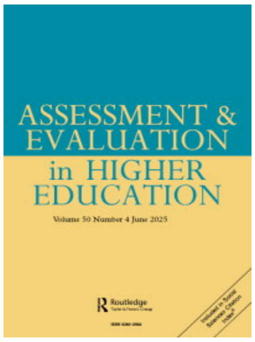

Taylor & Francis
Taylor & Francis Group

Assessment & Evaluation in Higher Education

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

How vulnerable are UK universities to cheating with new GenAI tools? A pragmatic risk assessment

Philip M. Newton

To cite this article: Philip M. Newton (16 Jun 2025): How vulnerable are UK universities to cheating with new GenAI tools? A pragmatic risk assessment, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2025.2511794

To link to this article: https://doi.org/10.1080/02602938.2025.2511794

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2025.2511794

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

How vulnerable are UK universities to cheating with new GenAI tools? A pragmatic risk assessment

Philip M. Newton

Swansea University Medical School, Swansea, Wales, United Kingdom

ABSTRACT
There has been considerable speculation about the risk that new generative AI tools like ChatGPT pose to higher education, particularly assessments and cheating. However it is unclear how much risk the UK higher education sector is exposed to. This survey study used a modified list experiment to evaluate that risk. Most students surveyed were using GenAI, and almost all were frequently assessed using methods that are vulnerable to cheating with GenAI (unsupervised online examinations, and essays). An estimated 22% of them reported cheating in academic year 2023/24 and almost all were confident that they understood the policy of their university. However the activities that they report using GenAI for were not always clearly identifiable as cheating. These findings suggest an urgent need for the use of more appropriate forms of summative assessment in UK higher education, and clarity over the policies and definitions used to support those assessments.

KEYWORDS
Artificial intelligence; assessment validity; academic integrity; cheating

Introduction

This study evaluates the risk that the UK higher education sector is exposed to from student cheating with new generative AI (GenAI) tools. The study uses anonymous survey methods (a list experiment) which allow participants to respond fully anonymously. The emergence of powerful new GenAI tools such as ChatGPT has generated concern about students using these tools to cheat, reflected in headlines from the UK press such as 'AI cheating is overwhelming the education system' (Naughton 2024) and 'ChatGPT-written essays swamp marking season' (Grove 2024). Others propose that these concerns are unfounded, for example papers showing that students using AI do not improve their performance (Smerdon 2024), and that inaccuracies in the output produced by ChatGPT are serious (Hicks, Humphries, and Slater 2024), and thus might negate concerns about cheating. Another response to the concerns is that, even if students did use ChatGPT on their assessments, it can be reliably detected in MCQ examinations (Sorenson and Hanson 2024) or in text form, even if paraphrased (Turnitin 2024).

The potential for cheating affects some assessment methods more than others. Multiple blinded studies show that ChatGPT is able to write essays and similar forms of asynchronous coursework to an equivalent or higher standard than university students, in ways that are difficult to detect by human markers (reviewed in Newton and Jones 2024). Even when markers suspect that they have identified text produced with GenAI, they are not confident proceeding with an

CONTACT Philip M. Newton p.newton@swansea.ac.uk Swansea University Medical School, Swansea SA2 8PP, Wales, United Kingdom.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02602938.2025.2511794.

academic misconduct investigation (Covington and Vruwink 2024). These challenges are compounded by the limitations of tools to identify text produced with GenAI. These tools are largely accurate, but they can be bypassed and there is no way to independently verify their output, so the risk of false positives seems unacceptably high (Perkins et al. 2024; Weber-Wulff et al. 2023).

Online unsupervised examinations are also an at-risk format. A student using a now-retired version of ChatGPT (GPT-4), would pass existing standardised MCQ-based examinations, even medical postgraduate level (Newton and Xiromeriti 2024). This level of performance is even better with the current (May 2025) free version of ChatGPT (GPT-4o), which can analyse images and search the internet (Newton et al. 2025).

