

Students' Perspective on AI-Supported Assessment of Open-Ended Questions in Higher Education

Daniel Braun¹ ^a, Patricia Rogetzer¹ ^b, Eva Stoica² and Henry Kurzhals³

¹Department of High-tech Business and Entrepreneurship, University of Twente, Enschede, The Netherlands

²Faculty of Electrical Engineering Mathematics and Computer Science, University of Twente, Enschede, The Netherlands

³Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

Keywords: Open-Ended Questions, Automated Grading, Artificial Intelligence, Education, User Perspective.


Abstract: Artificial Intelligence (AI) is widely used for the assessment of multiple-choice questions. There is an increasing effort to also use it for open-ended questions. While the use of AI can benefit the learning of students, e.g. by increasing the number of feedback moments, most applications focus on saving costs by reducing the need for manual assessment. The perspective of teachers on this kind of automation has been studied extensively, the student perspective, however, is still under-researched. This paper presents the results of two surveys and a series of interviews among students to identify their perspective on AI-supported assessment and elaborate on under which conditions they would accept such technology. The results show that the majority of students (more than 80%), is, under certain conditions, open to AI-supported assessment. Most importantly, they stress that humans should still be involved in the assessment (human-in-the-loop).


1 INTRODUCTION

Assessment is an important part of the learning process. Some argue it is even more influential on the success of learners than teaching (Medland, 2016). In higher education, multiple-choice and open-ended questions are used for both formative and summative assessment. Open questions have shown to be more effective in assessing student learning (Funk and Dickson, 2011) and support a “deep approach” to learning. Multiple-choice questions rather support a “surface approach” (Gibbs, 2006). Still, multiple-choice questions (MCQs) are used with increasing frequency. This is a trend that is amplified by the use of e-assessment. Some researchers fear students could fail to develop communicative competencies due to the increasing amount of MCQs (Paxton, 2000). One of the main reasons for their popularity is that the assessment of MCQs is time-efficient and scalable because their evaluation can be automated (Roberts, 2006). Therefore, especially in very accessible learning environments, like Massive Open Online Courses (MOOCs), MCQs are often used for practical reasons, although open questions could sup-

port student learning better. Automating or supporting the assessment of open-ended questions with Natural Language Processing (NLP) could increase their usage in such settings and not only improve the existing assessment but also help to create additional feedback moments for students, both of which can support the learning process and therefore contribute to achieving the Sustainable Development Goal (SDG) of quality education (SDG 4) (United Nations, 2016).

The current public and scholarly debate about AI and assessment is often focused on the implications that large language models like ChatGPT have on different forms of assessment and whether they will make them redundant. Often, these discussions solely consider summative assessment and fail to acknowledge the important role that formative assessment can have during the learning process. (Gilson et al., 2023; Choi et al., 2023) The limited amount of existing research on the automated assessment of open questions is almost exclusively focused on the reduction of teacher workload and does not consider the perspective of students and how AI-supported assessment could benefit their learning (see Section 2). Therefore, there is a gap when it comes to the perspective of students on AI-supported assessment. Considering their perspective when designing, implementing, and using such technology should not only be an ethical

^a  <https://orcid.org/0000-0001-8120-3368>

^b  <https://orcid.org/0000-0001-9582-6320>

imperative but is also crucial to the acceptance of the technology.

This paper presents the perspective of students on AI-supported assessment of short open-ended exam questions, based on surveys and semi-structured interviews involving a total of 81 students. 38 of the participants had previous experiences as assessors while being student teaching assistants, combining the perspective of students with the perspective of assessors.

The results show that students are open to the technology, as long as human assessors are still involved. Based on the findings, the paper presents guidelines for the design of NLP systems that support grading in a way that is perceived as fair and helpful by students. These guidelines can help to build trust in a system, which is crucial for its acceptance.

2 RELATED WORK

Baker et al. (2019) differentiate three categories of AI in education: learner-facing solutions, teacher-facing solutions, and system-facing solutions, that are used by administrative or managing staff. Tools that support assessment are usually teacher-facing solutions, but can also be learner-facing if they provide direct feedback, and system-facing, if they provide, e.g., analytics of assessment results. Even if a tool is not learner-facing, it is important to keep in mind that its application will have an impact on learners and their perspective should therefore be considered.

