Improvement of AI-assessment systems in grading open questions based on the teaching assistant's view

Author: Anastasia Coviliac

University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT

Artificial intelligence (AI) and automation technologies emerged and were extensively introduced in higher education in recent years. Even though a lot of research was done on the experimental level, limited literature examines and assesses the effects and coherence of applying different AI techniques in assessing open questions in the form of text, applying higher education requirements. The following research aspires to understand the practical adjustments of AI technologies in digital assessment at the university level, considering a teaching assistant view. To accomplish the scope, the systematic literature review is used to create an overview of AI assessment systems, their benefits, and possible limitations. Interviewees give supportive information on the literature findings and make suggestions for continuous improvement of AI-driven systems by concluding on the importance of human presence. Moreover, teaching assistants give insights on the acceptance level of AI integration in the grading process of the open questions. Because of the fast changing AI technologies and approaches, this paper emphasizes the importance of the teaching assistant collaboration with numerous algorithms in the context of the country, institution and willingness to adapt to the digital assessment.

Graduation Committee members:

Dr. Patricia Rogetzer
Dr. Daniel Braun

Keywords

Artificial Intelligence, Assessment, Open-ended questions, Teaching assistants, Grading system, Higher education

1. INTRODUCTION

Artificial intelligence (AI) and automation technologies emerged and were extensively introduced in higher education in recent years (Ouyang et al., 2022). However, in the educational field, AI is so far only frequently used in local and experimental contexts and not so much adopted at the educational system level (OECD, 2020). Therefore, a lot of research was made to integrate AI as a supportive tool helping teachers and examiners for the educational purpose. Gabriel et al. (2022) identified that there is increased demand on a smooth high-quality integration of AI technologies in higher education that require specific skills and knowledge. Syzdykbayeva et al. (2021) suggested that AI can significantly improve the education system due to its ability to optimize and automize a part of teachers' work, permitting them to spend more time on other professional activities that require human implication. An important and often overlooked consideration of AI in higher education is the student assessment, and specifically mechanisms and algorithms used for detecting correct or wrong student answers when it comes to grading open questions. Some researchers already experimented with evaluation tools, for example Aluthman (2016) developed the automated essay evaluation (AEE) system that immediately computes assessment, feedback, and automated scores of students in an online learning environment. Other researchers, such as Mubarak et al. (2020) created advanced AI algorithms, such as genetic algorithms for a predicted assessment and performance of students, with automated results identical with scores achieved via human assessment.

Despite all research done, limited literature examines and assesses the effects and contribution of applied AI assessment methods within the relationship with teaching assistants. The implication of collaborative assessments for open questions, between AI technologies and teaching assistants deserves to be explored further. Moreover, past screening criteria do not guarantee a completeness of all AI technologies examined. This challenge represents a research gap that is addressed in this thesis. The scope is to gain a better understanding of the range of AI techniques in assignment/exam assessment and grading on university level, taking into consideration a teaching assistant's view. Applied view on teaching assistants is important insight in answering research questions, as these persons represent a "bridge" between teachers and students. They know both sides of the spectrum, from the one hand, they can take the role of examiners and grade assignments/exams according to the examination standards and teacher's requirements, from the other hand they can be tutors for students and give more accurate feedback to them in the improvement of the grade, in comparison with teachers. This is due to their past positions as students, they know weak points of the subject, related to the study process and can give helpful information on how to achieve high results in passing the subjects, requiring specific knowledge and abilities.

1.1 Objective and research questions

Demand on providing open questions requiring argumentative answers to assess students' knowledge is increasing among universities (Alsanie et al., 2022). Despite this, statistics shows that student numbers will continue growing in the global demand for prosperous higher education (Monitor, 2018). It creates a motivation to implement AI technology that will optimize student assessment and support teachers and teaching assistants by taking away some repetitive and simple tasks—for instance, checking exam questions according to a specific grading rubric and assigning the obtained points for several groups of students. However, creating such an AI-driven grading tool creates challenges. The first is that scoring an open-ended answer in many cases is mainly subjective and does not have well-defined criteria based on all answer alternatives. The second is that scoring this type of answer requires AI advanced algorithms and complex morpho-syntactic structures checked along with the content of the assessed subject.

Additionally, determining appropriate AI technologies becomes difficult when factors such as different types of open questions and the limited number of algorithms used are changing per country, higher educational institution, and acceptance of technological evolution. This phenomenon directly influences the need of current universities to adapt to new AI assessment technologies. Although, this research aims to estimate AI technologies' opportunities to support assessing the assignment/exam open questions field and their limitations. It helps to define the initial research objective of providing recommendations on possible future collaboration between graders and AI-driven grading systems in the context of grading open questions. The research reviews the technical part of AI assessment evolution integrated in some countries. Additionally, the existing relationship between AI-driven grading environments and teaching assistants will be examined.

The main research question concerns the improvement of appropriate AI technologies in the higher educational format assessment system in response to teaching assistants' aspect, together with examining areas of internal features and possible future trends.

What are the suggested adjustments of existing AI assessment systems that should be considered while grading open questions, based on teaching assistants' perspective?

Three sub questions help us to answer the overall research question of this research project when it comes to grading open questions. Firstly, variety of existing AI assessment technologies should be listed in order to give an overview, it will be investigated via systematic literature review. Secondly, the awareness of possible benefits and limitations of examined AI applications in grading should be addressed. For this, three dimensions of strengths, weaknesses and future adaptations will be examined using the SpeedGrader and Remindo grading platforms commonly used at the University of Twente. Finally, the level of student acceptance of AI assessment integration will be investigated. The expert interview results will be used for answering over two sub questions. A sub-objective of this research is to give insights for further improvements of existed AI tools based on the University of Twente interviewed sample.

1. What are suited AI algorithms for assessing open questions in higher education in different countries?

2. What are the benefits and weaknesses of AI assessment systems applied at the University of Twente?

3.What attitude do teaching assistants have regarding using supporting AI system for grading open questions?

2. METHODOLOGY AND RESEARCH DESIGN

This research has a qualitative explorative research design. Hence, the data collection methods chosen are systematic literature review combined with semi-structured qualitative interviews. A systematic literature review permits collecting information from different sources in an unbiased and rigorous manner (Tranfield et al., 2003). This commonly used methodology includes five significant steps:

1. Research question formulation
2. Locating studies
3. Study selection and evaluation
4. Analysis and synthesis
5. Reporting and using the results

The literature review involves the researcher in a critical thinking process and provides a comprehensive understanding of the status quo (Xiao & Watson, 2019).

The semi-structured interviews are chosen as a data collection method to bridge the academic and practical perspectives on the researched topic, as well as to identify new concepts and draw conclusions based on the researched sample. Thus, the findings from the literature review will be supported and supplemented during the interviews. Since multiple AI application technologies in the higher education industry are already feasible and discussed via the systematic literature review section, the focus of this research will be on examining and carrying conclusions based on two AI assessment tools: SpeedGrader and Remindo, for giving general conclusions that can be implemented for other AI tools spread around the world. The targeted market for data collection is the University of Twente, located in the Netherlands. Thus, the people participating in this research have expertise in researched tools.

It is challenging to determine the appropriate amount of interviewers because the literature has not come to a consensus on this matter (Adams, 2015; Baker & Edwards, 2012). A reasonable number of interviews for qualitative research ranges between 5 and 20, depending on its scope (Adams, 2015; Baker & Edwards, 2012; Galvin, 2015). The overall number of interviews in this study was fixed at 18, taking into account the constraints suggested in the literature.

2.1 Literature search

In terms of collecting scientific literature and research papers, a systematic literature search (SLS) is carried out. This method contributes to developing a better understanding of the AI positioning for higher education assessment and created relationship between AI and teaching assistants. A systematic search was performed on the following electronic databases: Scopus, Web of Science, IEEE, Taylor & Francis, Wiley, and EBSCO in order to find the relevant articles and scientific publications.This selection of databases was determined, as they represent the largest and reliable sources and publisher databases.

The following keyword combinations based on the research questions are used: ‘artificial intelligence” OR “AI” OR “machine learning” AND “education”, “artificial intelligence” OR “AI” AND “assessment”, “grading open questions” AND “artificial intelligence” or “AI”, ‘teaching assistants” AND “higher education”, “business” AND “students” AND “assessment”, “open questions” AND “grading criteria”, “AI” AND “algorithms” and “higher education” , “AI” OR “Artificial Intelligence” AND “algorithms” AND “higher education” AND “assessment”, The libraries of Google Scholar, Research Gate, and the university library - FindUT assist for additional literature searches. As the study’s scope covers a quickly changing and upgrading AI technology in sector of higher education, mainly papers from 2018 onwards, written in English are reviewed to ensure the knowledge and data passed to the reader is not outdated. The quality of papers was assessed by four criteria: abstract, publication year, fit between research questions and research design, source reliability.