However, new GenAI tools clearly offer enormous benefits to learners in ways that previous technologies have not, perhaps most obviously in academic writing (Wang, Li, and Bonk 2024). Therefore, from a policy perspective it is now unclear what it means to 'cheat', and what is an acceptable use of these tools to support learning. For example, if a student produces an essay where an outline draft has been generated by GenAI, does that essay represent the learning of the student, or is it cheating? The UK Higher Education Policy Institute asked students whether they had ever engaged in a range of behaviours using GenAI. 88% of them report using GenAI for their assessments, while 54% reported that they were dissuaded from using the tools due to fear of being accused of cheating (Freeman 2025).

The focus here is on the UK but these are international issues. A survey of US students also found that students are commonly using GenAI tools but they are concerned about being accused of cheating (Baek, Tate, and Warschauer 2024), while another US survey found that the tasks that students are using GenAI for do not seem to clearly fall into, or out of, a description of academic misconduct, e.g. 'getting started', 'generating a first draft' or 'help with an assignment' (Golding et al. 2024). This uncertainty is reflected in international university policies regarding GenAI use for assessments. An analysis of policies from 20 of the world's leading universities found that they have struggled to keep up with the pace of change and most are still focused on the basic issue of originality (Luo 2024) rather than harnessing the benefits of GenAI in a way that allows a clear distinction between the work of the student and that of the technology.

Thus, despite the headlines, we do not fully understand the evidence 'on the ground'. Do the sensational headlines accurately reflect a situation where UK higher education is at serious risk from cheating enabled by GenAI, or are the concerns overblown? An estimation of risk is made by identifying both the hazard and the risk, where the hazard is the nature of a problem, and the risk is the likelihood of that problem occurring. The research described above indicates that, in the age of GenAI, unsupervised online examinations and asynchronous written coursework like essays are both significant hazards to assessment validity. However, to understand the impact of this hazard, and thus to make effective policy, we need data on the associated risk: how common is it for students in UK higher education to be assessed using these vulnerable methods? The need for the current survey arises in part from an absence of a centralised data source which could allow investigation into the assessment profiles used by different universities or disciplines. In addition, do UK students understand the university rules and regulations regarding what they can and cannot do with GenAI when completing their assessments? How many students report cheating using GenAI, and what are they doing?

A major concern when trying to develop evidence-based policies for assessment and academic integrity is the quality of that evidence base (Tight 2024). For example, when conducting research on cheating by university students, it is very common to use direct questioning surveys on convenience samples. This is when a sample of volunteer participants is asked 'have you ever cheated' or similar, with no incentive for students to participate, or tell the truth, and with limited or no efforts to understand whether the results obtained from this convenience sample are representative of the population. These methods are likely associated with under-reporting of cheating. This can be addressed by making efforts to ensure that the sample is representative of the population and by guaranteeing participant anonymity using indirect or incentivised questioning methods such as a list experiment (Newton 2023) (see methods for a full description of the list experiment methodology used here).

In one of the few published studies which have attempted to estimate the extent of cheating using ChatGPT, a study in Vietnam found that when students were asked directly whether they had cheated, 9.6% indicated that they had. When asked indirectly using a list experiment, this figure was 23.7%, suggesting that students were unwilling to directly admit to cheating (Nguyen and Goto 2024). Unfortunately, this remains one of very few studies which have attempted to quantify the percentage of students who have cheated using GenAI, perhaps in part due to the challenges in defining what cheating actually means. In the case of Nguyen and Goto, they simply asked students whether they had cheated and left it up to the students to decide on the meaning of the question. A similar approach is taken here.

Thus the research questions here are, all set in the context of the UK higher education:

1. How common is the use of assessment methods that are vulnerable to cheating with GenAI?
2. Are students using GenAI?
3. How confident are students that they understand the rules about what they can, and cannot do with GenAI on their assessments?
4. How common is cheating with GenAI?
5. How are students cheating using GenAI?

The study used two strategies to make it easier for participants to answer honestly and without fear of recrimination: a modified list experiment using the method of Nguyen and Goto (Nguyen and Goto 2024), combined with recruitment of UK university students from a paid anonymous online participant pool (Prolific.com). These participant pools can be used to obtain a large sample size quickly, with detailed demographic information that allows for the sample to be benchmarked to the population under study. Guaranteed anonymity facilitates asking sensitive questions and so they have been widely used to study issues of academic misconduct and GenAI use (e.g. Wang, Li, and Bonk 2024; Baek, Tate, and Warschauer 2024; Oliver et al. 2022; Stojanov, Hannawa, and Adam 2025).

