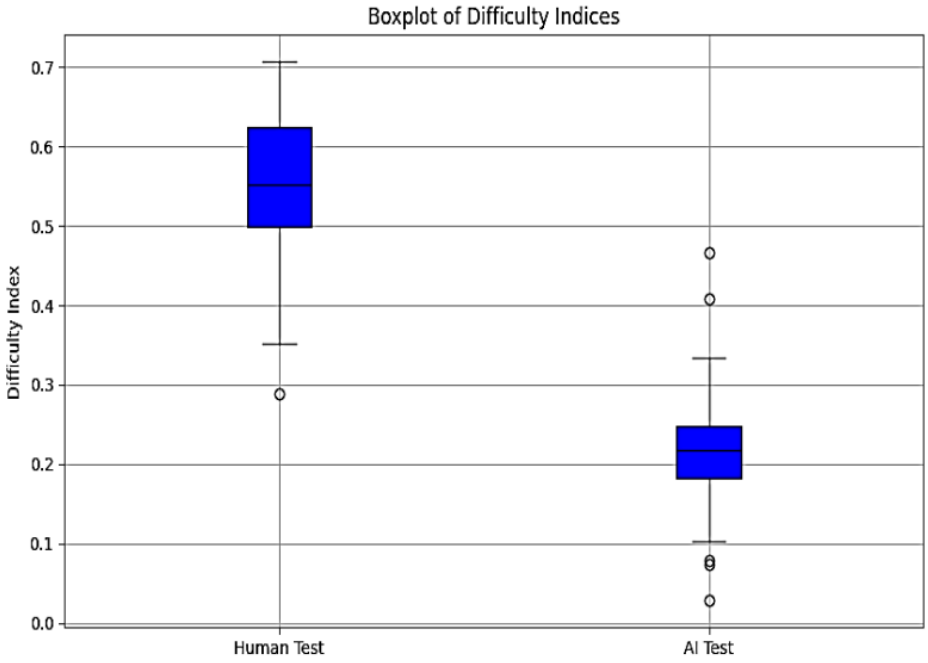

Figure 3. Comparing levels of difficulty between questions

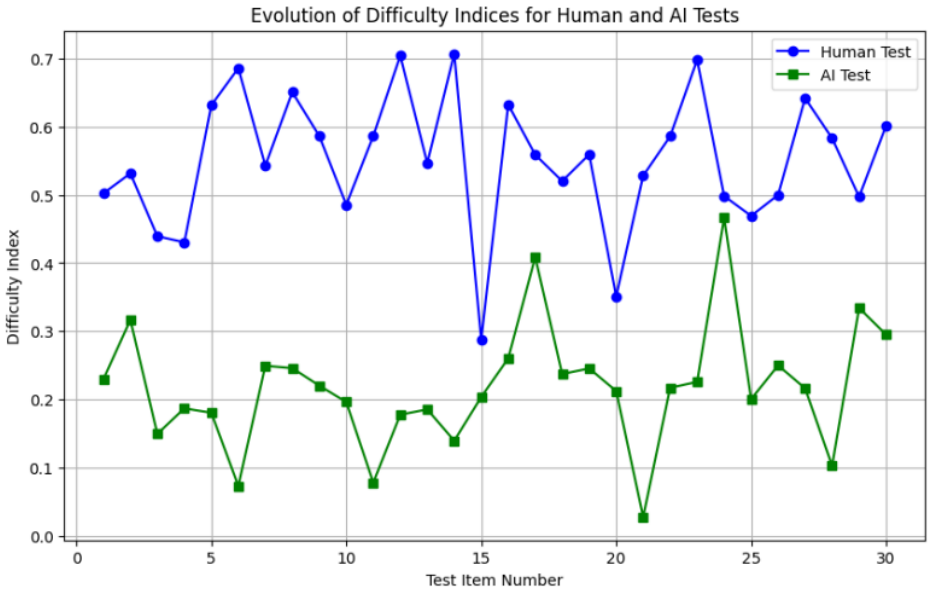

Figure 4. Fluctuation of difficulty in different test items

As for Test 1 (human-crafted), the mean P_i is 0.55, which is in the ideal range for balance assessment (0.30–0.70). In terms of item distribution, 43.3% of items belong to the good range (0.40–0.60); 13.3% of items belong to the easy range (> 0.70), meaning that only a few items are too easy; and 6.7% of items belong to the extremely easy range (< 0.30), meaning no friction when it comes to reading the content. As for Test 2 (AI-generated), the mean P_i is 0.22, indicating extreme item difficulty. In terms of item distribution, an alarming 83.3% of items were classified as too harsh (< 0.30). Only a few items were placed in the optimal or easy range, making the test less reliable and frustrating for participants. The results indicate that even though the human-written test maintained an acceptable level of item difficulty, the AI-prompted test was too difficult or too easy across the examined dimensions. A sizeable portion (83.3%) of the AI test items were dominantly (P_i < 0.30) labeled in the more complex category, likely frustrating to participants and resulting in disengagement. In comparison, the items in the human test were distributed more equitably and fairly, as nearly half (43.3%) had their optimal range (i.e., P_i between 0.40 and 0.60).

As shown in Figure 3 and Figure 4, a boxplot between the human and AI tests on the difficulty index of the questions can be seen. The human test is balanced, as most questions fall in the golden range. The difficulty indices of all items in the AI test are lower, suggesting that the items are difficult for many.

3.2.2 Discrimination index (D)

The discrimination index is a measure of an item that still can distinguish between high- and low-performing participants.

