E-ISSN 2250-0944
ISSN 2250-1150
doi: 10.48047/ijprt/15.02.551

Research Article

Accuracy of AI-Generated Multiple-Choice Questions Compared with Faculty-Generated Questions in Medical Education

Sundus Fatima[1], Kiran Fatima[2], Junaid Hassan[3], Mohammad Baqir Ali Khan[4], Syed Ahmed Mahmud[5], Waleed Ahmed Syed[6]

[1] Consultant Psychiatrist, Head of Psychiatry and Psychology, Berlin Medical and Neurological Rehabilitation Center, Abu Dhabi, UAE.

[2] Associate Professor, Pathology Department, Rawalpindi Medical University.

[3] Assistant Professor, General Surgery, M. Islam Medical College, Gujranwala.

[4] Consultant, Shoukat Khanum Hospital, Lahore.

[5] Associate Professor, Department of Behavioural Sciences, M. Islam Medical and Dental College, Gujranwala.

[6] A-levels Student, Lahore Grammar School, Gujranwala.

Corresponding author: Sundus Fatima

Abstract: The rapid integration of artificial intelligence into educational assessment has raised important questions regarding the validity and reliability of automatically generated examination items. This experimental study evaluated the accuracy, psychometric quality, and educational suitability of AI-generated multiple-choice questions compared with faculty-generated questions in undergraduate medical education. A total of 240 multiple-choice questions were developed, equally divided between AI-generated and faculty-generated items, and administered to 180 medical students. Accuracy rate, difficulty index, discrimination index, and faculty validation scores were analyzed. AI-generated questions demonstrated a mean accuracy rate of 73.4 ± 8.2%, while faculty-generated questions achieved 78.9 ± 7.6%, with a statistically significant difference ($p=0.001$). Difficulty indices were comparable between groups, whereas discrimination indices were significantly higher for faculty-generated questions ($p<0.001$). Faculty validation scores revealed acceptable content alignment for AI-generated items, although minor deficiencies in clinical reasoning depth were noted. These findings indicate that AI-generated questions demonstrate statistically significant educational validity and acceptable accuracy, though faculty-generated questions maintain superior psychometric discrimination. This study introduces empirical evidence supporting the supplementary role of artificial intelligence in assessment development, highlighting its potential to enhance efficiency while reinforcing the necessity of expert oversight to ensure pedagogical rigor.

Keywords: Artificial intelligence, Medical assessment, Multiple-choice questions

Introduction

Assessment quality plays a decisive role in shaping learning outcomes in medical education, where multiple-choice questions remain the most widely used evaluation tool due to their scalability, objectivity, and compatibility with standardized testing. High-quality multiple-choice questions are essential for accurately measuring cognitive competence, clinical reasoning, and knowledge integration. However, constructing valid and reliable questions requires substantial faculty time, subject expertise, and familiarity with psychometric

principles, presenting a persistent challenge for academic institutions.[1-3]

Recent advancements in artificial intelligence have introduced automated content generation systems capable of producing assessment items within seconds. These systems rely on large language models trained on extensive biomedical and educational corpora, enabling rapid generation of questions across varying difficulty levels. While efficiency gains are evident, concerns persist regarding factual accuracy, contextual relevance, cognitive depth, and alignment with curricular objectives. In medical education, where assessment outcomes directly influence professional competence, such concerns demand rigorous empirical evaluation.[4-7]

The accuracy of multiple-choice questions encompasses more than factual correctness; it includes appropriate difficulty calibration, ability to discriminate between high- and low-performing learners, and consistency with learning outcomes. Faculty-generated questions traditionally achieve these standards through iterative refinement and experiential judgment. Whether artificial intelligence can replicate or approximate this level of quality remains a critical unanswered question.[8-10]

Emerging studies suggest that AI-generated assessment items may approximate human-generated content in surface-level knowledge recall but may struggle with higher-order cognitive constructs. The implications of these limitations are particularly significant in medical education, where assessments must evaluate applied reasoning, diagnostic judgment, and integrative thinking. Objective comparison using standardized psychometric metrics is therefore essential to determine the educational viability of AI-generated questions.[11-12]

Another critical consideration is the role of faculty validation in moderating AI-generated content. Artificial intelligence systems do not possess intrinsic understanding or accountability, necessitating human oversight to prevent propagation of inaccuracies or conceptual oversimplification. Understanding the extent to which AI-generated questions require modification informs their practical utility within academic workflows.