In 2019, Zawacki-Richter et al. performed a “systematic review of research on artificial intelligence applications in higher education”, analysing 146 papers published between 2007 and 2018. Of these 146 publications, 13 are concerned with automated grading, citing “reducing costs and the time associated with [...] large-scale assessments” as the main benefit of the technology (Zawacki-Richter et al., 2019, p. 17). None of the publications on automated grading and only two out of 146 publications mentioned there discuss the ethical implications of the technology.

Sánchez-Prieto et al. (2020) developed a model to measure the acceptance of students for AI-based assessment based on the technology acceptance model (TAM) by Davis (1989). However, the items they propose for the measurement, like “I find AI-based systems easy to use.” or “My interaction with AI-based tools is clear and understandable”, are highly dependent on a concrete implementation or tool and are therefore more suitable to measure the acceptance for one concrete tool. In our research, we wanted to investigate under which preconditions students are open to AI-supported assessment. Such information cannot

be derived from the closed questions and statements suggested by Sánchez-Prieto et al. (2020). Therefore, we decided to use a more open approach.

The few existing studies that do take into account students’ perspective on AI-supported assessment, e.g. by Galassi and Vittorini (2021), Mirmotahari et al. (2019), and Scharber et al. (2008), do so based on concrete implementations. Common themes identified by all authors are that transparency and understanding how the systems work lead to higher acceptance.

Tan et al. (2022) found that a lack of trust in an automated grading process can increase the anxiety level before an assessment and thereby negatively influence the performance of students. This underlines the importance of considering the student perspective when designing and implementing AI-supported assessment technologies. Failing to do so is not only problematic from an ethical perspective but also with regard to student success.

3 APPROACH

To identify the perspective of students on AI-supported assessment of open-ended questions, we used two approaches. For more in-depth insights, we conducted semi-structured interviews with students. For a broader overview, we conducted two online surveys. In addition to the general perspective of students, we were particularly interested in three aspects:

- *Do students with teaching and grading experience have a different view on AI-supported assessment of open-ended questions than other students?* We hypothesised that students with grading experience would be more aware of the fallibility of human assessors and therefore more open to AI support. This question can help to differentiate whether students compare AI-supported assessment to an idealised notion of manual assessment they might have in mind or the real assessment process.
- *Are students from technical disciplines more reluctant to accept AI-supported grading?* We believed that students with a more technical study background, like computer science, could be more hesitant towards AI involvement in grading. Because, compared to less technical disciplines, like psychology, they could be more aware of the limitations of current NLP technologies.
- *In which parts of the assessment process is human involvement most important to students?* We are looking at scenarios in which the grading process

is supported by AI, rather than fully automated. Therefore, the question arises, in which parts of the process human involvement is most beneficial from the student's perspective. The answer to this question can help us in developing guidelines for building AI systems that are accepted by students.

3.1 Interviews

Twelve interviews were conducted with students that have previously or were currently working as teaching assistants, i.e., have experience in grading other students' work. The interview was semi-structured to give participants enough space to voice their opinions and present their perspectives.

3.2 Surveys

In addition to the interviews, two online surveys were conducted. The first survey was targeted at students with experience in teaching and assessment and used the same questions used during the interview (see Appendix 6.1). 26 students participated in this first survey. The second survey was conducted among a general population of students, i.e. students without assessment experience. The questions of the second survey can be found in Appendix 6.2. 43 students participated in the second survey.

4 RESULTS

This section will first present the results from the survey and interviews with students with assessment experience (Section 4.1) and then the results from the general student survey (Section 4.2).

4.1 Students with Assessment Experience

The combined 38 participants in the interviews and the first survey were all recruited from one Dutch university. Table 1 shows that most participants (~73%) come from technical study programs with high shares of computer science influence like business information technology or computer science itself.

4.1.1 Human-in-the-Loop

Half of the interviewees thought that humans-in-the-loop are very important and should "assess more than just the end result". For the other half, the necessary involvement strongly depends on the type of questions. In their opinion, answers which can be assessed

Table 1: Study programs of participants with teaching experience.

Study Program	# Participants
Business Information Technology	17
Computer Science	11
Industrial Engineering Management	3
Business Administration	3
Industrial Design Engineering	2
Communication Science	1
Philosophy	1

purely factually, like a maths equation, need less human involvement than more opinionated questions.

