Journal of Pedagogical Sociology and Psychology
Volume 7, Issue 3, 2025
https://doi.org/10.33902/jpsp.202536891

Research Article

Assessing the psychometric properties of AI-generated multiple-choice exams in a psychology subject

Jomar Saif P. Baudin

Faculty of Psychology Program, Social Sciences Department, College of Arts and Sciences, Southern Luzon State University, Lucban, Quezon, Philippines

Correspondence should be addressed to Jomar Saif P. Baudin jbaudin@slsu.edu.ph
Received 15 April 2025; Revised 10 July 2025; Accepted 21 July 2025

This study assessed the psychometric properties of AI-generated multiple-choice questions in undergraduate psychology education, specifically focusing on an Experimental Psychology course. Using a mixed-methods approach, we evaluated 80 multiple-choice questions created by ChatGPT-4 through expert content validation, administration to undergraduate psychology students, and comprehensive psychometric analysis. Results indicated that AI-generated items demonstrated reasonable content validity with strongest ratings for clarity and weakest for distractor quality. The assessment showed acceptable internal consistency and moderate test-retest reliability. Item difficulty analysis revealed a slight tendency toward easier items, with 58.75% of questions in the medium difficulty range. Discrimination indices were generally good, though 11.25% of items showed poor discrimination. Cognitive level analysis identified a significant imbalance in the distribution of items across Bloom's taxonomy, with 73.75% targeting lower-order thinking skills (remembering, understanding) and only 8.75% addressing higher-order skills (analyzing, evaluating). These findings suggest that while AI can generate psychometrically sound multiple-choice questions for psychology assessment, significant limitations exist in assessing higher-order cognitive skills. The results support a hybrid approach that leverages AI for generating fundamental knowledge assessment items while preserving human expertise for higher cognitive domain evaluation. This study contributes to the emerging understanding of AI applications in psychology education by providing domain-specific insights into the capabilities and limitations of AI-generated assessment tools.

Keywords: Artificial intelligence, assessment, chatgpt, cognitive levels, multiple-choice questions, psychology education, psychometric properties, reliability, validity

1. Introduction

The landscape of educational assessment is undergoing a substantial transformation with the emergence of artificial intelligence [AI] technologies. Multiple-choice questions [MCQs] remain one of the most widely used assessment formats in higher education due to their efficiency, objectivity, and capacity to evaluate a broad spectrum of cognitive domains (Doughty et al., 2024; Mistry et al., 2024). However, the creation of high-quality MCQs presents significant challenges for educators, requiring substantial time, expertise, and resources (Cheung et al., 2023; Mead & Zhou, 2023). The development process involves careful consideration of item validity, reliability, and alignment with learning objectives, tasks that traditionally demand extensive human effort.

Recent advancements in AI, particularly large language models [LLMs] such as ChatGPT, Google Bard (now Gemini), and GPT-4, have demonstrated promising capabilities in natural language processing and content generation (Al Mashagbeh et al., 2024; Doughty et al., 2024). These models can understand context, generate coherent text, and even mimic human-like reasoning to varying degrees. Such capabilities have sparked interest in leveraging AI for educational assessment, including the automated generation of MCQs. The potential benefits of AI-assisted assessment are substantial: reducing educator workload, increasing assessment frequency, standardizing question quality, and potentially mitigating human biases in test creation (Owan et al., 2023).

Despite these promising prospects, several critical questions remain about the psychometric properties of AI-generated assessments. The quality of any assessment instrument is fundamentally determined by its psychometric characteristics, reliability, validity, difficulty level, discrimination power, and overall measurement precision (Ali & Talat, 2024; Kıyak & Emekli, 2024). However, the extent to which AI-generated MCQs satisfy these essential criteria, particularly in specialized domains like psychology, remains insufficiently explored. Psychology, with its complex theoretical frameworks, nuanced concepts, and emphasis on critical thinking, presents unique challenges for automated assessment creation (Ngo et al., 2024).

The integration of AI into educational assessment raises additional concerns beyond purely psychometric considerations. Ethical implications related to academic integrity, the appropriate role of AI in education, and potential biases embedded within AI systems warrant careful examination (Owan et al., 2023). Moreover, the pedagogical alignment between AI-generated assessments and educational objectives particularly higher-order cognitive skills critical to psychology education, requires thorough investigation.

This study aims to comprehensively assess the psychometric properties of AI-generated multiple-choice exams in undergraduate psychology education. By examining key metrics including item validity, reliability, difficulty indices, discrimination power, and content alignment with learning objectives, this research will provide empirical evidence regarding the quality and educational utility of AI-generated assessments. Additionally, this study will investigate the comparative performance of different AI models in creating psychology-specific MCQs, offering insights into their relative strengths and limitations. Through this detailed psychometric analysis, the research seeks to inform evidence-based practices for integrating AI technologies into psychology assessment design, ultimately contributing to more efficient, effective, and equitable evaluation methods in higher education.

2. Literature Review

2.1. Evolution of AI in Educational Assessment

The application of artificial intelligence in educational assessment has evolved significantly over the past decade, transitioning from rule-based systems with limited capabilities to sophisticated neural network-based models capable of generating complex content. Early automated item generation [AIG] approaches relied on predefined templates and computational algorithms with minimal natural language understanding (Mead & Zhou, 2023). These systems, classified as "weak AIG" by Drasgow et al. (2006), lacked predictive control over item psychometrics and required substantial human intervention.

The emergence of advanced large language models has marked a paradigm shift in AI's potential for educational assessment. These models, trained on vast corpora of text data, can now generate content that increasingly resembles human-created materials in sophistication and nuance (Doughty et al., 2024; Mistry et al., 2024). Owan et al. (2023) characterized this evolution as progressing through three stages: (1) rule-based systems primarily for objective scoring, (2) machine learning algorithms for pattern recognition and feedback, and (3) current generative AI models capable of creating diverse assessment items and adaptive testing experiences.

The integration of AI into educational assessment offers numerous potential benefits. Al Mashagbeh et al. (2024) highlighted efficiency gains, noting that ChatGPT reduced MCQ creation time from 30-60 minutes per human-crafted question to approximately 5-15 minutes. Additionally, AI systems can generate large question banks for practice and formative assessment, potentially enhancing the breadth and frequency of student evaluation (Bitzenbauer, 2023). Kic-Drgas and Kılıçkaya (2024) further emphasized AI's capacity for rapid customization of questions to specific learning objectives and course content, allowing for more tailored assessment experiences.

