Open Access

Original Article

ChatGPT In Assessment: A Comparison of Artificial Intelligence (AI)-Generated and Human-Generated One Best-Type Multiple Choice Questions

Muhammad Moeed Haidar Naqvi, Masood Jawaid*, Syeda Tahseen Fatima, Asad Ali***, Tayyaba Azhar**

Department of Medical Education, Azra Naheed Medical College, The Superior University, Lahore Pakistan, *Department of Medical Education, Pharm Evo, Karachi Pakistan, **Department of Research and Biostatistics, University College of Medicine and Dentistry, University of Lahore, Pakistan, ***Department of Medical Education, Multan Medical and Dental College, Multan Pakistan

ABSTRACT

Objective: To compare the quality of Multiple-Choice Questions (MCQs) created by humans and Chat GPT using post-hoc item analysis.

Study Design: Quasi-Experimental Research Design.

Place and Duration of Study: Department of Medicine, Azra Naheed Medical College, The Superior University, Lahore Pakistan, from Feb to Aug 2024.

Methodology: A total of 100 participants were given two sets of 50 questions. One set was created using Chat GPT while the other was crafted by human teachers. Duration of generation of MCQs was noted. Post hoc item analysis was done (difficulty index, discrimination index and distractor analysis). Student scores were compared. Independent sample t-test was applied to compare mean difficulty index, discrimination index and student scores. Two proportions z-test to compare mean distractor efficiency. The results were reported as means, frequencies and percentages and analyzed using SPSS version 27.

Results: Total of 100 MCQs were evaluated. Chat GPT generated MCQs in 25 minutes, significantly faster than humans' 205 minutes. The Difficulty Index (Chat GPT: 30.3, humans: 34.7, p=0.39), Discrimination Index (Chat GPT: 0.21, humans: 0.19, p=0.514), student scores (Chat GPT: 19.4, humans: 18.46, p=0.125) and Distractor efficiency (Chat GPT: 75%, humans 69.5%, p=0.385) were statistically non-significant, indicating similar quality.

Conclusion: The quality of MCQs produced by Chat GPT and Human Educators is comparable; however, the duration required by Chat GPT is significantly less. A mixed approach, where subject matter experts qualitatively evaluate the MCQs generated by Chat GPT, can enhance the overall assessment quality.

Keywords: Artificial Intelligence, Educational Measurement, Educational Technology, Medical Education, Machine Learning.

How to Cite This Article: *Naqvi MMH, Jawaid M, Fatima ST, Ali A, Azhar T. ChatGPT in Assessment: A Comparison of Artificial Intelligence (AI)-Generated and Human-Generated One Best-Type Multiple Choice Questions. Pak Armed Forces Med J 2026; 76(6): 822-826.*

DOI: https://doi.org/10.51253/pafmj.v76i6.12930

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

University medical staff, burdened with responsibilities like patient care, teaching, evaluation, research, and administration, face a significant workload challenge, especially in maintaining rigorous standards in medical exams. Faculty although trained but still fall short of creating high quality MCQs.[1,2] Multiple-choice questions (MCQs) have long been essential for knowledge-based assessments due to their consistency and efficiency.[3,4] However, creating high-quality MCQs is time-consuming. Artificial intelligence (AI), including ChatGPT, is revolutionizing education by efficiently addressing this challenge.[5,6] These AI tools, can create high-quality MCQs, providing feedback to students, and even grading exams, eliminating medical staff's burden and improving assessments. It is with this technology that medical education is poised to receive a further boost.[7,8]

Correspondence: Dr Muhammad Moeed Haidar Naqvi, Department of Medical Education, Azra Naheed Medical College, Lahore Pakistan

Received: 27 Nov 2024; revision received: 03 Jan 2025; accepted: 11 Jan 2025

Cheung *et al.*, showed that the effectiveness of ChatGPT in developing MCQs was significant. ChatGPT developed 50 MCQs in only 20 minutes and 25 seconds, while human experts took 211 minutes and 33 seconds to develop them.[9] When comparing question quality, ChatGPT was found to be slightly inferior to humans only in the domain of relevance. Therefore, in the case of generated questions, ChatGPT was observed to have a greater variation in scores compared to human-generated questions, where scores were observed to have a higher degree of coherence with a comparatively smaller range of variation.[9] However, as Khilnani notes, while using ChatGPT may be helpful, it cannot be completely relied upon. Issues include the absence of medical background, inability to appreciate context, inability to check answers, and inability to create graphics. This is why input from medical experts is essential while crafting MCQs.[10] The objective of the study was to compare the quality of MCQs created by human teacher to those generated by Chat GPT using item analysis.