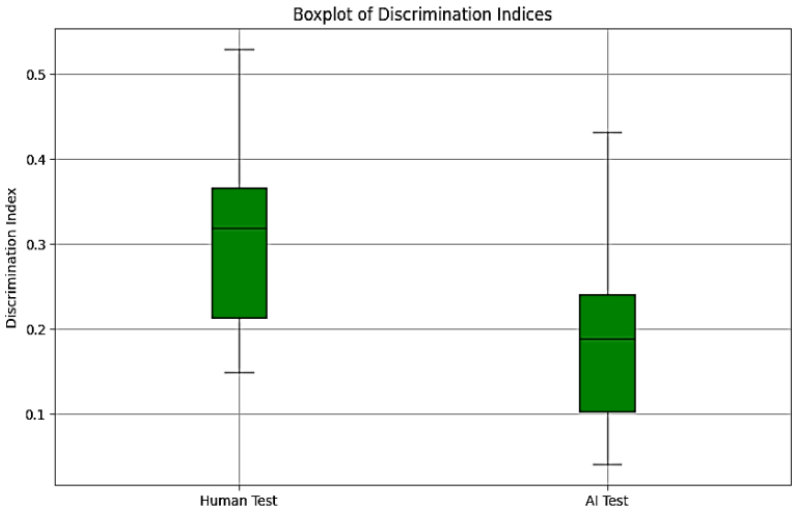

Figure 5. Performance of tests differentiating between students

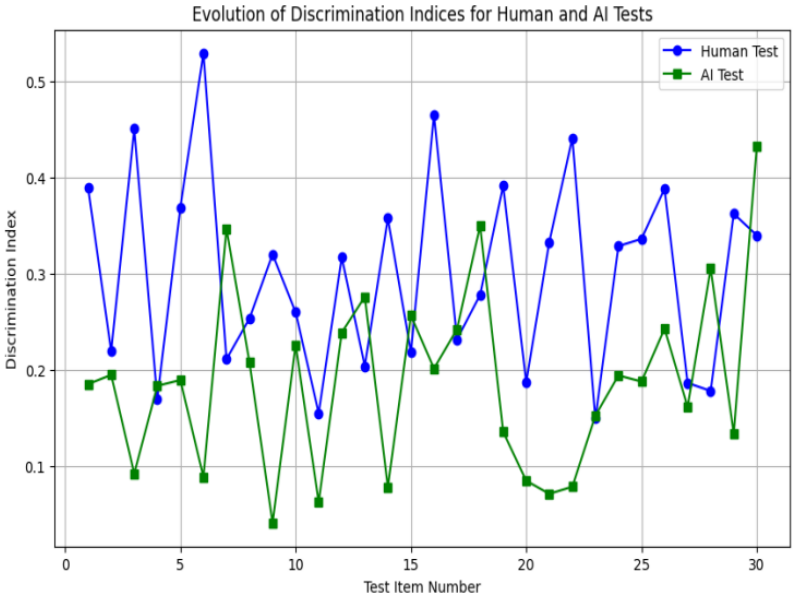

Figure 6. Tracking of questions separating high and low performers

As for Test 1 (human-crafted), the mean *D* is 0.31, aligning with the threshold for effective discrimination. In terms of item classification, 50% of the items fell into the category of *D* values ≥ 0.30, thus being classified as "well-discriminating" items. Only a few items showed very little discrimination power, which is desirable in psychometric terms and indicates close adherence to the principles of psychometrics. As for Test 2 (AI-generated), the mean *D* is 0.16, underscoring significant deficiencies in item quality. In terms of item classification, 76.7% of items were classified as problematic, i.e., having a discrimination index < 0.20. Therefore, these items did not discriminate adequately between participants at different skill levels. AI items' much lower discrimination ability indicates that the generative models are inadequate in generating diverse, nuanced, reliable, and valid measurements without further human expert refinement. A mean *D* of 0.31 on the human test indicates its power to separate high-performing from low-performing participants, a key aspect of finding where students lack understanding. The mean *D* of 0.16 on the AI test means that many of the questions on the test did not distinguish well enough between high- and low-ability students to be clinically useful, as shown in Figure 5.

Figure 6 measures how well the questions differentiate between high- and low-performing participants. The human test has higher discrimination values, showing it is better at assessing different skill levels. The AI test struggles to separate participant abilities, making it less effective for evaluation.