However, the implementation of AI in educational assessment faces several challenges. Mead and Zhou (2023) found that while 71-90% of GPT-3-generated items for a psychometrics certification exam were deemed useful, nearly half violated basic multiple-choice principles (e.g.,

having no correct key or multiple correct answers). Similarly, Ngo et al. (2024) reported that ChatGPT 3.5 achieved only 32% accuracy when generating immunology MCQs, with 43% requiring substantial modifications and 25% containing incorrect or misleading answers. These findings suggest that while AI demonstrates promising capabilities, significant quality control and human oversight remain necessary.

2.2. Comparative Performance of AI Models in MCQ Generation

Recent research has increasingly focused on comparing the performance of different AI models in generating high-quality MCQs across various disciplines. Mistry et al. (2024) conducted a rigorous evaluation of radiology board-style MCQs generated by GPT-4 and Llama 2, finding that GPT-4 significantly outperformed Llama 2 across all quality criteria (clarity, relevance, suitability, distractor quality, and rationale adequacy). Notably, GPT-4-generated questions matched the quality of established American College of Radiology Diagnostic Radiology In-Training (ACR DXIT) exam questions highlighting the substantial differences in accuracy between models, with GPT-4 achieving 100% accuracy in answers compared to Llama 2's 69%. In the medical domain, Al Mashagbeh et al. (2024) compared ChatGPT 3.5, ChatGPT 4, and Google Bard across diverse question types, finding that ChatGPT 4 consistently outperformed other models in accuracy for most question formats. Particularly notable were the accuracy rates for true/false questions and MCQs. This study also revealed variation in model performance based on question type, with all models struggling with calculation-based questions despite their sophisticated language capabilities.

Ahmed et al. (2024) provided further insights by comparing ChatGPT 3.5 and Google Bard in generating dental caries MCQs. While Bard demonstrated slightly higher accuracy, the difference was not statistically significant. Interestingly, the models exhibited distinct patterns of strengths and weaknesses: Bard generated questions at higher cognitive levels (analysis/synthesis) but frequently used absolute terms (e.g., "always," "never") that are considered flaws in MCQ design, whereas ChatGPT more commonly produced format errors, ambiguity, and repetition issues. Cheung et al. (2023) conducted a multinational prospective study comparing ChatGPT-generated medical graduate exam MCQs with those created by experienced human examiners. Their findings revealed no significant difference in overall quality scores between AI and human-made questions with AI scoring lower only in relevance. Notably, ChatGPT generated 50 MCQs in approximately 20 minutes, while humans required over 3.5 hours for the same task, a more than tenfold efficiency improvement. Perhaps most striking was that experienced assessors could not reliably distinguish between AI and human-generated questions in blind evaluations, suggesting that AI-generated content has achieved substantial parity with human work in certain contexts.

These comparative studies collectively indicate that more recent, sophisticated AI models (particularly GPT-4) demonstrate superior performance in MCQ generation compared to earlier versions or alternative models. However, they also highlight that AI performance varies considerably based on subject domain, question type, and specific quality criteria, underscoring the need for model selection and prompting strategies tailored to particular educational contexts.

2.3. Validity and Reliability

Validity, the extent to which an assessment measures what it purports to measure has been evaluated in various ways for AI-generated items. Sihite et al. (2023) examined ChatGPT 3.5-generated reading comprehension questions, finding that only 10 out of 30 questions met validity criteria when evaluated using Pearson correlation methods. Similarly, Nasution (2023) reported that 20 out of 21 AI-generated biology MCQs demonstrated statistical validity when tested with undergraduate students, while one question was invalid due to ambiguous phrasing. Reliability, which concerns the consistency of assessment results, has shown more promising outcomes. O (2024) compared AI-human-made and fully human-made test forms for a TESOL theory course, finding comparable reliability coefficients. Nasution (2023) reported a Cronbach's alpha of 0.655 for AI-generated biology MCQs after removing an invalid question, indicating fair reliability. Van

de Poll et al. (2022) evaluated AI-generated organizational assessment questionnaires, finding that with sample sizes of 100+ respondents, the instruments consistently met validity and reliability thresholds.

These findings suggest that while AI-generated items can achieve acceptable reliability levels, validity remains more challenging and often requires human intervention to ensure alignment with educational objectives and construct accuracy.

2.4. Difficulty and Discrimination

Item difficulty and discrimination power are crucial psychometric properties that determine an assessment's ability to differentiate between high and low-performing students. Nasution (2023) found that AI-generated biology MCQs exhibited varied difficulty levels: 9 easy questions, 10 medium, and 2 difficult. Discrimination indices similarly varied, with 4 poor, 9 adequate, and 8 good discriminators. This distribution suggests that while AI can generate items across the difficulty spectrum, they may not be optimally balanced without human curation. Doughty et al. (2024), in their study of AI-generated programming MCQs, identified specific quality issues that affect discrimination: 4.9% of AI-generated items had multiple correct answers (compared to 1.1% of human-created items), and 4.0% contained distractors that revealed the correct answer (versus 0.9% in human items). These flaws can substantially undermine an item's ability to discriminate between knowledge levels, requiring careful review and revision. O (2024) provided one of the few direct comparisons of difficulty and discrimination between AI-assisted and human-created items. For true-false questions, AI-human items showed higher difficulty (76.74% correct vs. 67.91%) but lower discrimination (0.38 vs. 0.49). Interestingly, for MCQs, AI-human items demonstrated better discrimination (80% "excellent" vs. 50% in human items) and more functional distractors (75% vs. 67.5%). These mixed results highlight the complex interaction between AI capabilities and question formats.

2.5. Cognitive Levels and Learning Objectives

The alignment of assessment items with desired cognitive levels, often categorized using Bloom's taxonomy, represents another important dimension of psychometric quality. Ahmed et al. (2024) found that Google Bard generated dental caries questions at higher cognitive levels compared to ChatGPT 3.5, which primarily produced knowledge/comprehension-level questions. Similarly, Ali and Talat (2024), in their systematic review, noted that most AI-generated MCQs targeted lower Bloom's taxonomy levels, with fewer questions assessing higher-order reasoning skills. Doughty et al. (2024) provided a notable exception to this pattern, reporting that AI-generated programming MCQs demonstrated superior alignment with learning objectives compared to human-created questions. This study employed a structured pipeline that classified learning objectives using Bloom's taxonomy and mapped them to appropriate MCQ types before generation, suggesting that proper prompting strategies may enhance cognitive alignment. Sihite et al. (2023) specifically examined ChatGPT 3.5's ability to generate higher-order thinking questions for reading comprehension, finding that 40% targeted analysis, 30% evaluation, and 30% creation, a promising distribution of higher cognitive levels. However, the validity issues noted earlier suggest that generating higher-order questions alone is insufficient; these questions must also accurately assess the intended constructs.