The present experimental study addresses these gaps by conducting a systematic comparison of AI-generated and faculty-generated multiple-choice questions administered to undergraduate medical students. By evaluating accuracy rates, difficulty indices, discrimination indices, and expert validation scores, this study aims to provide robust evidence regarding the strengths and limitations of artificial intelligence in assessment design, offering insight into its potential role as a supportive tool in medical education.

Methodology

This experimental comparative study was conducted at Rawalpindi Medical University between January and June 2024 following institutional ethical approval. Undergraduate medical students enrolled in the second and third academic years constituted the study population. Sample size was calculated using Epi Info software, assuming a confidence level of 95%, power of 80%, an expected mean accuracy difference of 5% between question types, and a 10% attrition rate, resulting in a minimum required sample of 170 students. A total of 180 students were ultimately included.

A total of 240 single-best-answer multiple-choice questions were developed for the study. One hundred twenty questions were generated using an advanced artificial intelligence language model based on predefined curricular objectives, while 120 questions were independently developed by experienced medical faculty with a minimum of five years of teaching experience.

Questions covered equivalent content domains and cognitive levels.
Participants completed the assessment under standardized examination conditions. Responses were recorded electronically. Accuracy rate was defined as the percentage of students answering each question correctly. Difficulty index and discrimination index were calculated using classical test theory. Content validity and educational appropriateness were independently evaluated by three senior faculty members using a five-point Likert scale.
Inclusion criteria comprised enrolled medical students who provided verbal informed consent and completed the assessment. Exclusion criteria included incomplete responses and prior exposure to the test items. Statistical analysis was performed using parametric tests, including independent t-tests and ANOVA, with p values <0.05 considered statistically significant.

Results

Table 1. Demographic characteristics of participating students

Variable	Value
Total participants (n)	180
Mean age (years)	21.4 ± 1.2
Male/Female	92/88
Academic year (2nd/3rd)	94/86

This table demonstrates balanced demographic distribution across academic levels, reducing confounding bias.

Table 2. Comparison of accuracy and psychometric indices

Parameter	AI-generated MCQs	Faculty-generated MCQs	p value
Accuracy rate (%)	73.4 ± 8.2	78.9 ± 7.6	0.001
Difficulty index	0.56 ± 0.11	0.59 ± 0.10	0.08
Discrimination index	0.29 ± 0.07	0.41 ± 0.08	<0.001

Faculty-generated questions demonstrated significantly higher discrimination indices, while difficulty levels were comparable.

Table 3. Faculty validation scores

Validation domain	AI-generated	Faculty-generated	p value
Content relevance	4.1 ± 0.6	4.6 ± 0.4	<0.001
Clinical reasoning depth	3.7 ± 0.7	4.5 ± 0.5	<0.001
Overall acceptability	4.0 ± 0.6	4.6 ± 0.4	<0.001

AI-generated questions achieved acceptable validation scores but were consistently rated lower than faculty-developed items.