As a guideline, Denyer and Tranfield (2003) suggestions applied for the articles screening procedure visualized in **Figure 1**. Inclusions and exclusion criteria were implemented for the final selection. Even though not all articles specifically met one of the research queries based on the keyword combinations, their abstract suggested that papers can be useful. Therefore, specific articles were not discarded immediately. As a result, the following inclusion criteria were related to the screening procedure:

- Considers AI technologies’s application not restricted to a specific country/continent
- Considers AI technologies used in higher education and school
- Considers AI systems applied in engineering, medicine studies

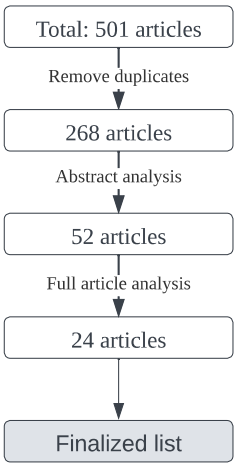

Figure 1. Article screening methodology (based on: Denyer and Transfield, 2003)

After the screening, 24 articles related to AI assessment systems were taken as a basis for the literature review section. Other included articles worked as supplement sources for establishing a clear overview of the researched topic.

2.2 Setup of interviews

The interviewees are chosen based on their selected study program and current job occupation, as well as their research interests and time availability. Interviewees have diversified study background and active teaching assistant positions with experience in grading open questions via Speedgrader/ Remindo assessment environments are among the selection criteria. As a research environment, the University of Twente was selected, in which teaching assistants took participance in this research, representing the research sample. The existing AI grading tools, such as SpeedGrader **(see Subsection 3.2.1)** and Remindo **(Subsection 3.2.2)** are widely used in the selected research institution in the context of grading open questions during exams and assignments. Student assistants are working in the study programmes of psychology, international business administration, business and IT, chemical engineering, civil engineering and industrial engineering and management.

As this research is conducted among students who simultaneously took a job position as teaching assistants during the post effect of a global pandemic, time availability, as well as health and safety protection of participants, are amongst the top

priorities. The interviews could efficiently be conducted in a hybrid environment, online or offline, due to the advances in online communication environments and technologies.

First, the interviewees are contacted via email containing the invitation to participate in the study research via the interview. They receive information regarding an overview of the research topic, including the scope, objective, and importance of the participant in the study. The list of the main questions to be asked during the interview (**see Appendix B**). Out of the 30 teaching assistants contacted with different program backgrounds, 18 agreed to contribute to the following research either offline or online.

To protect the interviewees' privacy, personal information such as their names, ages and gender is not shown. However, to demonstrate how they contribute to the following research subject, their current study and participants' jobs are included in the table provided in **Appendix C**. In the same table, a summary of interviewees' answers to the posed questions can be found.

3. CONCEPTUALIZATION AND THEORETICAL FRAMEWORK

3.1 AI and Machine learning algorithms

Tedre et al. (2021) refer to the roots of AI that belong to the fundamentals of higher education that were computed since the birth of AI in 1956 (McCorduck & Cfe, 2004). In higher educational institutions, the theory of AI focuses on mathematical building models based on algorithms that generalize unstructured data. There is an estimation of the fast growth of model-driven data analytics approaches that have a scope of guidance in the development, perception, and recognition of the existing algorithms (Luan et al., 2020). Over the past few decades, research on AI has advanced to intelligent computing technologies, such as intelligent tutoring systems (Nye, 2015), robotic systems (Anwar et al., 2019), and chatbots (Smutny & Schreiberova, 2020). Social media platforms such as Facebook, Twitter, online learning environments [e.g., MassiveOpen Online Courses (MOOCs)], intelligent tutoring systems (e.g., AutoTutor), learning management systems (LMSs), and sensors were outcomes of the breakthrough of researchers to the big data. Increased interest in AI permitted further development in assessment, tutoring, and continuous analysis of students within the digitalization direction.

3.1.1 AI algorithms used for assessment in higher education

One reason for introducing AI in higher education was the rapid progress in computer technologies that spread over all industries. Additionally, the COVID-19 infection increased demand for intelligent technologies allowing higher education institutions to switch teaching methods in the online environment (Agarwal, 2022). Higher education globally tried to seek online teaching methods while also introducing various algorithms for student assessment (Coman et al., 2020). Recently it permitted to gain much attention to AI development features that will automatically assess students' performances and automated assessment systems are considered one of the most promising applications in education (Akgun & Greenhow, 2021).

The essence of AI in grading open questions has been raised for a long time, Gardner et al. (2021) enumerated four leading Automated essay scores (AES) studied recently, components that can help in grading text questions. AES included ProjectEssay Grading, Intellimetric, Intelligence Essay Assessor and e-Rater. Enumerated AES engines represent the basics of AESs driven today. The change was made in terms of algorithmic sophistication, data capacity and processing efficiency. Nevertheless, numerous studies highlighted different approaches to using text detection and evaluation algorithms.

According to a recent study by Zawacki-Richter et al. (2019), AI algorithms can generate just-in-time feedback simultaneously with the student assessment. Rather than stop-and-test, algorithms developed the probability prediction in case of a student failing on assignments or exams or dropping out of a course with conspicuous accuracy. Another research conducted by Mujtaba et al. (2020) emphasized the importance of the assessment as it supports information on the individual ability, skills, and knowledge of a group of students. Computerized adaptive testing (CAT) represents automated assessments that serve questions based on a student's ability. CAT works on the principle of providing a question with medium difficulty. Depending on the response of a test taker, the CAT immediately proposes either an easier or a complicated query (Meijer & Nering, 1999).

AI assessment applications show the recent boom in the educational field, for example automated scoring engines were introduced by the most extensive online course providers such as Coursera and EdX with the scope of assessing the essay questions of hundreds of students (Murphy, 2019). Over 500 world-spread universities have used another tool, called "Gradescope" to develop and distribute scoring and numeric assessment (Blumenstyk, 2018). The tool was introduced by Singh et al. (2017), presenting an online system for handwritten homework assignments and exams. The applied method of the tool was flagging the wrong answers and marking the correct ones by providing feedback remarks. Generalized remarks could help a student eliminate repetitive mistakes for the following assessment tasks.

An additional automated scoring engine was mentioned by Perin & Lauterbach (2018) in their assessment of students' writing skills. Based on the Coh-Metrix measures, initially representing automated tools for theoretical and applied language processing within a written text. The Coh-Matrix text analyzer was designed from pre-existing automated systems, permitting the mental construction of relations between ideas in the text.

Another technique for the text questions assessment is Natural Language Processing (NLP) which was detailly reviewed by Zhai et al. (2022). NLP functions on the principle of underlying content or entire phrase patterns as units of analysis. NLP can be used appropriately for analyzing short constructed responses and is therefore appropriate in assessing students' explanations (Chowdhary, 2020). Another review conducted by Zhai et al. (2020) provided statistical evidence that more than ten algorithms have simultaneously been used for the evaluation of written responses to a scientific topic. Further, combined algorithms used simultaneously with different performance weights can achieve better accuracy results than individual algorithms (Wilson & Roscoe, 2021). For achieving a robust human-machine relationship, a model c-rater-ML was created, for assessing students' written argumentation and validation for the formative scientific works, with the support of vector regression (Lee et al., 2021). A systematic review on AI in the field of student assessment conducted by González-Calatayud et al. (2021) referred to a study conducted among engineering students which obtained personalized feedback and performance evaluation based on quantitative and qualitative information. The study's main objective was to continuously monitor and evaluate the artificial intelligence-based Student Diagnosis, Assistance, and Evaluation System (StuDiAsE).

Another application of AI in higher education assessment is exam evaluation; the tool of online proctoring (OP) is broadly considered (Coghlan et al., 2021). OP software uses an AI algorithm that analyzes exam recordings to identify suspicious examinee behaviors or items in their learning environment. According to Chin (2020), OP contains features potentially attractive for universities. The researcher claims that the OP software increases the accuracy of exam supervision and the detection of cheating.

The AI feedback and assessment framework was discussed by Mirchi et al. (2020) in the domain of medical students. The research investigated the creation of the Virtual Operative Assistant, which represented an educational feedback platform based on the medical student proficiency performance benchmarks. Another AI finding in the health science educational application was the development of a valid and reliable psychometric measurement system for assessing the perceived readiness of medical students in the context of study materials (Karaca et al., 2021). Another study aligns well with the AI algorithms with significantly improved tailored feedback; consequently, NLP was created to assess Entrustable Professional Activities (EPAs) and determine patterns in individual resident autonomy (Kirubarajan et al., 2021).

Automated assessment systems deal completely differently with correcting essays and open questions in comparison to numeric or multiple-choice questions with precise analysis on wrong or correct answers to the question. Despite numerous papers on AI applications and algorithms, there is still a growing potential to deal with the complexities of these algorithms used in the context of students' learning processes. Thus, in educational assessment, AI technologies demonstrate a promising future for assessing teachers in grading open questions and providing feedback explanations. Growing expectations are supported by continuous research on AI assessment development in the higher education industry, part of which was presented in this section.