Materials and methods

Participant pool

Participants were recruited using the online labour market 'Prolific.com' (Palan and Schitter 2018). Due to the anticipated length of the study (>5 min) two attention checks were included in this study, as detailed below in 'survey structure'. In accordance with Prolific.com policy (Prolific.com 2022), participants who failed both attention checks were not paid and their data were not included in the analysis.

Eligible participants were undergraduate university students studying in the United Kingdom, including those who had recently finished their finals, studying at year 2 or above. This last criterion was selected so that students would be more likely to have experienced summative assessment that counted towards their degree. The most up-to-date demographic profile of this group, at the time of participant recruitment, was obtained from the Higher Education Statistics Agency and was from 2021/2022 (HESA 2023).

Ethical approval

Approval was obtained from the Swansea University Medical School Research Ethics Committee, approval number 1 2024 9775 8913. All participants were informed that they provided informed consent by agreeing to continue past the Participant Information Sheet.

List experiment design

A full copy of both surveys is available as Supplementary Material S1. This was a survey based study, incorporating a modified list experiment in the style of Nguyen and Goto to address research question 4 (how common is cheating with GenAI) (Nguyen and Goto 2024). A modified list experiment incorporates both direct and indirect questioning of the sensitive question. The list experiment was designed as follows. Participants were split into two groups, a control group and an experimental group, each answering different surveys. Both surveys were identical for the content designed to answer the first three research questions. For the fourth and fifth research questions (how common is cheating with GenAI, and what does it look like), one group serves as a control group for the list experiment but is also then directly asked whether they have cheated (direct questioning for research question 4), and how (research question 5). The other group is the experimental indirect questioning group. For the list experiment itself, the control group is presented with a group of non-sensitive questions which have yes/no answers (e.g. whether they agree 'winter is my favourite season'). The experimental group is presented with the same questions, but with an additional, sensitive question added (e.g. 'have you cheated'). Both groups are asked only to answer how many of the questions they would answer 'yes' to, i.e. they do not answer each question individually. The difference between the average number of questions answered as 'yes' between the two groups allows for an estimate of the number of participants who have answered yes to the sensitive question, but in a way which transparently guarantees the anonymity of participants (Blair and Imai 2012; Nguyen and Goto 2024; Lépine, Treibich, and D'Exelle 2020).

Survey procedure

Recruitment took place in August 2024. Participants first accessed the survey by responding to the advert on Prolific.com which stated the eligibility criteria and informed the participants that the survey included attention checks and that they would not be paid if they failed the attention checks. Once the survey was accessed, a summary page restated the eligibility criteria and attention check policy. Then the Participant Information Sheet explained the purpose of the research, the risks of taking part, data protection details, and relevant contact details. There was then a confirmation page which restated the eligibility criteria for a third time and asked the participant to confirm they met the criteria. There was then an automated Captcha to screen out bots (e.g. 'select all squares that contain a bridge'), followed by automated Prolific ID confirmation. Participants were then asked a series of demographic questions to benchmark the sample against the population.

The survey then explained the vulnerable assessment methods separately (unsupervised online examinations, and asynchronous written coursework) and asked participants whether they had been assessed using those methods and, if so, to estimate approximately how much of their degree had been assessed in this way. This was followed by the first attention check question where participants were presented with a list of five fake GenAI tools and asked '*Please consider the list of AI tools listed below. Which of them have you used in the last 12 months? Check all that apply*'. Because all the tools are fake, the correct answer option was 'None of the above'. If a participant answered any differently, they failed the attention check. The next question gave participants a list of genuine GenAI tools and again asked which they had used in the last 12 months. Participants then rated their agreement with the statement *'I know the rules used by my university regarding what is allowed, and what is not allowed, when using AI tools like ChatGPT to help with assessments'* on a standard 5-point Likert scale. Following this was the second attention check, presented in the list experiment format. This check consisted of 6 statements, with the question '*Below is a list of statements. Please read each statement carefully and count how many of them are true for you. Enter that number in the box below. So for example, if three different statements are true* *for you, enter the number 3'.* Five of the statements were nonsensical false statements that could not be true, while one statement was 'I am a living human being'. Therefore the correct answer for this attention check was '1'.