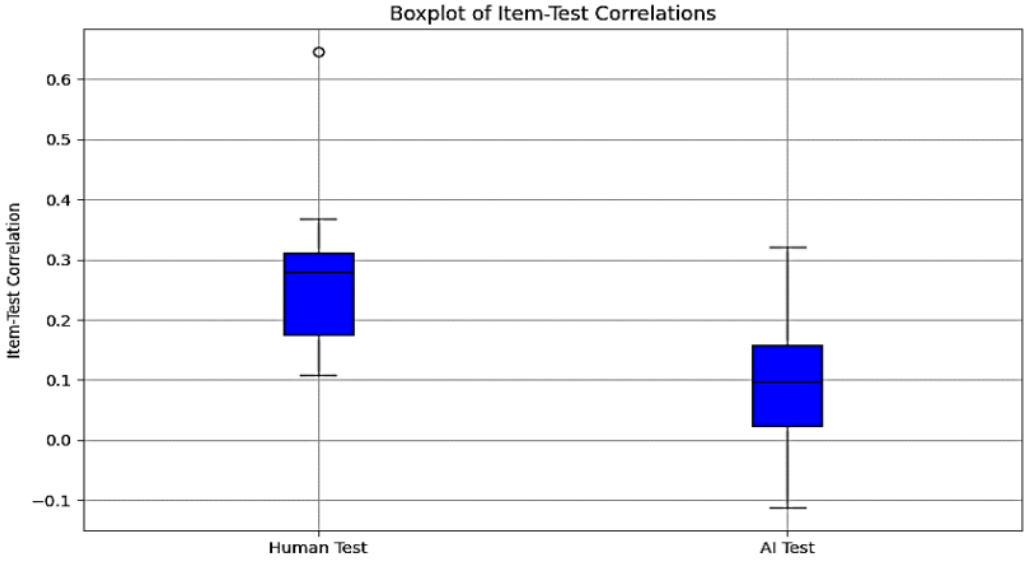

Figure 7. Consistency and quality of questions

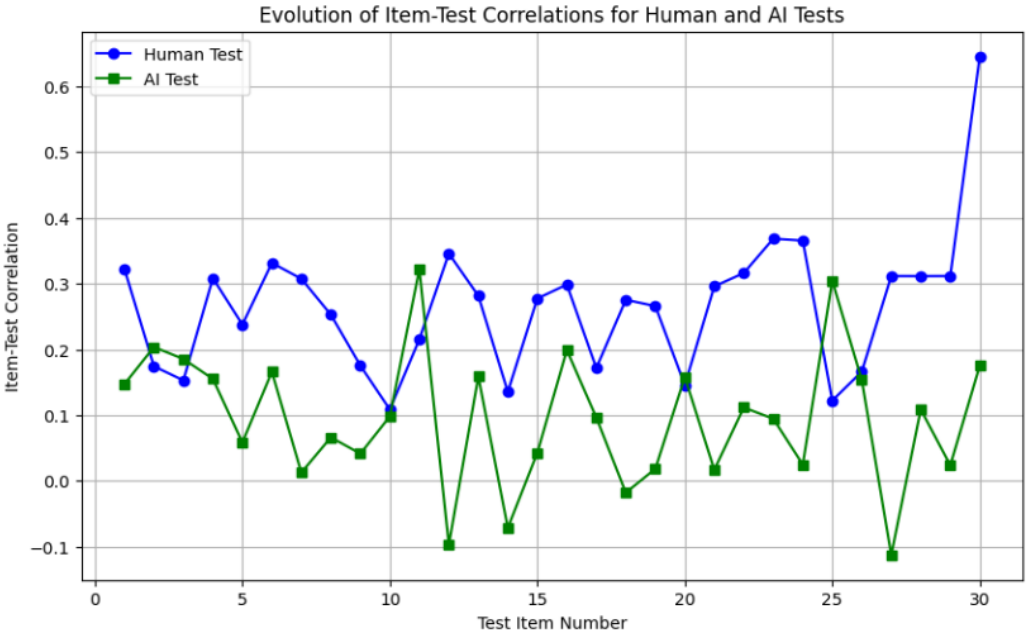

Figure 8. Alignment of each item with the overall test

3.2.3 Item-test correlation (Rit)

Item-test correlation assesses the correlation between individual items and overall performance by participants. As for Test 1 (human-crafted), the mean R_{it} is 0.26, indicating a relatively good correlation between specific items and the total test. 26.7% of items were classified as practical ($R_{it} \geq 0.30$) (subset of the most substantial items; others may only need minor changes). As for Test 2 (AI-generated), the mean R_{it} is 0.09, showing an extreme lack of fit between items and test. None of the items showed an acceptable correlation ($R_{it} \geq 0.30$), raising considerable doubts about the AI-constructed exam's content validity. The test's sharp distinction highlights the human-designed test's overall greater coherency and construct alignment, in contrast to the variable pattern observed in the AI-generated test. The average item-test correlation ($R_{it} = 0.26$) reflects a reasonable agreement between items and total performance on the human test. On the other hand, the AI test average R_{it} of only 0.09 signals poor alignment, meaning at least some of these questions were not appropriately measuring participant performance.

As shown in Figure 7 and Figure 8, item-test correlation represents the correlation between an individual question and test performance. More consistent correlations with the human test can be seen, which indicates well-formed and aligned questions. The correlations are lower for the AI test than for the human test, reflecting that the questions aren't aligned with overall scores.

3.3 Test-Level Reliability