3.1.2 AI applications in different countries in higher education

At some national levels, specific AI applications were implemented within the educational process to handle a global transformation of the education system. Barakina et al. (2021) reviewed existing methods used in the international environment. Considering the application of neural networks as AI techniques, China introduced these methods in the assessment of tests in the form of essays. Such AI-based neural networks, using deep learning algorithms (Chen, 2018), perform in the unpredictable way of criteria selection done by the network as a checkup and, as a result, unexpected grade results for the student. Barakina et al. (2021) mentioned that starting in 2018, Latin American states with local governance initiations started to use virtual assessment in national education widely. Uruguay adopted a learning solution that gives personalized feedback based on the AI analysis of the student work according to the student's knowledge level (Perera & Aboal, 2019), called the 'Mathematics Adaptive Platform.' A similar approach is used in countries such as Brazil with the 'MECFlix' system and Chile.

A representative case of AI in higher education as an emerging area shows Canada that becomes a global leader in implementing AI tools in combination with traditional learning methods, with growing AI hubs located in Montreal, Edmonton, and Toronto cities (Randhawa & Jackson, 2020). Australian universities adopted the OP system discussed in **Chapter 1**, providing convenience and flexibility for educational institutions (Selwyn et al., 2021). A study on the AI development industry in China (Wan, 2021) employed the automated speaking assessment system, with the core function of automized scoring and guiding student for their speaking performance improvement. Computational thinking (CT) was a separate concept that required attention from Chinese teachers. CT represented competency-based assessments to measure students' progress and judge the reliability of CT in education programs (Hsu et al., 2018). Indonesia created LINEN Assessment Apps (Sasmoko et al., 2021), a system acquired with the self-assessment concept for measuring Indonesian students' literacy, numeracy, and entrepreneurial mindset skills.

TU Delft, a technical University in the Netherlands (2021) has initiated an assessment within the Multi-Model Analytics AI model. Study programs examined the AI analysis of the cooperation process between students with the aim of providing feedback. AI system used sensors to percept the dialogue between groups and consequently created foresting models for each group member regarding their contribution and collaboration within the project work.

Asian countries and the American continent show positive progress in AI application adaptation. Also, in Europe, higher institutions manifest actions supporting IT (Information technologies) automation in higher education. The Netherlands launched a campaign for a fast AI promotion, and the acceleration plan involves intelligent digitalization of the teachers and students in higher institutions, supported by innovative tools (Dutch Government, 2021). It, therefore, seems as if the breakthrough of AI assessment algorithms depends on higher institutions' wish to adopt automized applications, as well as significant investments of a country in research and development in the AI assessment field within the educational context (Ye, 2022). Nevertheless, it remains a long way for AI to become ubiquitous in all educational institutions worldwide.

To conclude, there is an estimation of the fast growth of model-driven data analytics approaches that have a scope of guidance in the development, perception, and recognition of the existing algorithms (Luan et al., 2020). As seen in this section, introduced AI assessment systems have various algorithm approaches and goals. A big number of investigated research was done based on different assessment criteria goals, subject features examined and selected sample. The literature review showed that it is challenging to say what is a standardized and appropriate form of AI assessment system used worldwide. For establishing an appropriate AI assessment system of open questions within a higher educational institution, factors such as governmental support in AI development within the country, investment, willingness to have flexibility in the higher education and public behavior to accept the radical changes.

3.2 AI assessment systems

Online featured tools that handle the structured and administrative view of submissions and grading assignments or exams are essential for all universities. The selected environment is SpeedGrader, used in the Canvas environment that the University of Twente manages. Second is Remindo, using AI algothitms to come with efficient analysis of the digital exams provided.

3.2.1 SpeedGrader

It was chosen as an analyzed AI system environments because of contained AI algorithms that help in assessing teachers and teaching assistants in the grading process. Moreover, it contains several features, that represent internal capabilities that can be implemented for other AI grading systems.

SpeedGrader works on the principle of the direct submission of assignments on the web page and allows to assign personalized feedback directly on the user's screen (Oswal, 2019).

Consequently, the reader, in the form of a student or teacher, receives visual notifications if some action was made regarding its work. Mirmotahari et al. (2019) outlined the rubrics part of the SpeedGrader that permits teachers to set up personalized assessment templates according to teaching methods used during the lecture classes. A standardized form is afterward used by teaching assistants or teachers to assess students' work in the form of projects or assignments. SpeedGrader, as an assessing tool, has a "Save" button that permits to save standardized feedback template that later appears automatically in the comment box for the following students' assignments. This feature helps later to navigate through all submitted works and leave fastly recently saved feedback, or choose the comments saved previously by the teaching assistants and adapt them accordingly to the student performance. However, the weak point of this component is unpredictable activity and technical bugs (Mirmotahari et al., 2019) that led to uncertainty among students about whether their exam/assignment had been passed on or not. One cause of the mentioned technical defect of the "Save" button is the difficulty of simultaneously using the SpeedGrader tool assessment rubric by checkers for the same students.
Consequently, a time delay can occur in publishing the grades and feedback. Moreover, it causes time delays for examiners/graders as processing each student's work requires additional time to grade than it was planned from the assigned moment of teaching assistants to the task.
Assessing open questions among students via accessibility and optimization of the LMS itself, which SpeedGrader represents, becomes a big challenge in such shifting and rapidly changing web environments. As a result, SpeedGrader can be updated using AI algorithms that will permit a smooth and quick check of the student assignments in extensive numbering. For possible update implementations, the opinion of the current teaching assistant will be examined in **Section 4.** It will further give insights into the required adjustments of AI assessment systems that should take a goal to support teachers in the open question examination.

3.2.2 Remindo tool

The need for the digital assessment system has arisen in recent years, obtaining additional support due to COVID-19. Digital testing permitted to streamline the entire exam process. Many study programs at the University of Twente required to shift more to digital tools in examinations, using a system called Remindo (Heath et al. 2021). The tool was widely spread among other Dutch universities, including the University of Amsterdam, Leiden University, and Utrecht University. Remindo tool was used in the Utrecht University research, where Vida et al. (2021), defined Remindo Toets as a software product developed by Paragin. This Dutch education company provides an educational institution with a standardized platform to create, manage, grade, and review exams. The University of Twente provides guidelines for using the Remindo environment (see Figure 1). The **4th Stage** "Grade" of the cycle will be in detail considered, as it provides information on the assessment process and helps answer research questions based on the Remindo example. **Stage 5** is essential for AI assessment adjustments, that can have an opportunity to speed up the process of analyzing the examined questions.

Figure 2. Instructions of using Remindo tool in the assessment cycle (Source: University of Twente, 2022)

Open questions of an exam are assigned to different correctors (teachers or teaching assistants) for grading. The exam office sets a correction round up after the exam is finalized. Afterward, the corrector can see the assigned correction work. A filtered feature is available for the corrector, and the assigned person can decide to check all questions per one student or to check all answers per one question. The corrector gives a scoring criterion, and the Remindo exam environment automatically evaluates the total score. After submitting the results, it is impossible to access the correction round repetitively for the corrector. Thereby, it is possible for a teacher to edit the correction for an individual student by using the "Analysis page." An additional possibility is to enable the plagiarism check in case of student similarity work; it can be activated on the request of TAs or teachers.

Another research investigated by Bergmans et al. (2021) analyzed Remindo in combination with Proctorio and concluded that it is an easy-to-use system for students and teaching staff.

Similar to SpeedGrader, reviewed in **Subsection 3.2.1** Remindo should introduce adjustments in the program interface, that will permit to keep updated the AI system. Therefore, interviewed teaching assistants give relevant insights based on current Remindo utilization, that can be used for further AI-driven assessment creation.

4. INTERVIEW RESULTS

This section provides the descriptive statistics on the distribution of teaching assistants (TAs) participated in the research (see **Section 4.1**). Next, the benefits and weaknesses of assessment systems used at the University of Twente are listed (**Section 4.2** and **Section 4.3**). The opinion of TAs regarding the adaptation of AI assessment systems is included in creating insights for future research. The willingness of accept the AI-driven grading system is found in **Section 4.4.**

4.1 Demographics

This section analyses demographic characteristics of the respondents, for the research. In order to comply with statistical requirements, the study aimed to gather data from a minimum 5 till 20 respondents (Adams, 2015; Baker & Edwards, 2012; Galvin, 2015). The study results showed that data was collected from 18 TAs, whereof eight are IBA TAs, six BIT and four TAs from other studies, such as Technical Computer Science, Chemical Engineering, Communication Science and Psychology studies.

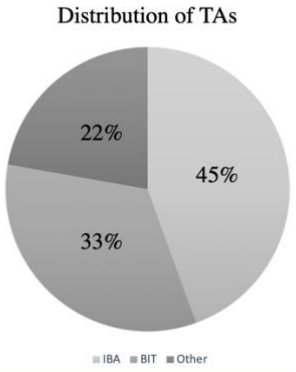

Figure 3. Distribution of TAs within study program

Figure 3 indicates the distribution of interviewers. Educational background, together with the experience in grading and assessment of students, are considered representative variables. The participants' age distribution, gender composition, and population group were not considered representative for answering research questions.