METHODOLOGY

The quasi-experimental study was conducted in the Department of General Medicine, Azra Naheed Medical College, Lahore Pakistan, a constituent college of The Superior University. The study was conducted from February 2024 to August 2024 under approval of Ethical Review Board of the University College of Medicine and Dentistry and the University of Lahore (ERC 130/23/12) and Azra Naheed Medical College, The Superior University, Lahore (ANMC/ERB/7921-24). The sample size of 77 was calculated using the WHO sample size calculator based on expected scores of 40% for AI-generated MCQs and 60% for human-generated MCQs derived by a study done by Cheung et al. Their study reported slightly lower performance in AI-generated MCQs, primarily due to variability in question quality and alignment with learning objectives.[9] The significance level was set at 5%, with a statistical power of 80% for the test. However, to ensure more robust findings a total of 100 final-year medical students was selected using a non-probability convenience sampling technique. Two assistant professors from the General Medicine department made the MCQs. The faculty members involved had five and six years of experience, had done Certificate in Health Professions Education and attended several workshops on MCQs development. The faculty members used Davidson's Principles and Practice of Medicine as the reference book. The principal investigator used Chat GPT to create MCQs.

Inclusion Criteria: Final-year medical students at Azra Naheed Medical College were included in the study.

Exclusion Criteria: Individuals who declined to participate and those who did not give consent were excluded.

MCQs were developed in two sets: 50 manually by faculty members and 50 generated using ChatGPT. (Figure) Both the MCQs were on the level of application and analysis of bloom's taxonomy. Human-generated MCQs (Paper I) were created manually without AI tools, while AI-generated MCQs (Paper II) utilized ChatGPT with topic-specific prompts to ensure unbiased content generation. Each MCQ had five options with one key and four distractors, which totals to 200 distractors in each set of 50 MCQs. The topics were from endocrinology, gastroenterology, and cardiology. The topics were kept the same for both type of MCQs. Distinct windows and a new account were used for each MCQ generation to eliminate reinforcement learning of Chat GPT. All questions by humans and ChatGPT, were retained. Pre hoc analysis was not done meaning no initial review or filtering process was done by subject matter experts aimed at identifying the poor question construction, mistakes, and relevance before testing them on students, this was done to determine the inherent capabilities of Chat GPT in MCQ generation. The study included a blinding strategy, ensuring neither students nor invigilators knew whether MCQs were AI or human generated. Half the students first attempted Paper I and then Paper II, while the other half reversed the order. Optical Mark Recognition (OMR) sheets were used to collect responses.

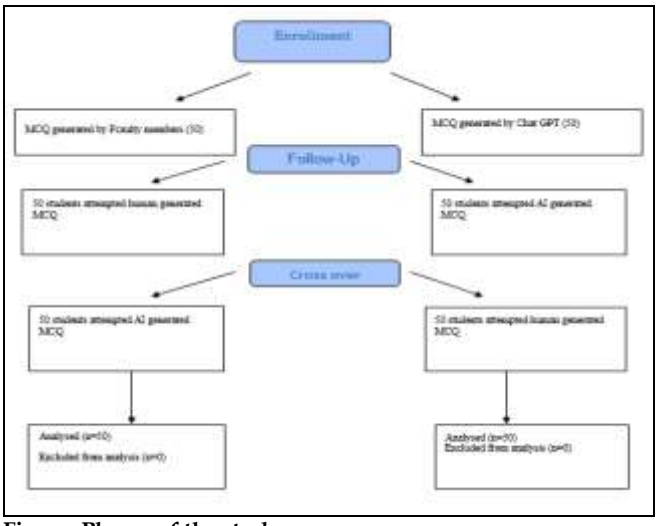

Figure: Phases of the study

The OMR sheets were scored using Zip Grade which also generated the item analysis report. Data was compiled and analyzed using Statistical Package for Social Sciences version 27.0. Numerical variables such as age, scores, difficulty and discrimination indices were summarized as mean ± standard deviation. Gender and distractor analysis was presented as frequencies and percentages. Independent sample t-tests were applied to compare scores, difficulty index and discrimination index between the two sets of papers. Z-test was applied to compare the mean distractor efficiency. The interpretation of difficulty index (P) is: P<30% denotes "Difficult," P between 30–70% is "Acceptable," and P>70% is "Easy." The discrimination index (D) is interpreted as: a negative D value identifies a defective item; D-value between 0–0.19 indicate "Poor discrimination"; D-values between 0.2–0.29 is "Acceptable discrimination"; D values between 0.3–