These findings collectively indicate that while AI models can generate questions across cognitive levels, they may require specific prompting and design strategies to reliably produce higher-order assessment items that maintain validity and discrimination power.

2.6. AI Integration in Psychology Education and Assessment

Psychology education presents unique challenges for AI-generated assessments due to the discipline's complex theoretical frameworks, emphasis on critical thinking, and integration of multiple knowledge domains (clinical, cognitive, developmental, social, etc.). While several studies have explored AI-generated MCQs in medical, language, and science education, research specific

to psychology assessment remains limited. The closest analogue comes from Mead and Zhou's (2023) study of AI-generated psychometrics certification exam items. Their findings indicated that while most items were deemed useful, nearly half violated multiple-choice principles, and very few were considered ready for use without editing. The authors noted particular challenges with psychological constructs that require precise operational definitions and nuanced understanding, characteristics central to psychology education.

Ngo et al. (2024) examined ChatGPT's performance in generating immunology MCQs, finding substantial limitations in handling complex conceptual material. Only 32% of questions were correct with appropriate explanations, while 25% contained incorrect or misleading information. Given psychology's similar complexity and conceptual nuance, these findings suggest potential challenges in generating valid psychology assessment items without significant human oversight. Bitzenbauer's (2023) pilot study of ChatGPT activities in physics education demonstrated positive student engagement and perception of AI-assisted learning. Students particularly valued the critical analysis of AI-generated content, suggesting potential applications for psychology education where critical evaluation of information sources is paramount. However, the study also revealed occasional inaccuracies in AI-generated content, highlighting the need for careful implementation in educational contexts. The cognitive alignment challenges identified by Ahmed et al. (2024) and Ali and Talat (2024) have particular relevance for psychology education, where higher-order skills like analysis, application, and evaluation are essential learning outcomes. If AI-generated assessments primarily target lower cognitive levels, they may inadequately assess the critical thinking and application skills central to psychology education.

Collectively, these studies suggest that while AI-generated assessments show promise for psychology education, they likely require discipline-specific prompting strategies, careful validation processes, and integration within a broader assessment framework that includes diverse evaluation methods.

The review of existing literature reveals several significant gaps in our understanding of AI-generated MCQs for psychology education and assessment. While studies have examined AI-generated MCQs across various disciplines (medicine, biology, language, programming), research specific to psychology assessment remains notably scarce. Psychology's unique combination of theoretical complexity, empirical methodology, and application to human behavior may present distinct challenges for AI-generated assessment. Many studies have focused on surface-level quality metrics (e.g., expert ratings of clarity, relevance) rather than comprehensive psychometric analysis. Fewer studies have rigorously examined properties such as internal consistency, test-retest reliability, criterion validity, and item response patterns when AI-generated assessments are administered to actual students. The capacity of AI to generate psychology questions that effectively assess higher-order thinking skills (application, analysis, evaluation) remains inadequately explored. Given psychology education's emphasis on critical thinking and application of theoretical knowledge to real-world contexts, this represents a crucial research gap. While some studies have compared different AI models, these comparisons have rarely included comprehensive psychometric evaluation, particularly for psychology content. Understanding which models produce psychometrically superior psychology assessment items would provide valuable guidance for educators. Research on optimal prompting strategies specifically for psychology assessment generation is virtually nonexistent. The effectiveness of different prompt structures, psychology-specific instructions, and model parameters for generating valid and reliable psychology MCQs requires systematic investigation. Few studies have analyzed how students interact with and perform on AI-generated psychology assessments compared to traditional assessments. Whether AI-generated items produce different response patterns, completion times, or learning outcomes remains largely unknown.

This research aims to address these significant gaps by conducting a comprehensive psychometric evaluation of AI-generated multiple-choice exams in undergraduate psychology education. By analyzing validity, reliability, difficulty, discrimination, and cognitive alignment, this study will provide empirical evidence regarding the quality and utility of AI-generated

psychology assessments. Additionally, by comparing different AI models and prompting strategies, the research will offer practical guidance for psychology educators seeking to integrate AI into their assessment practices. The significance of this research extends beyond psychology education, contributing to the broader understanding of AI's capabilities and limitations in educational assessment. As educational institutions increasingly consider adopting AI technologies, evidence-based research on their psychometric properties becomes essential for informed decision-making. This study will help establish best practices for the responsible integration of AI in assessment design, potentially enhancing efficiency while maintaining rigorous educational standards. By identifying specific strengths and limitations of AI-generated psychology assessments, this research will inform the development of hybrid assessment approaches that leverage AI capabilities while addressing their current limitations through strategic human oversight. Hence, this research aimed to assess the psychometric properties of AI-generated multiple-choice exam in undergraduate psychology education and evaluate its potential as valid and reliable assessment tools.

3. Methodology

3.1. Research Design

This study employed a mixed-methods research design combining quantitative psychometric analysis with qualitative expert evaluation to comprehensively assess the properties of AI-generated multiple-choice questions in psychology education. The quantitative component involved statistical analysis of psychometric properties (reliability coefficients, difficulty indices, and discrimination indices), while the qualitative component incorporated expert ratings and cognitive level classification. This design allowed for triangulation of findings, providing a more complete understanding of AI-generated assessment quality than either approach alone would offer.

3.2. Participants

3.2.1. Student participants

The study involved undergraduate psychology students (N = 120) enrolled in the Experimental Psychology course at a Southern Luzon State University, Lucban, Quezon, Philippines. Participants were recruited through convenience sampling from two class sections, with a balanced distribution. The sample included 73 female (60.8%) and 47 male (39.2%) students, with a mean age of 20.4 years (SD = 1.86). Participation was voluntary, and students received course credit for their involvement. The sample size was determined based on recommendations for psychometric validation studies, which suggest a minimum of 100 participants for reliable estimation of item parameters (Kyriazos, 2018).

3.2.2. Expert evaluators

Five psychology faculty members with expertise in assessment design served as expert evaluators for content validity analysis. These evaluators had an average of 12.3 years (SD = 4.2) of teaching experience in psychology and had previously participated in test development for departmental or standardized assessments. None of the experts had prior experience working with AI-generated assessment items to avoid potential bias in their evaluations.

3.3. Data Collection

3.3.1 Materials

Three distinct sets of psychology MCQs were developed for this study:

AI-generated items: A set of 80 MCQs was generated using ChatGPT-4 (October 2024 version). The prompt provided to the AI system included: (a) the course syllabus and learning objectives, (b) specific instructions to create psychology MCQs with four answer options (one correct and three

plausible distractors), and (c) uploaded content from the four preliminary chapters of the Experimental Psychology course (see Appendix 1). The prompt instructed the AI to create questions across different cognitive levels according to Bloom's taxonomy. All AI-generated items were used as produced by the system without human editing to preserve the authentic AI output for evaluation.