The data relating to the experience of TAs in grading systems such as SpeedGrader and Remindo is equally. The sample included nine respondents, eight from the IBA program and one BIT student who are acknowledged in both grading environments used in assessing assignments and exams by the University of Twente. The other nine TAs used the SpeedGrader tool without Remindo corrector mode.

This descriptive information should be considered while making conclusions, as the sample of TAs who used SpeeGrader is doubled compared to the number of TAs who performed job tasks in the Remindo environment.

4.2 Benefits of the SpeedGrader and Remindo

This section will emphasize the results of data collected from TAs, precisely the advantages of SpeedGrader and Remindo, representing the assessment systems.

Explaining the first feature benefit of SpeedGrader, most interview respondents preferred the "Save" button, which allows saving comments. TA has only to explain the answer once, indicating the distribution points. Afterwards, the same TA or other TAs in the role of correctors use the previously saved comments and apply them to various checked assignments/exams. Consequently, this SpeedGrader feature permits to make time-efficient and repetitive tasks for superfluous. Also, the "Save" button element can be seen as an overview of teachers' grading templates and provides ideas for TAs on how to proceed with grading. When teacher works with the" "Save" button, they show a grading rubric that represent a guidens for teaching assistants. Another advantage mentioned by the interviewed sample was an easy way of distributing the points among the exercises of the checked assignments/exams. It can simultaneously post feedback from TA in the form of a comment, together with the grade received by the student. The option "Hidden" permits postponing the publication moment of the feedback and grades till all students' works are not finished in grading. In case if TA has doubts regarding the grading allocation of a student's assignment/exam, the "Hidden" option allows for double-checking of exercises and TA assessment by the teacher, with the possibility to change grading outcomes. Another benefit discussed by interviewed TAs was accessibility and navigation through all students' works, favorably storing all responses in one place. The list of students is publicly available in the SpeedGrader environment, and TAs can easily click on the specific person/group to which they were assigned for the grading procedure. Therefore, a clear overview of the database of students leads to a user-friendly setting of SpeedGrader, according to the TAs.

Discussion of benefits observed in the Remindo tool are limited due to a restricted sample who were experienced in grading using the Remindo environment. TAs pointed out that it is an optimized platform, with the standardized digital exam format among students. Analyzing TAs' responses, further benefits of Remindo were established, such as an already included grading scheme that increases checker time efficiency for grader. Grading process can be divided into two possibilities for checkers, either to check students per one question, or to check all questions per student. The first option of checking one question for all students can permit to get into the question and increases the productivity of a checker due to standardized and repetitive question assessment. Besides mentioned advantages of Remindo, additional benefits included automated calculation of the total points per student. Questions are highlighted in green when they are already graded; at one time, all questions that still require grading are highlighted in red. This characteristic efficiently works for TAs who keep track of the total number of students' works and contributes to not mistakenly miss some students' exercises.

4.3 Weaknesses of SpeedGrader and Remindo

After conducting the interviews, the observations showed that SpeedGrader and Remindo contain several technical drawbacks that TAs observed and would like to see improved. Starting with the SpeedGrader tool, the "Save" button mentioned in **Section 4.2** under benefits has a reverse side of the coin. If TAs every time use this feature for saving their variants of comments, it results in a vast list of similar comments that cannot be deleted and modified afterward. Consequently, it creates a mess of comments, and TAs get confused about the appropriate comment for a specific student's answer. Besides this, the functionality of SpeedGrader regarding opening the multiple screens simultaneously is limited. For TAs, it is inconvenient to open the grading template provided by the teacher and multiple documents that should be graded within the SpeedGrader system at the same time. As a result, TAs are required to use the second screen opened with a solution file, which leads to repetitive tasks conducted with mistakes, hence, to decreased productivity and focus of TAs, who have to switch every time between screens while grading assignments/exams.

Another technical difficulty occurs when two or more TAs try to check the same student's work at the same time by checking different questions provided in the content of an assignment/exam. At this moment, the SpeedGrader system is crashing and is not saving answers of both correctors, thus creating time delays and the need for a repetitive check of the same assignments/exams. TAs must communicate with each other for updating messages to specify the time distributed among them for checking assignments/exams to avoid the situation mentioned above. SpeedGrader also shows lags when a TA provides a long comment with detailed feedback. Some TAs specified that there is an unclear overview of which questions/students have already been graded. The dropdown search is currently available but searching for a specific student is time-consuming. Hence, SpeedGrader is an online tool; assignments/exams are not always available, and there is a need for manual download and export of different file formats attached to assignments/exams. File formats can contain Excel spreadsheets, RStudio scripts, SPSS code, and laboratory works that the TA should check, but it requires additional time to open these files.

Along with the mentioned drawbacks, there is also missing overview in submitted assignments/exams of teams/pairs working together (group submission). For project groups, TA does not have a full access to the assignments' properties. Therefore, sometimes TA faces the situation to manually align grading for each student, which leads to reduced time efficiency. A further inconvenience is the absence of a "Flag" function for the question and the corresponding student's answer, which can be unclear for the TA and requires attention from other TAs or teachers. Apart from this, a plagiarism check is not introduced automatically in the opened submitted student works interface. while checking. The interviewed TAs suggested that the SpeedGrader system should incorporate a feature for plagiarism check; they suggested the version of matching between students' answers and the TAs solution manual. TAs emphasize that it is impossible to remember 200 students and all their answers; in this condition, cheating is simple.

Remindo received less critisim from TAs, as they find Remindo more time-efficient than SpeedGrader. However, there are some

imperfections identified by TAs. As a result, the absence of the "Save" button of the comments is the first drawback that should be adjusted and taken from the SpeedGrader tool. For some TAs, Remindo shows a confusing interface, providing the possibility to grade per question/student. It requires some time adaptation, as the Remindo environment's structure is not intuitive enough. One TA provided a student perspective on the Remindo exam environment and gave a critical view of the drawbacks: There is no grammar checker was provided during the exam. Due to limited time during the exam duration, students are stressed and do not have time to check all their answers for grammar and spelling. It should be noted that the structure of the Remindo exam environment is unclear for TAs as well as for students. Thus, students cannot see the structure of the exam from the beginning, with all questions that include several subpoints. In addition to this, some exams structures do not allow the students the possibility to move between questions.

It is important to emphasize that all specified drawbacks were analyzed from the TA's point of view, meaning that knowledge of all possible interface characteristics can be limited due to the teacher's desire to protect all assessments from fraud.

4.4 AI assessment system adaptation in higher education

After completing the semi-structured interviews with the participants, it was identified that majority TAs admit that the AI tool to be introduced would be seen as additional help for their work (**see Figure 4**). The arguments of time efficiency supported the benefits of AI involvement in the process of grading. An algorithm can execute repetitive tasks faster than a human. Additionally, the number of students complaining related to the delayed exam results will be eliminated.

In the context of grading open questions, interviewers showed response variations. From the perspective of being a student, four interviewees expressed the wish to be graded by humans when it comes to open questions. Argumentations defined the importance of human input in the assessment process. AI might have a more challenging time "reading between the lines" or understanding metaphors and jargon words used in the answer of a student. The AI interpretation of the student's expressing thoughts can be biased and inconsistent. However, the majority supported adjusting the AI assessment system in the grading process. They emphasize that AI can check students' work and outline keywords or synthetics of the sentence essential for the correct/wrong answer. After that, a recap should be done by the corrector, as a teacher or teaching assistant who will evaluate the AI grading outcomes. In this case, if AI grading were accurate according to the assessment rubric, teachers would spend less time checking it. In the opposite scenario, if the AI assessment system made errors and was not advanced in accepting sensitive answers, it will create additional time that teaching assistants either teachers should recompile and redo the assessment procedures, resulting in time delays.

When a human is grading, it permits the discussion of the comments or the grade afterward, as the humans know why they evaluated work in a specific way. There are concerns that AI technical characteristics can either speed up the process of checking answers or make it longer, based on the number of errors and complexity the algorithm deals with. The other 13 interviewees think in conditionally accepting the AI supporting tool. In the case of AI proper work, the grading process will be faster, making the tool necessary, especially for a large class of students. However, they find grading open questions subjective even for TAs, as TAs have a specific template of the correct answers to take away bias. Sometimes students provide answer that is not included in the template; TA then has to judge himself on how to give points. On the contrary, AI supported tools, can have a few variants and templates that in the end will give 0 points for the student's answer.

From the perspective of technical studies, such as Computer Science and Chemical Engineering, the interviewees emphasized that their majority of students usually requires collecting the data from laboratory experiments or writing different variants of a code with the same output. From this point of view, using AI in evaluating this type of questions can result in a counterproductive and biased outcome. However, they accept the replacement of humans to AI tools in tasks related to the theoretical part of the technical subjects, such as concept explanation and standardized measurement procedures.