List experiment procedure

The next item in the survey was the list experiment. Both groups received 6 identical statements that could be true or false. The questions were designed to be uncontroversial (e.g. 'I wear glasses or contact lenses for reading') to avoid 'design effects'.(i.e. when the presence of the sensitive question affects participant responses to the other questions (Blair and Imai 2012)). The list of statements for the experimental (indirect questioning) group contained an extra statement; '*I used AI tools to cheat on any of my university assessments in the last 12 months*'. The control (direct questioning) group was given a range of answer options to choose from, 0 to 6. For the experimental (indirect questioning) group the range was the same except that '0 or 7' was used instead of '0'. Combining these options prevents floor/ceiling effects wherein participants in the experimental group reveal their answer to the cheating question if the correct answer was 0 or 7 (Lépine, Treibich, and D'Exelle 2020). For the experimental group, this was the final survey question. The control group were then asked to reconsider the list of AI tools presented earlier and then whether they had used these tools to cheat on their academic studies in the last 12 months, with answer options of yes, no, or 'I used them, but I don't know if it was cheating'. They were then asked a free text question *'If you did use tools like ChatGPT to cheat, can you briefly describe what you did?'.* Both groups were then debriefed and the contact details of the principal investigator were restated. Participants who passed the attention checks were paid. Those who failed two attention checks were rejected (not paid) and additional participants recruited (though see 'pilot phase'). Participants were paid 9.00 GBP per hour, which was the 'good' payment level recommended by Prolific at the time of the study.

Pilot phase

An initial pilot phase collected data from 100 participants in each group. The recruitment target was set to 116 on the basis that some may fail both attention checks. The pilot highlighted two issues. The first was that, due to a technical error, the '0' option was not visible for the list experiment question. The second issue was that many participants failed the second attention check by selecting '5' as the option. The 5th question was the only true statement, suggesting that participants had misinterpreted the question and had entered the number of *the* correct statement rather than the number *of* correct statements. One participant directly messaged to apologise that they had incorrectly selected '5' on this basis but had only realised this once they have moved to the next question. These two issues were corrected for the main data collection phase. A further adjustment was that the second attention check list question was reworded to make it clearer and provide an example ('if three statements are true for you, then select 3) as well as label statements with letters next to them. These corrections meant pilot data could not be used for the main list experiment analysis but were included in the analysis for research questions 1–3. Participants who selected '5' for the second attention check were considered to have passed this check.

Phase 1. Approximately 300 participants were recruited to both the control and the experimental survey. These were recruited in batches of ~100 participants, but with each open separately to prevent duplicate submission. Once a batch of 100 had been recruited to one survey, the submissions were screened for attention check failure, and these were rejected. This survey was then closed and the other survey opened, having excluded submissions from those who had already participated. This staggered collection was designed to ensure that the data collection for both groups was carried out at approximately the same time (recruitment of each batch of 100 participants was completed in a few hours) and to facilitate random allocation of participants to both groups. These are important components of effective list experiment design (Blair and Imai 2012; Lépine, Treibich, and D'Exelle 2020).

Phase 2. The demographics of study participants recruited so far were screened against the benchmark HESA data. Most categories were within ~5% of the benchmark, apart from the age brackets, where younger/older brackets were under/overrepresented by ~ 30 percentage points. Therefore, phase 2 focused on recruiting students from the '20 and under' bracket and this was set as a screening criterion on Prolific. Unfortunately this group appeared to also be under-represented on Prolific. Another ~100 participants were eventually recruited but this took a few days and plateaued before recruitment targets were met.

Phase 3. The pool of eligible undergraduates to complete phase 2 was exhausted, but the sample size remained very modest compared to Nguyen and Goto meaning the analysis would be underpowered. Therefore, an additional 225 participants were added to each group with no restrictions on age.

