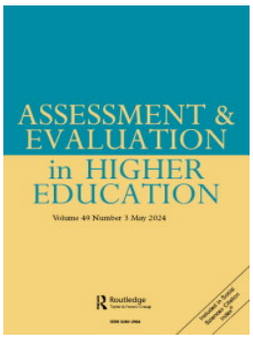

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: www.tandfonline.com/journals/caeh20

Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices

Alexandra Farazouli, Teresa Cerratto-Pargman, Klara Bolander-Laksov & Cormac McGrath

To cite this article: Alexandra Farazouli, Teresa Cerratto-Pargman, Klara Bolander-Laksov & Cormac McGrath (2024) Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices, Assessment & Evaluation in Higher Education, 49:3, 363-375, DOI: 10.1080/02602938.2023.2241676

To link to this article: https://doi.org/10.1080/02602938.2023.2241676

Published online: 01 Aug 2023.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
2024, VOL. 49, NO. 3, 363–375
https://doi.org/10.1080/02602938.2023.2241676

Routledge
Taylor & Francis Group

Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices

Alexandra Farazouli, Teresa Cerratto-Pargman, Klara Bolander-Laksov and Cormac McGrath

Stockholm University, Stockholm, Sweden

ABSTRACT

AI chatbots have recently fuelled debate regarding education practices in higher education institutions worldwide. Focusing on Generative AI and ChatGPT in particular, our study examines how AI chatbots impact university teachers' assessment practices, exploring teachers' perceptions about how ChatGPT performs in response to home examination prompts in undergraduate contexts. University teachers ($n=24$) from four different departments in humanities and social sciences participated in Turing Test-inspired experiments, where they blindly assessed student and ChatGPT-written responses to home examination questions. Additionally, we conducted semi-structured interviews in focus groups with the same teachers examining their reflections about the quality of the texts they assessed. Regarding chatbot-generated texts, we found a passing rate range across the cohort (37.5−85.7%) and a chatbot-written suspicion range (14–23%). Regarding the student-written texts, we identified patterns of downgrading, suggesting that teachers were more critical when grading student-written texts. Drawing on post-phenomenology and mediation theory, we discuss AI chatbots as a potentially disruptive technology in higher education practices.

KEYWORDS

AI-chatbots; assessment; higher education; home examination; Turing test

Introduction

The launch of ChatGPT in November 2022 sparked debate about the impact of AI chatbots in education. Countless media articles, online discussions and forums discussed issues related to plagiarism and academic integrity, reporting incidents of students' use of ChatGPT for their examination (Kirshner 2023; Jane 2023) as well as several attempts to develop AI-written-text detectors (Bowman 2023). Several institutions announced bans on ChatGPT (Shen-Berro 2023), issued a series of guidelines (The McGraw Center for Teaching & Learning, Princeton University 2023) and held online events to inform and support teachers.

CONTACT Alexandra Farazouli alexandra.farazouli@edu.su.se

Recent studies speculate that ChatGPT and similar AI-Chatbots may lead to fundamental changes in education and in assessment and evaluation practices in particular (Zhai 2022). Several studies piloting ChatGPT highlight the benefits that AI chatbots could possibly bring to education, suggesting new opportunities and modes of learning (Mollick and Mollick 2022; Nikolic et al. 2023). However, such studies suggest that empirical research on the topic is required to further investigate the impact, challenges and risks of AI chatbots in educational settings. This study explores empirically the impact of ChatGPT on teachers' assessment practice.

Research on assessment and evaluation in higher education has examined issues regarding e-assessment (Bearman, Nieminen, and Ajjawi 2023), online examinations (Dawson, Nicola-Richmond, and Partridge 2023), academic integrity issues related to online examinations (Huber et al. 2023) and teachers' perceptions of e-assessments (Mimirinis 2019), indicating that further attention and research is needed to explore the impact of technological innovations in assessment practices. To date, the field has not examined issues related to Generative Artificial Intelligence (GAI) and AI chatbots' impact on assessment in greater detail. Consequently, the aim of this paper is to explore the potential impact of AI chatbots on assessment practices in higher education. To that end, we identify ChatGPT as a case of a state-of-the-art AI chatbot. We refer to the first ChatGPT version issued in November 2022 including all updates until March 2023. The research questions guiding this study are:

RQ1: 'How well does ChatGPT 'perform' in home examination at the undergraduate level?'
RQ2: 'What is the mediating role of ChatGPT on teachers' assessment practices?'.

We focus on the disciplines of philosophy, law, sociology and education where home examinations requiring students to produce long-format text answers are common practice. We consider home examination as full or partial mandatory fulfilment of the course final grade. We designed a two-step study, which included an experimental part inspired by the Turing Test (TT) and follow-up semi-structured focus group interviews. We adopt a post-phenomenological lens (Rosenberger and Verbeek 2015), to discuss ChatGPT's mediating role on teachers' assessment practices and better situate its potential impact on higher education.