Factor such as the possibility of earning money via taking on TA tasks is crucial for students who start work as teaching assistants. The interview finding mentions that most TAs fear that future AI investigations can cut off part of their salary. Despite the repetitiveness of tasks and routine work while grading assignments/exams, they find it a significant argument against the adaptation of AI assessment. Moreover, they find TA work as a personal and professional growth that can positively contribute to their social and professional skills in their career path.

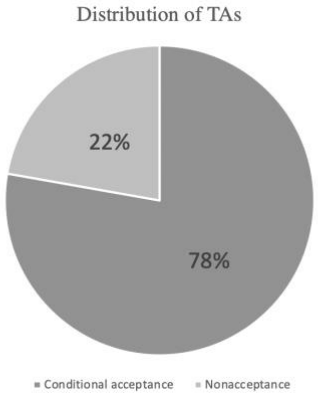

Figure 4 Distribution of TAs regarding acceptance of AI in grading

In conclusion, TAs will need time to adapt to the AI assessment introduction comprehensively. They do not think that in the next ten years, AI-driven assessment models will be developed in such a way that a total TA replacement will occur. Their attitude can be suspicious because they fear being replaced by AI-driven assessment systems. However, they understand that implementing the technologies in the grading process can have a list of advantages for the entire education optimization, including grading text open questions.

5. DISCUSSION AND INSIGHTS

5.1 General insights

The discussion section finding from the literature review on the AI-driven grading systems emphasized that there is no standardized form of the algorithm used in each country for each institution regarding grading open questions. It is essential to consider what approaches and allowed methods are implemented in the government for improving AI assessment systems. Also, literature suggested the importance of the suggested AI-driven assessment of open questions, which should be based on the student's ability, skills, and personal characteristics. After elaborating on the interview results, the practical study supported the idea of the need for personalized feedback. Interviewed TAs

said that the current AI level performs standardized responses and feedback variants. From one point of view, each student is treated equally; from another point, an AI system can fail a student since he uniquely formulates his thoughts, but the meaning of his answers fits the boundaries of the correct response variant. Zhai et al. (2020) explained that more than ten algorithms have simultaneously been used to evaluate written responses to a scientific topic, leading to increased accuracy of algorithms. The interview findings suggested that features from SpeedGrader can be combined with the Remindo aspects and environment to improve the grading AI system currently used at the University of Twente.

Further, the strengths and weaknesses were implied to create future opportunities to answer the possible AI adaptation research question in grading open questions.

Using the literature review and data collected from interviews, one of the major strengths of using AI-driven assessment systems is the time reduction of grading. In the theoretical part, some advanced AI-assessment techniques were enumerated that can fully replace human presence in the grading process. However, in the practical part, TAs expect the AI's ability to identify the keywords, highlight them, and speed up the process of TA for grading. Furthermore, it can count the total score instead of the corrector and make a straightforward interface of grading statistics, meaning how many assignments/exams are graded and still wait to be graded. Another strength of AI-driven grading mentioned in the research is a standardized and unbiased way of grading persons. TAs emphasize that sometimes they can evaluate student works based on their mood, tiredness, and personal attitude for the student. For example, when a student is not performing well in other subjects, and TA is informed about it, he can judge the student and have lower expectations for the grading performance in a particular subject when the TA grades, than in reality.

From the weak points, interviews suggested that their AI-driven assessment systems can include a limited variant of possible responses within the algorithm. The literature review did not show the specific data related to the capabilities of the number of variants included within one algorithm. Still, the research suggested that the more algorithms are included, the better the accuracy of the grading process of the text open questions. Another drawback of AI-driven assessment is the absence of a review session where students can go and ask questions. Interviews mentioned that if the AI grading system gives errors in the grading outcomes and the student will notice that, the teacher or teaching assistant will be responsible for it.

5.2 Possible adjustments

When identifying adjustments, the accuracy concept should be evidential. There should not be a difference in grading results between humans and AI. Therefore, due to the subjective scaling measures of the assessment of open questions, the enormous number of algorithms used within one system is the first adjustment that should be implemented. The subsequent suggested adjustment is the double-check opportunity for TAs. The AI-driven system should not be alone; there should be a constant collaboration between TAs and the AI-driven assessment system; both parts can perform combined work and achieve significant efficiency and productivity. AI system can allocate the points and give a standardized variant of feedback per question, while TA can undertake elaboration on more detailed and personal feedback in case some grading issues arise.

Based on the benefits and weaknesses of SpeedGrader and Remindo tools, adjusted technical improvements can be done for optimizing grading process. The SpeedGrader limitedness of the interface, to open multiple files on the same screen gave insights to integrate interface with more screens available. SpeedGrader adjustments should be implemented in the direction of the interface similarity of Remindo. Another feature suggested, can be taken from the “Comment” function of the Microsoft Word. It can permit specific feedback of a section where student made a mistake. As TAs found the absence of the “marking/flagging” option, it can be included in the updated version, with the purpose of further discussion of the questions together with the teacher. Next insight that can increase the SpeedGrader efficiency as a grading tool, can be an automated plagiarism check. TAs by checking numerous assignments/exams or project works can miss similar works between students/groups. Moreover, sometimes the content of assignments/exams is stable through years, and it is difficult for the TA to observe patterns of plagiarism from different years’ works. TA having two positions of student and teacher is limited in the data access of the assignments/exams used before. Therefore, TA cannot see the answers similarity and a teacher would have additional workload to check this afterwards.

5.3 Academic contribution

Past studies mostly emphasized the appliance of AI in education from learning process context (Goksel & Bozkurt, 2019). However, AI utilization is still an emerging topic in higher education research (Ouyang et al., 2022). This is especially relevant for studies of AI assessment outside the text open questions. For this reason, research on assessment criteria within AI is still relatively rare and even more so in combination with country’s higher education’s dimensions. For these reasons, this study will add to the existing literature by summarizing possible internationalized AI assessment techniques, exploring teaching assistant individual opinion on how AI influence higher education and how AI can assist professors in grading. Moreover, TAs give relevant insights of two commonly used AI tools used the context of the University of Twente grading environment, SpeedGrader and Remindo. These tools are widely used outside of the institution and can work as a start point for other researchers from different universities, to implement and update their personalized grading systems within AI automatization implementation.

Identifying human collaboration within technical patterns will create opportunities to further analyze and test specific algorithms in the context of the country, institution and willingness to adapt to the digital assessment. Additionally, teaching assistants' valuable contributions can be reused and developed for the future promising opportunities of AI assessment systems, as the current literature reviewed showed limited number of the research done from the teaching assistant perspective.

5.4 Practical contribution

The paper will provide new insights into technical AI differences between countries at the higher educational level. Internationalized methods will permit higher educational institutions, identifying appropriate AI methods in dependence of the algorithm applied and type of questions asked during the tests. Teaching assistant perspective and opinion, regarding the use of AI model assigned to help grading text questions can be used as a starting point to understand behavior attitude of both groups, teachers, and students. Teaching assistants represent middle persons and can give objective insights on acceptance of AI implementation from student perspective as well as from graders position. SpeedGrader and Remindo as examined assessment systems, now have a list of benefits and drawbacks that can be used for continuous development of the systems. The

weaknesses mentioned should be a start point for continuous AI adjustments that can be reused for a variety of AI assessment systems created afterwards. Moreover, this paper can then provide indications to better guide teachers in creating AI assessment tools, that will support them and teacher assistants in grading open questions.

6. CONCLUSION AND LIMITATIONS

In this paper we answered the research question: "What are the suggested adjustments of existing AI assessment systems that should considered while grading open questions based on teaching assistants' perspective? The purpose of it was to establish the support of AI technologies for teachers and teaching assistants in the assessment process. Adjustments were based on the relationship between strengths, weaknesses, and possible opportunities of current AI-driven systems. Implementing an AI algorithm on the theoretical question can be more accessible, which requires a specific base of concept knowledge from the student that can be assessed using keywords. However, then work assumes required personalized data that should be fulfilled, such as IT code for resolving a mathematical problem or a laboratory observation based on the experiments, AI has limits in the perception of what is correct and what is wrong. Even the use of figures of speech, for instance, metaphor, hyperbole, or paradox, can lead to AI recognition resulting in disproportionate weight in favor of or against the student's answer. Adaptation suggestions were based on recommendations of analyzed SpeedGrader and Remindo grading systems considering the teaching assistant perspective. The research showed that AI systems should focus on automatically giving standardized and detailed feedback based on students' answers while TAs assess the system. The additional search was done with the qualitative analysis of 18 responses from the TAs. They outlined the benefits of AI; however, they also pointed out potential risks that should be considered before implementing the tool on the institutional level. The research found that accepting AI will represent an incremental change in the educational system. Students do not show complete trust in intelligent technologies that will assess them. Changing their behavior that will lead to a full acceptance of AI-driven assessment opportunities require time, investment, and experiments. Moreover, the examined literature review showed that a limited number of papers were investigated for AI assessment systems at the national higher education level. It concludes that most nations are skeptical about giving complete trust to AI-driven assessment in the academic fields.