Cronbach's α was used to assess the internal consistency for both tests. As for Test 1 (human-crafted), Cronbach's α is 0.752, above the threshold for acceptability in educational assessments. This indicates stable and reliable test characteristics with stable item characteristics. As for Test 2 (AI-generated), Cronbach's α is -0.1, an indication of a significant systemic defect with the AI-generated test, where the output from the responses could not show similarity or consistency. It means an AI-generated test is going to get a backward reliability score, which indicates intrinsic issues with item alignment and grading models. In contrast, while still standard in terms of use for the classroom, results generated from the human-built exam are regular, run-of-the-mill.

3.4 Sensitivity Analysis

A post-hoc analysis of variance (ANOVA) revealed no significant score differences by gender ($F = 0.43$, $p = 0.51$) or age group ($F = 1.12$, $p = 0.34$), confirming demographic neutrality. Sensitivity tests excluding outliers (top/bottom 5% scores) yielded consistent results, reinforcing the robustness of the CTT metrics.

4. Summary of Results

From this analysis, it can be seen that human tests are well ahead in all metrics when compared to AI tests. On the other hand, the human-designed test got the perfect mix in the level of difficulty and discrimination indices, correct relationships of difficulty indices with both right and wrong responses of item-test correlation, and reliability of the final test. In contrast, these defects were presented in the AI-generated test in terms of a distorted item difficulty distribution, low discrimination, poor coverage of constructs, and extremely low-reliability capabilities.

The results remind us of AI's inability to create readable, high-quality assessments without human oversight and the need for pedagogical knowledge in test design. Although AI also has the promise to enhance human creativity through other tools, its construction requires constrained development and combination to reach the targets for academically valid evaluation.