Analysis (research questions 1–4)

Quantitative data are reported using descriptive statistics, apart from the percentage of students who reported cheating. This was estimated from the list experiment using the same analysis reported by Nguyen and Goto (Nguyen and Goto 2024): the mean number of 'yes' answers was calculated from each list, and then the mean from the control list was subtracted from that of the experimental list to give the proportion of respondents who answered 'yes' to the question about cheating. 34 participants in the control group answered '0' to the control questions whereas only one answered 6, therefore all participants in the experimental list who answered '0 or 7' were conservatively coded to '0' for the analysis.

Research question 5. Analysis of cheating behaviours

The survey asked participants *'If you did use tools like ChatGPT to cheat, can you briefly describe what you did?'*. 179 participants entered useable text in this section ('useable' meaning that they wrote more than 'N/A' or similar). 61 of these were students who had answered 'yes' when asked if they had cheated. The remaining students answered this question even though they had previously responded that they were 'unsure' (50) or 'no' (68) when asked if they had cheated. The full set of responses is available in Supplementary Material S2. The responses were analysed thematically using the six step protocol of Braun and Clarke (Braun and Clarke 2006). The researcher read through the data and then generated a series of 18 codes. The data were then re-read and each code was condensed into nine themes. Where a comment indicated >1 theme, then it was coded to the theme which represented the biggest component of the comment.

Results

Benchmarking of sample

RQ1 How common is the use of assessment methods that are vulnerable to cheating with GenAI??

91.1% of students (from a sample of 1484) reported that they were being assessed using 'written coursework such as essays', and that an average of 57.5% of their degree had been assessed this way in the last 12 months, averaging to 52.2% for the entire sample. 54.6% of students reported that they had been assessed using unsupervised online examinations, and that an average of 50.5% of their degree had been assessed this way during the last 12 months, averaging to 27.5% for the entire sample.

RQ2 Are students using GenAI?

81.4% of students (*N* = 1484) reported using at least one GenAI tool. ChatGPT was the most common (72.6%), followed by Google Gemini/Bard (17.8%), Co-Pilot (13.9%), Claude.AI (5.1%) and Pi.AI (1%). 18.6% of students responded, 'None of the above'.

RQ3 How confident are students that they understand the rules about what they can, and cannot, do with GenAI on their assessments?

Students (*N* = 1484) were asked to rate their agreement with the phrase *'I know the rules used by my university regarding what is allowed, and what is not allowed, when using AI tools like ChatGPT to help with assessments'*, using a 5-point Likert scale. 85.3% agreed with statement (53.6% completely, 31.7% somewhat), while 9.8% disagreed (6.4% somewhat, 3.4% completely). 4.9% neither agreed nor disagreed.

RQ4 How common is cheating with GenAI?

This was measured using both direct questioning, and indirect questioning *via* the modified list experiment (Nguyen and Goto 2024). In the direct questioning sample, 8.87% (*N* = 733) answered yes to the question *'Did you use these tools to cheat in your academic studies in the last 12 months?'*, and 9.4% answered *'I used them but I don't know if it was cheating'*. Analysing the data only from those who reported that they had used one or more of the listed GenAI tools (*N* = 590), the figures were 11.0% who reported cheating and 11.4% who were unsure. The estimate from the indirect questioning method was that 21.9% of students had answered 'yes' to the question '*I used AI tools to cheat on any of my university assessments in the last 12 months*', rising to 26.4% amongst users. Data are summarised in Table 2. Since younger students are under-represented in the sample, and the limited research on cheating in UK universities suggests that cheating is more common in younger students (e.g. Newstead, Franklyn-Stokes, and Armstead 1996), the direct questioning analysis was undertaken again but weighting the different age brackets accordingly. This returned an overall estimate of cheating percentage of 9.20%, compared to 8.87% for the unweighted average, suggesting that the imbalance of age brackets in the sample slightly underestimates the percentage of cheating in the population.

RQ5 How are students cheating using GenAI?

Responses were short; the average being 18.2 words, with a total corpus of 3264 words across 178 comments from 61/65 students who answered 'yes' to having cheated, along with 50 who were unsure and 67 who said they had not cheated. Each theme is illustrated below with 1–2 sample comments for the larger themes followed in brackets by whether that student identified their behaviour as cheating, not cheating, or unsure. Every comment is from a different participant. Comments are presented verbatim.

Plan (44 Comments)

Table 1. Demographic benchmarking of population, sample and subsamples.