The additional search was done with the qualitative analysis of 18 responses from the TAs. They outlined the benefits of AI; however, they also pointed out potential risks that should be considered before implementing the tool on the institutional level. The research found that accepting AI will represent an incremental change in the educational system. Students do not show complete trust in intelligent technologies that will assess them. Changing their behavior that will lead to a full acceptance of AI-driven assessment opportunities require time, investment, and experiments. Moreover, the examined literature review showed that a limited number of papers were investigated for AI assessment systems at the national higher education level. It concludes that most nations are skeptical about giving complete trust to AI-driven assessment in the academic fields.

6.1 Limitations

It is difficult to predict how AI will evolve in the upcoming years and what trends will emerge in its application for higher education. It is essential to mention that AI requires much research before its implementation in assessing the open-ended questions. Although the literature review for the current study is from 2018 onwards, it can be outdated at the moment of the new technological boom or trend that arose in the domain of intelligent technologies. Another issue is the inability to demonstrate the impact of different applied AI technologies in the context of the University of Twente. AI grading systems such as SpeedGrader and Remindo can become permanently closed for public use due to the outdated algorithms and methodologies used within the tools. Another issue was raised by representatives and their opinion regarding the adaptation and willingness to use AI in higher education, and the level of acceptance can be biased due to the sample selected within one higher education institution. Finally, the stated benefits and weaknesses can be biased and associated with obstacles for the researchers who did not experience the same situation as TAs while grading via SpeedGrader and Remindo.

7. ACKNOWLEDGMENTS

I would like to express my deepest appreciation to Dr. P. B. Rogetzer and Dr. Rer. Nat. D. Braun for supervising and supporting me throughout this research. Especially, to my first supervisor for guiding me during this research with practical feedback and for the clear explanations during the discussion sessions. Moreover, I would like to thank all the respondents for their time and valuable input provided during the interview sessions, which helped me to come up with clear research conclusions and finalize this research. I would like to express my gratitude to the University of Twente IBA advisors, who contributed for students throughout the three years bachelor. Finally, I'd like to express my appreciation to the reader who took the time to read my research.

8. REFERENCES

1. Adams, W. C. (2015). Conducting Semi-Structured Interviews. *In Handbook of Practical Program Evaluation: Fourth Edition* (pp. 492–505). Wiley Blackwell. https://doi.org/10.1002/9781119171386.ch19

2. Agarwal, P., Swami, S., & Malhotra, S. K. (2022). Artificial Intelligence Adoption in the Post COVID-19 New-Normal and Role of Smart Technologies in Transforming Business: a Review. *Journal of Science and Technology Policy Management*., Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JSTPM-08-2021-0122

3. Akgun, S., & Greenhow, C. (2021). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. AI and Ethics, Issue, Number, 1-10.https://doi.org/10.1007/s43681-021-00096-7

4. Alsanie, W., Alkanhal, M. I., Alhamadi, M., & Alqabbany, A. O. (2022). Automatic scoring of arabic essays over three linguistic levels. *Progress in Artificial Intelligence*, *11*(1), 1-13. DOI: https://doi.org/10.1007/s13748-021-00257-z

5. Aluthman, E. S. (2016). The effect of using automated essay evaluation on esl undergraduate students' writing skill. *International Journal of English Linguistics*, 6(5), 54. https://doi.org/10.5539/ijel. v6n5p54

6. Anwar, S., Bascou, N. A., Menekse, M., and Kardgar, A. (2019). A systematic review of studies on educational robotics.J. *Pre-College Eng. Educ. Res.* (J-PEER)9,19–42. doi: 10.7771/2157-9288.1223

7. Baker, S. E., Edwards, R., & Doidge, M. (2012). *How many qualitative interviews is enough?: Expert voices and early career reflections on sampling and cases in qualitative research.*

8. Barakina, E. Y., Popova, A. V., Gorokhova, S. S., & Voskovskaya, A. S. (2021). Digital Technologies and Artificial Intelligence Technologies in Education. *European Journal of Contemporary Education*, 10(2), 285-296. DOI: 10.13187/ejced.2021.2.285

9. Blumenstyk, G. (2018). Can artificial intelligence make teaching more personal. *The Chronicle of Higher Education*, 1-3. Https://www.chronicle.com/article/CanArtificialIntelligence/243023 (date of last access 20.05.2022)

10. Bywater, J. P., Chiu, J. L., Hong, J., & Sankarana-ayanan, V. (2019). The teacher responding tool: Scaffolding the teacher practice of responding to student ideas in mathematics classrooms. Computers & Education, 139, 16–30. https://doi.org/10.1016/J.COMPEDU.2019.05.004

11. Chen, 2018 –Chen, S. (2018). China's schools are quietly using AI to mark students' essays... but do the robots make the grade? https://www.scmp.com/news/china/society/article/2147833/chinas-schools-are-quietly-using-ai-mark-students-essays-do (date of last access: 09.05.2022)

12. Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75,264–75,278. https://doi.org/10.1109/ACCESS.2020.2988510

13. Chen, P., Lu, Y., Zheng, V. W., Chen, X., & Yang, B. (2018). KnowEdu: A system to construct a knowledge graph for education. *IEEE Access*, 6, 31,553–31,563. https://doi.org/10.1109/ACCESS.2018.2839607

14. Chin, M. (2020). Exam anxiety: How remote test-proctoring is creeping students out. The Verge. https://www.theverge.com/2020/4/29/21232777/examity-remote-test-proctoring-online-class-education. (date of last access 22.05.2022)

15. Coghlan, S., Miller, T., & Paterson, J. (2021). Good proctor or "big brother"? Ethics of online exam supervision technologies. *Philosophy & Technology, 34(4)*, 1581-1606. https://doi.org/10.1007/s13347-021-00476-1

16. Chowdhary, K.R. (2020). Natural Language Processing. In: Fundamentals of Artificial Intelligence. Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_19

17. Coman, C., Țîru, L. G., Meseșan-Schmitz, L., Stanciu, C., & Bularca, M. C. (2020). Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective. *Sustainability*, *12*(24), 10367. https://doi.org/10.3390/su122410367

18. Dutch Government (2021). *The Dutch digitalization strategy 2021* url: https://www.nederlanddigitaal.nl/binaries/nederlanddigitaal-nl/documenten/publicaties/2021/06/22/the-dutch-digitalisation-strategy-2021-eng/210621-min-ezk-digitaliseringstrategie-en-v03.pdf (date of last access 16.06.2022)

19. Joseph, M. C. (2014). Becoming a Qualitative Researcher: A Narrative Account of Conducting My First Qualitative Study Involving In-depth Life History Interviews. *Online Journal of Social Sciences Research*, *3*(1), 10-20. http://onlineresearchjournals.org/JSS/pdf/2014/jan/Molefe.pdf

20. Gabriel F. , Marrone R., Van Sebille I., Kovanovic V. & Laat de M.(2022) Digital education strategies around the world: practices and policies, *Irish Educational Studies, 41:1*, 85-106, https://doi.org/10.1080/03323315.2021.2022513

21. Galvin, R. (2015). How many interviews are enough? Do qualitative interviews in building energy consumption research produce reliable knowledge? *Journal of Building Engineering, 1, 2–12.* https://doi.org/10.1016/j.jobe.2014.12.001

22. Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning*, 37(5), 1207-1216. DOI: 10.1111/jcal.12577

23. Goksel, N., & Bozkurt, A. (2019). Artificial intelligence in education: Current insights and future perspectives. In *Handbook of Research on Learning in the Age of Transhumanism* (pp. 224-236). IGI Global. https://doi.org/10.4018/978-1-5225-8431-5.ch014

24. González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, 11(12), 5467 https://doi.org/10.3390/app11125467

25. Hsu, T. C., Chang, S. C., & Hung, Y. T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education*, 126, 296–310 https://doi.org/10.1016/j.compedu.2018.07.004

26. Jopp, R., & Cohen, J. (2020). Choose your own assessment–assessment choice for students in online higher education. *Teaching in Higher Education*, 1-18. DOI: 10.1080/13562517.2020.1742680

27. Karaca, O., Çalışkan, S. A., & Demir, K. (2021). Medical artificial intelligence readiness scale for medical students (MAIRS-MS)–development, validity and reliability study. *BMC medical education*, 21(1), 1-9. https://doi.org/10.1186/s12909-021-02546-6

28. Kirubarajan, A., Young, D., Khan, S., Crasto, N., Sobel, M., & Sussman, D. (2021). Artificial Intelligence and Surgical Education: A Systematic Scoping Review of Interventions. *Journal of Surgical Education.* https://doi.org/10.1016/j.jsurg.2021.09.012

29. Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J., Ogata, H., ... & Tsai, C. C. (2020). Challenges and future directions of big data and artificial intelligence in education. *Frontiers in psychology*, *11*, 580820 https://doi.org/10.3389/fpsyg.2020.580820

30. McCorduck, P., & Cfe, C. (2004). Machines who think: A personal inquiry into the history and prospects of artificial intelligence. *CRC Press*. https://www.jstor.org/stable/24966276 (last date of access 04.05.2022)

31. Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & Del Maestro, R. F. (2020). The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PloS one*, 15(2), e0229596. https://doi.org/10.1371/journal.pone.0229596

32. Mirmotahari, O., Berg, Y., Fremstad, E., & Damsa, C. (2019). Student engagement by employing student peer reviews with criteria-based assessment. *In 2019 IEEE*

Global Engineering Education Conference (EDUCON) (pp. 1152-1157). https://doi.org/10.1109/EDUCON.2019.8725174

33. Monitor, I. C. E. F. (2018). Study projects dramatic growth for global higher education through 2040. *Retrieved July, 20*, 2020. URL: https://monitor.icef.com/2018/10/study-projects-dramatic-growth-global-higher-education-2040/ date of last access (16.06.2022)

34. Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments.* https://doi.org/10.1080/10494820.2020.1727529

35. Mujtaba, D. F., & Mahapatra, N. R. (2020, December). Artificial Intelligence in Computerized Adaptive Testing. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 649-654). https://doi.org/10.1109/CSCI51800.2020.00116.