0.39 is "Good discrimination"; and D values exceeding 0.4 equals "Excellent discrimination". A distractor analysis to assess the efficiency of incorrect answer choices. A *p*-value of ≤0.05 was considered statistically significant.

RESULTS

Total of 100 students were included in the study. The mean age of the participants was 23.1+1.02 year with minimum and maximum value of 21 and 26. Most respondents were between 22 and 24 years old. There were 45% males and 55% females.

Duration of work by Chat GPT was 25 minutes and Human Teachers took 205 minutes.

While comparing the student performance on Chat GPT and human-created MCQs. Total score for each paper was 50. The mean score for Chat GPT was 19.4±4.71 and for humans 18.46±4.54. However, the t-test value (-1.54) and *p*-value (0.125) indicate no statistically significant difference, suggesting similar performance impact from both question sets.

Difficulty index of human-created MCQs show a balanced difficulty spread: 44% are "Difficult", 44% were "Acceptable", and 12% were "Easy". The data for Chat GPT paper indicated that 60% of the questions fall into the "Difficult" category, suggesting a significant proportion of challenging questions. Meanwhile, 28% of the questions were considered "Acceptable," and only 12% were deemed "Easy." This distribution highlights that most questions are of a higher difficulty, which could reflect the difficult nature of the MCQs as the questions were on the Bloom's level of application and analysis. The mean difficulty index for human-generated questions (34.7±27.1) was marginally higher than for Chat GPT questions (30.3±24.7). With a t-value of -0.85 and *p*-value of 0.39, the difference was not statistically significant, confirming comparable difficulty levels in both papers.

In terms of discrimination index 48% of human-generated MCQs exhibit "Poor discrimination", and 6% show negative discrimination (misleading items). Only 18% were "Acceptable", 22% were "Good", and 6% were "Excellent". These results reveal the need to revise poorly performing items to improve the discriminative ability of human-made tests. The Discrimination Index analysis of Chat GPT made MCQs showed that 44% exhibit "Poor discrimination", 34% were "Acceptable", 16% were "Good", and 6% were "Excellent". Nearly half of the questions show poor differentiation between high and low-performing students, signaling a need for improvement to enhance assessment validity. While comparing mean discrimination index between the two groups there was a slightly higher mean Discrimination Index for Chat GPT (0.21±0.10) compared to humans (0.19±0.13). The t-value of (0.65) and the non-significant *p*-value (0.514) indicates similar discriminative capacity.

Table: Summary of Key Findings

Index Type	Metric/Category	Chat GPT (n=50)	Human (n=50)
	Mean ± SD	30.3±24.7	34.7±27.1
	t value	-0.85	
	p-value (CI)	0.39 (-14.7 to 5.86)	
	Difficult (<30%)	30(60.0%)	22(44.0%)
	Acceptable (30 to 70%)	14(28.0%)	22(44.0%)
	Easy (>70%)	6(12.0%)	6(12.0%)
	Mean	0.21±0.1	0.19±0.13
	t value	0.65	
	p-value (CI)	0.514 (-0.032 to 0.064)	
	Negative (wrong item)	0 (0.0%)	3(6.0%)
	Poor discrimination (0 to 0.19)	22(44.0%)	24(48.0%)
	Acceptable (0.2 to 0.29)	17(34.0%)	9(18.0%)
	Good discrimination (0.3 to 0.39)	8(16.0%)	11(22.0%)
	Excellent (>0.4)	3(6.0%)	3(6.0%)
Student Scores Comparison	n	100	100
	Mean	19.4 ± 4.71	18.46 ± 4.54
	t value	-1.54	
	p-value (CI)	0.125 (-2.30 to 0.28)	
Distractor Analysis	n	200	200
	Mean distractor efficiency	75%	69.5%
	p-value	0.385	
Duration of generation of MCQs		25 minutes	205 minutes

Distractor efficiency was analyzed to assess the functionality of distractors. In Chat GPT-generated MCQs, 50 out of 200 distractors were non-functional, resulting in a mean distractor efficiency (MDE) of 75%. In contrast, human-created questions had 61 non-functional distractors, yielding an MDE of 69.5%. *p*-value of 0.385 was statistically non-significant. The findings suggest that Chat GPT-generated more functional distractors and is slightly more effective.