Generative AI and the emergence of AI-chatbots

GAI refers to a type of AI that is based on unsupervised machine learning models, pre-trained on certain datasets that can generate new content such as text, images and sound. Large Language Models (LLMs) are a type of artificial neural network created to process and generate natural language text (Wei et al. 2022). LLMs use deep learning algorithms to learn the patterns and structures of language from massive amounts of textual data, and then use that knowledge to generate new text based on prompts or inputs (Jurafsky and Martin 2023).

AI Chatbots date back to the 1960s with MITs first chatbot ELIZA that could simulate a conversation. With the introduction of transformers' architecture in 2017 (Vaswani et al. 2017), there has been a significant development in the field of natural language processing (NLP). This new architecture of LLMs was built to efficiently process large amounts of textual data and perform a wide range of complex language tasks. Thereafter, OpenAI, which is an AI research company founded in 2015, released the first generative pretrained transformer (GPT) model, which was trained on a massive corpus of text data and could generate text in a variety of styles and genres. This was followed in 2019 by GPT-2 (Radford et al. 2019) which was an even larger and more powerful model with the ability to produce more human-like text. This model was followed by GPT-3 (Brown et al. 2020) in 2020 including 175 billion parameters, and introducing the concept of *few-shot learning*, in which the model could be trained to perform new tasks with just a few examples of labelled data.

ChatGPT, developed by OpenAI, was launched as a variant of GPT-3 specifically designed for conversational applications. Subsequent versions were updated in early 2023 and evolved to GPT-3.5 and GPT-4 (OpenAI 2023). Other AI chatbots attracting the interest of the educational research community include Google's Bard and BigScience Bloom.

In spring 2023, OpenAI published a report launching GPT-4 using a simulated bar examination to test its performance, scoring around 10% of top scorers. However, recent studies indicate that there is a need for empirical research to examine these developments in context where we can gain insights about the impact they have on practices (Mollick and Mollick 2022; Nikolic et al. 2023).

AI chatbots in education

AI Chatbots have been one of the main focus areas of research on AI in education and technology-enhanced learning (Bahja, Hammad, and Hassouna 2019), often associated with self-regulated learning (Maldonado-Mahauad et al. 2022) and intelligent tutoring systems (Mirzababaei and Pammer-Schindler 2022). Previous studies examining AI chatbots' strengths and weaknesses report that availability and usefulness are among the most critical success factors, while poor content, lack of legal frameworks and poor conversation design are the main reasons for the failures of the systems in practice (Janssen, Grützner, and Breitner 2021; Kasneci et al. 2023). Further research has examined the potential of AI chatbots to enhance education and the strategies education stakeholders need to develop to tackle the risks of possible harm and disruptions (Tlili et al. 2023).

Several studies have been conducted implementing TT-inspired designs, in order to test whether GPTs and chatbots can perform certain tasks as well as humans. Such studies have explored the capabilities of GPT-3 in writing essays (Elkins and Chun 2020) and engineering prompts (Choi et al. 2023), discussing the capabilities and flaws of GPT-3 when manipulating its parameters, the prompts, reporting that the quality of the essays could be high and often imaginative and innovative (Elkins and Chun 2020; Choi et al. 2023). A number of studies have also inquired whether GPT-3 could author an academic paper with minimum human involvement (Zhai 2022).

Several studies have been carried out aiming to explore ChatGPT's capacity in responding to examination prompts in higher education. One study examining how ChatGPT would perform in law school examinations by testing its answers to four authentic examinations reported that ChatGPT responses passed the examination with an average grade of C+ (Choi et al. 2023). Another study tested ChatGPT in several open-ended questions from an MBA course in operations management, where the chatbot response scored B to B- (Terwiesch 2023). A recent study exploring how ChatGPT performs in engineering education examinations suggests that ChatGPT generates passable responses in certain subjects and excellent responses in certain types of examination (Nikolic et al. 2023).

Hwang and Chang (2021), in their review study examining the trends in the literature regarding chatbots in education, identified that the majority of studies adopted quantitative research designs and focused on students' learning, while teachers' perceptions and education activities such as assessment are overlooked. Studies on AI chatbots in education also suggest that further conceptual and empirical research on the field of assessment and evaluation of student knowledge in higher education needs to be conducted to better understand what changes AI chatbots might introduce to current assessment practices (Nikolic et al. 2023).

As future versions of ChatGPT and other AI chatbots are emerging, capable of generating passable responses to many types of assessments, it is important to conduct studies that provide empirical evidence. Our study aims to explore the impact such AI chatbots may have on university teachers' assessment practices.

A post-phenomenological lens on technological mediations in education

Situating assessment practice into the socio-material world of education (Sørensen 2009) requires us to consider that digital technologies configure and shape contemporary assessment practices (e.g. mediated by digital examination software). The way digital technologies are designed, used and presented in the public discourse have implications for teachers' assessment practices, as they potentially transform how the teachers grade students' work and consider what assessment entails in the new technological landscape.