			List		Cheating?		
Criterion	Popn	Sample	Control	Exp	Yes	No	Unsure
N	1099485	1484	616	640	65	599	69
Female	56.8	56.0	56.8	55.9	52.3	55.8	63.8
Male	43.0	42.0	41.2	42.2	46.2	41.9	34.8
Non-binary	0.2	2.0	1.9	1.9	1.5	2.3	1.4
20≤	51.5	23.5	23.7	25.3	26.2	21.2	26.1
21–24	33.2	33.6	33.3	33.4	32.3	32.7	40.6
25–29	6.0	17.7	17.0	18.4	13.8	17.4	15.9
30≥	9.3	25.3	26.0	22.8	27.7	28.7	17.4
White	71.0	62.4	60.1	64.5	47.7	63.1	59.4
Black	8.0	15.5	16.9	14.7	18.5	15.5	18.8
Asian	14.0	16.4	18.5	15.2	26.2	16.2	18.8
Mixed	5.0	4.4	3.6	4.2	7.7	4.0	2.9
Other	2.0	1.3	1.0	1.4	0.0	1.2	0.0
Not known		0.0	0.0	0.0	0.0	0.0	0.0
Science	47.8	58.3	58.9	57.3	46.9	59.0	65.9
Non-Science	52.2	41.7	41.1	42.7	53.1	41.0	34.1
Home	82.4	86.2	86.2	85.6	78.5	88.0	84.1
International	17.6	13.3	13.6	13.4	21.5	12.0	14.5
Unsure		0.5	0.2	0.9	0.0	0.0	1.4

Population data were from the UK Higher Education Statistics Agency 2021/2022 report, using undergraduate non-first year full time students (HESA 2023). The 'cheating' sample represents the list experiment control group, broken down according to their answers on the direct questioning re: cheating. The 'all' sample, and the 'cheating' sample both include participants from the pilot phase whose data could not be used for the list experiment (see text) but whose responses were analysed for the remaining research questions. The science/non-science distinction was made using HECOS code breakdown from HESA, with the geography code split 3:1 science:non-science since this reflects the split in the HESA data.

Table 2. Estimates of the percentage of students who have cheated using GenAI in the last 12 months.

	Direct questioning	Indirect questioning
All	8.87% (1.05)	21.92% (6.48)
GenAI users only	11.00% (1.29)	26.40% (7.23)

Direct questioning groups were asked about their use of tools like ChatGPT and were then asked '*Did you use these tools to cheat in your academic studies in the last 12 months*?' Indirect questioning groups were asked the same question *via* a list experiment (see methods). Figures are given as actual (direct questioning) or estimated (indirect questioning) percentages with the standard error (SE).

This involved getting started, identifying the structure and plan of an assignment, generating ideas or inspiration

> By technicality, it's cheating. But for me, I value my studies and the actual learning, but struggle with certain aspects of the work that I don't receive academic support for. After discussing with my lecturers and not receiving adequate adjustments, I used AI to turn my own research notes and ideas into a first draft of essays and then edited those to be fluently my own work and ideas. I'm just dyslexic and struggle with making first drafts, but besides the initial write up everything else was my own work. (Cheating)

> I'd ask it to write an essay plan. I tend to procrastinate until the day before the deadline so I have ChatGPT decide how much time I have left to dedicate to each part of the essay. I have to make sense of a topic but then I write in my own words (Not cheating)

Write (29 Comments)

This involved getting GenAI to generate assignment content in a way that indicated the output from the GenAI would be the main content submitted by the students

'Write my stuff lmao' [laughing my ass off] (Cheating)

'I put the required pdf in chatgpt and asked to write my essay' (Cheating)

Explain (26 Comments)

This including giving examples, condensing and summarising content

'I used ChatGPT to help me to understand the material, and how I should go ahead with working out the problem that was issued in the question' (Unsure).

'Asked questions when I was stucked on a particular questions, needed to be sure I was getting it right' (Cheating).

Proof (16 Comments).

This included paraphrasing, synonyms, but not rewriting sections

'*to make my writing flow better and spruce up my language a bit'* (Not cheating)

Rewrite (15 Comments)

This involved taking assignment content and giving it to GenAI to improve, beyond simple proofreading

'I never used it to cheat, but I used it for paraphrasing my essay and rephrasing statements in a different way aside how it was stated word for word by the author' (Not cheating).

