Received: 21 January 2024 | Accepted: 14 March 2025

DOI: 10.1111/bjet.13585

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

The impact of generative AI on academic integrity of authentic assessments within a higher education context

Alexander K. Kofinas[1] | Crystal Han-Huei Tsay[2] | David Pike[1]

[1]Graduate School of Business, University of Bedfordshire, Luton, UK

[2]Executive Business Centre, University of Greenwich, London, UK

Correspondence
Alexander K. Kofinas, Graduate School of Business, University of Bedfordshire, University Square, Luton LU1 3JU, UK.
Email: alexander.kofinas@beds.ac.uk

Generative AI (hereinafter GenAI) technology, such as ChatGPT, is already influencing the higher education sector. In this work, we focused on the impact of GenAI on the academic integrity of assessments within higher education institutions, as GenAI can be used to circumvent assessment approaches within the sector, compromising their quality. The purpose of our research was threefold: first, to determine the extent to which the use of GenAI can be detected via the marking and moderation process; second, to understand whether the presence of GenAI affects the marking process; and finally, to establish whether authentic assessments can safeguard academic integrity. We used a series of experiments in the context of two UK-based universities to examine these issues. Our findings indicate that markers, in general, are not able to distinguish assessments that have had GenAI input from assessments that did not, even though the presence of GenAI affects the way markers approach the marking process. Our findings also suggest that the level of authenticity in an assessment has no impact on the ability to safeguard against or detect GenAI usage in assessment creation. In conclusion, we suggest that current approaches to assessments in higher education are susceptible to GenAI manipulation and that the higher education sector cannot rely on authentic assessments alone to control the impact of GenAI on academic integrity. Thus, we recommend giving more critical attention to assessment design and placing more emphasis on assessments that rely on social experiential learning and are performative rather than output-based and asynchronously written.

KEYWORDS

academic integrity, authentic assessments, experiment, generative AI

Practitioner notes

What is already known about this topic

- GenAI has enabled students to complete higher education assessments quickly and with good quality, leading to challenges in academic integrity.
- GenAI has transformed the requirements and considerations in assessment design in higher education.
- Authentic assessments are seen as a prominent way to tackle the GenAI challenge.

What this paper adds

- We provide quantitative and qualitative experimental evidence suggesting that GenAI can generate authentic assessments that pass the scrutiny of experienced academics.
- We demonstrate how the use of authentic assessments alone does not protect the academic integrity of students in higher education.
- Our qualitative analysis indicates that markers may generate false positive and false negative results if they suspect GenAI tampering in an assessment. Thus, students' learning is not assessed correctly.

Implications for practice and/or policy

- When universities and national organisations design policies regarding GenAI, authentic assessments are not the panacea; the focus must remain on assessment design.
- Assessments of learning need to shift from assessing output to focusing on process and relevance to the workplace. That would mean a paradigmatic shift from written assessments to synchronous interpersonal assessments.
- The move away from written assessments has implications that are far reaching for the academy if written assessments cannot be trusted as a reliable indicator for and of learning.

INTRODUCTION

ChatGPT rose to prominence in media headlines in late 2022 and within days acquired a million users (Marr, 2023). OpenAI's ChatGPT is an example of a generative artificial intelligence (GenAI) system. GenAI systems offer free and paid access, and anyone with internet access and the inclination can utilise such services to produce content. GenAI systems take instructions, also known as prompts, from users and respond by generating text, images, or other artefacts. Users can interact and engage with such systems to clarify questions and develop the responses the system produces. Recently, ChatGPT entered public discourse for three reasons: first, it is highly integrated into systems users are familiar with, such as web browsers, search engines, and word editors (Pastis, 2023); second, ChatGPT's responses have given the impression to users that there is genuine intelligence in the system (Deng & Lin, 2022); and third, GenAI systems can appear to be experts on any subject because of the comprehensive way they respond to questions and prompts (Herbold et al., 2023).

The use of GenAI involves a range of ethical issues (Cotton et al., 2023). As GenAI systems continue to develop at a very fast pace, they often outpace efforts to define clear ethical boundaries for appropriate use, providing an opportunity for users to adopt a consequentialist approach (Sinnott-Armstrong, 2003) in the usage of GenAI. The consequentialist approach is an ethical framework that evaluates the morality of actions based on their outcomes or consequences. In the case of higher education assessments, students weigh the potential positive outcomes (eg, efficiency and ease) against the possible negative consequences (eg, getting caught for academic dishonesty). If they determine that the benefits outweigh the risks, they may choose to use GenAI in ways that compromise academic integrity. Circumventing the assessment process in that way can undermine expected standards and norms in academic writing and assessment (Cronan et al., 2018; Sutherland-Smith, 2008).

Assessment is a vital part of learning and teaching in higher education, and as educators, we are mindful of ensuring that students are engaging honestly with formative and summative submissions in a way that facilitates learning. Universities have moved from relying on human markers to determine authorship of assessments to using antiplagiarism software such as Turnitin (Rolfe, 2011). However, with the rise of essay mills, ghostwriting services and contract cheating, it has become more challenging to determine the authorship of assessments (Bartlett, 2009). Essay mills and ghostwriting services allow students to outsource their work to third parties, who produce original content on their behalf, making it nearly impossible to detect their inputs through traditional plagiarism detection tools (Sweeney, 2023). This type of contract cheating undermines the academic process, as it circumvents the need for students to engage with the material and learn. The recent emergence of GenAI further complicates this issue, as GenAI can produce high-quality, seemingly original work that is difficult to distinguish from student-authored content (Spennemann et al., 2024). If educators cannot reliably determine the authorship of an assessment, the value of higher education degrees can be undermined, posing a serious threat to the higher education sector.

There is an ongoing debate about whether GenAI is fully capable of producing an assessment (Kocoń et al., 2023); however, there is evidence to suggest that written types of assessments commonly used in institutions, such as reports and essays and even take-home exams, can be vulnerable to the use of GenAI (Scarfe et al., 2024). These vulnerabilities are the result of two qualities that GenAI possesses: the ability to produce unique text rapidly (Kolade et al., 2024; Rudolph et al., 2023) and the ability to produce outputs that appear to be of high quality (Herbold et al., 2023). The question for universities and educators is how to balance the potential benefits of GenAI, such as its ability to help students understand their assessment, to outline what might be written, and to brainstorm ideas (Sok & Heng, 2023), against inappropriate uses such as cheating on an exam (Giannos & Delardas, 2023) or on other assessments (Ventayen, 2023) or asking GenAI to devise ways of circumventing detection (Spennemann et al., 2024). One potential answer offered in the literature is to return to invigilated exams, a suggestion that is not favoured by many educators, as exams can be seen as a rather inauthentic assessment of learning, even though that format would enhance the academic integrity of the work produced (Bower et al., 2024; Forsyth, 2022). Another suggestion has been to develop assessments that are more authentic, that are more closely aligned to the world of work and to the kind of problems students face after graduation (Ellis et al., 2020; Sotiriadou et al., 2020).

Within this context, this work set out to explore how GenAI influences the outcomes of assessments. We focused on the impact of GenAI on the assessment marking process in two different UK higher education institutions (HEIs) within the field of business studies. Our study was exploratory, and we used a range of assessments widely used by academics across the world to determine the answers to three research questions:

> **RQ1**: Can assessment markers differentiate between human-authored assessments, GenAI-modified assessments, and GenAI-generated assessments?
> **RQ2**: Does the presence of GenAI have the potential to influence the marking and moderation process?
> **RQ3**: Would different levels of authenticity in assessments have an impact on the ease of detection of GenAI usage?

Participants in the study marked a range of assessments, with three versions: (a) assessments genuinely produced by students, which had already been marked with grades agreed upon, even though the particular submissions and their corresponding grades were initially unknown to the participants; (b) versions of the same assessments modified by GenAI with the tool ChatGPT 3.5; (c) assessments solely produced using GenAI that followed the requirements of the assessment brief. Our results indicate that GenAI can support students in circumventing the necessary effort and learning required to produce an assessment. Such efforts would not be reliably detected by human markers, but the knowledge that GenAI may have been used can affect markers' behaviour and judgement. Finally, we determined that the level of authenticity in the assessment has had a very limited impact in terms of GenAI usage detection.

The paper comprises five sections. The first is a literature review on GenAI and the importance of assessment in higher education, followed by a section that details the mixed methods research design. The third section comprises findings that aim to provide preliminary answers to our questions, followed by a discussion regarding our findings. In the concluding section, we provide insights on the impact of GenAI on the marking process and recommendations for action in terms of assessment design.