36. Murphy, R. F. (2019). Artificial intelligence applications to support K-12 teachers and teaching. *Rand Corporation*, *10* https://doi.org/10.7249/PE315

37. Nye, B. D. (2015). Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education*, *25*(2), 177-203. https://doi.org/10.1007/s40593-014-0028-6

38. OECD. 2020. PISA 2018 Results (Volume V): Effective Policies, Successful Schools, PISA, OECD *Publishing, Paris.* https://doi.org/10.1787/ca768d40-en

39. Oswal, S. K. (2019). Disability, ICT and eLearning platforms: Faculty-facing embedded work tools in learning management systems. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 105-111). https://doi.org/10.1145/3308561.3355620

40. Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*, 1-33 https://doi.org/10.1007/s10639-022-10925-9

41. Perera, M., & Aboal, D. (2018). The impact of a Mathematics Computer-Assisted Learning Platform on Students' Mathematics Test Scores. *Fundación CEIBAL*. https://hdl.handle.net/20.500.12381/415 (last date of access 29.05.2022)

42. Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, 28(1), 56-78. https://doi.org/10.1007/s40593-016-0122-z

43. Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied psychological measurement*, *23*(3), 187-194. https://journals.sagepub.com/doi/pdf/10.1177/01466219922031310?casa_token=nz_iUzBLvugAAAAA:6Bf7Okh4xXQGkqrGb2Q5vFPY7R6T8-6JkggmiE0n0apgdP8nK-zfwJwi6FI3GWxGIhgGvFwJfRM (last date of access 16.05.2022)

44. Randhawa, G. K., & Jackson, M. (2020, January). The role of artificial intelligence in learning and professional development for healthcare professionals. *In Healthcare management forum* (Vol. 33, No. 1, pp. 19-24). https://doi.org/10.1177/0840470419869032

45. Sasmoko, G. N., Indrianti, Y., & Ronald Hermanus, D. (2021, December). LINEN ASSESSMENT APPS: Artificial Intelligence Integration to develop Android-based applications that measure the" literacy, numeracy, entrepreneurial mindset" capacity of students in Indonesia. In *2021 4th Artificial Intelligence and Cloud Computing Conference* (pp. 160-164). https://doi.org/10.1145/3508259.3508295

46. Selwyn, N., O'Neill, C., Smith, G., Andrejevic, M., & Gu, X. (2021). A necessary evil? The rise of online exam proctoring in Australian universities. *Media International Australia*,107, 2411–2502. https://doi.org/10.1177/1329878X211005862

47. Singh, A., Karayev, S., Gutowski, K., & Abbeel, P. (2017, April). Gradescope: a fast, flexible, and fair system for scalable

assessment of handwritten work. In *Proceedings of the fourth (2017) acm conference on learning scale* (pp. 81-88). http://dx.doi.org/10.1145/3051457.3051466

48. Smith, A., Leeman-Munk, S., Shelton, A., Mott, B., Wiebe, E., & Lester, J. (2019). A multimodal assessment framework for integrating student writing and drawing in elementary science learn- ing. *IEEE Transactions on Learning Technologies*, 12(1), 3–15. https://doi.org/10.1109/TLT.2018.2799871

49. Smutny, P., and Schreiberova, P. (2020). Chatbots for learning: a review of educational chatbots for the facebook messenger. *Computers & Educ*ation 151:103862. https://doi.org/10.1016/j.compedu.2020.103862

50. Syzdykbayeva, A., Baikulova, A., & Kerimbayeva, R. (2021). Introduction of Artificial Intelligence as the Basis of Modern Online Education on the Example of Higher Education. In *2021 IEEE International Conference on Smart Information Systems and Technologies* (SIST) (pp. 1-8). IEEE. https://doi.org/10.1109/SIST50301.2021.9465974.

51. Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I., & Pears, A. (2021). Teaching Machine Learning in K–12 Classroom: Pedagogical and Technological Trajectories for Artificial Intelligence Education. *IEEE Access*, 9, 110558-110572. https://doi.org/*10.1109/ACCESS.2021.3097962*

52. Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, 14(3), 207-222. https://doi.org/10.1111/1467-8551.00375

53. TU Delft University (2021). AI in education – the possibilities are promising. URL: https://www.tudelft.nl/teachingacademy/newsletters/the-educator-june-2021/ai-in-education-the-possibilities-are-promising (date of last access: 12.06.2022)

54. University of Twente (2022). Grading in Remindo – Stepwise. URL: https://www.utwente.nl/en/educational-systems/about-the-applications/remindo/remindo-4-grading/grading-interface/ (date of last access: 16.06.2022)

55. Wan, X. (2021). A study on the current development of Artificial Intelligence in education industry in China. In *2021 7th International Conference on Education and Training Technologies* (pp. 28-35). https://doi.org/10.1145/3463531.3463536

56. Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1), 87-125. https://doi.org/10.1177/0735633119830764

57. Xiao, Y., & Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. In *Journal of Planning Education and Research* (Vol. 39, Issue 1, pp. 93–112). SAGE Publications Inc. https://doi.org/10.1177/0739456X17723971

58. Xu, W., & Ouyang, F. (2021). A systematic review of AI role in the educational system based on a proposed conceptual framework. *Education and Information Technologies*, 1-29. https://doi.org/10.1007/s10639-021-10774-y

59. Zawacki-Richter, O., Marín, V. I., Bond, M., & 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. https://doi.org/10.1186/s41239-019-0171-0

60. Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching.* https://doi.org/10.1002/tea.21773

61. Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, 56(1), 111–151 https://doi.org/10.1080/03057267.2020.1735757

9. APPENDICIES

Appendix A – List of abbreviations

LIST OF ABBREVIATIONS

AI	Artificial Intelligence	**LMS:**	Learning Management Systems
AES	Automated essay scores	**NLP**	Natural Language Processing
AEE	Automated essay evaluation	**OP**	Online Procotring
CAT	Computerized adaptive testing	**SLS:**	Systematic Literature Search
CT:	Computational thinking	**TA:**	Teaching Assistant
EPA:	Entrustable Professional Activities		
IT:	Information Technology		

Appendix B – Interview questions

1. How much time do you spend manually grading open question (text question/ explanation of a relationship) provided on the exam per student?

2. From a student's perspective, in the near future (10 years) would you like to be graded during the exam (with text open questions) only by an AI tool, by humans or do you prefer a hybrid approach? Justify your answer.

3. From a TA perspective, do you find Speedgrader a useful tool to effectively and efficiently grade open questions and save time for teachers (compared to manual grading)?

4. What features would you like to add in Speedgrader in order to make this tool more efficient for grading open questions (statistics/information systems/ theoretical subject)?

5. What benefits and weaknesses does a grading system like Remindo have?

6. In grading exam with text open question (concept explanation), do you find it necessary to use an additional AI tool that will speed up the process of grading?

7. Do you think the use of an AI tool can outsource some of your (TA) tasks and do automized and repetitive work for you?