Not cheat (14 Comments)

Comments which simply stated that the participant had not used GenAI to cheat

Coding (9 comments)

Participants specifically described using GenAI for coding tasks.

Other (6 Comments)

Unrelated to academic assignments (e.g. writing marketplace ads, poetry for pleasure)

Not use (4 Comments)

Comments stating that the participant had not used GenAI at all.

Discussion

The findings reported here suggest that higher education providers in the United Kingdom are exposed to a high risk of cheating using new GenAI tools. Most students were using these tools, and almost all had been assessed using methods which are vulnerable to cheating with these tools. A substantial majority of students (85.3%) were confident that they knew the rules regarding (un)acceptable use of AI in assessments. The numbers of students who reported having cheated was approximately 1-in-5. However, the behaviours they report as cheating were varied and it is unclear whether these would constitute cheating under conventional interpretations of assessment security and integrity.

A December 2024 survey of UK University students for the HEPI website found 80% of students agreed that their institution has a clear policy on GenAI use in assessments (Freeman 2025), very similar to the August 2024 data reported here, and substantially higher than the figure of 64% reported by HEPI in 2023 (Freeman, 'Provide or Punish?', 2024). This apparent high level of confidence is perhaps surprising given the speed and scale of developments in the technology and the challenges, identified here, of a lack of clarity over what is cheating when using GenAI. Very many of the behaviours identified by students in the current study could reasonably be interpreted as legitimate use of GenAI for academic means. Earlier research on plagiarism suggested that self-reported confidence does not equate to actual understanding: students significantly over-estimated their confidence and underestimated the penalties for academic misconduct (Newton 2016) and new students have a very limited understanding of the current conventions regarding referencing, plagiarism and cheating (Locquiao and Ives 2024). Given that there is limited agreement on what constitutes cheating with GenAI (Golding et al. 2024; Freeman, 'New HEPI Policy Note', 2024), it seems reasonable to hypothesise that the self-reported confidence found here is similarly misplaced, and this exacerbates the risk to UK higher education. This issue is reflected in the aforecited HEPI study which acknowledge that students appear confident that their institution has a clear policy, there recommends an urgent development of those policies so that there is a clear distinction between acceptable, encouraged uses of GenAI to support learning, and assessments where GenAI use is not permitted so that learning can be meaningfully certified in its absence (Freeman 2025).

However, just because students can cheat does not mean that they are. The estimates of self-reported cheating behaviour here range from 8.87% to 26.4% depending on the definition and the measure. These estimates are close to those identified in a methodologically similar study of undergraduates in Vietnam (Nguyen and Goto 2024) and will likely be interpreted in different ways. The number reported here does not seem to represent a 'homework apocalypse' as has been predicted (Mollick 2023), but the numbers of students affected is still very large. The very latest HESA data indicate that there are 2.94 million students studying in the United Kingdom, of whom ~1.1 million are non-first year undergraduates from the population sampled here (Higher Education Statistics Agency 2024). The lowest estimate of cheating from this sample (8.87%) would still represent ~100,000 non-first year undergraduate students cheating using GenAI. The highest estimate (extrapolating from the indirect questioning estimate, across all students in the UK) would give almost 650,000 students cheating.

One of the challenges in making evidence-based policy on assessment, cheating and academic integrity is the quality of that evidence, often based on direct-questioning surveys with small convenience samples (Tight 2024; Newton 2023). The present study used two methods aimed at increasing the rigour of the survey responses: indirect questioning and the use of an anonymous, paid participant pool (Prolific). However a higher estimation of cheating was still found by using indirect questioning, suggesting that the anonymity promised by Prolific is insufficient to reassure students that their responses to direct questions cannot be traced, even though participants on Prolific generally give honest answers where there is no motivation not to do so (Wiese 2023). These data further underscore the need for rigour when undertaking survey-based research on academic misconduct. However, one of the main advantages of the method used here, the anonymity of the participants, is also a limitation because it precludes any direct *post hoc* analysis of the data, for example of the demographic or other features of the participants who admitted to cheating. There are some methods available for indirect inference of these features from list experiments (Blair and Imai 2012; Nguyen and Goto 2024), but they require a larger sample size to be meaningful. This challenge is compounded by the modest numbers of participants admitting to cheating, even with the list experiment estimate. This can be partially addressed by using the modified method as deployed here, wherein the control group were also subject to direct questioning and thus it is straightforward to identify those features, but the low estimate of cheating found by the direct questioning means that such *post hoc* analysis of the current data set would be unreliable.