LITERATURE REVIEW

Generative artificial intelligence and its impact on higher education

For the general public, and many academics who were not involved in AI-related activities, awareness of the potential benefits and risks of GenAI began with the release of OpenAI's GPT 3.5 product in November 2022 (OpenAI, 2024). The system had a fixed dataset that was set back to September 2021, and the initial abilities of GPT 3.5 were enough to provoke much academic debate (Cao et al., 2023; Wu, 2024) and to promote the development of this study. GenAI's development quickly upended efforts by academics to ensure academic integrity and the honest production of students' assessments (Rahman & Watanobe, 2023). The problem of interference by a third party is not new; however, in the past the operators of services were human and there were cost implications for students (Bartlett, 2009; Medway et al., 2018). However, GenAI assessments can now be generated for free, or very cheaply. Unlike lecturers and higher education support systems, GenAI is available 24 hours a day and can provide responses to questions and quickly produce a fully fleshed-out and plausible-looking assessment (Aydın & Karaarslan, 2023; Dowling & Lucey, 2023). For some students, GenAI appears to possess an air of intellectual authority (Firth et al., 2024; Sok & Heng, 2023; Sullivan et al., 2023; Sweeney, 2023), and for students who are experiencing a change in the way they are learning (eg, studying in the UK), there may be a temptation to use GenAI-generated text, as they may perceive it as superior to the prose they could write themselves. However, proving that a student purchased an assessment or used GenAI is difficult, and academics in our respective institutions have resorted to reviewing their in-person interactions and perceptions of students' academic abilities or conducting vivas to identify misconduct.

If a student elects to utilise GenAI to help write or fully develop an assessment, what are the potential risks aside from violating academic integrity rules? GenAI writing presents five main challenges. First, the GenAI systems' underlying training that produces responses to users may exhibit biases, including political (Rozado, 2023), racial and cultural (Ray, 2023), and gender (Gross, 2023) biases. Second, many systems are trained to absorb data from online sources, and these sources may contain data that are not representative or the framing of that data can lead to lawsuits (Appel et al., 2023). Under-representative data refer to sources are often disproportionately populated by English-language and Western-centric content. Even when non-Western languages are included, the data tends to be limited in volume, context and domain diversity. This underrepresentation creates a gap in the model's ability to understand, generate, or process text that reflects the cultural and linguistic nuances of these communities (Guo et al., 2024). Third, GenAI can ‘hallucinate’, or make ‘mistakes’ with data and facts (Kocoń et al., 2023; Rudolph et al., 2023). Fourth, and connected to the third point, like all computer systems, GenAI responds to inputs from humans, and the language used by users is not always precise and clear. Users must curate their prompts carefully to get the information they require from a GenAI system (Giray, 2023; Lo, 2023). Finally, not all students utilise appropriate levels of autonomy (Carnell, 2016) and criticality when examining information sources and data. This refers to the earlier point around intellectual authority and a deferral of responsibility to someone or something else to complete a task. To compound the problem, the tools academics typically rely upon to check for use of GenAI in assessment writing can be biased and are considered inaccurate (Sullivan et al., 2023).

In the current UK socioeconomic climate, students face very high pressures: family life and caring responsibilities, attending university to study, often work responsibilities and higher costs of living. Under pressure, students may not be aware of or may choose to ignore institutional policies and guidance regarding academic practice (and malpractice). It used to be cost-prohibitive for students to seek the services of a third-party author; now free GenAI services or paid services with minimal cost are available. Thus, there is a strong temptation for students to adopt a consequentialist approach (Sinnott-Armstrong, 2003) to complete their assessments, leading to students prioritising task completion over falling foul of the institution's rules on academic practice. Thus, GenAI's expedient and proficient creation of an assessment (Ventayen, 2023) presents an existential threat to universities and their assessment processes.

Importance of assessment in higher education

Assessments are an important component of higher education and represent a prime measurement of students' competence and learning (Cilliers et al., 2012). Assessments fall into two categories: formative and summative. Formative assessments encompass any assessment for learning that aids students' understanding of assessment requirements (Brown, 2005; Newton, 2007). Summative assessments measure learning (Broadbent et al., 2018; Kofinas, 2018) and are directly linked to certification and progress (Newton, 2007). Student engagement with a combination of formative and summative assessments can help educators evaluate the level of learning that occurs within the classroom.

In the higher education setting, students participate in formative activities within a module of study to engage with content to learn and/or achieve a higher grade in a summative assessment (Kolb & Kolb, 2005; Lu et al., 2003). Students are often strategic and focus on summative assessments, a phenomenon that Biggs (2003) labelled backwash. If summative assessments are disproportionately influential in driving students' learning (Brown, 2005; Holmes, 2015; Iannone & Simpson, 2016; Raupach et al., 2013), intelligent assessment design would facilitate learning via summative assessments and would use formative assessments in a manner that aligns closely with the summative (Broadbent et al., 2018; López-Pastor et al., 2013; Trotter, 2006). Biggs (2003) proposed in his constructive alignment model that the backwash phenomenon should be considered in assessment design and that educators should develop their modules starting from the summative assessments and working their way backwards towards the learning objectives of the module.

As a consequence of this constructive alignment, an educator should link all aspects of learning in a module to the summative assessment while ensuring that this connection is communicated clearly. A well-designed assessment approach interweaves formative and summative assessments carefully to ensure that the learning happens in a thoughtful, well-scaffolded manner to take advantage of the backwash and to enhance the deep learning of the students (Biggs, 2003; Pereira et al., 2015). Consequently, the quality of assessment determines whether the students learn what they are meant to learn, and the academic integrity of the assessments becomes a crucial academic quality issue (Glendinning, 2022; Harlen, 2004). HEIs have implemented various measures to promote academic integrity in their summative assessments, including codes of conduct, academic integrity policies and other educational programmes (Cotton et al., 2023; Harlen, 2004; Morris, 2018).

Given the importance of assessment in higher education, GenAI poses two challenges for assessment design: first, it provides the possibility for students to develop assessments directly rather than going through an iterative developmental process (Cotton et al., 2023; Rasul et al., 2023). Second, GenAI can quickly produce outputs that appear credible and original, even though there may be factual mistakes, a phenomenon known as hallucination (Beutel et al., 2023). GenAI has made it very difficult to ascertain when an assessment that is submitted is an original assessment that demonstrates the student's level of learning. Our first two research questions explore this issue of GenAI detection that academics face and the impact it may have in the grading process.

The response to the GenAI challenge from organisations such as the Quality Assurance Agency for Higher Education (2024) and Advance HE (2024) has been that universities need to move towards more innovative, authentic assessments. Authentic assessments link learning to the actual world of work and consequently provide learning experiences in activities, tasks and ideas that are directly applicable to the workplace (Lund, 1997; Mueller, 2005). Hobbins et al. (2022) suggested that authentic assessments need to meet four dimensions: realism, cognitive challenge, evaluative judgement criteria, and evaluative judgement feedback. Kaider et al. (2017) suggested that an assessment is authentic when it provides proximity to the world of business as well as an authentic task as the foundation for the assessment. Authentic assessments are considered harder to circumvent and have been considered a potential way to respond to the GenAI challenge (Ellis et al., 2020; Sotiriadou et al., 2020).

Alternatively, if higher education providers cannot control access to GenAI, they may be able to direct this educational technology towards productive usage by encouraging students to use GenAI in their assessment preparation. That would be in accord with the emerging view among scholars that students should know how to use GenAI, as it is becoming a critical employability skill (Pagani et al., 2023). Authentic assessments alongside other types of assessments, such as other interactive oral assessments (Krautloher, 2024; Ward et al., 2023) and role-play–type assessments are seen as potential types of assessment that would support the usage of GenAI as a learning tool while safeguarding the academic integrity of the assessment.

In this work, we focused on a range of authentic assessments to test whether the authenticity of assessment would assist the educators in identifying GenAI-generated text and thus facilitate the graders' evaluation of the quality of the assessments produced.

METHODOLOGY

This exploratory research work involved two medium-sized post-92 business schools within the UK. The research team worked at these two business schools; thus, convenience dictated the choice of these settings. The assessments we have chosen to run our experiments are widely used in the higher education sector. The processes for marking and moderation align with sector norms, with first marking and then moderation being the standard process for all undergraduate marking.

To answer the research questions, we conducted a two-phased within-subjects experimental design to compare AI writing to human writing. We chose the within-subjects experimental design to achieve reduced variability in marking standards and efficiency while being mindful of order effects (Charness et al., 2012).