The socio-material nature of technological artefacts and their capacity to alter, transform and change educational practices need to be further explored (Cerratto-Pargman, Knutsson, and Karlström 2015; Cerratto Pargman and Jahnke 2019). To do that, we turn to the post-phenomenological approach (Rosenberger and Verbeek 2015) which provides us with a framework to explore and better understand the potentially mediating role of AI chatbots on university teachers' assessment practices. The post-phenomenological approach is grounded in the philosophy of technology research and mediation theory (Ihde 1993), which focuses on studying the relations that develop or regress between humans, technology and the world. It emphasises in particular how technological artefacts impact human perceptions (i.e. experience) and actions (i.e. praxis). Technological developments are within this tradition studied as the starting point for further analysis of how artefacts mediate human experiences and practices, and therefore such approaches are combined with empirical investigations (Verbeek 2016). Emphasis is put on technological mediations to discern the role of technologies in shaping human perceptions and actions, and, more specifically, moral actions and decisions, interpreting science, perceiving art and experiencing media (Verbeek 2006).

Post-phenomenological approaches in education have inspired, for example, the study of students' experiences of using their own devices in the classroom (Aagaard 2015), virtual reality mediations roles in learning environments (Voordijk and Vahdatikhaki 2022) and the materiality of learning (Sørensen 2009). We adopted in this study a post-phenomenological approach to understanding the mediating role of ChatGPT on teachers' assessment practices.

Method

We designed a study that consisted of a Turing test-inspired assessment, where university teachers blindly assessed student and bot-written responses to home examination questions. We collected four main sources of data: (i) grading sheets that were completed by the participants, where they assigned grades to the texts following each department's grading format; (ii) interview data collected during four semi-structured focus-group interviews with the participants discussing their reflections on the quality of the texts they assessed; (iii) field notes taken during the experiment and the interviews; and (iv) participants' notes on the texts they assessed.

All participants were informed about the aim of the study and asked to sign consent forms before participating. The participants were informed that their participation in the study required them to be part of a workshop on assessment where they would be asked to assign a score and assess the quality of several responses to a home examinations/assignments from their department. They were informed that their participation is voluntary and they can opt-out at any time without specifying the reasons. The study aligns with the National Ethical Review Board and university regulations for ethical research.

Context and participants

Twenty-four university teachers from four departments—philosophy (eight), education (four), sociology (six) and law (six)—were included in our study. They had different positions; PhD

students (six), assistant professors (twelve) and professors (six), with 1 to 36 years of teaching experience The age range of the participants was between 23 and 69 years; 10 females and 14 males participated.

The participants were recruited in collaboration with contact persons at each department. All the participants had earlier shown interest in AI chatbots and had varying experiences of using chatbots.

Procedure

Outline of the Turing test experiment

The Turing test (TT) was established in 1950 by Alan Turing as a method of testing whether a 'machine can think'. This design enabled us to study how teachers assessed AI chatbot and student-written texts assuming the potential use of ChatGPT and to examine teachers' experiences of assessment in the potential presence of ChatGPT.

We designed a TT experiment adapted to each of the four departments, where the participants were asked to assess five to six responses to home examination questions from an introductory course at the undergraduate level from their department. Three of the texts were student responses previously graded as excellent (A), good (C) or adequate (E). Two of the course leaders, who also acted in recruiting teachers and validating the material used in the test, participated only in the focus group interviews ($n = 24$). The participants in the TT part of the study ($n = 22$) were given a grading sheet, a description of the assignment, and six responses.

We used three ChatGPT-generated texts, and manipulated them in three levels: ChatGPT0, ChatGPT1 and ChatGPT2. ChatGPT0 refers to verbatim ChatGPT output where the description of the examination question was used as a prompt. For ChatGPT1 texts we prompted the chatbot with the English translation of the description of the examination question when it was written in Swedish (law, sociology, education) and asked it to add references to the literature. ChatGPT2 refers to a combination of multiple ChatGPT outputs. To generate these outputs, we requested longer texts referring to the minimum and maximum word limit of the examination, along with specific citations and references linked to the course literature. In addition, we crafted different questions and prompts where parts of the assignment were run separately. All three types of texts (ChatGPT0-2) were transferred to Word files and followed the format instructions described in the examinations.

We included three variants of chatbot texts in order to examine how convincing were the responses ChatGPT could generate in response to different levels of prompt engineering and output manipulation. In order to ensure the soundness of the material prepared for the TT experiment in terms of relevant and adequate responses to the home examination questions, we shared the ChatGPT-generated texts with subject experts. The director of studies in the department of education suggested that the ChatGPT0 response in their case was of very low quality and, therefore, we did not include it in the TT experiment. All ChatGPT responses were checked for plagiarism using the universities plagiarism detection software.

Plagiarism check

The results from the plagiarism check indicated that only 4 out of 14 ChatGPT texts were flagged as plagiarised (5–11%). However, the detection software identified plagiarised text located in the references lists, repetitions of the question prompts and references to factual knowledge or established theory, which would not be considered plagiarised text in a realistic assessment setting.

Focus group interviews. Following the TT experiment, we conducted semi-structured focus group interviews with the participants to explore how they engaged with the texts and how they reasoned about their assessments. During the interviews, the participants shared and discussed their reflections on the texts with two members of

our research team present. We asked the participants to discuss how they assessed the texts in terms of quality. Later, after we revealed the texts' authors—that is some of the texts were written by students and some others by AI chatbots—we asked the participants to share their reflections about whether they could discern them and what made them suspect chatbot presence in a text. The interviews were audio recorded and transcribed and then analysed by the first author in consultation with the other authors.