In the survey intended for teaching assistants, participants could rate how important they believe human input to be in the grading process on a Likert scale from 1 (not at all) to 10 (very much). More than 80% of respondents believe human input to be very important (7 and above).

When asked about the benefits and drawbacks of human assessment, the twelve aspects shown in Table 2 were repeatedly mentioned in interviews and the survey. In general, participants believe that human assessors are better capable of understanding and interpreting answers and their nuances. More specifically, participants believe that only human assessors can provide valuable feedback and consider consequential errors. Another advantage that students see in human assessment is the connection they have with the assessor. Partially, this is seen as a possibility to be assessed more favourably based on the personal relationship. At the same time, students see the influence of emotions and a lack of consistency as disadvantages of human assessment. Finally, participants think that only human assessors can identify flawed questions in the assessment process, e.g., if the scores for one question are consistently low, even for students that scored well in all other questions.

Having been assessors themselves, participants of the first survey and interviewees were also able to provide downsides to human assessment. Several participants mentioned that (especially student) assessors sometimes lack the necessary knowledge to perform a solid assessment and that vague assessment criteria can amplify the problem. Consistency between assessors, but also over time, was also seen as an issue of manual assessment. In this context, participants also mentioned specifically the influence of emotions and "correction fatigue". On a more technical level, almost a third of the participants mentioned that problems with tools that are being used during manual assessment can negatively influence the process.

Table 2: Benefits and drawbacks of manual assessment.

Benefits	Drawbacks
Better understanding of responses	Vague criteria
Interpretation	Limited knowledge
Consideration of consequential errors	Issues with tools
Provide feedback	Differences between assessors
Student-teacher relationship	Consistency issues
Identify flawed questions	Influence of emotions and “correction fatigue”

4.1.2 Potential of AI

Overall, participants were very open to the idea of AI-supported assessment. More than 80% of the interviewees and 84% of participants in the survey agreed that AI could help in the assessment. Participants saw the potential for a range of tasks to be automated, from checking the work of human assessors, taking over repetitive tasks, and increasing consistency, to having a pre-screening to identify completely right or wrong answers. Some participants suggested that the time they could potentially save through such automation could be spent on aspects that cannot be automated, like giving feedback and discussing results. When asked about other potential advantages of AI, 75% of the interviewees and 92% of the survey participants mentioned faster grading. While 92% of the interviewees believed that such a tool would be generally helpful in supporting them, only 38% of the interviewees mentioned that as a potential advantage (see Table 3).

Asked about disadvantages, teaching assistants were concerned that an AI tool, especially when newly introduced, could mean additional work for them, like providing training data and double-checking the results of the system. Some were worried that, over time, teaching assistants could stop verifying the results of the AI and use them unchecked.

4.2 General Student Body

The general student survey was completed by 43 individuals. Most of them (~ 86%) did not come from technical disciplines (see Table 4). However, only 11% of the participants felt like not knowing much about AI. At the same time, 75% of the participants were not aware that tools are being developed for automated assessment in higher education. When asked to name such tools, only one participant provided an answer.

83% of the participants (30 out of 36 who responded to this question) think AI could be used as part of the assessment process, however, almost all agree, only under the constraint that a human is still involved. Other, less frequently mentioned, con-

straints included the difficulty of the question and whether it is a purely factual or more open, interpretable question. In an MCQ, asking whether participants would prefer their next exam to be graded by only a human, only an AI, or with a hybrid approach, 86% out of 43 respondents chose the hybrid approach, 13.5% only human, and 0.5% only AI.

To find out where students think an AI could be applied successfully in the grading process, participants were asked which parts they think should be occupied by humans and which could be occupied by an AI. The most common theme was that questions that “leave room for interpretation” should be assessed only by humans, while closed questions could be assessed by an AI. Generally, mainly two points were seen as a big advantage of AIs by the participants, speed and consistency.

5 DISCUSSION

Overall, the results show that both, students in general and students with teaching experience, are very open to AI-supported assessment of open-ended questions. However, for both groups, it is an important prerequisite that AI is only used as a support tool and that teachers are still involved in the assessment.

When asked about the advantages and disadvantages of humans and AIs as assessors, it was indeed visible that students with teaching experience seemed to be more aware of the fallibility of human assessors, by pointing out aspects like correction fatigue, knowledge issues and many more, while students without teaching experience only saw a lack of consistency as a potential issue. However, this awareness did not lead to a higher acceptance of AI in comparison to the general student population. Within the groups, we did not see an influence of the study program on acceptance.