Participants

Two undergraduate degree programmes, the Bachelor of Science in Business Management and the Bachelor of Arts in Business Purchasing and Supply Chain Management, were identified through the research team's professional networks. Upon receiving module leaders' agreement, one module at each undergraduate academic level (ie, levels 4 to 6) within each programme was selected to generate writing samples. Through module teams' recommendations, two pairs of academics from each programme who were familiar with the modules (eg, through teaching, marking, module/programme leadership) but were not involved in marking the selected writing samples were invited to mark the samples based on the established assessment rubrics. Thus, in total, four pairs of academics, totalling eight, participated in the experiment. They read the participant information sheets and gave consent. Participants' full-time work experience in higher education ranged from 3 to 15 years. Participants were advised that they should not discuss the individual details of assessments outside of the research interviews for the duration of the experiment.

Each pair of academics was asked to evaluate three types of assessments with no prior grade knowledge: (a) existing unaltered student assessments; (b) assessments entirely authored using GenAI; and (c) GenAI-modified assessments (assessments where we modified sections of the assessment with content generated by GENAI software). ChatGPT 3.5 was used for this experiment, and a protocol was followed in developing each assessment. In the two-phased experiment, the four pairs of participating markers were first asked to mark randomly coded writing samples across three undergraduate academic levels independently. Then each pair of markers met with a member of the research team to reach an agreement on the marks, followed by researcher debrief and interview. The research team then used the agreed marks to provide descriptive statistics in answering the research questions, supplemented by interview data. The descriptive statistics from the experimental design, together with the textual information obtained from the interviews, formed our mixed-methods research design. Such a design allows for a more comprehensive understanding of a phenomenon and validation and corroboration of findings across different types of data

(Cohen et al., 2011; Turner & Turner, 2009). For example, qualitative insights gained from our interview data explained anomalies that emerged in the quantitative data. Given that we explored new phenomena, this design was particularly beneficial in helping us explore in-depth answers to our research questions.

Ethical considerations

Institutional ethics permission was sought individually from each of the two universities and was obtained prior to data collection. When participants were recruited, they were informed that they would be evaluating writing samples, some of which had been modified using GenAI. However, we did not specify which samples had undergone GenAI treatment, nor did we disclose the exact nature of the GenAI modifications for each assessment.

We had to address several ethical considerations. The recruitment of participants followed a convenient sampling approach where we approached module leaders to provide us access to appropriate members of the teams. The study required GenAI-driven modification of existing student assessments, which were anonymised and modified as required. With regard to the marking process, we followed standard marking procedures without any modification or deviation, and we had an early brief with all staff involved in our experiments and a debrief after the marking but before the interviews where we revealed all aspects of the experiment as a stimulus for the discussion.

Writing sample selection

The writing samples at each module contained work of three types. The first type consisted of three pieces of work written and marked at three grade bands (below pass mark ≤40%; 50%–60%; and first-class, ie, around 70% or more) by past students; these stood for the human-authored assessments, as they had been marked and had a healthy Turnitin similarity score. In the second group, GenAI (specifically ChatGPT 3.5) was used to modify the same three assessments, generating GenAI-modified versions of these assessments. In the third group, based on assessment briefs in each of the chosen modules, the research team created one assessment that was solely GenAI-authored. To achieve this, the research team took an assessment brief and used prompts to generate a piece of work within 1½ hours, simulating a 'mischievous' student. Appendix A shows an example of prompts used in a Level 6 assessment.

Therefore, for each programme, seven writing samples (three human-authored, three GenAI-modified, and one GenAI-authored) were prepared for each undergraduate academic level, resulting in a total of 21 writing samples, which were then randomly coded. We used a simple coding scheme, with HE1 and HE2 for each university and Levels 4–6 to indicate the year. We then used letters to indicate the version of the assignment. For example, HE1-L4-J denoted the assessment we chose for L4 from the first HEI and version J of that assessment. All assessment variations were coded and presented in that manner.

The research team carefully selected assessments that represented a variety of approaches but had an explicit output, such as essays, reports and case study analyses, where GenAI could have an impact. Thus, we avoided assessments that included class presentations or exams, deeming that GenAI would probably have limited impact. Furthermore, to ensure that participating markers evaluated the writing quality of each sample reliably, each assessment was anonymised, so that it had no information that could identify the original author or a third party. We ensured that the assessment was not involved in any prior academic misconduct cases and did not score high on Turnitin similarity or AI indices. In addition, the writing samples were reformatted and were similar in terms of topic and length.

We used two different ways to map out the authenticity of each assessment. First, Hobbins et al. (2022) suggested that authentic assessments need to meet four dimensions: realism, cognitive challenge, evaluative judgement criteria and evaluative judgement feedback. Table 1 maps our assessments based on these criteria.

Our sample had three assessments that were moderate/low in terms of authenticity, one that was medium and two that were medium/high.

To validate this evaluation, we further mapped the six assessments against a second framework suggested by Kaider et al. (2017). They suggested that we can determine the level of authenticity on two dimensions: in terms of learning activities that approximate activities in the world of work (authentic activity) as well as the level of proximity the students experience to the world of work. Table 2 reallocates assessments based on the Kaider et al. (2017) matrix.

Based on this framework, we had two assessments of high authenticity, two of medium and two of low, thus ensuring a nice distribution of assessments. Unfortunately, none of these assessments scored very high in terms of business proximity, meaning they were not conducted in the actual workplace or within professional environments.

Looking at the results of the mapping across both frameworks, it is clear that assessments HEI1-L4 and HEI2-L5 scored high on authenticity, while assessments HEI2-L4 and HEI2-L6 scored low. Thus, we were reasonably assured that our assessments represented a range of high and low authenticity assessments.

Data collection procedure

The two-phased experiment consists of (1) marking and (2) mark reconciliation, debrief, and interview. A day to a week passed between Phase 1 and 2 to allow participants time to mark. In Phase 1 marking, markers were emailed three assessment briefs and marking criteria for L4, L5, and L6 assessments, from which the research team selected 12 assessments (four per level) that were a mix of human-authored, GenAI-augmented, and GenAI-generated samples. Each sample had a randomly assigned code, leaving limited room for 'mark guessing'. As an independent task, each marker was asked to assign grades against the marking criteria using a provided mark sheet. Once 12 assessments were marked, each marker was sent the remaining nine assessments to a total of 21 assessments. The Phase 1 marking procedure took the markers about 3 hours.

In Phase 2, four semi-structured in-depth interviews were conducted, each lasting approximately 2 hours (see Appendix B for the interview protocol). Initially, researchers merged the mark entry sheets collected from each pair of markers. A researcher in the interview then facilitated discussions between the markers to reach a grade consensus on all 21 writing samples while providing a justification for the agreed grade. If there was a significant difference in their scores, the markers reviewed the assessment criteria and discussed the specific elements of the writing samples that led to these discrepancies. After discussing and reconciling their differences, the markers agreed on a final mark that reflected their consensus, ensuring fairness and accuracy in grading. Unlike mark agreements in real-world scenarios, where a third marker or moderator would be assigned if the initial markers could not agree, our experiment did not involve a third marker.

After reaching an agreement on marks, the interviewer conducted a debriefing, revealing the identities of the assessments, including the original marks and feedback given to the human submissions, the GenAI-augmented samples, and the GenAI-generated samples. Participants were then asked to evaluate the similarities and differences between the

TABLE 1 Module assessment selection based on the authentic assessment tool developed by Hobbins et al. (2022).

Institution and academic level	Realism assessment engages students with problems or important questions relevant to everyday life beyond the classroom	Cognitive challenge categories of cognitive processes related to using, modifying or rebuilding knowledge into something new	Evaluative judgement criteria assessment provides opportunities for students to critically judge their own performance based on clear expectations	Evaluative judgement feedback assessment engages students with meaningful feedback to allow for improvement
HEI1-L4	Moderate	Moderate	High	Moderate
HEI2-L4	Moderate	Moderate	Moderate	Low
HEI1-L5	Moderate	Moderate	Moderate	Moderate
HEI2-L5	High	Moderate	Moderate	Moderate
HEI1-L6	Low	Moderate	Low	Moderate
HEI2-L6	Low	Moderate	Moderate	Low

TABLE 2 Module assessment selection based on Kaider et al. (2017) authenticity-proximity framework.