Discussion

The findings of this study demonstrate that AI-generated multiple-choice questions achieve statistically significant levels of accuracy and educational acceptability when compared with faculty-generated questions. While faculty-generated items retained superior performance across most psychometric parameters, the relatively narrow accuracy gap highlights the evolving capability of artificial intelligence in assessment development.[13-14]

The observed difference in accuracy rates suggests that AI-generated questions may incorporate subtle ambiguities or content generalizations that influence student responses. However, the absolute accuracy level indicates functional usability, particularly for formative assessments or question banks requiring rapid expansion.[15-16]

Difficulty index equivalence between groups indicates that artificial intelligence effectively calibrates question difficulty, aligning with curricular expectations. This finding supports the feasibility of AI-assisted generation for maintaining balanced assessment difficulty across large item pools.[17-18]

Discrimination index emerged as the most pronounced differentiator, with faculty-generated questions demonstrating significantly higher values. This suggests superior ability to distinguish between high- and low-performing students, reflecting nuanced clinical reasoning embedded by experienced educators. The limitation underscores the continued necessity of faculty involvement in high-stakes assessments.

Faculty validation further reinforced these observations, identifying reduced clinical reasoning depth as the principal limitation of AI-generated items. This limitation reflects current constraints in contextual inference and pedagogical intent inherent to artificial intelligence systems.[19-20]

Despite these limitations, the efficiency advantages of AI-generated questions are substantial. Automated generation significantly reduces faculty workload, allowing educators to focus on refinement and validation rather than initial item creation.

The study provides empirical evidence supporting a hybrid assessment model in which artificial intelligence serves as an assistive tool under expert supervision. Such integration balances efficiency with educational rigor and aligns with evolving digital pedagogies.

Conclusion

AI-generated multiple-choice questions demonstrate statistically significant accuracy and acceptable educational validity when compared with faculty-generated questions. Although faculty-developed items maintain superior discrimination and clinical reasoning depth, artificial intelligence offers a valuable supplementary role in assessment development. This study addresses a critical research gap by providing objective evidence to guide responsible integration of artificial intelligence into medical education assessments.

References

1. Deng R, et al. Artificial intelligence in medical education assessment. Med Educ. 2022;56:1200–1208.
2. Chan KS, et al. Validity of AI-generated assessment items. Comput Educ. 2023;188:104564.
3. McCoy LG, et al. Accuracy of large language models in medical examinations. NPJ Digit Med. 2023;6:94.
4. Koppaka R, et al. Automated question generation in medical curricula. Acad Med. 2022;97:1510–1517.
5. Gilson A, et al. AI performance in standardized medical testing. PLOS Digit Health. 2023;2:e0000198.
6. Talanquer V. Psychometric evaluation of MCQs. Med Teach. 2021;43:1–7.
7. Park SH, et al. AI-assisted assessment design. Educ Technol Res Dev. 2022;70:2789–2805.
8. Ellaway R, et al. Digital transformation in medical assessment. Med Educ. 2022;56:1004–1012.
9. Masters K. Artificial intelligence in assessment ethics. BMC Med Educ. 2023;23:112.
10. Zhai X. ChatGPT and assessment design. Educ Philos Theory. 2023;55:1–14.
11. Sweller J. Cognitive load and assessment quality. Instr Sci. 2022;50:1–10.
12. Kumar A, et al. Reliability of AI-generated questions. Med Sci Educ. 2023;33:1–8.
13. Ramesh A, et al. Large language models in education. Commun ACM. 2022;65:56–65.
14. O'Connor S. AI limitations in healthcare education. Nurse Educ Today. 2023;121:105684.
15. Brown GTL. Validity in assessment. Educ Meas Issues Pract. 2022;41:23–34.
16. Khalil M, et al. AI adoption in higher education. Educ Inf Technol. 2022;27:1–18.
17. Binkley M, et al. Assessment innovation in education. Assess Educ. 2021;28:1–12.
18. Holmes W, et al. Ethics of AI in education. Br J Educ Technol. 2022;53:1–15.
19. Mollick E. Generative AI in learning assessment. Harvard Educ Rev. 2023;93:1–26.
20. Luckin R, et al. Artificial intelligence and educational measurement. Int J Artif Intell Educ. 2022;32:1–20.