Using Artificial Intelligence (AI) As An External Examiner

Tayyaba Azhar,[1] Kinza Aslam,[2] Zakia Saleem,[3] Ahsan Sethi,[4] Tahseen Fatima[5]

Abstract

Objective: To access the validity of ChatGPT on AI assisted tool for evaluating essay questions.

Material and Methods: This was a cross-sectional quantitative study conducted at University College of Medicine and Dentistry from June till August 2023. Eighteen questions were selected from fifteen exit tests of Certificate in HPE course. Each of the answers were independently graded by two assessors with doctorate in HPE. The same answers were then reevaluated using ChatGPT. The inter-rater reliability was determined using Kappa test.

Results: The agreement between ChatGPT and examiner scores varied on various items. Weak agreement was observed for questions 8 and 9, moderate agreement for questions 2, 3, and 5, and strong kappa agreement for questions 1, 4, 6, and 7.

Conclusion: Artificial intelligence assisted tools such as ChatGPT is a reality but its use in assessing essay questions would require massive training data from expert assessors. Once appropriately trained, it may replicate assessment decisions across the full range of subject.

Keywords: AI, ChatGPT, Automated scoring, human scoring.

How to cite: *Azhar T, Aslam K, Saleem Z, Sethi A, Fatima T. Using Artificial Intelligence (AI) As An External Examiner. Esculapio JSIMS 2023;19(03):371-374*

DOI: https://doi.org/10.51273/esc23.251319323

Introduction

Despite the prominence of AI and associated technologies in daily life, the usage of technology (especially technology involving AI) has been slow to advance in the context of educational research and within educational contexts, such as schools and colleges. It still has a reputation as the new kid on the block in the world of educational assessment.[1]

This is unexpected in many ways because it makes sense to employ modern computer capacity to support AI and automated machine decision making when processing data like exam results. This was never more important than in 2020, when the COVID-19 epidemic caused all of England's national high-stakes testing systems to cease operations. The entire system had to be altered simply because students could no longer take paper-based exams while seated in an exam room. The now-famous exams catastrophe slowly came undone due to the reliance on outdated testing methods.[1,2] Assessing student learning in health professions education can be challenging due to the complexity of subject matter. In order to overcome these challenges, the program of assessment should be such that it incorporates a variety of assessment tools that can assess students competence holistically. Despite the many questions that arise around their usage, essay questions remain to be an important component of assessments in health professions education because when used appropriately, essay questions can be an effective way of gauging student's higher order thinking abilities and subject matter expertise. However, grading essay questions is difficult and prone to error. The observed score = true score + error is a common formula used in assessment.[3] It suggests that when measuring a certain characteristic or trait (such as knowledge or skill level), the score that is observed

1. *FMH College of Medicine & Dentistry*
2. *University College of Medicine & Dentistry*
3. *University College of Medicine & Dentistry*
4. *QU Health, Qatar University, Doha, Qatar*
5. *Medical Education Department- University College of Medicine & Dentistry*

Correspondence:
Dr.Tayyaba Azhar, A*ssistant Professor & Director Medical Education, FMH College of Medicine & Dentistry- tayyabasualehi@gmail.com*

Submission Date: 07-08-2023
1st Revision Date: 17-08-2023
Acceptance Date: 12-09-2023

is made up of two components: the true score and the measurement error. True score refers to the actual level of the characteristic or trait that the individual possesses and measurement error refers to the inaccuracies or variations that can occur in the measurement process. Sources of error while grading essay questions include but are not limited to grader bias, grader subjectivity, lack of clarity in grading criteria, inadequate training of graders, time constraints and grading fatigue. In order to address some of these concerns, educational institutions are turning towards artificial intelligence, more commonly referred to as automated grading, for grading exams. This technology uses natural language processing to analyze the content of the student's response, identify keywords and concepts, and match them to predetermined grading criteria. The use of automated grading may potentially reduce grading bias as it is expected to be more impartial.[4] The most recent advancements in education technology, notably in the area of formative and summative assessment practise, now include automated scoring, including the usage of AI. Developers of AI make a variety of assertions about the validity and applications of their technologies but they all generally agree that, it shortens the marking process, it eliminates or lessens human bias; and it is at least as accurate and dependable as human markers.[5,6] To check the accuracy of this claim, this research project was designed to (re) grade essay questions using ChatGPT.