	Proximity			
	Learning experiences that occur in real workplaces and professional contexts, in online or live complex simulated workplace environments, and those that enable students to interact directly with industry practitioners or community members on work-related activities			
Authenticity Learning activities and assessments require students to work on problems, processes and projects that they may encounter in their professions and produce artefacts reflecting professional practice		Low proximity	Medium proximity	High proximity
	High authenticity	Active case studies and scenarios HEI1-L4 and HEI2-L5		
	Medium authenticity	HEI1-L6 and HEI1-L5		
	Low authenticity	Critical literature review HEI2-L4 and HEI2-L6		

original assessments and the GenAI-augmented and GenAI-generated assessments. They were also asked to share their usual approaches to identifying academic misconduct and their views on the future use of GenAI tools in higher education. In summary, the participants' overall impressions, grading differences and reflections were gathered. The process was carefully designed to minimise order effects and sources of bias (Charness et al., 2012) while ensuring academic integrity. All participants were thoroughly briefed and debriefed.

Data analysis

Quantitative data, in the form of marks given to each assessment, were first collected from each of the eight participants during Phase 1. During Phase 2, the research team used the agreed grades from reconciliation as descriptive statistics to contribute towards answering the first two research questions. Quantitative data were analysed using standard descriptive statistics, including the calculation of means and differences in marks between human-authored, GenAI-modified and GenAI-generated assessments.

Qualitative data were gathered from interview notes taken during the debriefing sessions and follow-up interviews with participants. The data were analysed using thematic analysis, a method that identifies themes and patterns to understand how markers' awareness of GenAI's involvement influenced their grading. The recorded interview data were transcribed and systematically analysed following the steps outlined by Boyatzis (1998). These steps included familiarising oneself with the data through repeated readings of the transcriptions, generating initial codes by identifying and labelling relevant segments of text, and then examining these codes to detect patterns and broader themes. These themes were reviewed, refined and related to our research questions, both inductively from the raw data and deductively from existing pedagogic literature. Table 3 provides an example of how codes emerged from interview extracts and were subsequently developed into themes.

TABLE 3 Example of thematic coding.

Interview extract	Codes	Theme
There's **not a single grammatical mistake**, not single punctuation [mistake] or single spelling is usually in American English	Perfection and lack of errors	Characteristics of AI-generated texts
I did catch a student using ChatGPT and how I did it was that I asked ChatGPT the question and it came back with **a particular form of words**. It is not a silver bullet and I found those form of words in the conclusions of four essays	Uniformity in structure and style	
I normally tend to check at least a few references and if I checked a couple of a few of those and I see that there is no, **it doesn't exist** out, then would have raised this as GPT to use	Referencing style and authenticity	
The type of the **language** that they use in terms of writing the, here I think it'll be very, very **generic**	Generic and descriptive language	
What AI is doing is **use lot of words**. As though you know **in the same sentence, what after what, after what meaning the same thing and then leading to a conclusion**. Umm. That's helpful this pattern	Repetitiveness and over-descriptiveness	

RESULTS

To answer RQ1, before the researcher debriefing, the four pairs of markers were informed that some of the samples were generated or modified using GenAI. They were then asked to 'guess' (after the reconciliation of marks had been completed) which those were. Most markers were able to 'sense' AI-modified samples but were much less able to differentiate between human-authored or GenAI-authored coursework with confidence and accuracy, demonstrating the sophistication of AI's output. As shown in Tables 4 and 5, among the four pairs, the accuracy of guessing ranged between 33.3% (7 out of 21) and 85.7% (18 out of 21).

Pair 2 had 33% accuracy in detecting AI versus non-AI work. One participant in Pair 2 mentioned they tried to spot potential AI-influenced work at the beginning of marking but gave up as they continued. This was not only due to the difficulty in discerning AI from non-AI work, but also because the quality of GenAI work may be dependent on "prompt engineering".

> It is difficult…. It's not easy to confirm because when you're using AI, it's really important to say how we train it. If we give the AI more specific guidance or we clearly tell it what would be our expectation and we give AI more like a learning materials or paper, …we ask it to construct a better quality one. It could be, but yeah, it depends.

Yet, with the research team's facilitation, Pair 2 was able to identify GPT-modified work at Level 4. One participant noted:

> The AI work is, … I mean, the piece is a little bit better [in terms of the writing quality and writing skills]. Maybe HEI2-L4-E is written by AI, and I suppose HEI2-L4-A is written by an AI.

Pair 1 had 85.7% accuracy in detecting AI versus human work. One participant in Pair 1 was the module leader of Level 4 and 6 modules from which the writing samples were derived. They were very familiar with the design of these modules' assessments and had taught small cohorts of about 20 students in each module. As this pair was interviewed, both participants reported that the 'writing style' helped them differentiate between AI and human work. They both said that every writer had a distinct, nonperfect writing style:

> The sort of everyone in their writing style makes the same kind of errors. Everyone has a distinct referencing style, whether they're using a tool or not. They have a very distinct way of doing that, and that's quite easy to spot as well.

Both participants described GenAI as not seeming to have a writing style, or in other words, GenAI was "too perfect":

> Human work, …. they're gonna do with it what they feel. It's their own style, their own way, or how they make sense of it. Styles of paragraphs are usually around the same if it's fully written by AI that they should be quite on the genius.… Whereas I feel the ChatGPT tends to follow a very strict structure…. There's not a single grammatical mistake, not single punctuation [mistake]…. Spelling is usually in American English, which doesn't really necessarily mean anything, because you have such a diverse cohort, but usually tends to ring bells, and they see American English a lot.

TABLE 4 Writing sample identities and markers' predictions in Higher Education Institution 1.

				Pair 1	Pair 2
Academic level	Grade band	Writing sample code	Sample identity: Original (O), AI-modified (M), AI-generated (G)	Guess of ChatGPT: Yes (Y), No (N), Unsure (U)	Guess of ChatGPT: Yes (Y), No (N), Unsure (U)
Level 4	≦40	HE1-L4-Z	O	N	N
		HE1-L4-F	M	Y	Y
	50–59	HE1-L4-J	O	N	N
		HE1-L4-Q	M	Y	Y
	≧70	HE1-L4-R	O	N	N
		HE1-L4-E	M	Y	Y
		HE1-L4-A	G	Y	Y
Level 5	≦40	HE1-L5-A	O	N	U
		HE1-L5-G	M	Y	U
	50–59	HE1-L5-K	O	N	U
		HE1-L5-R	M	U	U
	≧70	HE1-L5-S	O	N	U
		HE1-L5-F	M	U	U
		HE1-L5-B	G	Y	U
Level 6	≦40	HE1-L6-B	O	N	U
		HE1-L6-H	M	Y	U
	50–59	HE1-L6-L	O	N	U
		HE1-L6-S	M	Y	U
	≧70	HE1-L6-T	O	N	U
		HE1-L6-G	M	Y	U
		HE1-L6-C	G	U	U
Accuracy of guess percentage				85.70%	33.30%

TABLE 5 Writing sample identities and markers' predictions in Higher Education Institution 2.

				Pair 3	Pair 4
Academic level	**Grade band**	**Writing sample code**	**Sample identity: Original (O), AI-modified (M), AI-generated (G)**	**Guess of ChatGPT: Yes (Y), No (N), Unsure (U)**	**Guess of ChatGPT: Yes (Y), No (N), Unsure (U)**
Level 4	≦40	HE2-L4-K	O	N	Y
		HE2-L4-H	M	Y	N
	50–59	HE2-L4-C	O	N	N
		HE2-L4-L	M	Y	Y
	≧70	HE2-L4-N	O	U	N
		HE2-L4-G	M	U	Y
		HE2-L4-Z	G	N	Y
Level 5	≦40	HE2-L5-Z	O	Y	Y
		HE2-L5-L	M	N	N
	50–59	HE2-L5-N	O	U	U
		HE2-L5-Y	M	U	U
	≧70	HE2-L5-T	O	U	U
		HE2-L5-B	M	U	U
		HE2-L5-S	G	Y	U
Level 6	≦40	HE2-L6-D	O	N	Y
		HE2-L6-G	M	Y	N
	50–59	HE2-L6-H	O	N	N
		HE2-L6-O	M	Y	Y
	≧70	HE2-L6-Q	O	Y	Y
		HE2-L6-L	M	Y	N
		HE2-L6-A	G	N	N
Accuracy of guess percentage				42.90%	33.3%

We speculate that there were both false positive and false negative cases. False positives occurred when markers incorrectly identified original, unaltered student submissions as being influenced by GenAI, at times mistakenly flagging genuine work as dishonest. This happened when markers answered “Y” or “U” to “O” samples in Tables 4 and 5, resulting in six incidents in our experiment. This implies that students who are not cheating have been penalised on suspicion alone.

Conversely, false negatives occurred when assessments that have been modified or generated by GenAI were not recognised as such and were mistakenly considered as original student work. It appears in seven cases where markers answered “U” and “N” to “M” and “G” samples in Tables 4 and 5. This occurrence highlights the challenges in accurately detecting GenAI involvement, potentially leading to unjust penalties for honest students and failing to identify actual instances of GenAI use.