8. What grading tasks will not be able to be done by an AI system alone without your assessment?

9. Do you think in near future (10 -20 years) an AI tool can totally replace teaching assistants in higher education?

Appendix C – Interview results in chronological order

Interviewee	*Average Time spent grading one exam open question, per student (Statistics/Information Systems subject)*	*Willingness to be graded by AI/human from student view*	*Benefits and weaknesses of SpeedGrader, from TA view*	*Benefits and drawbacks of Remindo*	*Necessity to use AI in grading open questions, from TA view*	*Features of AI in grading open questions*	*Replacement of TAs in near future by AI*
1st TA **International Business Administration studies**	7 minutes per student	Human Reasons: • The possibility to ask questions, comment on the assignment/exam • Statistics can be graded by AI, but theoretical subjects not	Benefits: Speedgrader allows to save comments (copy paste task) Weaknesses: • Repetitive tasks and mistakes can be done by AI • 2 and more persons can't work on the same student at the same time (system is crashing and is not saving answers) • Too many variants of comments that can't be deleted and modified, after "save" bottun	Remindo is efficient Weaknesses: To add saving comments button, same as in SpeedGrader	Not necessary, but useful for TAs	**Saves time** **Avoids repetitive mistakes,** due to human factors (tiredness) AI **can't find the reason of the mistakes** made (in Statistics), TA knows better as he/she was in the role of student AI algorithm will identify the mistake fastly but will **not provide feedback**	No, AI can be used as an extra tool. Change requires time and adaptation from all

2nd TA **International Business Administration studies**	5 minutes per student	Hybrid Reasons: • Human **brain better understands deviations in open** questions • AI tool makes automatization easy	Benefits: Speedgrader saves pre-made comments, which apply to a wide variety of students Weaknesses: **Overview lost in submitting assignments/exams due to teams/pairs working together** **Align grading in project groups leads to reduced time efficiency**	No experience in grading via Remindo	Not necessary, but helpful	AI tool which recognizes key words and concepts **Speeds up the process of reading unnecessary information in answers** **Recognition of the concepts that are not defined by the original name, but have the same purpose** Not able to provide argumentation and give own opinion (unstandardized)	If AI is focused on standardized answers, then no

3rd TA **International Business Administration studies**	15 minutes per student	Hybrid Reasons: • AI tool speeds up the grading process • Human-made errors are avoided • **Systematic or model errors occur, that can be compensated by humans**	Benefits: • SpeedGrader posts grades and comments at the same time Weaknesses: • SpeedGrader is lagging when writing long comments • Inconvenient to open grading scheme and multiple documents that should be graded	Benefits: • Questions highlighted in green, when they are already graded • System calculated total points per student Weaknesses: • Remindo system opens window asking to put "0" points, when you still don't know how many points should be assigned	It if works, then it can be useful. Problematic	Open questions are open for interpretation, **the error of an AI would be high** **Human has to proofread the AI graded open questions, it will not save time** **Some scales and measures when grading project are subjective, it can't be automized** AI tool can be useful in grading type of questions with right and wrong format	AI can replace most of the tasks of Tas, but not all of them

4th TA **International Business Administration studies**	15 minutes per student	Hybrid Reasons: • AI tool to have grading process more consistent • Depends on the level of AI • Human touch for the grading is still necessary	Benefits: • SpeedGrader is quite efficient Weaknesses: • Another screen open with solution file • • "Save comment" function is not working well • Grading simultaneously is not possible • Functionality of SpeedGrader regarding opening of other screens is quite limited, interface of more screens can be integrated more ins SpeedGrader	Bnefits: • Remindo is a good tool • No need to change anything while selecting grading options Weaknesses: • Confusing interface (grade per question, grade per student) • No "save comment" issue	Not sure, but TAs are wasting a lot of time on doing basic tasks	Open questions require accuracy Handwriting in some online assignments can be challenging for AI tool A lot of nuances in grading, can't be robotic	Unrealistically ambitious AI can replace half of hours

5th TA **Chemical Engineering studies**	20-30 minutes per student for lab journals (no experience in grading Statistics and Information Systems subject)	Hybrid Reasons: • Grading with AI but under humans' control • By making clear in between scientific results (exact number, procedure, theory) and answers on their own perspective	Benefits: • No difference between grading by hand and looking in grading manual • Provides some ideas how to comment, using samples evaluation from teacher or other TAs Weaknesses: • **SpeedGrader should provide system of matching between students answers and TAs manual**	No experience in grading	Depends on the situation Recognition of similarity between students' answers or structure of journals can speed up grading process If reignition went wrong, due to difference in handwritten answers of different students, it will be time consuming to fix it manually	AI can be used in checking the structure of lab journal, results, background knowledge Questions requiring interpretation and explanation it is better to grade manually.	In some tasks

6th TA **Industrial Engineering and Management studies**	5 minutes per student	Hybrid But not as split up 50% teacher grades and 50% AI	Benefits: Useful Weaknesses: • **Assignments/** **Exams are not available every time, there is need for manual download and export of files** • SpeedGrader should be adjusted to look similar as Remindo	Optimized platform	Not necessary Can create several issues Depends how the system is trained, what kind of algorithms are used **The accuracy of the grade: Can create difference in grading between human and AI**	**In fields such as arts, literature difficult to implement AI** **Using metaphor, jargon or phrases, AI can percept it incorrectly** Due to student stress and rush, some signs can be skipped in equations/sentences, AI can be confused and give 0 points	No trust People need to adapt to the change

7th TA **Business and Information Technology studies**	10 minutes per student	Hybrid Reasons: • No trust in AI in open questions, subjective • No 100% trust in TAs, as they don't want to understand what the student wanted to say • Double check from professor	Benefits: • User friendly • All students are available, and we can click on specific person Weaknesses: • Incorporate feature for plagiarism • TAs can't remember 200 people and all student answers, cheating is very simple	No experience in Remindo checking Weaknesses from student perspective: • No Grammar checker provided during the exams • No possibility to move between questions • To see the structure of the exam, from the beginning, with questions that include subpoints	Not necessary Answers are subjective	Students can provide answers not included in the TA's template, TA has to judge himself how to allocate points, discussing with teacher AI tool can give 0 points in the end Psychology fields no possible to use AI tool In BIT perspective, the programming code can be subjective, depending on the student's method to write it	I don't think Open questions are difficult to grade, no trust in AI, sometimes disagreement with TAs

8th TA **Business and Information Technology studies**	10 minutes per student	Human Reasons: • Human input is valuable • Possibility of the feedback • AI might have a harder time "reading between lines"	Benefits: Having all responses in one place Weaknesses: • Unclear overview of which questions/students have been graded • Currently, the dropdown search is available and search for a certain student is time-consuming • No example answer to the question, shown alongside for each question	Benefits: Grading criteria is included, it can be more time efficient Weaknesses: The structure of the environment is not being intuitive enough	Depends on the situation AI can on one hand speed up the process or make it longer, based on how it is configured	AI can make automized and repetitive tasks AI can't give relevant feedback	I don't think

9th TA **International Business Administration studies**	10 minutes per student	Human Reasons: • AI would not be able to comprehend the way each student expresses his/her thoughts • It might fail amazing idea, because of unstandardized answer	Benefits: • Provides quick overview of the grading progress of a group/student • Offer simple access to any attached files Weaknesses: Not clearly defined grading ranges	Benefits: • A clear list highlighting student works that need to be graded • The format of the exam is kept • Teachers can write tips on grading each exercise right above the comment section Weaknesses: No clear overview of the grades and not yet graded exams	No, I find it ineffective	AI can't verify Excel files attached, complicated calculations supported by open questions	No Humans can't be replaced by AI

10th TA **International Business Administration studies**	10-20 minutes per student	Human Reasons: • AI can replace human when it comes to grading Multiple Choice questions • Human is more able to adapt to different answers	Benefits: • Saves a lot of time, due to "Safe answer" option • You only must explain the answer ones and then reuse it Weaknesses: • No possibility to work at the same time multiple people • If there is a question you are not sure about, a marking option can be included for farther discussion with the teacher	No experience in grading via Remindo	No, it is not necessary TAs have a nice side job with a lot of flexibility	AI is not useful in grading open questions Statistics is an exception, as there is usually only one correct answer in the end of the open question Exam reviews require TAs, AI will not be able to explain the allocation of points and possible improvement for the student	The technology will be there to replace TAs and it will be possible But not prefererable

11th TA **Psychology studies**	10-15 minutes per student	Hybrid Reason: • AI can be quicker than humans, so students receive grades in a faster way	Benefits: • A great oversight of question and answer of the student • Easy give amount of points Weaknesses: • Not flag function for the question which is unclear for TA • To introduce "comment" function similar to Microsoft Word for specific feedback	Benefits: Possibility to see the correct answer next to the student's answer Possibility to check per question instead per student, permits to get into the question No weakness points	Yes, it can be very handy	AI can filter certain words which are used, so it will permit for TA and teacher to see if student mentioned certain concept Time efficient AI can't provide feedback in a constructive manner	It probably could do in some way, but TA has a lot of functions – for example answering questions during the lectures and tutorials It doesn't seem that AI will do these tasks in near future

12th TA **International Business Administrationstudies**	For Business Law and Entrepreneurship 15 minutes For Statistics 5	Hybrid Reasons: • TA sometimes don't like students • AI can be quicker than humans, in case if it is advacned system • How good is the AI clarification	Benefits: Making a preparation thing efficiently, can reduce time Weaknesses: • Calculating the score is done manually • Can't make notes for the students as highlighting **Suggestion: provide a clear guideline of use of SpeedGrader for TAs**	Benefits: • For exam check you can access anywhere every time • Students have bad handwring, so digital assessment is a plus • Grading is easy, example is provided No weakness points	Yes, on advanced level of AI it can be very handy of checking key words	AI can filter certain words which are used, but it is not needed AI can be more standardized, TA can be relaxed TAs have a nice job, for international students to earn some money AI be limited in word detection, it can allocate points not so correctly Sometimes the grading is done not for full points 0.5, 0.7	It probably could do in some way, but I am not sure if it should **TA personal learnings and growth**