5. Discussion

5.1 Strengths of Human-Crafted Tests

The results of this study underscore the distinct advantages of human expertise in crafting MCQs. The strengths of the human-designed test include the following key aspects:

a) Balanced distribution of item difficulty

The human test was appropriately split among challenging, medium, and manageable items, with many more items in the ideal range (P_i) having a value between 0.40 and 0.60. This balance allowed the test to range across a continuum of competencies and to discriminate between different levels of skills.

b) Internal consistency

Cronbach's α score of 0.752 passes the threshold for educational level reliability, indicating a high degree of internal consistency for the human-designed test. This shows that the content of individual items is well aligned with the examination's core educational goals and reinforces the view that the test is a valid measure of personal knowledge.

c) Effective discrimination

In the human-crafted test, the mean discrimination index (D) of the items was 0.31, and 50% of the items were categorized as "well-discriminating." This strong discrimination power between good and bad students shows the fine calibration of the test designer, who designed the items keeping in mind the learners' performance.

Such strengths demonstrate the value of human insight in creating psychometrically valid, scalable, and results-aligned assessments. Psychometric principles were applied with caution to ensure test validity and reliability, rendering them useful for the meaningful evaluation of learners. The test created by humans showed important strengths in terms of balanced P_i (P_i mean = 0.55) and high internal consistency (α = 0.752). These strengths illustrate the careful use of psychometric principles to produce a reliable and valid assessment that can be used in a pedagogical context.

5.2 Weaknesses in AI-Generated Tests

Although AI tools such as ChatGPT can create test items more efficiently than ever, the current study also found significant challenges to implementing psychometric principles (Naseer et al., 2024; Yunjiu et al., 2022):

a) Excessive item difficulty

The AI-generated test had roughly three-quarters (83.3%) of the items labeled as very hard ($P_i < 0.30$). This led to an unsatisfactory distribution that annoyed participants and created the very test that could not properly gauge the skill of all the participants. The overall score for item difficulty was low (P_i = 0.22), which illustrates the difficulty of AI in accurately calibrating content.

b) Result analysis evaluation

As for the item-test agreement (R_{it}) and internal consistency (Cronbach's α) in the AI test, this disjointedness indicates that AI was unable to build items that would successfully link the construct, at the test level, to item specifications.

c) Lack of educational subtlety

This imbalance between recall and application versus higher-order thinking was a natural outcome of the low and generic human prompting that was driving the generation of AI materials. This limitation is important for pedagogical settings because contextualized and learner-centered assessments are essential for their validity.

These conclusions indicate that even if such AI-assistant assessments can provide a framework upon which to base an assessment, using them alone does not have the refinements needed for high-stakes assessments. These tests are in danger of reducing the quality and integrity of tests without human oversight.

The AI-generated test's weak discrimination (D = 0.16) and negative reliability (α = -0.1) suggest fundamental flaws in item construction. Unlike human designers, ChatGPT lacks pedagogical intentionality; it cannot calibrate difficulty to the audience's knowledge level or align distractors with common misconceptions. For example, AI items often overemphasized recall (e.g., "What is the formula for X?") rather than application, reducing discriminatory power. These findings in medical education reveal greater severity in mathematics, where contextual nuance is critical.

5.3 Potential for Hybrid Collaboration

Although this has its pitfalls, with human minds being a necessity in the application of AI technology for educational assessment design, it illustrates the possibility of a collaborative hybrid, using the benefits of both—AI writing tools and human brain creativity (Whitehill & LoCasale-Crouch, 2023). An interim solution could augment human speed and productivity by creating drafts of tests that a human then reviews and verifies. Educators can review AI-generated items for psychometric and pedagogical properties (e.g., balance of difficulty, sound discrimination, etc.).

a) Enhanced input with explicit psychometric principles

In engineering prompts, strict psychometric guidelines (e.g., ideal range of difficulties and discrimination thresholds) must be set to calibrate these trials with little or no human inspection. The output that comes out of these AI tools has to be trained and steered towards producing educationally relevant results.

b) Reviews and calibration by educators

Test drafts require careful scrutiny by experts in a given academic discipline. This enables educators to calibrate items to improve their fit with syllabus objectives, eliminate poorly worded items, and recalibrate item difficulty and discrimination. The iterations allow us to ensure that the output meets both psychometric and pedagogic expectations (Milić et al., 2024; Yunjiu et al., 2022; Zakareya et al., 2024).

c) AI as an assistant, not a replacement

ChatGPT and other similar AI tools should be embraced as auxiliaries to the human teacher, not independent replacements. While AI may churn out products quickly, the contextual, pedagogical, and psychological knowledge that only experienced educators have access to will always need to be used to keep testing the high-quality and high-functioning products they need to be (Agarwal et al., 2023).

d) Mixed approach for optimization and quality

A hybrid model can offer a middle-ground solution and leverage the speed and cost savings of generative AI but adding human scrutiny to offer a more scalable option for assessment generation. This might allow educators to produce preliminary questions, which can then be fine-tuned, tested and psychometrically validated by AI.