Drawing on Graneheim and Lundman (2004) work we divided the text data into meaning units including participants' reflections on their assessments during the TT experiment. These units then formed condensed meaning units and through abstraction created several codes which then grouped into thematic categories and finally formed our main themes. By combining the data from the grading sheets, the participants' notes on the texts they assessed along with our field notes we aimed at validating our findings.

Findings

Participants' assessment of AI chatbot responses

According to the grades assigned by the participants on ChatGPT responses (Table 1), the chatbot achieved a passing grade between 37.5% (education) and 85.7% (philosophy) across the different types of ChatGPT responses (Table 2). A minority of ChatGPT texts were awarded a failing grade (37%) while E and C were the most frequent passing grades. The most manipulated versions of the chatbot's outputs (ChatGPT1-2) achieved the highest grades.

Table 1. Participants'(P1–22) assessments of ChatGPT responses per department (*: suspicion of chatbot-written text).

	Philosophy						
	P1	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*
ChatGPT0	C	C	D	*F	D	E	*F
ChatGPT1	*C	B	C	C	D	D	E
ChatGPT2	B	D	D	*F	B	D	C
	Law						
	P8	*P9*	*P10*	*P11*	*P12*		
ChatGPT0	E	F	F	*F	F		
ChatGPT1	F	C	E	*E	F		
ChatGPT2	F	C	F	E	F		
	Sociology						
	P13	*P14*	*P15*	*P16*	*P17*	*P18*	
ChatGPT0	F	F	F	E	F	E	
ChatGPT1	*C	F	E	E	E	D	
ChatGPT2	*B	E	A	*F	B	B	
	Education						
	P19	*P20*	*P21*	*P22*			
ChatGPT1	C	*F	*F	E			
ChatGPT2	F	F	E	F			

Table 2. Participants' perceptions of the performance of ChatGPT responses.

	Likelihood to pass (%)	*Likelihood to fail (%)*	*Likelihood to be suspected (%)*	*Likelihood to not be suspected (%)*
ChatGPT0	44	56	17	83
ChatGPT1	77	23	23	77
ChatGPT2	64	36	14	86

Participants' perceptions of the quality of the texts

With regards to the strengths of ChatGPT texts, the participants highlighted that the quality of language was very high with an absence of typographical errors while following good logical structure, syntax and grammar rules. The answers were reported to be very precise, addressing all the points required from the description of the examination, and concise, providing a list of arguments and keeping the content succinct. Several of the chatbot responses included unusual or novel arguments and statements which in certain cases were positively perceived as creative and innovative:

> *"I think that text number five was more interesting!"* (teacher, law)

> *"Yes, text number five was good. I think it included more aspects than any of them."* (teacher, law)

Focusing on the weaknesses of ChatGPT texts, participants' assessment awarding low or failing grades was based on several criteria such as argumentation strategy, use of references and relevance to the content of the course. Several chatbot texts lacked clear arguments in answering the questions or provided unclear 'lines of thought':

> *"(The text) kept repeating the same empty statement over and over. And it was like the statement didn't say much. It did not give much as far as an exposition of a philosophical view, but it kept repeating itself over and over."* (teacher, philosophy)

In several cases, the participants reported that the texts did not adequately engage with the course literature. Instances including irrelevant content to the examination question, lack of examples elaborating on an argument or referring to specific content from the course, along with limited engagement of the texts with a given context or hypothetical situation were also some of the main weaknesses of chatbot responses.

Suspected indicators of AI chatbot activity

In certain cases, participants suspected the text might be written by a chatbot and not a student. The perceived suspicion was between 14 and 23% across the three types of manipulated ChatGPT texts (see Table 2 for details).

Table 1 indicates cases where the teacher suspected AI chatbot activity, highlighting that, in most of the cases, teachers failed texts when suspecting non-human texts. However, there are also cases where although the teacher suspected AI activity in the composition of the text they assigned passing grades, such as Bs and Cs. During the interviews, the participants brought up that if they suspected AI chatbot-written texts in the future, they would probably flag them for plagiarism check or ask for advice from the university support services.

There were several descriptive characteristics of the chatbot-generated texts indicating ChatGPT or 'non-real-student' activity connected to nonsensical statements, lack of personal opinion or emotional expression. The participants suspected ChatGPT text when they identified several words which either sounded out of context or strange to them:

> *"and just like the way they repeat the question… sounded to me like it's someone trying to fake… to fake being a human."* (teacher, sociology)

Instances of texts characterised as non-human contained non-sensical statements about factual knowledge, strange use of synonyms and perceived translations of English terms, and repetitions of the prompt of the examination question:

> *"The example also makes no sense… if you know Kepler's Laws of Planetary Motion, the example makes absolutely no sense. It says (that) these two laws are logically reducible from Kepler's observations. That's complete nonsense!"* (teacher, philosophy)

Several teachers who had previous experience in generating text *via* ChatGPT could also recognise the 'writing style' of ChatGPT using certain linking words, structure and vocabulary. Other

instances of 'strange' text included referring to repeated content and arguments, different terminology than the one that teachers typically use in the course material and lectures and a lack of citations and references:

> *"Both (texts)use social capital, but in a way that they weren't taught in the class and also isn't really common in the reading. It is not wrong, but it's certainly suspicious."* (teacher, sociology)

Texts described as impersonal and inconclusive, including vague arguments, 'empty statements', too simple content and not expressing a point of view, were also suspected by the participants as ChatGPT written texts. Resemblance in the content and arguments among the three types of ChatGPT text was another indicator of AI activity.