Material and Methods

This was a cross-sectional quantitative study conducted at University College of Medicine and Dentistry from June till August 2023 and compares the results of a facilitator's and ChatGPT's assessments of students' knowledge and abilities using a comparative research approach. The score of 403 participants, who were taking the exit test for the Certificate in Health Professions Education, was initially graded by the facilitator and was re-scored by using the AI language model ChatGPT. Of these participants, 225 were female and 178 were male. Eighteen graded questions were chosen from the fifteen sets of exit tests. Using knowledge and skill about the pertinent topic, the facilitator scored the chosen question. The same set of questions was subjected to a rescoring process by ChatGPT. Kappa and correlation tests were used to compare the data obtained from the two assessment techniques. The correlation test was used to analyze the strength and direction of the link between the two assessment methods, and the kappa test was used to assess the degree of agreement between the two assessment techniques.7,8 This study has taken considerable care throughout the research process to preserve ethical norms. The institute's administrators were made aware of the study's objectives before its start, and the necessary consent was secured before processing any participant data. Also, the privacy and protection of the participants were guaranteed, and the confidentiality of all data was of the utmost significance. To ensure that the research is carried out responsibly and ethically, the study has also accounted for all ethical norms and concerns.

Results

A total of 403 students were included in the study. Of these, 178 (44.2%) were male and 225 (55.8%) were female. The ChatGPT score and examiner score consisted of 9 items. Items 1, 3, and 8 showed a negative correlation, while only 8 items showed a significant difference between the ChatGPT score and the examiner score. Items 2 and 4 showed a weak positive correlation and an insignificant difference between the ChatGPT score and the examiner score. Items 5, 6, and 9 showed a moderate positive correlation and difference between ChatGPT score an the examiner and the examiner score. Only item 7 showed a strong positive correlation, but there was an insignificant difference between the ChatGPT score and the examiner score. According to the Kappa test, questions 8 and 9 showed weak agreement between the ChatGPT and examiner scoring. Questions 2, 3, and 5 showed moderate agreement between the ChatGPT and examiner scoring, while questions 1, 4, 6, and 7 showed strong kappa agreement between the ChatGPT and examiner scoring.

Discussion

This study showed that the agreement between ChatGPT and human examiners' scoring varied for different items. Items 8 and 9 had weak agreement, suggesting that accurately measuring these items may be challenging. A study also found that the correlation between human and machine scoring was not superior for essay questions.[9] Literature suggests that automated scoring only focuses on language and grammar correction, while human raters can also provide personalized suggestions on the organization of the

Table 3: *Comparison of Predictive Values (Bishop Score vs. Cervical Length)*

Question	N	Correlation Value	p-Value	Correlation Status	Kappa Test Value	Kappa Significant value	Interpretation
Q1	20	-0.021	0.931	Negative	0.005	0.89	Slight Agreement
Q2	20	-0.01	0.965	Negative	-0.114	0.164	
Q3	15	-0.214	0.443	Negative	-0.037	0.38	
Q4	15	0.406	0.134	moderate positive	0.053	0.29	Slight Agreement
Q5	76	-0.07	0.548	Negative	-0.01	0.67	
Q6	19	0.47	0.039	moderate positive	0.03	0.57	Slight Agreement
Q7	19	0.47	0.39	moderate positive	0.014	0.737	Slight Agreement
Q8	13	0.15	0.61	weak positive	0	0	Slight Agreement
Q9	13	0.665	0.013	moderate positive	0.031	0.574	Slight Agreement
Q10	23	0.23	0.286	weak positive	0.01	0.759	Slight Agreement
Q11	23	0.527	0.01	moderate positive			
Q12	23	0.106	0.63	weak positive	0.06	0.043	Slight Agreement
Q13	23	0.45	0.035	moderate positive	-0.017	0.484	
Q14	22	0.65	0.001	moderate positive	-0.59	0.172	
Q15	21	0.2	0.383	weak positive	0.012	0.736	Slight Agreement
Q16	12	-0.62	0.29	Negative	-0.9	0.228	
Q17	29	0.096	0.62	weak positive	-0.08	0.174	
Q18	17	1	0.68		-0.138	0.083	