5.4 Recommendations for Practice

For educators, a hybrid workflow is recommended:

a) AI drafting: Use ChatGPT to generate initial item pools (e.g., generate 20 MCQs on calculus fundamentals).

b) Human refinement: Adjust difficulty by modifying distractors (e.g., adding plausible errors) and ensure alignment with learning objectives.

For developers, integrating psychometric guardrails (e.g., auto-rejecting items with $P_i < 0.3$ or $D < 0.2$) could improve output quality. Institutions should train teachers in prompt engineering (e.g., specifying Bloom's taxonomy levels) to optimize AI utility.

5.5 Implications for Future Assessment Design

The study suggests that strengthening a cooperative use of technology for education protection is the realization of the different elements. AI tools can provide assistance to the human experience as people make fast, consistent decisions (with the necessary caution) in moderation. The results evidence the possibility of embedding AI within human-centered pedagogies and suggest design artifacts for the implementation of human-centered future learning technologies at scale through best practices in psychometrics and pedagogy. In this way, the best (hybrid) world approach gives power to educators and developers to steer the quality and penetration of assessments in every dimension of learning (Hambleton & Jones, 1993).

6. Recommendations

6.1 For Systems that Evaluate Using AI

AI tools such as ChatGPT have the potential to support the rapid generation of test items. However, their effective deployment in educational contexts requires significant improvements and enhancements to ensure alignment with pedagogical goals and educational quality standards. These recommendations address four principles to strengthen when designing and developing systems aimed at providing AI-generated feedback (Milić et al., 2024; Naseer et al., 2024; Nasution, 2023; Whitehill & LoCasale-Crouch, 2023; Yunjiu et al., 2022):

a) Integrate user-based feedback loops

- Give immediate feedback loops for users (i.e., educators and test administrators) to report all items that need to be flagged, clarify what a test may measure, and make improvements for each item.
- It is possible to feed AI over time and experience to create more relevant and psychometrically robust context items in the subsequent versions of AI.

b) Incorporate statistical standards into algorithm development

- Build psychometry right into the AI algorithms from the inception of the design. Specific item parameters (difficulty thresholds of $P_i = 0.30 - 0.70$) can be defined to filter out concept-irrelevant random test questions and/or questions that are beyond the difficulty range.
- Establish cut-offs for the discrimination indices for items to one that truly serves the purpose of discrimination, that is, to discriminate between high- and low-performing students. Therefore, systems can deliver better quality items, as there is less need for post-analysis interference, and metric compliance is automated.

c) Improve contextual and pedagogical calibration

- On-board domain-specific educative material to augment AI training datasets to generate item content tailored to multiple pedagogical needs.
- Combine contextual variables (subject alignment, field relevance, and cognitive skill targeting) to vary content types and gain test-based utility in education.

d) Develop validation dashboards for educators

- Develop intuitive front ends for teachers to check, edit, and adapt AI-generated material in real time.

Provide psychometric indices (e.g., difficulty, discrimination, and reliability) that users can study and use to make empirical data-driven decisions about fine-tuning an assessment before it is administered.