Faculty-created items: A parallel set of 80 MCQs was developed by two psychology faculty members (not among the expert evaluators) who regularly teach psychology course. These instructors received the same course materials and content domain specifications as those provided to the AI system. Faculty members independently created 20 questions each, which were then combined into a single set.

Validated psychology test bank items: A reference set of 60 MCQs was selected from established psychology test banks with known psychometric properties. These items matched the content domains and approximate difficulty levels of the other two sets, providing a benchmark for comparison.

3.3.2. Assessment instruments

Content Validity Rating Form: Expert evaluators used a structured rating form to assess each MCQ across five dimensions using a 5-point Likert scale (1 = poor, 5 = excellent): (a) alignment with learning objectives, (b) conceptual accuracy, (c) clarity of wording, (d) quality of distractors, and (e) appropriateness for undergraduate psychology education. The form also included space for qualitative comments on each item.

Cognitive Level Classification Guide*:* Based on the revised Bloom's taxonomy (Anderson & Krathwohl, 2001), this guide provided descriptions and examples of questions at each cognitive level (remembering, understanding, applying, analyzing, evaluating, and creating). Expert evaluators used this guide to classify each MCQ according to its predominant cognitive level.

Student Assessment Forms*:* Two parallel assessment forms (A and B) were created, each containing 90 MCQs (30 from each item source: AI-generated, faculty-created, and test bank randomly distributed throughout the test). This design allowed for comparison of item performance across sources while controlling for potential order effects.

Demographics and Feedback Questionnaire*:* Students completed a brief questionnaire collecting demographic information and their perceptions of the assessment experience. This questionnaire was administered after the completion of the second assessment to avoid influencing test performance.

3.3.3. Item generation and preparation

The AI-generated items were obtained through a series of prompting sessions with ChatGPT-4, with each session focusing on a specific content domain. All interactions with the AI system were documented, including the exact prompts used and any system constraints encountered. Faculty-created items were developed over a two-week period, with instructors following the same content specifications as provided to the AI system. Faculty members were not exposed to the AI-generated questions before creating their own to prevent contamination. Test bank items were selected from published resources based on their content relevance, known psychometric properties, and usage history in undergraduate psychology courses. All items were formatted identically in the assessment forms to avoid providing visual cues regarding item source. Items were assigned numeric codes to facilitate blind evaluation and statistical analysis.

3.3.4. Expert evaluation

Expert evaluators received an orientation session explaining the evaluation criteria and procedures. They were not informed about the source of individual items (AI-generated, faculty-created, or test bank) to reduce potential bias. Each expert independently rated all 80 MCQs using

the Content Validity Rating Form and classified each question according to Bloom's taxonomy using the Cognitive Level Classification Guide. After completing individual ratings, experts participated in a consensus meeting to discuss items with substantial rating discrepancies (defined as a range of ≥2 points on any dimension). While perfect consensus was not required, this discussion helped clarify evaluation criteria and resolve major disagreements.

3.3.5. Student Assessment Administration

The study employed a counterbalanced design where half the student participants (n = 60) completed Form A followed by Form B two weeks later, while the other half (n = 60) completed the forms in reverse order. This design controlled for potential order effects while enabling test-retest reliability analysis. Assessments were administered during regular class sessions using a secure online testing platform. Students had 75 minutes to complete each 80-item assessment form, which was determined to be sufficient based on pilot testing. After completing both assessment forms, students responded to the demographics and feedback questionnaire.

3.4. Data Analysis

3.4.1. Content validity analysis

Intraclass correlation coefficients [ICCs] were calculated to assess inter-rater reliability among expert evaluators. Mean ratings for each evaluation dimension were computed and compared across the three item sources (AI-generated, faculty-created, test bank) using one-way ANOVA with post-hoc comparisons. Content validity indices [CVIs] were calculated for each item by determining the proportion of experts rating the item as 4 or 5 on relevant dimensions. Items with CVI < 0.80 were flagged for further analysis. Qualitative comments from expert evaluators were thematically analyzed to identify common strengths and weaknesses of AI-generated items.

3.4.2. Reliability analysis

Internal consistency reliability was assessed using Cronbach's alpha for each assessment form and for item subsets based on source and content domain. Test-retest reliability was calculated using Pearson correlation coefficients between performances on identical items across the two testing sessions. Standard error of measurement [SEM] was computed to estimate the precision of assessment scores. Item-total correlations were calculated to identify potentially problematic items ($r < .20$) that might reduce overall reliability.

3.4.3. Item difficulty and discrimination analysis

Item difficulty indices (p-values) were calculated as the proportion of students answering each item correctly. Item discrimination indices were computed using point-biserial correlations between item performance and total test score. Distractor analysis examined the frequency distribution of responses across all options to identify non-functioning distractors (selected by <5% of students). Comparative analyses were conducted to identify statistically significant differences in difficulty and discrimination indices across the three item sources using repeated measures ANOVA.

3.4.4. Cognitive level analysis

The distribution of cognitive levels (according to Bloom's taxonomy) was compared across item sources using chi-square analysis to determine whether AI-generated items demonstrated a different cognitive profile than faculty-created or test bank items. The relationship between cognitive level and psychometric properties (difficulty, discrimination) was examined using correlation analysis to identify potential patterns specific to AI-generated items.

3.5. Ethical Considerations

This study strictly followed the university's ethical guidelines. All student participants provided informed consent prior to involvement. Data were collected and stored securely with identifying

information removed from the analysis dataset. Students were informed that their performance on the research assessments would not affect their course grades, and they were provided with a debriefing session after study completion. Additionally, to maintain student trust and perceptions of fairness, this study ensured transparency about the use of AI in test development. Participants were informed that some assessment items were AI-generated and that all items underwent rigorous expert validation and quality checks. This approach helped assure students that AI-assisted assessments were credible, fair, and aligned with academic integrity standards.

3.6. Limitations

Several methodological limitations should be acknowledged. First, the study focused on a single AI model (ChatGPT-4) and a specific prompt structure, which limits generalizability to other AI systems or prompting strategies. Second, the sample was drawn from a single university, potentially restricting the applicability of findings to other educational contexts or student populations. Third, the assessment was conducted in a controlled research environment, which may not fully represent authentic classroom assessment conditions. Finally, while efforts were made to match content and difficulty across item sources, inherent differences in topic selection and question framing may have influenced the comparative analyses.