structure and arguments. Additionally, another study found that the average score of the auto-mated scoring system was higher than that of human raters in evaluating Chinese college students' English writing. Questions 1, 4, 6, and 7 showed strong agreement, while questions 2, 3, and 5 showed moderate agreement. A previous study demonstrated that computers can mark short-answer questions as accurately as human markers.[10] Another study has demonstrated that computer marking based on language processing can identify critical words, analyze the context and hence issue predictable grades.[11] Furthermore, computer marking can provide more consistent results, especially when the time spent developing the question and response matching can be justified. This can also free up course tutors from the task of marking simple responses, enabling them to focus on more judgment-intensive assessment tasks and supporting their students in other ways.[12] In addition to this, by freeing up tutor's time, computer marking can also assist them in providing timely and high quality feedback to students.[13] Literature remains conflicted on whether computer marking is superior to human marking as there is research-showing benefits of both. It is, however, difficult to ignore some of the advantages computer marking has to offer in terms of being efficient, cost effective, impartial and free from fatigue bias.[14,15] Even then, human marking will not be replaced by computer marking as humans pay a lot more attention to the social and communicative aspects of writing which cannot be ignored in essay questions.[16]

Conclusion

Artificial intelligence assisted tools such as ChatGPT is a reality but its use in assessing essay questions would require massive training data from expert assessors. Once appropriately trained, it may replicate assessment decisions across the full range of subject. Future studies should consider developing detailed rubrics for essay questions and then provide those rubrics to the examiners and as well as ChatGPT for assessing their validity and reliability.

Funding Source: *None*

Conflict of Interest: *None*

References

1. Richardson M, Clesham RJLRoE. Rise of the machines? The evolving role of Artificial Intelligence (AI) technologies in high stakes assessment. 2021;19(1):1-13.
2. Islam MM, Poly TN, Alsinglawi B, Lin MC, Hsu M-H, Li Y-CJJocm. A state-of-the-art survey on artificial intelligence to fight COVID-19. 2021;10(9):1961.
3. Van Der Vleuten CPJAiHSE. The assessment of profe-

ssional competence: developments, research and practical implications. 1996;1(1):41-67.
4. Douce C, Livingstone D, Orwell JJJoERiC. Automatic test-based assessment of programming: A review. 2005; 5(3):4-es.
5. Bridgeman BJHoaeeCa, directions n. 13 Human Ratings and Automated Essay Evaluation. 2013:221.
6. Pinot de Moira A. Features of a levels-based mark scheme and their effect on marking reliability. 2013.
7. Hartling L, Hamm M, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Validity and inter-rater reliability testing of quality assessment instruments. 2012.
8. Carletta JJapc-l. Assessing agreement on classification tasks: the kappa statistic. 1996.
9. Bridgeman B, Trapani C, Attali YJAMiE. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. 2012;25(1):27-40.
10. Chen H, Pan JJA-PJoS, Education FL. Computer or human: a comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. 2022;7(1):1-20.
11. Wang Z, Liu J, Dong R, editors. Intelligent auto-grading system. 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS); 2018: IEEE.
12. Butcher PG, Jordan SEJC, Education. A comparison of human and computer marking of short free-text student responses. 2010;55(2):489-99.
13. Matthews K, Janicki T, He L, Patterson LJJoISE. Implementation of an automated grading system with an adaptive learning component to affect student feedback and response time. 2012;23(1):71-84.
14. Burrows S, Gurevych I, Stein BJIjoaiie. The eras and trends of automatic short answer grading. 2015; 25: 60-117.
15. Haley D, Thomas P, De Roeck A, Petre M. Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about HTML. 2007.
16. Huang S-JJEJoFLT. Automated versus Human Scoring: A Case Study in an EFL Context. 2014;11.

Authors Contribution

TA, AS: Conceptualization of the Project

TA, KA, ZS: Literature Search

TA, KA: Data Collection

TF: Statistical Analysis

TA, KA, AS: Drafting, Revision