Participants' assessment of student responses

From the analysis of the grading sheets filled out by the participants, we identify that the teachers tended to be more critical towards students-written texts. Table 3 presents the grades that the participants awarded to student responses previously graded with A, C or E. It illustrates that in most of the cases across the four departments there are examples of downgrading and failing student responses, with those previously assessed as excellent (A) or very good (C) often downgraded. The teachers rarely awarded a high grade and seemed to set higher standards for a passing grade.

Participants' perceptions of the quality of the texts

Overall, the participants were more critical towards student responses with respect to structural and aesthetical flaws. Among the reasons for assigning a low grade or failing a response were incidents of repetitive statements, incoherency in content and meaning of the text, where students mixed accurate and inaccurate claims, and shortness in length. Issues related to the suggested use of references, such as lack of citations and references, and engagement with relevant literature which was not listed in the course readings, were additional flaws leading the teacher to assign a lower grade:

> *"Here there is something odd. Something is off here, right? It doesn't really capture the language right"* (teacher, sociology, on text A)

Suspected indicators of AI chatbot activity

Several of the participants suspected that some of the student texts assessed were bot-written either because they assumed that some errors and inaccuracies were not human-like or because they thought that a response was 'too good to be written by a real student'. More specifically, some of the student texts were falsely labelled as chatbot-written because they cited 'made-up' references, engaged with literature not listed in the course readings or were partly irrelevant to the context of the question or the scientific field:

> *"It's weird that they have all these YouTube videos as references. And when the main textbook for this course… they call it philosophy of language, and it's philosophy of science!"* (teacher, philosophy)

> *"Yeah. Just, you know, putting together a reference that's not real. And apparently, students are capable of that, too!"* (teacher, philosophy)

Additionally, a number of responses were not perceived as student-written because they contained inaccurate or irrelevant content and provoked a 'non-human feeling':

> *"I mean, everything is there. There's like references, there is an answer to the question. It's kind of like for me it was just (some) strange phrases, sometimes empty sentences, but it used a lot of concepts in a good way."* (teacher, sociology)

Table 3. Participants' (P1–22) assessment of student responses per department.

	Philosophy						
	P1	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*
Text A	A	C	C	F	D	B	B
Text C	E	B	B	D	D	D	E
Text E	B	D	A	C	D	C	E
	Law						
	P8	P9	P10	P11	P12		
Text A	C	A	C	A	A		
Text C	A	C	A	C	C		
Text E	A	C	C	C	A		
	Sociology						
	P13	P14	P15	P16	P17	*P18*	
Text A	C	C	C	D	C	B	
Text C	B	D	D	C	C	D	
Text E	B	E	F	D	D	E	
	Education						
	P19	P20	P21	P22			
Text A	E	E	F	C			
Text C	C	F	F	E			
Text E	E	E	E	F			

Text A: student text previously awarded an A.
Text C: student text previously awarded a C.
Text E: student text previously awarded an E.

Texts which were very well-written in terms of structure and flow of arguments, language—including no typological or grammatical errors—and content, including rich content related to the suggested readings, and accurate statements of factual knowledge, were also perceived as non-human written.

Discussion

Our aim for this study was to examine emerging technologies (in this case AI chatbots) in relation to university teachers' assessment practices. With regards to RQ1: 'How does ChatGPT 'perform' in home examinations at the undergraduate level?' our findings, support previous studies on GPT performance (Elkins and Chun 2020; Choi et al. 2023; Terwiesch 2023), and show that ChatGPT achieved a high passing grade rate of more than 66% in home examination questions in the fields of humanities, social sciences and law. Competence in language, precision, consistency and creativity were among the strengths of the ChatGPT responses that the participants highlighted. The weak points identified were inadequate argumentation, lack of references from the course literature and unrelated to the course content responses. However, such flaws could also be found in student texts. Several aspects of the texts led the participants to 'flag' and suspect that the text was not written by a human, such as phrases and words which were nonsensical and vague arguments which were repeated across the chatbot texts.

With regards to RQ2: 'What is the mediating role of ChatGPT in teachers' assessment practices?' we identify that, overall, participants seemed to adopt a more critical stance when assessing student texts. In several cases, the participants downgraded previous student responses as they did not accept mistakes such as repetitions, short answers and lack of engagement with course literature. Our findings show that participants seemed to raise their standards for awarding high and passing grades as in rare circumstances they awarded an A and in total failed more than 9% of student responses which had previously received passing grades. Certain characteristics of student texts were perceived as chatbot-written by the participants; for instance, responses

which were either very well-written or referred to irrelevant content and literature, assuming that such responses would not be written by students.