4. Results

4.1. Content Validity Analysis

The content validity of the AI-generated multiple-choice questions for Experimental Psychology was evaluated by five expert faculty members using the Content Validity Rating Form. Table 1 presents the mean ratings across five dimensions.

Table 1
Mean Expert Ratings of Content Validity Dimensions for AI-Generated MCQs

Dimension	*Mean Rating (1-5)*	*SD*	*Content Validity Index*
Alignment with learning objectives	4.12	0.68	0.84
Conceptual accuracy	3.94	0.87	0.78
Clarity of wording	4.21	0.63	0.88
Quality of distractors	3.62	0.91	0.72
Overall appropriateness	4.08	0.72	0.82

As shown in Table 1, expert evaluators rated the AI-generated items highest on clarity of wording (M = 4.21, SD = 0.63, CVI = 0.88) and alignment with learning objectives (M = 4.12, SD = 0.68, CVI = 0.84). The items received moderate ratings for overall appropriateness (M = 4.08, SD = 0.72, CVI = 0.82) and conceptual accuracy (M = 3.94, SD = 0.87, CVI = 0.78). The lowest ratings were for quality of distractors (M = 3.62, SD = 0.91, CVI = 0.72), suggesting this as an area of relative weakness in the AI-generated questions.

Inter-rater reliability among the five expert evaluators was moderate to high, with intraclass correlation coefficients (ICCs) ranging from 0.68 to 0.83 across the evaluation dimensions.

Qualitative analysis of expert comments revealed several patterns regarding AI-generated items. Strengths consistently noted included clear question stems, appropriate use of psychological terminology, and good coverage of core experimental psychology concepts (research methods, ethics, sampling techniques). Frequently cited weaknesses included occasional issues with distractor plausibility and some conceptual oversimplifications.

Content analysis by topic area indicated that the AI-generated questions covered key domains of experimental psychology with the following distribution: scientific principles (24%), research ethics (18%), research methods (28%), and sampling techniques (30%). This distribution was judged by experts to adequately reflect the relative importance of these topics in an experimental psychology course.

Expert evaluators identified 14 questions (17.5%) that required substantial revision due to potential conceptual inaccuracies or ambiguities. These primarily involved items related to research ethics scenarios and the application of statistical concepts.

4.2. Reliability Analysis

To assess the reliability of the AI-generated assessment, both internal consistency and test-retest reliability were calculated. Table 2 presents Cronbach's alpha coefficients for the complete assessment and for content domain subscales.

Table 2
Internal Consistency Reliability Coefficients for AI-Generated Assessment

Item Set	*Cronbach's Alpha*	*95% CI*
Complete assessment (all items)	0.78	[0.72, 0.84]
Scientific principles items	0.72	[0.65, 0.79]
Research ethics items	0.81	[0.75, 0.87]
Research methods items	0.76	[0.68, 0.83]
Sampling techniques items	0.75	[0.68, 0.82]

As presented in Table 2, the complete AI-generated assessment demonstrated acceptable internal consistency (α = .78, 95% CI [0.72, 0.84]). Among the content domains, research ethics items showed the highest reliability (α = .81, 95% CI [0.75, 0.87]), followed by research methods items (α = .76, 95% CI [0.68, 0.83]) and sampling techniques items (α = .75, 95% CI [0.68, 0.82]). Scientific principles items demonstrated the lowest, though still acceptable, reliability (α = .72, 95% CI [0.65, 0.79]).

Test-retest reliability was calculated using Pearson correlation coefficients between performances on identical items across the two testing sessions (separated by two weeks). The overall test-retest reliability coefficient was r = .74 (p < .001), indicating moderate stability over time.

Item analysis revealed that 6 items (7.5%) had low item-total correlations (r < .20), suggesting these items may be measuring constructs different from the rest of the test. Four of these problematic items were in the research methods domain, and two were in the sampling techniques domain.

The standard error of measurement (SEM) for the complete assessment was 2.84 points, indicating that a student's true score would fall within ±5.57 points (95% confidence interval) of their observed score.

4.3. Item Difficulty and Discrimination Analysis

Item difficulty indices (p-values) and discrimination indices (point-biserial correlations) were calculated for all AI-generated items. Table 3 summarizes the distribution of items across difficulty and discrimination categories.

As shown in Table 3, the majority of AI-generated items (47 items, 58.75%) fell within the medium difficulty range (.25 ≤ p ≤ .75), while 22 items (27.5%) were classified as easy (p > .75) and 11 items (13.75%) as difficult (p < .25). Regarding discrimination indices, 18 items (22.5%) demonstrated excellent discrimination (rpb > 0.40), 32 items (40%) showed good discrimination (0.30 ≤ rpb ≤ 0.40), 21 items (26.25%) exhibited fair discrimination (0.20 ≤ rpb < 0.30), and 9 items (11.25%) had poor discrimination (rpb < 0.20).

The mean difficulty index across all items was 0.62 (SD = 0.18), indicating a moderate level of difficulty overall. The mean discrimination index was 0.32 (SD = 0.12), suggesting good average discriminating power.

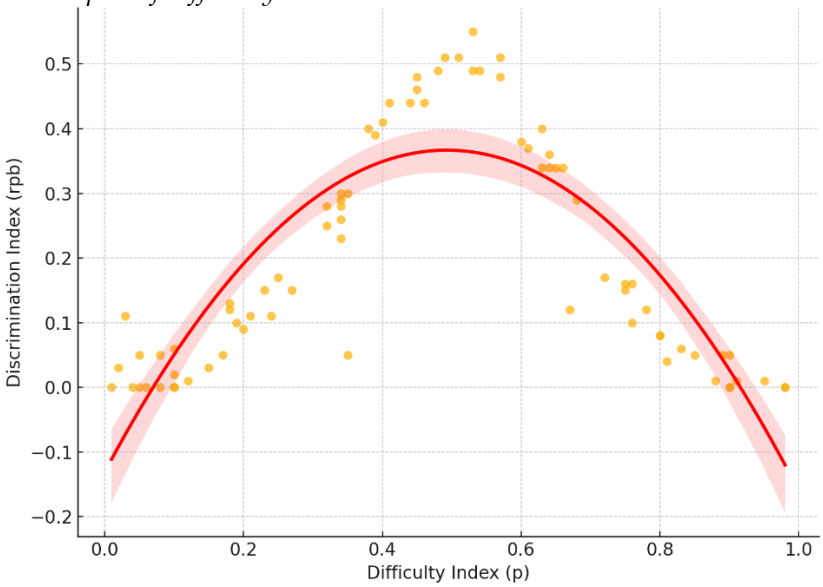
Distractor analysis revealed that 18 items (22.5%) contained at least one non-functioning distractor (selected by fewer than 5% of students). The most common issue was the presence of distractors that were either too implausible or too similar to the correct answer.