When asked about the advantages and disadvantages of AI in assessment, it was visible that many of the common prejudices towards AI, both negative and positive, were present in the responses. For example, almost all students attributed objectiveness to AI as-

Table 3: Benefits and drawbacks of AI-supported Assessment.

Most frequently mentioned	Interviews	Survey
Faster grading (+)	9/12 (75%)	24/26 (92%)
Support work (+)	11/12 (92%)	10/26 (38%)
Cause additional work (-)	8/12 (66%)	17/26 (65%)
Unchecked usage of results (-)	8/12 (66%)	2/26 (8%)

Table 4: Study programs of participants in the general student survey.

Study Program	# Participants
Business Administration	27
Engineering	3
Teaching Profession	3
Computer Science	2
Social Studies	2
Law	1
Medicine	1
Psychology	1
Sports Management	1
Society and Technology	1
Ecosystem Management	1

assessment, although AI systems can be biased in many ways (Mehrabi et al., 2021). On the other hand, some properties that can potentially be offered by an AI, like providing explanations for an assessment, were seen as exclusively achievable through manual assessment.

When asking teachers about the biggest potential of AI in assessment, they often think of a separation based on the quality of the response, i.e., an AI can quickly identify answers that are very good or very bad, leaving only the more ambiguous cases for the human assessor. The students, however, were solely focusing on the type of question that is asked. They see the potential for AI mainly for questions that are “easier”, more “objective” and less “subjective”, as well as more “factual” and less “interpretable”. These properties are connected to questions that can be used to test the lowest two categories in Bloom’s Taxonomy of educational objectives, “knowledge” and “comprehension” (or “remember” and “understand” in the revised version), because for these kinds of questions, less variance in the answers would be expected, compared to questions higher up in the taxonomy, that involve more creativity in the answering process (Forehand, 2010).

5.1 Limitations

The work has limitations influencing the generalisability and the general explanatory power of the results, originating from the selection of participants and the nature of the surveys themselves:

- All students with assessment experience interviewed and surveyed were recruited from the same university, introducing a bias towards the assessment customs at this university.
- By design, the study was very open as to what is considered “AI-supported assessment” in order to be able to catch the general attitude of students, independent of concrete systems or approaches. However, this limits the comparability of results since students might have different kinds of systems in mind.
- In general, students were asked about their opinion in an open, vague scenario. It is old folk wisdom that “talk is cheap”, so the opinion of students could change once they face the prospect of really being assessed supported by an AI system in a relevant examination.
- The majority of participants had a background in business or technology. Although our work implies that there is a significant difference between technical and non-technical students in the survey, it could still be that students from a humanities background have different views.

6 CONCLUSION

This paper presents the perspective of students on AI-supported assessment of open-ended questions in higher education. The results from two surveys and a series of interviews among students with and without teaching experience show that the vast majority of students (over 80%) is, in general, open to the idea of an AI being involved in the assessment of their work and can see potential benefits in it. Based on their perspectives, three essential requirements for AI-supported grading tools can be derived that are key to the acceptance of students:

1. There should always be a human-in-the-loop overseeing the assessment.
2. An explanation should be provided for each assessment.
3. It should be assessed individually for each question whether it is appropriate to use AI support

for the assessment. A guideline for the appropriateness can be Bloom's taxonomy.

With this work, we want to encourage researchers and developers working in the field of AI-supported assessment to give more consideration to the perspective of students. The development and implementation of tools that affect students and their assessment should always be critically accompanied by them. The findings presented in this paper can be a first guideline on how to design systems in a student-friendly way.