Based on these empirical observations, we argue that there is a need to clarify that, although we did not ask the participants to identify the chatbot-written responses—we asked them to give scores and evaluate the quality of the responses during the TT experiment—most of the participants stated in the focus group interviews that they were constantly conscious of the potential presence of a ChatGPT-generated text which affected their assessment. We noticed that participants perceived that the evaluation of the responses required them to distinguish between student and chatbot texts. This led them to suspect and 'flag' several responses as AI-generated texts with flaws that were comparable or identical to those in both chatbot and student responses. We acknowledge the tougher grades could be a result of a Hawthorne effect in the experimental setting (McCarney et al. 2007), and we do not aim to generalise our findings. Instead, our study attempts to examine the teachers' re-actions and perceptions, and the implications for their assessment practices in the potential presence of emerging technological artefacts, such as ChatGPT.

In line with previous literature on assessment in higher education, assessment designs might vary a lot across disciplines (Fernández-Ruiz, Panadero, and García-Pérez 2021) as well as assessment styles and evaluation strategies among university teachers (Fernández Ruiz et al. 2022). We found significant differences in assessment among the departments. We identified that the law department was less prone to downgrade and fail ChatGPT and student responses compared to the other departments. One possible explanation for that could be related to the different assessment cultures of the disciplines and the content evaluated, as in the case of legal education the examination questions were designed to test factual knowledge and reasoning, while in the case of education, the examination required students to reflect on given readings and synthesise an opinion. In the case of sociology, we observed that high grades were assigned to the most manipulated versions of ChatGPT responses (ChatGPT1 and ChatGPT2), which could possibly relate to the nature of the examination questions requiring the students to elaborate their answers as closely as possible to specific theories and concepts. However, since our study did not aim at comparing the different departments but rather examined teachers' assessment practices, we suggest that future studies considering the different cultures and scientific traits in science, humanities and social sciences could further explore the impact of AI-chatbots in assessment.

ChatGPT mediating role in assessment practices

Drawing on post-phenomenological approaches (Ihde 1993; Aagaard 2015; Rosenberger and Verbeek 2015; Verbeek 2016) we argue that the presence of AI chatbots like ChatGPT in society may have impacted teachers' perception of student texts. The presence of AI chatbots may prompt teachers to ask "who has written the text?" and thereby question students' authorship, potentially reinforcing mistrust at the core of teacher–student relationship. The presence of AI chatbots may also disrupt teachers' assessment practices and question their inherent trust in the students writing their home examinations by themselves.

Our study shows that ChatGPT may have an impact on university teachers' assessment practices in amplifying their criticality and suspicion and challenging their trust in students' texts. The participants' suspicion of ChatGPT-generated text has an impact on their perceptions of the quality of the texts assessed as we observe participants' criticality is amplified. Such changes in the perception of the quality of the text are reflected in downgrading. We identified that participants were prone to assess differently than they usually did because they were aware of the potential use of ChatGPT in the texts presented to them.

We observed that participants assumed that certain student-written texts were bot-written, perceiving flaws in student-written texts as non-human. As such, ChatGPT, mediated, altered and re-shaped teachers' perceptions about what technology (AI chatbot) can achieve in this context and how students perform. This observation is in line with Verbeek's argument that even without

interacting with a technological artefact, its presence in society could influence behaviour (Verbeek 2011). Issues related to academic integrity and authenticity in authorship are central in assessment research in higher education (Huber et al. 2023), and therefore we suggest that future studies should explore such issues in relation to GAI and AI chatbots in the context of higher education.

Limitations

We acknowledge that the material crafted to be used in the TT experiment is to a great extent historical by using the ChatGPT March 2023 version, which was powered by GPT-3.5. However, although Open AI's launch of GPT-4 in the spring of 2023 claimed a significant increase in its performance compared to GPT-3.5 (OpenAI 2023), a recent study on educational research examining ChatGPT reported that in certain cases GPT-3 generated more accurate responses and, in other cases, the increase in GPT-4 performance was limited to 2% in physics examinations (Nikolic et al. 2023). Although the current paid version of ChatGPT is powered by GPT-4, ChatGPT's version used in our study is still available as a free option and likely to be used by many students.

This study allowed us to examine the potentially mediating role of chatbots in university teachers' assessment practices as it combines different forms of data and an experimental setting, with a close-to-realistic environment situation. Although the participants did not know whether all, several or none of the responses were student-written or partly or completely ChatGPT-written, we recognise that the experimental setting along with the aim of the study might have influenced them in terms of assuming ChatGPT presence in the texts. We acknowledge that issues of differences in assessment styles and disciplines' assessment cultures along with the experimental setting might have affected the results of the grading sheets, and therefore we further explored teachers' assessment reasonings using the focus group interviews. Follow-up and further empirical studies examining AI-chatbots' uptake in education and assessment practices are also needed, including multiple stakeholders in higher education institutions. As GAI technologies are fast-evolving, future studies including multiple variants of chatbots and uses in higher education would benefit the technology-enhanced teaching and learning research field.