Table 3
Distribution of AI-Generated Items by Difficulty and Discrimination Indices

Classification	*Number of Items*	*Percentage*
Difficulty Index		
Easy ($p > .75$)	22	27.5%
Medium ($.25 \leq p \leq .75$)	47	58.75%
Difficult ($p < .25$)	11	13.75%
Discrimination Index		
Excellent (rpb > 0.40)	18	22.5%
Good (0.30 ≤ rpb ≤ 0.40)	32	40%
Fair (0.20 ≤ rpb < 0.30)	21	26.25%
Poor (rpb < 0.20)	9	11.25%

Figure 1 illustrates the relationship between item difficulty and discrimination indices for AI-generated items. The scatter plot reveals a curvilinear relationship typical of well-functioning test items, with items of moderate difficulty (p ≈ 0.5) showing the highest discrimination indices.

Figure 1
Scatter plot of difficulty and discrimination index



4.4. Cognitive Level Analysis

Expert evaluators classified each AI-generated question according to Bloom's revised taxonomy. Table 4 presents the distribution of items across cognitive levels.

Table 4
Distribution of AI-Generated Items across Bloom's Taxonomy Levels

Cognitive Level	*Number of Items*	*Percentage*
Remembering	31	38.75%
Understanding	28	35%
Applying	14	17.5%
Analyzing	5	6.25%
Evaluating	2	2.5%
Creating	0	0%

Table 4 shows that the majority of AI-generated questions were classified at lower cognitive levels, with 31 items (38.75%) at the remembering level and 28 items (35%) at the understanding level. Together, these lower-order thinking skills accounted for 73.75% of all questions. Medium-level cognitive questions (applying) comprised 14 items (17.5%), while higher-order thinking

questions were minimally represented: 5 items (6.25%) at the analyzing level, 2 items (2.5%) at the evaluating level, and no items (0%) at the creating level.

Chi-square analysis revealed a significant difference in the distribution of cognitive levels between AI-generated items and ideal proportions for undergraduate assessment (χ^2 = 18.42, df = 5, p < .01). The AI-generated items showed overrepresentation at lower cognitive levels (remembering, understanding) and underrepresentation at higher levels (analyzing, evaluating, creating).

Analysis of the relationship between cognitive level and psychometric properties revealed that items at the "applying" level had the highest mean discrimination index (M = 0.38, SD = 0.09), while items at the "remembering" level had the lowest (M = 0.28, SD = 0.13). This difference was statistically significant ($F(3, 76) = 4.83$, $p < .01$).

5. Discussion

5.1. Content Validity of AI-Generated Psychology MCQs

The results of the content validity analysis provide cautiously optimistic support for the use of AI-generated multiple-choice questions in undergraduate psychology assessment. With mean expert ratings ranging from 3.62 to 4.21 across dimensions and content validity indices between 0.72 and 0.88, the AI-generated items demonstrated reasonable alignment with learning objectives and appropriate conceptual coverage for an experimental psychology course. These findings align with Doughty et al. (2024), who reported that 81.7% of AI-generated programming MCQs passed all quality criteria, though our results show somewhat lower overall validity indices.

However, the identification of 17.5% of items requiring substantial revision highlights important limitations in the AI's ability to consistently generate valid psychology assessment items. The lower ratings for "quality of distractors" suggest that plausible but incorrect options remain challenging for AI systems to generate, consistent with the findings of Mead and Zhou (2023), who reported that 46.7-48.9% of AI-generated certification exam items violated multiple-choice principles. Expert qualitative feedback highlighted the AI's strength in creating clearly worded questions and appropriate use of psychological terminology, suggesting that language models have effectively incorporated domain-specific vocabulary. This aligns with Mistry et al.'s (2024) observation that GPT-4 demonstrated sophisticated domain knowledge in generating medical assessment items. However, the conceptual oversimplifications noted by our experts echo concerns raised by Ngo et al. (2024), who found that 25% of ChatGPT-generated immunology questions contained incorrect or misleading information.

The topical distribution of AI-generated questions (scientific principles, research ethics, research methods, and sampling techniques) demonstrated adequate coverage of experimental psychology domains, suggesting that the AI model successfully extracted the conceptual structure of the discipline from its training data. This balanced coverage is promising for the potential use of AI in comprehensive course assessment, though the need for human verification remains evident in the proportion of questions requiring revision. Beyond quantitative psychometric indices, a qualitative comparison suggests that while AI-generated items demonstrate clear structure and coverage of core concepts, they often lack the creative nuance and contextual depth found in faculty-created items. This aligns with expert comments noting that human-crafted questions tend to embed subtle scenario variations and discipline-specific perspectives that AI outputs may not fully replicate without targeted refinement. Given the consistent issues with distractor plausibility, it would be beneficial to implement a targeted post-generation human review process that focuses specifically on evaluating and refining distractors before final item use. Additionally, future studies could explore automated strategies, such as adversarial prompting or using secondary AI checks, to systematically test and improve the plausibility of distractors during item generation.

5.2. Reliability of AI-Generated Psychology Assessments

The reliability analysis yielded generally favorable results, with an overall Cronbach's alpha of .78 for the complete assessment, indicating acceptable internal consistency. This is comparable to O's (2024) finding of a .78 reliability coefficient for AI-human-made test forms in TESOL education, suggesting that AI-generated psychology assessments can achieve reliability standards similar to those in other disciplines. The domain-specific alpha coefficients further indicate consistent performance across content areas, with research ethics items showing the highest internal consistency. The test-retest reliability coefficient demonstrated moderate stability over time, which is promising for the potential use of AI-generated items in summative assessment where score consistency is important. However, this coefficient is somewhat lower than typically desired for high-stakes testing, suggesting that further refinement of AI-generated items may be necessary for such applications. The standard error of measurement indicates reasonable precision for classroom assessment purposes but may be too large for standardized testing contexts.

The identification of items with low item-total correlations (7.5% of the total) suggests that while the vast majority of AI-generated items contributed coherently to the measurement of relevant psychological constructs, a small proportion assessed distinct or unrelated content. This percentage is lower than the 25% of items found to be problematic in Nasution's (2023) evaluation of AI-generated biology MCQs, suggesting potentially better performance in psychology content. The concentration of problematic items in research methods and sampling techniques domains may reflect challenges in the AI's conceptualization of methodological nuances versus more straightforward factual content.