To answer RQ2, in Tables 6 and 7 we compared the mark difference between the original samples and the GenAI-modified samples. It was found that the GenAI-augmented samples only enhanced the original student-written work at certain grade bands. There seemed to be consistency in improvement at the lower second grade band (nine out of 12 cases, increasing by two to eight percentage points) and some mark improvement at the below-pass-mark

TABLE 6 Comparisons of marks awarded to human-authored work and ChatGPT modified work in Higher Education Institution 1.

				Pair 1		Pair 2	
Grade band	Academic level	Writing sample code	Sample identity: Original (O), AI-modified (M)	Remarks	Mark difference (M) − (O)	Remarks	Mark difference (M) − (O)
≦40	Level 4	HE1-L4-Z	O	38	0 ∆	40	0 ∆
		HE1-L4-F	M	38		40	
	Level 5	HE1-L5-A	O	38	20 ↑	45	0 ∆
		HE1-L5-G	M	58		45	
	Level 6	HE1-L6-B	O	35	5 ↑	38	0 ∆
		HE1-L6-H	M	40		38	
50–59	Level 4	HE1-L4-J	O	50	2 ↑	55	5 ↑
		HE1-L4-Q	M	52		60	
	Level 5	HE1-L5-K	O	52	8 ↑	52	3 ↑
		HE1-L5-R	M	60		55	
	Level 6	HE1-L6-L	O	45	3 ↑	35	7 ↓
		HE1-L6-S	M	48		28	
≧70	Level 4	HE1-L4-R	O	70	0 ∆	55	3 ↑
		HE1-L4-E	M	70		58	
	Level 5	HE1-L5-S	O	70	0 ∆	60	8 ↑
		HE1-L5-F	M	70		68	
	Level 6	HE1-L6-T	O	70	0 ∆	60	0 ∆
		HE1-L6-G	M	70		60	

TABLE 7 Comparisons of marks awarded to human-authored work and ChatGPT-modified work in Higher Education Institution 2.

				Pair 3		Pair 4	
Grade band	Academic level	Writing sample code	Sample identity: Original (O), AI-modified (M)	Remarks	Mark difference (M) − (O)	Remarks	Mark difference (M) − (O)
≦40	Level 4	HE2-L4-K	O	35	10 ↑	42	0 Δ
		HE2-L4-H	M	45		42	
	Level 5	HE2-L5-Z	O	52	3 ↑	48	4 ↑
		HE2-L5-L	M	55		52	
	Level 6	HE2-L6-D	O	42	0 Δ	38	0 Δ
		HE2-L6-G	M	42		38	
50–59	Level 4	HE2-L4-C	O	55	3 ↓	55	7 ↑
		HE2-L4-L	M	52		62	
	Level 5	HE2-L5-N	O	45	3 ↑	48	3 ↓
		HE2-L5-Y	M	48		45	
	Level 6	HE2-L6-H	O	45	7 ↑	42	6 ↑
		HE2-L6-O	M	52		48	
≧70	Level 4	HE2-L4-N	O	58	7 ↑	62	10 ↓
		HE2-L4-G	M	65		52	
	Level 5	HE2-L5-T	O	58	0 Δ	52	3 ↑
		HE2-L5-B	M	58		55	
	Level 6	HE2-L6-Q	O	58	10 ↓	62	10 ↓
		HE2-L6-L	M	48		52	

grade band (five out of 12 cases, increasing by three to 20 percentage points), but not at the first-class grade band. There was also a pattern of consistency at the below-pass-mark grade band. In seven out of 12 cases, markers awarded the same grade for the original submission and the GPT-modified sample. One participant in Pair 4 explained that because the structure in the human work (HEI1-L4-K) was poor, even with GenAI intervention, it did not improve much in structure or academic style. Similarly, both participants in Pair 3 stated that in several cases, the original submission sample and the GPT-augmented sample appeared very similar.

To our surprise, almost all pairs graded the writing samples of originally first-class grades much lower (except Pair 1). One participant in Pair 3 admitted:

> Some of this work is presented really nicely, and if people are not marking consistently, it's going to be very easy to look at something and go: That's got all the component parts, and it's ticking all the boxes from the assessment because the buzzwords are there. Because I think, you know, partly … when we're marking … there are things that we're hoping to see. If at the level you see things and tick them off, your mental list goes there. That said, that's there. You're kind of missing the substance behind what's actually been written.

We found that high-scoring samples seem to be adversely affected by false positives. In three out of 12 cases (primarily involving pair 3 and pair 4) at the first-class grade band in both Tables 6 and 7, markers assigned scores that were 10 percentage points lower on GenAI-modified assessments compared to the original submissions. Additionally, pairs 2, 3 and 4 awarded lower scores to original assessments that were initially awarded 70 and above, reducing them to the 60s or even lower. We suspect that the presence of GenAI indeed influenced the markers' grading process. Considering that the markers were aware that some of the samples they were marking had been generated by GenAI, it seems possible that markers may have subconsciously deducted marks from work they perceived as "too perfect."

Six GenAI-authored samples (ie, HE1-L4-A, HE1-L5-B, HE1-L6-C, HE2-L4-Z, HE2-L5-S, and HE2-L6-A) were generated from scratch by the research team using GenAI (see Tables 4 and 5). From the four pairs of participants' marking, the marks ranged between 55 and 70, or the upper-second level. We suspect that, as the assessments were selected for relative ease in using GenAI (eg, ChatGPT), students were not required to draw on learning experiences that occur in real workplaces and professional contexts, making it easier for them to use GenAI to complete an assessment. Since each of these six writing samples was generated within 90 minutes, GenAI appears to be a convenient, low-cost and powerful tool that may be misused by a 'mischievous' student without prior knowledge of a subject to pass assessments.

DISCUSSION

In response to RQ1, our results strongly indicate that assessments were easily compromised using GenAI, and many of our markers found it challenging to identify which were human-authored, which were GenAI-modified, and which were GenAI-authored. While the literature remains ambivalent with regard to the impact of GenAI on assessments (Bower et al., 2024), our findings suggest that there is an impact and that it is challenging to identify GenAI usage in assessments.

As revealed in our data analysis, although marked improvement at the lower grade bands from GenAI-augmented samples was limited, in higher grade bands, mischievous students—despite the concerns raised around GenAI competence (Kocoń et al., 2023; Rudolph et al., 2023)—would benefit from using GenAI to generate completely "original" work because of its low cost, high speed and relatively hard-to-detect features. Our results also indicated that GenAI-generated assessments, with adequate knowledge of prompt engineering (Giray, 2023; Lo, 2023), would easily assist a student in passing a module without having to engage in deep learning. This raises concerns about current approaches to assessment.

A common response might be to suggest that software exists to detect AI-generated content. However, using such tools can be problematic, as markers may not easily assess the level of knowledge demonstrated by the student (Gorichanaz, 2023) without further assessment. In fact, in many universities, current guidance regarding the usage of GenAI suggests that for all assessments that score high on AI detection, students would need to undergo an oral examination, making the marking process far more onerous and time-consuming. Extrapolating from this, if many written assessment submissions require a second layer of oral examinations to prove their academic integrity, we could argue that such assessments should be replaced altogether with face-to-face types of assessments based on performance or oral presentation. There are many interesting approaches to such assessments; for example, interactive oral assessments (Krautloher, 2024; Ward et al., 2023) and role-play–type assessments that allow for thorough evaluation of the student, thereby overcoming the GenAI challenge.

In response to RQ2, our findings indicated that the existence of GenAI influenced the marking process for staff, as they began second-guessing the authorship and attempted to spot GenAI-modified and GenAI-authored work. That led to unusually large grade disparities between first and second markers, as it appears that in some cases, markers may have subconsciously or consciously marked down work they suspected to be GenAI-modified or authored. This dynamic is noteworthy and indicates that the presence and awareness of GenAI have impacted the way academics mark and reconcile grades. Another interesting consequence of the markers' awareness of GenAI was that all human-authored work originally graded as first class was graded below first class, and in several cases, the two markers had significantly different grades. As indicated in our findings, good work was often viewed as work done by GenAI.

It is worth noting that as GenAI technology becomes more advanced, many of the criteria that our more successful markers (pair 1, for example) used to identify GenAI-modified and GenAI-authored assessments will not stand the test of time. For example, a couple of interviewees noted disjointedness in GenAI-manipulated work; this was due to the limitations in ChatGPT 3.5, which did not allow manipulation of text longer than 700 words. Another pair of markers highlighted the flow of language that GenAI tends to generate, an issue that could be easily overcome by smarter prompt engineering.