6.2 For Educational Institutions

Institutions, a significant factor, play a huge part in how well AI-driven assessment tools can be adopted. Hence,

institutions should take a proper step in alleviating concerns and providing teachers with tools in using systems like these so that positives and negatives of this sort of technology can be maximized and minimized. The recommendations below are institutional in nature, suggesting pathways through which these outcomes might be achieved (Hambleton & Jones, 1993; Kurdi et al., 2020):

a) Make use of AI tools trained by the people: educators

- Provide in-service training sessions on the merits and demerits of AI tools. Such training should highlight key psychometric principles and provide tangible examples for incorporating AI-produced content into valid assessments guided by pedagogical principles.
- Provide instructors with the necessary skills to operate AI platforms adequately, evaluate system outputs, and implement revisions.

b) Encourage co-designing activities

- Interdisciplinary partnerships for developing AI-driven assessments among educationists, data scientists, and instructional designers can help ensure the systems created meet the technical and pedagogical aspects needed.
- Design tests that leave the voices of the teachers and contextualized cues to the cue of design but are the scalable part of the AI systems.

c) Getting tests by using hybrid approaches

- Redirect policies to view AI as a complement rather than a substitute. These types of hybrid models will have to conform to labor division; AI writes the dull work, and teachers have to write and edit content for quality, relevance and diversity. As for the implementation of test methods, human specialists examine and attest to drafts of tests and readiness prior to execution.

d) Promoting AI use case research within institutions

- Enable systematic research to investigate new ways to use AI in the assessment process. The results of such studies can inform best practices, tailor tools to suit teachers and students better and guide the future development of AI-assisted educational systems.
- Bring in platforms and tie-ups with AI developers to adapt solutions for the right institutional objectives and eventual student profile.

This provides a rough framework for the implementation quality and governance of AI-based assessment systems, which institutions can use while adopting them more widely. At a higher level, the goal is to harness the speed and scalability of technology while eliminating components that undermine the concept validity of assessments. By reimagining a future for accessibility, efficiency, and practical deployment with innovative use of AI and strategic services to harness and scale human ability, educational institutions can co-create new assessment approaches.

7. Conclusions

This study systematically examined the psychometric properties of human- and AI-generated MCQs based on the CTT statistics, showing substantial quality, usability, and educational value differences between the two methods. In particular, the human-made test performed well in item difficulty, discrimination, and internal reliability, with a Cronbach's α of 0.752. In contrast, the test based on AI showed serious deficiencies, including over-difficulty, poor discrimination, misalignment with teaching goals, and unsatisfactory reliability (-0.1). Such results reiterate the necessity of human intuition and pedagogical experience in creating a robust test a robust test series. Even though these AI-created tests have potential as an initial structure, the existing technological applications cannot be implemented independently because the required sensitivity to the corresponding context and the adjustment of the content concerning psychometric principles are lacking. All measures favored tasks designed by humans over those generated by AI, such as the balance of difficulty (P_i mean = 0.55 vs. 0.22), discrimination (D mean = 0.31 vs. 0.16), and reliability (α = 0.752 vs. -0.1) of tests. This implies that despite the advantages of AI-assisted language and assessment systems above, they cannot replace human intervention and expertise, especially in designing pragmatic assessments, even at non-proficient levels.

ChatGPT and other AI tools promise a future of academic content production that is more computationally efficient and more scalable. However, their application in assessment development remains optimistic by university-industry requirements. Before AI can become a useful tool in this regard, it needs to be sharpened to tackle some significant shortcomings. These include embedding psychometric principles in algorithm design, a higher level of context sensitivity in provision of output, and a provision of valid, reliable and actionable output. Hence, AI in education needs to be viewed as an enhancement to the human process, not as a standalone. One example is the automation of mundane processes such as content creation by all forms of AI so that an educator can spend time and energy on high-level processes like strategic planning, curricular development, and student-level task support. A possible long-term development in adaptive AI technologies is the personalization of assessments, making questions adaptive not only to a student's performance level but also to the difficulty and content areas most relevant to them. Therefore, human judgment can be combined with AI. Human academia

teaches us the psychology of the learner, how to create the course, and the nuances of pedagogy, while AI architecture provides speed, scalability, and data-driven adaptability. When put together, each relies on the other to counter its inherent weaknesses, resulting in assessments that are (ideally) both efficient and meaningful.

This collaboration can be operationalized in co-design paradigms, in which students generate initial drafts of tests that teachers edit to fit syllabus goals and vice versa (in the case of AI-assisted writing). Models like these offer a middle ground: the consistency and scalability of automation with the contextual richness of supervised human evaluation make them appropriate frameworks for holistic and comprehensive assessments. The application of AI in educational assessment is a work-in-progress frontier that has the power to change the design, implementation, and evaluation of learning. But to fulfill this potential, the drive must be actively offset towards a darker future of technology efficiency against a brighter future of pedagogical integrity. The key to liberating the full potential of AI to help teachers and students be the best they can be is its use as a collaborator rather than a replacement for educators or technologists. Ultimately, combining human perspective with algorithmic analytics can reshape educational assessments in a scalable, fair, and impactful manner, establishing a more adaptive and student-centered educational ecosystem for the future.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Agarwal, M., Sharma, P., & Goswami, A. (2023). Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus*, *15*(6), e40977. https://doi.org/10.7759/cureus.40977.