5.3. Item Difficulty and Discrimination Properties

The analysis of item difficulty and discrimination indices revealed a generally favorable distribution of psychometric properties among AI-generated psychology MCQs. With 58.75% of items falling in the medium difficulty range and 62.5% showing good to excellent discrimination, the majority of AI-generated items demonstrated appropriate characteristics for classroom assessment. The mean difficulty index of 0.62 indicates a slight tendency toward easier items, which may be appropriate for undergraduate formative assessment but potentially insufficient for advanced course examinations. The curvilinear relationship observed between item difficulty and discrimination aligns with classical test theory expectations, suggesting that AI-generated items follow typical psychometric patterns. This finding is consistent with O's (2024) comparison of AI-human-made and human-made test forms, which found comparable psychometric properties between the two sources despite some differences in difficulty levels. The presence of non-functioning distractors in 22.5% of items represents a notable limitation in the AI's ability to generate plausible incorrect options that effectively discriminate between knowledge levels. This finding echoes Ahmed et al.'s (2024) observation that AI-generated dental caries MCQs frequently contained format errors (62% for ChatGPT) and logic issues with distractors, suggesting a common challenge across disciplines. The tendency to create either implausible distractors or options too similar to the correct answer indicates a need for human review and refinement in this aspect of MCQ design. The proportion of items with poor discrimination (11.25%) is relatively low compared to Nasution's (2023) finding that 19% of AI-generated biology MCQs had poor discrimination, suggesting potential advantages in psychology content generation. However, this still represents a substantial number of items that would require revision before use in formal assessment.

5.4. Cognitive Level Distribution and Alignment

The cognitive level analysis revealed a significant imbalance in the distribution of AI-generated items across Bloom's taxonomy, with 73.75% of questions targeting lower-order thinking skills (remembering and understanding) and only 8.75% addressing higher-order skills (analyzing and evaluating). This finding aligns with Ali and Talat's (2024) systematic review, which noted that

most AI-generated MCQs targeted lower Bloom's taxonomy levels, with fewer questions assessing higher-order reasoning skills. The complete absence of items at the "creating" level and near-absence at the "evaluating" level (2.5%) highlight a significant limitation in the AI's current capacity to generate assessment items for the highest cognitive domains. This finding is particularly relevant for psychology education, where critical thinking and evaluative reasoning are essential disciplinary competencies. Unlike Sihite et al.'s (2023) finding that ChatGPT 3.5 could generate a substantial proportion of higher-order thinking questions (40% analysis, 30% evaluation, 30% creation), our results suggest more limited capabilities in the experimental psychology domain.

The relationship between cognitive level and discrimination indices with "applying" level items showing the highest discrimination suggests that mid-level cognitive questions may be optimal for distinguishing between student knowledge levels in psychology assessment. This finding has implications for the strategic use of AI-generated items, potentially focusing AI generation on middle-tier cognitive questions where it performs well while reserving human expertise for creating higher-order assessment items. One likely cause for this cognitive level imbalance may be limitations in the initial prompt structure, which emphasized coverage of fundamental concepts but did not provide sufficiently detailed instructions for generating complex scenario-based or evaluative questions. This reflects a broader constraint of large language models, which tend to default to factual recall unless specifically guided otherwise. Addressing this limitation may require more sophisticated prompt engineering that explicitly instructs the AI to produce higher-order questions, such as case analyses, critical evaluations, or application scenarios.

5.5. Implications for Psychology Education and Assessment

The findings of this study have several important implications for the integration of AI-generated MCQs in psychology education. First, while AI can produce reasonably valid and reliable assessment items, human expert review remains essential, particularly for verifying conceptual accuracy and enhancing distractor quality. This aligns with Cheung et al.'s (2023) conclusion that AI-generated medical graduate exam MCQs required human oversight despite comparable overall quality to human-created items. Second, AI-generated psychology assessments demonstrate particular strength in creating clearly worded, terminologically appropriate items at the remembering and understanding levels. This suggests that AI could effectively supplement human assessment development for formative assessment, knowledge checks, and practice materials, potentially freeing educator time for developing higher-order assessment tasks. Third, the cognitive level imbalance indicates that current AI models have limited capacity to assess the full spectrum of learning objectives in psychology education. Educators implementing AI-generated items should be mindful of this limitation and either provide specific cognitive-level prompting (as suggested by Doughty et al., 2024) or supplement AI-generated items with human-created questions targeting higher cognitive domains. Fourth, the reliability data suggest that AI-generated psychology assessments can achieve acceptable internal consistency and moderate stability over time, making them potentially suitable for classroom assessment but requiring caution for high-stakes applications. The standard error of measurement indicates reasonable precision for formative purposes but may be insufficient for summative evaluation without item refinement.

To mitigate this gap, educators may consider developing hybrid human-AI item generation workflows where AI is used primarily to draft foundational knowledge questions, while expert faculty refine these drafts or create complementary items that target higher-order skills. Additionally, iterative prompt refinement, such as embedding explicit instructions for generating analysis and evaluation items, could help balance the cognitive distribution in future AI-generated assessments. The integration of AI-generated items alongside faculty-created and test bank questions also raises important considerations for test fairness. Consistency in style, cognitive demand, and contextual framing must be carefully managed to ensure that students are assessed

equitably, regardless of item source. This highlights the need for systematic review to harmonize AI and human-generated content within a single assessment instrument.

5.6. Limitations and Future Research Directions

Several limitations of this study should be acknowledged. First, the research focused on a single AI model (ChatGPT-4) and a specific prompt structure, which limits generalizability to other AI systems or prompting strategies. Future research should compare multiple AI models and explore optimal prompting techniques specifically for psychology assessment generation.

Second, the study examined MCQs for a specific undergraduate psychology course (Experimental Psychology), and findings may not generalize to other psychology subfields with different conceptual structures and cognitive demands. Additional research across diverse psychology domains would enhance understanding of AI capabilities in this discipline.

Third, while the methodology included both expert evaluation and student performance data, it did not directly compare AI-generated items with human-created items on the same topics and cognitive levels. Future studies should implement more controlled comparisons to isolate the effects of item source on psychometric properties.

Fourth, the current study did not explore the potential benefits of human-AI collaboration in item generation, where human experts refine AI-generated items. Research on such hybrid approaches could yield insights into optimal workflows for assessment development.

Fifth, this study did not investigate students' perceptions of and experiences with AI-generated assessments. Future research should examine how students interact with these items and whether they perceive differences between AI and human-created assessments.

6. Conclusion

This study provides empirical evidence regarding the psychometric properties of AI-generated multiple-choice questions in undergraduate psychology education. The findings suggest that AI can generate MCQs with reasonable content validity, acceptable reliability, and generally appropriate difficulty and discrimination characteristics, but with significant limitations in assessing higher-order cognitive skills. The results highlight both the potential of AI as a supplementary tool in psychology assessment development and the continued importance of human expertise in ensuring comprehensive evaluation of psychological knowledge and skills.