Acknowledgements

The authors would like to extend their gratitude to the Swedish Higher Education Research Network (SHERN), who provided invaluable feedback on a previous version of the article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Marianne and Marcus Wallenberg Foundation—WASP-HS—The Wallenberg AI, Autonomous Systems and Software Program—Humanities and Society (Grant number: 2020.0138).

ORCID

Alexandra Farazouli http://orcid.org/0000-0001-7601-3850
Teresa Cerratto-Pargman http://orcid.org/0000-0001-6389-0467
Klara Bolander-Laksov http://orcid.org/0000-0002-3345-3810
Cormac McGrath http://orcid.org/0000-0002-8215-3646

References

Aagaard, J. 2015. "Drawn to Distraction: A Qualitative Study of Off-Task Use of Educational Technology." *Computers & Education* 87 (September): 90–97. doi:10.1016/j.compedu.2015.03.010.

Bahja, M., R. Hammad, and M. Hassouna. 2019. "Talk2Learn: A Framework for Chatbot Learning." In *Transforming Learning with Meaningful Technologies*, edited by Maren Scheffel, Julien Broisin, Viktoria Pammer-Schindler, Andri Ioannou, and Jan Schneider, 582–586. Cham: Springer International Publishing. doi:10.1007/978-3-030-29736-7_44.

Bearman, M., J. H. Nieminen, and R. Ajjawi. 2023. "Designing Assessment in a Digital World: An Organising Framework." *Assessment & Evaluation in Higher Education* 48 (3): 291–304. doi:10.1080/02602938.2022.2069674.

Bowman, E. 2023. "A College Student Created an App That Can Tell Whether AI Wrote an Essay." *NPR*, January 9, 2023, sec. Technology. https://www.npr.org/2023/01/09/1147549845/gptzero-ai-chatgpt-edward-tian-plagiarism.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in neural information processing systems* 33: 1877–1901. doi:10.48550/arXiv.2005.14165.

Cerratto Pargman, T., and I. Jahnke. 2019. "Introduction to Emergent Practices and Material Conditions in Learning and Teaching with Technologies." In *Emergent Practices and Material Conditions in Learning and Teaching with Technologies*, edited by Teresa Cerratto Pargman and Isa Jahnke, 3–20. Cham: Springer International Publishing. doi:10.1007/978-3-030-10764-2_1.

Cerratto-Pargman, T., O. Knutsson, and P. Karlström. 2015. "Materiality of Online Students' Peer-Review Activities in Higher Education." https://www.semanticscholar.org/paper/Materiality-of-Online-Students%27-Peer-Review-in-Pargman-Knutsson/443f3043a94be285acc84f021d18df1a05b5070d.

Choi, J. H., K. E. Hickman, A. B. Monahan, and Daniel B. Schwarcz. 2023. "ChatGPT Goes to Law School." doi:10.2139/ssrn.4335905.

Dawson, P., K. Nicola-Richmond, and H. Partridge. 2023. "Beyond Open Book versus Closed Book: A Taxonomy of Restrictions in Online Examinations." *Assessment & Evaluation in Higher Education*. doi:10.1080/02602938.2023.2209298.

Elkins, K., and J. Chun. 2020. "Can GPT-3 Pass a Writer's Turing Test?" *Journal of Cultural Analytics* 5 (2):1–16. doi:10.22148/001c.17212.

Fernández Ruiz, J., E. Panadero, D. García Pérez, and L. Pinedo. 2022. "Assessment Design Decisions in Practice: Profile Identification in Approaches to Assessment Design." *Assessment & Evaluation in Higher Education* 47 (4): 606–621. doi:10.1080/02602938.2021.1937512.

Fernández-Ruiz, J., E. Panadero, and D. García-Pérez. 2021. "Assessment from a Disciplinary Approach: Design and Implementation in Three Undergraduate Programmes." *Assessment in Education: Principles, Policy & Practice* 28 (5–6): 703–723. doi:10.1080/0969594X.2021.1999210.

Graneheim, U. H., and B. Lundman. 2004. "Qualitative Content Analysis in Nursing Research: Concepts, Procedures and Measures to Achieve Trustworthiness." *Nurse Education Today* 24 (2): 105–112. doi:10.1016/j.nedt.2003.10.001.

Huber, E., L. Harris, S. Wright, A. White, C. Raduescu, S. Zeivots, A. Cram, and A. Brodzeli. 2023. "Towards a Framework for Designing and Evaluating Online Assessments in Business Education." *Assessment & Evaluation in Higher Education* 0 (0): 1–15. doi:10.1080/02602938.2023.2183487.

Hwang, G. J., and C. Y. Chang. 2021. "A Review of Opportunities and Challenges of Chatbots in Education." *Interactive Learning Environments* 0 (0): 1–14. doi:10.1080/10494820.2021.1952615.

Ihde, D. 1993. *Postphenomenology: Essays in the Postmodern Context*. Evanston, Ill: Northwestern University Press.

Jane, E. 2023. "CheatGPT?" *Openforum* (blog). February 20, 2023. https://www.openforum.com.au/cheatgpt-2/.