Ali, F. A., Sharif, S. Y., Ata, M., Patel, N., Rafay, M., Syed, H. R., & Iqbal, S. P. (2024). Chat GPT Develops Multiple Choice Questions (MCQs) for postgraduate specialty assessment – A reality or a myth? *Pak. J. Neurol. Surg.*, *28*(1), 142-149. https://doi.org/10.36552/pjns.v28i1.963.

Cook, D. A. & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am. J. Med.*, *119*(2), 166.e7-166.e16. https://doi.org/10.1016/j.amjmed.2005.10.036.

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

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educ. Meas. Issues Pract.*, *12*(3), 38-47.

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

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

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *Int. J. Artif. Intell. Educ.*, *30*, 121-204. https://doi.org/10.1007/s40593-019-00186-y.

Lee, G. & Kim, H. Y. (2024). Human vs. AI: The battle for authenticity in fashion design and consumer response. *J. Retail. Consum. Serv.*, *77*, 103690. https://doi.org/10.1016/j.jretconser.2023.103690.

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

Milić, B., Spajić, J., Bošković, D., Mitrović, K., & Lalić, D. (2024). AI VS. Human designers: Evaluating the effectiveness of AI-generated visual content in digital marketing. *In The Symposium GRID 2024, Novi Sad, Serbia*, 387-394. https://doi.org/10.24867/GRID-2024-p42.

Narayanan, S., Kommuri, V. S., Subramanian, N. S., Bijlani, K., & Nair, N. C. (2017). Unsupervised learning of question difficulty levels using assessment responses. *In Computational Science and Its Applications – ICCSA 2017, Trieste, Italy*, 543-552. https://doi.org/10.1007/978-3-319-62392-4_39.

Naseer, M. A., Nasir, Y., Tabassum, A., & Ali, S. (2024). ChatGPT-4 versus human generated multiple choice questions - A study from a medical college in Pakistan. *JSHMDC*, *5*(2), 58-64. https://doi.org/10.53685/jshmdc.v5i2.253.

Nasution, N. E. A. (2023). Using artificial intelligence to create biology multiple choice questions for higher

education. *Agric. Environ. Educ.*, *2*(1), em002. https://doi.org/10.29333/agrenvedu/13071.

Ohmoto, Y. (2024). Estimation of ICAP states based on interaction data during collaborative learning. *J. Educ. Data Min.*, *16*(2), 149-176.

Partchev, I. (2020). Diagnosing a 12-item dataset of Raven Matrices: With dexter. *J. Intell.*, *8*(2), 21. https://doi.org/10.3390/jintelligence8020021.

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

Taib, F. & Yusoff, M. S. B. (2014). Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. *J. Taibah Univ. Med. Sci.*, *9*(2), 110-114. https://doi.org/10.1016/j.jtumed.2013.12.002.

Tran, N., Pierce, B., Litman, D., Correnti, R., & Matsumura, L. C. (2024). Multi-dimensional performance analysis of large language models for classroom discussion assessment. *J. Educ. Data Min.*, *16*(2), 304-335. https://doi.org/10.5281/zenodo.14549071.

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

Whitehill, J. & LoCasale-Crouch, J. (2023). Automated evaluation of classroom instructional support with LLMs and BoWs: Connecting global predictions to specific feedback. *J. Educ. Data Min.*, *16*(1), 34-60. https://doi.org/10.5281/zenodo.10974824.

Yang, N. (2025). The impact of GPT models on education: Enhancing learning outcomes and addressing challenges. *ITM Web of Conf.*, *70*, 04007. https://doi.org/10.1051/itmconf/20257004007.

Yunjiu, L., Wei, W., & Zheng, Y. (2022). Artificial intelligence-generated and human expert-designed vocabulary tests: A comparative study. *Sage Open*, *12*(1), 1-12. https://doi.org/10.1177/21582440221082130.

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