In response to RQ3, it seems that moderate/high-level authentic assessments are neither a shield for academic integrity nor an immediate solution to the GenAI challenge. As we saw in the methodology section, some assessments were of low authenticity, and others were of high authenticity; however, this made very little difference in detecting GenAI usage. All nine assessments were relatively easy to reproduce or modify using GenAI, and the markers could not readily identify the difference between human-authored and GenAI-augmented or GenAI-generated variations. Thus, it is doubtful whether authentic assessments are the panacea that the Quality Assurance Agency for Higher Education (2024) and Advance HE (2023) suggest when dealing with GenAI.

Authentic assessments pose certain challenges (Ajjawi et al., 2020; Ellis et al., 2020) that the rhetoric around assessments does not seem to consider. Here we will focus on three immediate challenges posed by authentic assessments.

The first challenge relates to the **authenticity concept itself**. In a sense, all assessments are authentic—they provide a measurement of learning that should, by definition, be relevant in the workplace (Kaider et al., 2017). Thus, when suggesting that authentic assessments may help us manage the GenAI challenge, we need to be clearer about what type of authentic assessments we are considering.

The second challenge relates to the **complexity of authentic assessments**. Authentic assessments tend to be more complex, especially those that score high on authenticity, proximity, or both (Kaider et al., 2017). It is generally easier to manage the creation and execution of an exam (low authenticity, low proximity) than to arrange and manage a live project (medium/high authenticity, medium proximity). The former is predictable, with parameters known in advance, while the latter is highly unpredictable, with more diverse intended learning outcomes and higher risks due to the involvement of additional stakeholders. In general, the more authentic an assessment, the more time-consuming and complex it is to manage, particularly with high student volumes (Ajjawi et al., 2020; Ashford-Rowe et al., 2014).

The third challenge is the **speed of GenAI evolution**. Authentic assessments by themselves cannot provide enough of a safeguard against using GenAI to circumvent assessments (Ellis et al., 2020); the GenAI technology is evolving too rapidly, putting at risk increasingly complex types of assessment. Consider the introduction of similarity detection services (such as Turnitin), which led to the development of essay mills and other third-party customised writing services. GenAI may further escalate the continuous efforts by HEIs to protect their assessments' academic integrity.

CONCLUSIONS

Our work indicates that if we are to mitigate the risks posed by GenAI, we may have to reconsider our approach to assessment design. Our experimental results suggest that with GenAI tools, it is becoming challenging to use written types of assessment to evaluate students, and academics will be affected by their judgments of students' GenAI usage when marking such work, thus leading to doubts regarding the reliability of the marking and moderation process. This weakens output-based assessments and leads us to advocate for more performative and interactive types of assessments in lieu of written assessments.

As shown in our findings, while authentic assessments are often offered in the literature as the solution and can measure multidimensional outcomes and provide a semblance of real-world relevance, they are not immune to the challenges GenAI poses. Intelligent assessment design may still be important in safeguarding academic integrity, regardless of the degree of authenticity and workplace proximity an assessment may demonstrate (Ajjawi et al., 2020).

Thus, managing the impact of GenAI requires a paradigm shift in assessment philosophy and design (Adiguzel et al., 2023; Gorichanaz, 2023). Instead of viewing GenAI solely as a challenge or threat to academic integrity and assessment design, if we are intelligent in our assessment design, we can convert GenAI into a tool for further learning (Salinas-Navarro et al., 2024). For example, GenAI can be used to simulate real-world professional environments where students must solve complex, authentic problems. This includes tasks like designing marketing campaigns with AI-generated consumer data, conducting AI-facilitated business negotiations or analysing AI-driven case studies tailored to specific industries. Furthermore, GenAI can generate adaptive learning content, offering students iterative feedback on projects, such as refining an AI-drafted research proposal or creating multiple drafts of a business report based on AI-provided critiques. These applications transform GenAI into an active partner in experiential and collaborative learning rather than a passive content generator.

However, this would mean abandoning some current assessment practices, which may have become obsolete. Our experiment was exploratory, and the sample was small;

however, our results raise doubts about the nature of knowledge and learning and the value of written assessments. Perhaps learning and knowledge creation are socialised, situated in people and their interactions and as such experiential and tacit (Tsoukas, 1996; von Krogh & Roos, 1995). The advent of GenAI has made this perspective even more pronounced by putting a premium on socialised learning, as explicit knowledge and learning can be rapidly reproduced and reconfigured. Thus, what HEIs should assess is knowledge application and socialised knowledge which, likewise, is socially situated and experiential. That would mean a shift towards an assessment paradigm where we assess knowledge as generated in social settings, synchronous, experiential learning which is situated within social networks and social interactions (Kofinas & Tsay, 2021). Kofinas and Tsay (2021) in their work emphasised that large classes function as a social microcosm, leveraging diverse interactions and weak ties to foster socialised experiential learning. GenAI can enable rich learning experiences by facilitating collaborative projects where students can leverage GenAI simulated market trends and information to foster both individual and collective problem-solving skills. However, what would be assessed would be the ability of students to perform and communicate what they learned in a synchronous, interpersonal manner.

Moving away from assessing learning and knowledge using written work and static artefacts could have far-reaching ramifications, as they are currently an important and significant portion of the full range of higher education assessment types. Thus, more research is needed to evaluate the relevance of written assessments in a GenAI-dominated world and the nature of learning students should be engaging in, which would embrace GenAI as a tool. The ramifications on other modes of learning, such as online learning, distance learning, and other writing-dominated approaches to higher education, can be severe. A move away from text as an indicator of knowledge and learning has wider ramifications in the realm of research, as currently the main currency of research tends to be written published work such as journal articles and books, and with GenAI the value of such outputs becomes increasingly questionable. These are all fruitful areas for further research in this field as GenAI will continue impacting higher education for many years to come.

FUNDING INFORMATION

There was no funding support for this research.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

None.

ORCID

Alexander K. Kofinas https://orcid.org/0000-0003-4577-2883

REFERENCES

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, *15*(3), ep429.

Advance HE. (2023). *Higher education in the era of AI*. https://www.advance-he.ac.uk/news-and-views/higher-education-era-ai

Advance HE. (2024). *Authentic assessment in the era of AI*. https://advance-he.ac.uk/membership/all-member-benefit-projects/Authentic-Assessment-in-the-era-of-AI

Ajjawi, R., Tai, J., Huu Nghia, T. L., Boud, D., Johnson, L., & Patrick, C.-J. (2020). Aligning assessment with the needs of work-integrated learning: The challenges of authentic assessment in a complex context. *Assessment and Evaluation in Higher Education*, *45*(2), 304–316.

Appel, G., Neelbauer, J., & Schweidel, D. A. (2023, April 7). Generative AI has an intellectual property problem. *Harvard Business Review*. https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, *39*(2), 205–222. https://doi.org/10.1080/02602938.2013.819566

Aydın, Ö., & Karaarslan, E. (2023). Is ChatGPT leading generative AI? What is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*, *11*(3), 118–134.

Bartlett, T. (2009, March 20). Cheating goes global as essay mills multiply. *The Chronicle of Higher Education*, *55*(28). https://www.chronicle.com/article/cheating-goes-global-as-essay-mills-multiply/

Beutel, G., Geerits, E., & Kielstein, J. T. (2023). Artificial hallucination: GPT on LSD? *Critical Care*, *27*(1), 148. https://doi.org/10.1186/s13054-023-04425-6

Biggs, J. (2003). Aligning teaching for constructing learning. *Higher Education Academy*,*1*(4), 1–4.

Bower, M., Torrington, J., Lai, J. W., Petocz, P., & Alfano, M. (2024). How should we change teaching and assessment in response to increasingly powerful generative artificial intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies*, *29*, 15403–15439. https://doi.org/10.1007/s10639-023-12405-0

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage.

Broadbent, J., Panadero, E., & Boud, D. (2018). Implementing summative assessment with a formative flavour: A case study in a large class. *Assessment & Evaluation in Higher Education*, *43*(2), 307–322.

Brown, S. (2005). Assessment for learning. *Learning and Teaching in Higher Education*, *1*, 81–89.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv Preprint arXiv*, 2303.04226.

Carnell, B. (2016). Aiming for autonomy: Formative peer assessment in a final-year undergraduate course. *Assessment & Evaluation in Higher Education*, *41*(8), 1269–1283.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8.