Janssen, A., L. Grützner, and M. Breitner. 2021. "Why Do Chatbots Fail? A Critical Success Factors Analysis." https://www.semanticscholar.org/paper/Why-do-Chatbots-fail-A-Critical-Success-Factors-Janssen-Gr%C3%BCtzner/5cb8bcd29aca1849bdb0e22972d6b1cb6f70979f.

Jurafsky, D., and J. H. Martin. 2023. "Speech and Language Processing." Accessed March 25, 2023. https://web.stanford.edu/~jurafsky/slp3/.

Kasneci, E., K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." Learning and Individual Differences, 103: 102274. doi:10.1016/j.lindif.2023.102274.

Kirshner, S. 2023. "Education in the Age of ChatGPT." *Openforum* (blog). March 24, 2023. https://www.openforum.com.au/education-in-the-age-of-chatgpt/.

Maldonado-Mahauad, J., M. Pérez-Sanagustín, J. Carvallo-Vega, E. Narvaez, and M. Calle. 2022. "Miranda: A Chatbot for Supporting Self-Regulated Learning." In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, edited by Isabel Hilliger, Pedro J. Muñoz-Merino, Tinne De Laet, Alejandro Ortega-Arranz, and Tracie Farrell, 455–462. Cham: Springer International Publishing. doi:10.1007/978-3-031-16290-9_36.

McCarney, R., J. Warner, S. Iliffe, R. van Haselen, M. Griffin, and P. Fisher. 2007. "The Hawthorne Effect: A Randomised, Controlled Trial." *BMC Medical Research Methodology* 7 (1): 30. doi:10.1186/1471-2288-7-30.

Mimirinis, M. 2019. "Qualitative Differences in Academics' Conceptions of e-Assessment." *Assessment & Evaluation in Higher Education* 44 (2): 233–248. doi:10.1080/02602938.2018.1493087.

Mirzababaei, B., and V. Pammer-Schindler. 2022. "An Educational Conversational Agent for GDPR." In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, edited by Isabel Hilliger, Pedro J. Muñoz-Merino, Tinne De Laet, Alejandro Ortega-Arranz, and Tracie Farrell, 470–476. Cham: Springer International Publishing. doi:10.1007/978-3-031-16290-9_38.

Mollick, E. R., and L. Mollick. 2022. "New Modes of Learning Enabled by AI Chatbots: Three Methods and Assignments." SSRN Scholarly Paper. Rochester, NY. doi:10.2139/ssrn.4300783.

Nikolic, S., S. Daniel, R. Haque, M. Belkina, G. M. Hassan, S. Grundy, S. Lyden, P. Neal, and C. Sandison. 2023. "ChatGPT versus Engineering Education Assessment: A Multidisciplinary and Multi-Institutional Benchmarking and Analysis of This Generative Artificial Intelligence Tool to Investigate Assessment Integrity." *European Journal of Engineering Education* 48 (4): 559–614. doi:10.1080/03043797.2023.2213169.

OpenAI. 2023. "GPT-4 Technical Report." doi:10.48550/arXiv.2303.08774.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Rosenberger, R., and P. P. Verbeek. 2015. *Postphenomenological Investigations: Essays on Human–Technology Relations*. Lanham, MD: Lexington Books.

Shen-Berro, J. 2023. "New York City Schools Blocked ChatGPT. Here's What Other Large Districts Are Doing." Chalkbeat. January 6, 2023. https://www.chalkbeat.org/2023/1/6/23543039/chatgpt-school-districts-ban-block-artificial-intelligence-open-ai.

Sørensen, E. 2009. *The Materiality of Learning: Technology and Knowledge in Educational Practice. Learning in Doing: Social, Cognitive and Computational Perspectives*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511576362.

Terwiesch, C. 2023. *Would Chat GPT3 Get a Wharton MBA?* Pennsylvania: Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania.

The McGraw Center for Teaching & Learning, Princeton University. 2023. *Guidance on AI/ChatGPT*. Princeton: McGraw Center for Teaching and Learning. https://mcgraw.princeton.edu/guidance-aichatgpt.

Tlili, A., B. Shehata, M. Agyemang Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang. 2023. "What If the Devil is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots in Education." *Smart Learning Environments* 10 (1): 15. doi:10.1186/s40561-023-00237-x.

Vaswani, A., S. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Verbeek, P. P. 2006. "Materializing Morality: Design Ethics and Technological Mediation." *Science, Technology, & Human Values* 31 (3): 361–380. doi:10.1177/0162243905285847.

Verbeek, P. P. 2011. "*Moralizing Technology: Understanding and Designing the Morality of Things*."

Verbeek, P. P. 2016. "Toward a Theory of Technological Mediation: A Program for Postphenomenological Research." 189–204.

Voordijk, H., and F. Vahdatikhaki. 2022. "Virtual Reality Learning Environments and Technological Mediation in Construction Practice." *European Journal of Engineering Education* 47 (2): 259–273. doi:10.1080/03043797.2020.1795085.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, et al. 2022. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research, August*. https://openreview.net/forum?id=yzkSU5zdwD.

Zhai, X. 2022. "ChatGPT User Experience: Implications for Education." SSRN Scholarly Paper. Rochester, NY. doi:10.2139/ssrn.4312418.