The study contributes to the emerging understanding of AI applications in educational assessment by providing domain-specific insights for psychology education, where complex conceptual understanding and critical thinking are particularly valued. As AI technology continues to evolve, ongoing research on its educational applications will be essential for developing evidence-based practices that leverage technological capabilities while maintaining rigorous academic standards. Future research and practice should explore refined prompt engineering strategies and collaborative human-AI workflows to enhance the AI's capacity for producing higher-order thinking items. By combining AI's efficiency with human expertise in crafting complex, discipline-specific assessments, the field can better address current limitations in cognitive alignment. Future implementations should also examine the fairness and coherence of combining AI-generated and human-crafted items in the same assessment, to avoid unintended biases or inconsistencies that could affect test validity and student perceptions. For psychology educators, a balanced approach that strategically integrates AI-generated assessment items while preserving human oversight appears to be the most promising path forward.

Data availability: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declaration of interest: No conflict of interest is declared by author.

Ethics statement: All participants provided informed consent prior to their involvement in the study. They were informed about the study's purpose, procedures, and their right to withdraw at any time without consequence.

Funding: No funding was obtained for this study.

References

Ahmed, W. M., Azhari, A. A., Alfaraj, A., Alhamadani, A., Zhang, M., & Lu, C.-T. (2024). The quality of ai-generated dental caries multiple choice questions: A comparative analysis of chatgpt and google bard language models. *Heliyon*, *10*(7), e28198. https://doi.org/10.1016/j.heliyon.2024.e28198

Al Mashagbeh, M., Dardas, L., Alzaben, H., & Alkhayat, A. (2024). Comparative analysis of artificial intelligence-driven assistance in diverse educational queries: ChatGPT vs. Google Bard. *Frontiers in Education*, *9*, 1429324. https://doi.org/10.3389/feduc.2024.1429324

Ali, F., & Talat, H. (2024). Ai integration in mcq development: Assessing quality in medical education: a systematic review. *Life and Science*, *5*(3), 14. https://doi.org/10.37185/LnS.1.1.643

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.

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, *15*(3), ep430. https://doi.org/10.30935/cedtech/13176

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T.-H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong, Singapore, Ireland, and the United Kingdom). *PLOS ONE*, *18*(8), e0290691. https://doi.org/10.1371/journal.pone.0290691

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2024). A comparative study of ai-generated (GPT-4) and human-crafted mcqs in programming education. In N. Herbert & C. Seton (Eds.), *Proceedings of the 26th Australasian Computing Education Conference* (pp. 114-123). Association for Computing Machinery. https://doi.org/10.1145/3636243.3636256

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In S. H. Irvine, & P. C. Kyllonen (Eds.), *Educational measurement* (4th ed., pp. 471-516). American Council on Education.

Kic-Drgas, J., & Kılıçkaya, F. (2024). Using Artificial Intelligence (Ai) to create language exam questions: A case study. *XLinguae*, *17*(1), 20–33. https://doi.org/10.18355/XL.2024.17.01.02

Kıyak, Y. S., & Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. *Postgraduate Medical Journal*, *100*(1189), 858–865. https://doi.org/10.1093/postmj/qgae065

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (Efa, cfa) and sem in general. *Psychology*, *09*(08), 2207–2230. https://doi.org/10.4236/psych.2018.98126

Mead, A. D., & Zhou, C. (2024). Evaluating the quality of ai-generated items for a certification exam. *Journal of Applied Testing Technology*. Advance online publication. https://jattjournal.net/index.php/atp/article/view/173204

Mistry, N. P., Saeed, H., Rafique, S., Le, T., Obaid, H., & Adams, S. J. (2024). Large language models as tools to generate radiology board-style multiple-choice questions. *Academic Radiology*, *31*(9), 3872–3878. https://doi.org/10.1016/j.acra.2024.06.046

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

Ngo, A., Gupta, S., Perrine, O., Reddy, R., Ershadi, S., & Remick, D. (2024). ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Academic Pathology*, *11*(1), 100099. https://doi.org/10.1016/j.acpath.2023.100099

O, K.-M. (2024). A comparative study of AI-human-made and human-made test forms for a university TESOL theory course. *Language Testing in Asia*, *14*(1), 19. https://doi.org/10.1186/s40468-024-00291-3

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Sihite, M. R., Meisuri, M., & Sibarani, B. (2023). Examining the validity and reliability of Chatgpt 3. 5-generated reading comprehension questions for academic texts. *Randwick International of Education and Linguistics Science Journal*, *4*(4), 937–944. https://doi.org/10.47175/rielsj.v4i4.835

Van de Poll, J., Yong, Y., & van de Weijer, N. (2022). *Validity of AI-generated one-off questionnaires for organizational transformation. International Journal of Economics, Business and Management Research, 6*(2), 122–130. https://doi.org/10.51505/IJEBMR.2022.6208

Appendix 1. Prompt Used to Generate AI Multiple-Choice Questions

The following prompt was used to generate the AI-created multiple-choice questions for the Experimental Psychology examination:
(After uploading the specific module of topics to the chatbot) I need you to create multiple-choice questions for a college-level Experimental Psychology exam (Prelim Examination for the 2nd Semester, AY 2024-2025). Please generate 80 high-quality multiple-choice questions with the following specifications:

Topic Distribution:
- 20 questions on scientific principles and the scientific method in psychology
- 15 questions on research ethics in psychology
- 25 questions on qualitative research methods (naturalistic observation, case studies, archival research, etc.)
- 20 questions on sampling techniques and methodologies

For each question:
1. Create a clear stem that addresses a specific concept or application in experimental psychology
2. Provide four answer options (a, b, c, d) with only one correct answer
3. Ensure that incorrect options (distractors) are plausible but clearly incorrect
4. Include some application and scenario-based questions, not just factual recall
5. Format each question with the stem followed by four lettered options

The questions should be appropriate for undergraduate psychology students who have completed half a semester of an Experimental Psychology course. Include questions that assess understanding of key concepts, application of research principles, and critical thinking about psychological methods.

The final exam document should include:
- An appropriate header with course information (PSY06 -- EXPERIMENTAL PSYCHOLOGY (LECTURE))
- Exam identification (Prelim Examination, 2nd Semester, AY 2024-2025)
- Sequential numbering of all questions
- A footer section for faculty information

Please create questions that vary in difficulty, with approximately:
- 25% easy questions (basic recall and comprehension)
- 60% moderate questions (application and analysis)
- 15% challenging questions (evaluation and synthesis)