Cilliers, F. J., Schuwirth, L. W., Herman, N., Adendorff, H. J., & van der Vleuten, C. P. (2012). A model of the pre-assessment learning effects of summative assessment in medical education. *Advances in Health Sciences Education*, *17*(1), 39–53.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education*. Routledge.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239.

Cronan, T. P., Mullins, J. K., & Douglas, D. E. (2018). Further understanding factors that explain freshman business students' academic integrity intention and behavior: Plagiarism and sharing homework. *Journal of Business Ethics*, *147*, 197–220.

Deng, J., & Lin, Y. (2022). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, *53*, 103662.

Ellis, C., Van Haeringen, K., Harper, R., Bretag, T., Zucker, I., McBride, S., Rozenberg, P., Newton, P., & Saddiqui, S. (2020). Does authentic assessment assure academic integrity? Evidence from contract cheating data. *Higher Education Research and Development*, *39*(3), 454–469. https://doi.org/10.1080/07294360.2019.1680956

Firth, D. R., Derendinger, M., & Triche, J. (2024). Cheating better with ChatGPT: A framework for teaching students when to use ChatGPT and other generative AI bots. *Information Systems Education Journal*, *22*(3), 22–60. https://doi.org/10.62273/BZSU7160

Forsyth, R. (2022). *Confident assessment in higher education*. Sage.

Giannos, P., & Delardas, O. (2023). Performance of ChatGPT on UK standardized admission tests: Insights from the BMAT, TMUA, LNAT, and TSA examinations. *JMIR Medical Education*, *9*(1), e47737. https://doi.org/10.2196/47737

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Glendinning, I. (2022). Aligning academic quality and standards with academic integrity. In G. J. Curtis, J. Seeland, B. M. Stoesz, J. Clare, S. E. Eaton, & K. Rundle (Eds.), *Contract cheating in higher education: Global perspectives on theory, practice, and policy* (pp. 199–218). Springer. https://doi.org/10.1007/978-3-031-12680-2_14

Gorichanaz, T. (2023). Accused: How students respond to allegations of using ChatGPT on assessments. *Learning: Research and Practice*, *9*(2), 183–196. https://doi.org/10.1080/23735082.2023.2254787

Gross, N. (2023). What ChatGPT tells us about gender: A cautionary tale about performativity and gender biases in AI. *Social Sciences*, *12*(8), 435.

Guo, H., Venkit, P.N., Jang, E., Srinath, M., Zhang, W., Mingole, B., Gupta, V., Varshney, K.R., Sundar, S.S. and Yadav, A., 2024. Hey GPT, Can You be More Racist? Analysis from Crowdsourced Attempts to Elicit Biased Content from Generative AI. arXiv preprint arXiv:2410.15467.

Harlen, W. (2004). A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes. *EPPI-Center*. https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/ass_rv4.pdf?ver=2006-03-02-124724-997

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*, 18617. https://doi.org/10.1038/s41598-023-45644-9

Hobbins, J., Kerrigan, B., Farjam, N., Fisher, A., Houston, E., & Ritchie, K. (2022). Does a classroom-based curriculum offer authentic assessments? A strategy to uncover their prevalence and incorporate opportunities for authenticity. *Assessment & Evaluation in Higher Education*, *47*(8), 1259–1273.

Holmes, N. (2015). Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module. *Assessment & Evaluation in Higher Education*, *40*(1), 1–14. https://doi.org/10.1080/02602938.2014.881978

Iannone, P., & Simpson, A. (2016). University students' perceptions of summative assessment: The role of context. *Journal of Further and Higher Education*, *41*, 785–801.

Kaider, F., Hains-Wesson, R., & Young, K. J. (2017). Practical typology of authentic work-integrated learning activities and assessments. *Asia-Pacific Journal of Cooperative Education*, *18*(2), 153–165.

Kocoń, A., Cichecki, T., Kaszyca, A., Kochanek, M., Szydło, R., Baran, R., Bielaniewicz, M., Gruza, M., Janz, K., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kofinas, A. (2018). Managing the sublime aesthetic when communicating an assessment regime: The Burkean pendulum. *Management Learning*, *49*(2), 204–221. https://doi.org/10.1177/1350507617738864

Kofinas, A. K., & Tsay, C. H. H. (2021). In favor of large classes: A social networks perspective on experiential learning. *Journal of Management Education*, *45*(5), 760–785. https://doi.org/10.1177/1052562921102281 9

Kolade, O., Owoseni, A., & Egbetokun, A. (2024). Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework. *Heliyon*, *10*(4), e25953.

Kolb, A. Y., & Kolb, D. A. (2005). Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of Management Learning & Education*, *4*(2), 193–212. https://doi.org/10.5465/amle.2005.17268566

Krautloher, A. (2024). Improving assessment equity using interactive oral assessments. *Journal of University Teaching & Learning Practice*, *21*(4),1-17. https://doi.org/10.53761/4hg1me11

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *Journal of Academic Librarianship*, *49*(4), 102720. https://doi.org/10.1016/j.acalib.2023.102720

López-Pastor, V. M., Pintor, P., Muros, B., & Webb, G. (2013). Formative assessment strategies and their effect on student performance and on student and tutor workload: The results of research projects undertaken in preparation for greater convergence of universities in Spain within the European Higher Education Area (EHEA). *Journal of Further and Higher Education*, *37*(2), 163–180. https://doi.org/10.1080/0309877X.2011.644780

Lu, J., Yu, C.-S., & Liu, C. (2003). Learning style, learning patterns, and learning performance in a WebCT-based MIS course. *Information & Management*, *40*(6), 497–507. https://doi.org/10.1016/S0378-7206(02)00064-2

Lund, J. (1997). Authentic assessment: Its development & applications. *Journal of Physical Education, Recreation & Dance*, *68*(7), 25–28.

Marr, B. (2023, May 19). A short history of ChatGPT: How we got to where we are today. *Forbes*. https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-ChatGPT-how-we-got-to-where-we-are-today/

Medway, D., Roper, S., & Gillooly, L. (2018). Contract cheating in UK higher education: A covert investigation of essay mills. *British Educational Research Journal*, *44*(3), 393–418. https://doi.org/10.1002/berj.3335

Morris, E. (2018). Academic integrity matters: Five considerations for addressing contract cheating. *International Journal for Educational Integrity*, *14*(1), 15. https://doi.org/10.1007/s40979-018-0038-5

Mueller, J. (2005). The authentic assessment toolbox: Enhancing student learning through online faculty development. *Journal of Online Learning and Teaching*, *1*(1), 1–7.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

OpenAI. (2024). *OpenAI API*. https://platform.openai.com/docs/models

Pagani, R. N., de Sá, C. P., Corsi, A., & de Souza, F. F. (2023). AI and employability: Challenges and solutions from this technology transfer. In M. D. Lytras, A. A. Housawi, & B. S. Alsaywid (Eds.), *Smart cities and digital transformation: Empowering communities, limitless innovation, sustainable development and the next generation* (pp. 253–284). Emerald.

Pastis, S. (2023). 3 easy steps to start using ChatGPT, Google Bard, and Bing Chat A.I. tools right now. *Fortune*. https://fortune.com/2023/07/21/how-to-use-ai-bots-3-easy-steps-ChatGPT-google-bard-bing-chat/

Pereira, D., Flores, M. A., & Niklasson, L. (2015). Assessment revisited: A review of research in assessment and evaluation in higher education. *Assessment & Evaluation in Higher Education*, *41*(7), 1008–1032. https://doi.org/10.1080/02602938.2015.1055233

Quality Assurance Agency for Higher Education. (2024). *Reconsidering assessment for the Chat GPT era: QAA advice on developing sustainable assessment strategies*. https://www.qaa.ac.uk/docs/qaa/members/reconsidering-assessment-for-the-chat-gpt-era.pdf?sfvrsn=38d3af81_6

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, *13*(9), 5783.

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning & Teaching*, *6*(1),41–56. https://doi.org/10.37074/jalt.2023.6.1.29

Raupach, T., Brown, J., Anders, S., Hasenfuss, G., & Harendza, S. (2013). Summative assessments are more powerful drivers of student learning than resource intensive teaching formats. *BMC Medicine*, *11*(1), 61–71. https://doi.org/10.1186/1741-7015-11-61

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rolfe, V. (2011). Can Turnitin be used to provide instant formative feedback? *British Journal of Educational Technology*, *42*(4), 701–710. https://doi.org/10.1111/j.1467-8535.2010.01091.x

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, *12*(3), 148. https://doi.org/10.3390/socsci12030148

Rudolph, J., Tan, J., & Tan, E. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, *6*(2), 342–363.

Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024). Using generative artificial intelligence tools to explain and enhance experiential learning for authentic assessment. *Educational Sciences*, *14*(1), 83. https://doi.org/10.3390/educsci14010083

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing Test" case study. *PLoS One*, *19*(6), e0305354. https://doi.org/10.1371/journal.pone.0305354

Sinnott-Armstrong, W. (2003). Consequentialism. In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/consequentialism/

Sok, S., & Heng, K. (2023). ChatGPT for education and research: A review of benefits and risks. *SSRN Electronic Journal*, *3*(1), 110–121. https://doi.org/10.2139/ssrn.4378735

Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education*, *45*(11), 2132–2148. https://doi.org/10.1080/03075079.2019.1582015

Spennemann, D. H. R., Biles, J., Brown, L., Ireland, M. F., Longmore, L., Singh, C. L., Wallis, A., & Ward, C. (2024). ChatGPT giving advice on how to cheat in university assignments: How workable are its suggestions? *Interactive Technology and Smart Education*, *21*(4), 690–707. https://doi.org/10.1108/ITSE-10-2023-0195

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*, *6*(1), 31–40.

Sutherland-Smith, W. (2008). *Plagiarism, the Internet, and student learning: Improving academic integrity*. Routledge. https://doi.org/10.4324/9780203928370

Sweeney, S. (2023). Who wrote this? Essay mills and assessment—Considerations regarding contract cheating and AI in higher education. *The International Journal of Management Education*, *21*(2), 100818. https://doi.org/10.1016/j.ijme.2023.100818

Trotter, E. (2006). Student perceptions of continuous summative assessment. *Assessment & Evaluation in Higher Education*, *31*(5), 505–521. https://doi.org/10.1080/02602930600679506

Tsoukas, H. (1996). The firm as a distributed knowledge system: A constructionist approach. *Strategic Management Journal*, *17*(S2), 11–25. https://doi.org/10.1002/smj.4250171104

Turner, P., & Turner, S. (2009). Triangulation in practice. *Virtual Reality*, *13*, 171–181. https://doi.org/10.1007/s10055-009-0117-2

Ventayen, R. J. M. (2023). *ChatGPT by OpenAI: Students' viewpoint on cheating using artificial intelligence-based application.* Available at SSRN 4361548.

von Krogh, G., & Roos, J. (1995). *Organisational epistemology*. Macmillan. https://doi.org/10.1007/978-1-349-24034-0

Ward, M., O'Riordan, F., Logan-Fleming, D., Cooke, D., Concannon-Gibney, T., Efthymiou, M., & Watkins, N. (2023). Interactive oral assessment case studies: An innovative, academically rigorous, authentic assessment approach. *Innovations in Education and Teaching International*, *61*(5), 930–947. https://doi.org/10.1080/14703297.2023.2251967

Wu, Y. (2024). Evaluating ChatGPT: Strengths and limitations in NLP problem solving. *Highlights in Science, Engineering and Technology*, *94*, 319–325.

How to cite this article: Kofinas, A. K., Tsay, C.-H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*, *00*, 1–28. https://doi.org/10.1111/bjet.13585

APPENDIX A

SAMPLE PROMPTS USED IN A LEVEL 6 ASSESSMENT

This Level 6 assessment is a 2000-word critical review that requires a student to review one paper from the List of Articles for Critical Review on the virtual learning environment. The list contains 18 papers. Students are advised not to select a paper at random, but to consider their reasons for selection and develop them in the review. Students are advised to make use of the concepts and notions taught throughout the module. Students should be citing a wide range of academic literature, both from the tutorials and their own research.

Prompt	Researcher question	Result
1	Would you please summarise this assessment brief and the marking criteria?	GenAI had a clear understanding of this assessment
2	Select one of the articles from the list below for your review process. Do not select at random; go through them and select one. Develop/justify your reasons for selection	GenAI selected an article and provided reasons
3	That's very good. Thank you. Can you create an outline for the critical review of this article?	GenAI created a clear structure for the outline
4	In order to create a 2000-word critical review, can you suggest a word length for each section in the outline above?	GenAI provided a recommended length for each section
5	Thank you. Please indicate the chosen article and write a 200-word introduction for the critical review of [article reference]. In the introduction, please justify the selection of article and signpost the structure of the critical review	GenAI did what it was asked and generated a 208-word introduction
6	That's excellent. Would you please write according to the outline you made and create a summary section of the article for 400 words?	GenAI responded with 437 words

Prompt	Researcher question	Result
7	According to the outline you created, would you please write a section of critical analysis for 600 words? Consider different viewpoints and use third person in writing. For each issue identified, develop arguments and use a wide range of academic literature to support	GenAI initially generated 607 words and stopped because of word count. Then the researcher asked it to continue
8	Continue please	GenAI continued
9	Would you please provide full references for the articles you cited?	GenAI came up with full references using the Harvard Referencing Style and reminded the researcher to present references according to the institutional requirements
10	For the Critical Analysis section, would you paraphrase and come up with appropriate subheadings?	This time GenAI only came up with 400 words, but it should be acceptable as the researcher will take the subheadings and add back to the GPT response to Prompt 7
11	According to the outline you created, would you please write a section of "Discussion of Sustainable Development Elements" (400 words). In this section, please add subheadings where appropriate. Consider different viewpoints and use third person in writing. For each issue identified, develop arguments and use a wide range of academic literature to support	GenAI generated a response with 439 words. Headings were provided and in-text citations included
12	Would you please provide full references for the articles you cited?	GenAI did as requested
13	According to the outline you created, would you please write a section of Insights and Recommendations (300 words). In this section, please add subheadings where appropriate. Consider different viewpoints and use third person in writing. For each issue identified, develop arguments and use a wide range of academic literature to support	GenAI did as requested and generated 287 words
14	Thanks. If there are any new citations, please provide full references	GenAI did as requested
15	Please write a conclusion (100 words) for this critical review	After the conclusion, GPT came up with a critical review of 2058 words, which is perfect
16	(Cut and paste all the references that GenAI has generated in this chat) Would you please reorganise the references here to make sure they are ordered and in compliance with the Harvard referencing standards?	GenAI did as requested

APPENDIX B

INTERVIEW PROTOCOL

Section 1: Questions about your experience and yourself (10 minutes)

5 minutes: Meet and greet. Explain the purpose of the research based on Appendix A, Participant Information Sheet. Ask participants to give consent if they haven't.

1. Which department do you work in?
2. How long have you been working in higher Education? ________ year(s) _________ month(s)
3. How long have you taught or moderated for [module/unit name]?
4. What is your primary connection to [module/unit name]?

Section 2: Researcher facilitation of moderation (60 minutes)

1. Reveal the mark entry sheet where marks of 21 writing samples were assigned by each participant.
2. Ask each participant to justify marks assigned and reach agree marks for all writing samples. Questions may include
   a. Using the marking criteria, can you explain as you examined the assignment how you arrived at the grade for the assessment?

Section 3: Researcher debriefs and interview (50 minutes)

1. Present the full set of assessments to the participants and the original marks and feedback given to the human submissions.
2. Explain which versions of assessments were GenAI generated.
3. Evaluate how the marking results were different. NB: It is important to remind participants that the objective was not to trick them, but rather to see if GenAI could generate a reasonable approximation of an assessment.
   a. First, compare the original marks and the marks agreed upon by two participant markers on human-written assessments vis-à-vis the ChatGPT-generated assessment. Ask participants why they got it right (or wrong). Were there any patterns that made the GenAI-generated assessment recognisable?
   b. When comparing the originally graded assessment and the grade you gave, are you prepared to maintain the same grade? (and if so why?) (Repeat for each level of study)
   c. Compare marks between the human and AI-generated assessments. Discuss any grading differences.

   Suggested line of questioning:

   “Comparing [assessment code] and [assessment code], in what ways is the GenAI written assessment better or worse than the human written assessment?”

   “If your grade for the GenAI assessment [assessment code] was lower, what changes would have led to a higher grade? In other words, how could GenAI perform better?”

   “You gave a higher grade on the GenAI assessment [assessment code] than the human written assessment [assessment code]. What differences prompted you to give a higher grade?”

   “Does GenAI do a better job at [academic level x/year group x]?”

d. The future use of AI-assisted tools in higher education
   (i) What is your view of GenAI tools and their role in higher education?
   (ii) Are there particular types of assessment you would consider to be more vulnerable to GenAI usage?
   (iii) In what ways could assessments be redesigned to mitigate the risk of GenAI usage?
   (iv) In what ways do you see GenAI writing tools being used to improve students' assessment literacies? For example, where a student couldn't understand how an assessment might be written.

Thank participants and assure them that no personal identifying information will be used when presenting research findings.