DISCUSSION

The findings of this study underscore the transformative potential of AI, specifically ChatGPT, in medical education assessments. ChatGPT rapidly generates MCQs, reducing the time and effort required by educators.[11-13] This efficiency allows educators to focus on patient care, clinical studies, and advanced teaching methodologies. However, concerns persist about the educational depth and alignment of AI-generated content with medical curricula's nuanced requirements.[14-16]

One key result is the higher difficulty level of ChatGPT-generated questions compared to human-

created ones. ChatGPT produced more "Difficult" questions, consistent with studies highlighting AI's tendency to create complex content.[17] While this showcases AI's capability to produce professional-grade assessments, it also emphasizes the need for calibration to match examinees' abilities. Prior studies confirm that AI must be programmed to balance complexity and accessibility.[9]

The discrimination index revealed substantial proportions of "Poor discrimination" items in both human and AI-generated questions, reflecting challenges in differentiating high- and low-performing students. However, ChatGPT-generated questions showed slightly higher proportions of "Good" and "Excellent" discrimination indices, suggesting an edge in capturing performance variations. This finding aligns with research supporting AI's capacity to produce assessments with valid discriminatory abilities.[18-20]

The lack of significant differences in overall scores between human and ChatGPT-generated questions reinforces AI's potential as a reliable adjunct in assessments.[6] However, human oversight remains essential to address inconsistencies, ensuring AI-generated questions align with clinical and ethical standards, particularly in high-stakes medical training.[21-22]

Another promising aspect is AI's scalability. ChatGPT can process large datasets, identify trends in student performance, and generate adaptive tests targeting weaknesses.[21] Despite this, limitations persist in fully personalized learning due to AI's inability to adapt to learning styles and emotional states.[18]

The hypothesis that AI-generated MCQs are comparable to human-created ones in difficulty, discrimination ability, and performance outcomes is supported by these findings. AI is a powerful tool that, and when integrated effectively, enhances assessments without compromising quality. Future research should refine AI-generated content and explore its impact on critical thinking and decision-making skills.[23] With strategic integration, AI can support educators in sustaining high standards in medical education while optimizing time and resources.[24]

ACKNOWLEDGMENT

Special thanks to Prof. Dr. Atif Qureshi, Prof. Dr. Muhammad Shoaib, and Dr. Zaid Tayyab for their support in the MCQ development.

CONCLUSION

AI-generated questions displayed greater complexity and discrimination, while human questions were balanced in difficulty. The integration of artificial intelligence with human oversight can enhance assessment methods, promoting equity, improving quality, and the evaluation of students' problem-solving skills.

Conflict of Interest: None.

Funding Source: None.

Authors' Contribution

Following authors have made substantial contributions to the manuscript as under:

MMHN & MJ: Data acquisition, data analysis, critical review, approval of the final version to be published.

STF & AA: Study design, data interpretation, drafting the manuscript, critical review, approval of the final version to be published.

TA: Conception, data acquisition, drafting the manuscript, approval of the final version to be published.

Authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

REFERENCES

1. Nassar A, Reid S, Kahnamoui K, Tuma F, Waheed A, McConnell M. Burnout among academic clinicians as it correlates with workload and demographic variables. Behav Sci 2020; 10(6): 94. https://doi.org/10.3390/bs10060094
2. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. physicians' working hours and lowers their career satisfaction. Int J Health Serv 2014; 44(4): 635-642. https://doi.org/10.2190/hs.44.4.a
3. Collignon SE, Chacko J, Wydick Martin M. An alternative multiple-choice question format to guide feedback using student self-assessment of knowledge. Decis Sci J Innov Educ 2020; 18(3): 456–480. https://psycnet.apa.org/doi/10.1111/dsji.12213
4. Siddiqui S. Categorized and correlated multiple-choice questions: A tool for assessing comprehensive physics knowledge of students. Educ Sci 2022; 12(9): 575. https://doi.org/10.3390/educsci12090575
5. Mondal H, Marndi G, Behera JK, Mondal S. ChatGPT for teachers: Practical examples for utilizing artificial intelligence for educational purposes. Indian J Vasc Endovasc Surg 2023; 10: 200-205. https://doi.org/10.4103/ijves.ijves_37_23
6. Nasution N. Using artificial intelligence to create biology multiple-choice questions for higher education. Agric Environ Educ 2023; 2(1): 4–8. https://doi.org/10.29333/agrenvedu/13071
7. Klang E, Portugez S, Gross R, Kassif Lerner R, Brenner A, Gilboa M, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ 2023; 23: 772. https://doi.org/10.1186/s12909-023-04752-w
8. Leo J, Kurdi G, Matentzoglu N, Parsia B, Sattler U, Forge S, et al. Ontology-based generation of medical, multi-term MCQs. Int J Artif Intell Educ 2019; 29: 145–188. https://doi.org/10.1007/s40593-018-00172-w

9. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). PLoS One 2023; 18(8): e0290691. https://doi.org/10.1371/journal.pone.0290691
10. Khilnani AK. Potential of large language model (ChatGPT) in constructing multiple-choice questions. GAIMS J Med Sci 2023; 3(2): 1–3. https://doi.org/10.5281/zenodo.7751267
11. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. Rev Esp Educ Med 2023; 4(3): 98-103. https://doi.org/10.6018/edumed.587451
12. Olney AM. Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. *Proceedings of the AIEDLLM1 Workshop: Empowering Education with LLMs*; 2023 Jul 7; Tokyo, Japan. Memphis, TN: Institute for Intelligent Systems, University of Memphis. 2023. Available from: https://olney.ai/category/2023/07/07/mcq.html
13. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus 2023; 15(6): e40977. https://doi.org/10.7759/cureus.40977
14. Ilgaz HB, Çelik Z. The significance of artificial intelligence platforms in anatomy education: An experience with ChatGPT and Google Bard. Cureus 2023; 15(9): e45301. https://doi.org/10.7759/cureus.45301
15. Rao C, Prasad HLK, Sajitha K, Permi H, Shetty J. Item analysis of multiple-choice questions: Assessing an assessment tool in medical students. Int J Educ Psychol Res 2016; 2: 201. https://doi.org/10.4103/2395-2296.189670
16. Javaid M, Haleem A, Singh RP, Khan S, Khan IH. Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. BenchCouncil Trans Benchmarks Stand Eval 2023; 3(2): 100115. https://doi.org/10.1016/j.tbench.2023.100115
17. Aya AH, Khalifa M, Abdelsalheen M, Afifi S. A systematic review of coastal vulnerability assessment methods in the Egyptian context: The case study. HBRC J 2023; 19(1): 161–181. https://doi.org/10.1080/16874048.2023.2241964
18. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019; 5(2): e16048. https://doi.org/10.2196/16048
19. Chin C, Osborne J. Students' questions: A potential resource for teaching and learning science. Stud Sci Educ 2008; 44(1): 1–39. https://doi.org/10.1080/03057260701828101
20. Halagatti M, Gadag S, Mahantshetti S, Hiremath CV, Tharkude D, Banakar V. Artificial intelligence: the new tool of disruption in educational performance assessment. In: Smart Analytics, Artificial Intelligence and Sustainable Performance Management in a Global Digitalised Economy. Vol. 110A. Emerald Group Publishing Limited; 2023: 261-287. https://EconPapers.repec.org/RePEc:eme:csefzz:s1569-37592023000110a014
21. United Nations Educational, Scientific and Cultural Organization (UNESCO). Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development. Paris: UNESCO; 2019. https://unesdoc.unesco.org/ark:/48223/pf0000366994
22. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq 2023; rs.3.rs-2566942. https://doi.org/10.21203/rs.3.rs-2566942/v1
23. Haleem A, Javaid M, Qadri MA, Suman R. Understanding the role of digital technologies in education: A review. Sustain Oper Comput 2022; 3: 275–285. https://doi.org/10.1016/j.susoc.2022.05.004
24. Choi W. Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: A study using MCQs. BMC Med Educ 2023; 23: 864. https://doi.org/10.1186/s12909-023-04832-x