REFERENCES

- Baker, T., Smith, L., and Anissa, N. (2019). Educ-ai-tion rebooted? exploring the future of artificial intelligence in schools and colleges. Technical report, nesta foundation.
- Choi, J. H., Hickman, K. E., Monahan, A., and Schwarcz, D. (2023). Chatgpt goes to law school. *SSRN*.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.
- Forehand, M. (2010). Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4):47–56.
- Funk, S. C. and Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology*, 38(4):273–277.
- Galassi, A. and Vittorini, P. (2021). Automated feedback to students in data science assignments: Improved implementation and results. In *CHIItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, CHIItaly '21, New York, NY, USA. Association for Computing Machinery.
- Gibbs, G. (2006). Why assessment is changing. In *Innovative assessment in higher education*, pages 31–42. Routledge.
- Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., and Chartash, D. (2023). How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*, 9:e45312.
- Medland, E. (2016). Assessment in higher education: drivers, barriers and directions for change in the uk. *Assessment & Evaluation in Higher Education*, 41(1):81–96.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.
- Mirmotahari, O., Berg, Y., Gjessing, S., Fremstad, E., and Damsa, C. (2019). A case-study of automated feedback assessment. In *2019 IEEE Global Engineering Education Conference (EDUCON)*, pages 1190–1197.
- Paxton, M. (2000). A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education*, 25(2):109–119.
- Roberts, T. S. (2006). The use of multiple choice tests for formative and summative assessment. In *Proceedings of the 8th Australasian Conference on Computing Education-Volume 52*, pages 175–180.
- Sánchez-Prieto, J. C., Cruz-Benito, J., Therón Sánchez, R., García Peñalvo, F. J., et al. (2020). Assessed by machines: development of a tam-based tool to measure ai-based assessment acceptance among students. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(4):80.
- Scharber, C., Dexter, S., and Riedel, E. (2008). Students' experiences with an automated essay scorer. *Journal of Technology, Learning, and Assessment*, 7(1).
- Tan, S. H. S., Thibault, G., Chew, A. C. Y., and Rajalingam, P. (2022). Enabling open-ended questions in team-based learning using automated marking: Impact on student achievement, learning and engagement. *Journal of Computer Assisted Learning*, page 1–13.
- United Nations (2016). Transforming our world: The 2030 agenda for sustainable development.
- Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):1–27.

APPENDIX

This appendix contains the questions that were used for the survey of students with teaching experience and as a guideline for the semi-structured interviews (Appendix 6.1), as well as the questions of the second survey among the general student population (Appendix 6.2).

Survey I

- What is your study?
- How important do you think human input is for grading open-ended questions?
 - Why do you think that is important?
 - Can you name a few benefits of having a teacher/student grading an open-ended question instead of a machine/tool?
 - Are there any issues that can arise when assessing open-ended questions?
- Do you believe that automation can help with the assessment of open-ended questions? (Why/Why not?)

- Context: Let's suppose that an AI-supported grading tool is being implemented at your university. The purpose of this tool will be to support examiners in the grading of open-ended questions.
 - What might be a benefit of such a tool?
 - And what do you think the drawbacks/challenges would be with the AI-supported grading tool?
 - Any way of overcoming the challenges?
 - Do you believe that having such a tool would be easier to implement and use for particular subjects? Why?
- From your perspective what would be the general attitude of teaching assistants when this tool would be implemented?
 - Do you think that students who did not work as teaching assistants might perceive this differently? Why/Why not?
 - Would having higher transparency, for example, a clear explanation of the algorithms, help the TAs or the students perceive the AI-supported tool in a different way?
- Do you think there are better alternatives than AI to support the teachers/TAs in the grading of open-ended questions?
- of other steps you know). If you choose for/can imagine a hybrid approach, which steps of the grading process of open questions should be occupied by humans (teacher) and which by Artificial Intelligence?
- Can you imagine Artificial Intelligence-based tools grading multiple-choice questions only, open questions only, both, or none of them?
- Can you imagine Artificial Intelligence grading open questions for exams in higher education? If yes, with the support of humans or not?
- Are there cases where you can imagine Artificial Intelligence doing the grading on its own?
- Which part of grading do you think can only be occupied by humans (teachers) and not by Artificial Intelligence?
- Which part of grading do you think can only be occupied by Artificial Intelligence and not by humans (teachers)?

Survey II

- What is your study?
- Please react to the following statement by using the Likert scale: I have the feeling that I understand what Artificial Intelligence is and that I grasp the concept behind it.
- Do you know that Artificial Intelligence-based tools exist/are being developed for grading open questions of exams in higher education?
- Which AI-based tools for grading in higher education do you know?
- Do you think Artificial Intelligence should only support teachers when grading exams in higher education but not do it alone?
- Do you think Artificial Intelligence-based tools should be part of the grading process of exams in higher education?
- Would you rather want your next exam being graded by Artificial Intelligence only, human (teacher) only, or a hybrid approach?
- The parts of the grading process are: checking (first examiner), checking (second examiner), and adding the points for the final score (also think