There are urgent policy questions for the UK higher education sector; how to define and deliver assessments, and associated policies, which allow a meaningful certification of the learning of the student, either with, or without, GenAI. The qualitative analysis of open-ended questions undertaken here revealed a profound disconnect between simple ideas of 'cheating' and the more complex, nuanced uses for GenAI which blur the line between using GenAI for academic support versus misconduct, mirroring results found in other countries (e.g. Baek, Tate, and Warschauer 2024; Golding et al. 2024; Nguyen and Goto 2024). This urgency is exacerbated by the continued and very rapid development of these tools. In January 2025, Chinese company DeepSeek launched a new GenAI chatbot which claimed a more efficient development process and a free, open-source availability model with a performance that rivalled ChatGPT. Deepseek is poised to be updated again (Baptista et al. 2025), contributing to an 'arms race' between GenAI companies which means that, by the time you read this, there will almost certainly have been more developments, and so the policy issues even more urgent.

In summary, asynchronous written coursework and unsupervised online examinations are commonly used in UK higher education. These methods are vulnerable to cheating with GenAI, which is largely undetectable and allows students to achieve a high grade. Most students are using GenAI, and so there are serious questions about the use of these assessment methods as valid ways to certify the learning of students. There is an urgent need for the sector to develop more appropriate summative assessments in the age of GenAI, and for appropriate policies to support the use of those assessments.

Acknowledgements

The author wishes to thank Professor Melisa Wallace for proof-reading the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Philip M. Newton http://orcid.org/0000-0002-5272-7979

References

Baek, Clare, Tamara Tate, and Mark Warschauer. 2024. "ChatGPT Seems Too Good to Be True": College Students' Use and Perceptions of Generative AI." *Computers and Education: Artificial Intelligence* 7 (December): 100294. doi:10.1016/j.caeai.2024.100294.

Baptista, Eduardo, Julie Zhu, Fanny Potkin, and Julie Zhu. 2025. 'DeepSeek Rushes to Launch New AI Model as China Goes All In'. *Reuters*, 26 February 2025, sec. Artificial Intelligence. https://www.reuters.com/technology/artificial-intelligence/deepseek-rushes-launch-new-ai-model-china-goes-all-2025-02-25/.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20 (1): 47–77. doi:10.1093/pan/mpr048.

Braun, Virginia, and Victoria Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3 (2): 77–101. doi:10.1191/1478088706qp063oa.

Covington, Natalie V., and Olivia Vruwink. 2024. "ChatGPT in Undergraduate Education: Performance of GPT-3.5 and Identification of AI-Generated Text in Introductory Neuroscience." *International Journal of Artificial Intelligence in Education*. doi:10.1007/s40593-024-00427-9.

Freeman, Josh. 2024. "New HEPI Policy Note Finds More than Half of Students Have Used Generative AI for Help on Assessments – but Only 5% Likely to Be Using AI to Cheat." https://www.hepi.ac.uk/2024/02/01/new-hepi-policy-note-finds-more-than-half-of-students-have-used-generative-ai-for-help-on-assessments-but-only-5-likely-to-be-using-ai-to-cheat/.

Freeman, Josh. 2024. "Provide or Punish? Students' Views on Generative AI in Higher Education." https://www.hepi.ac.uk/2024/02/01/provide-or-punish-students-views-on-generative-ai-in-higher-education/.

Freeman, Josh. 2025. "Student Generative AI Survey 2025." Higher Education Policy Institute. https://www.hepi.ac.uk/2025/02/26/student-generative-ai-survey-2025/.

Golding, Jonathan M., Anne Lippert, Jeffrey S. Neuschatz, Ilyssa Salomon, and Kelly Burke. 2024. "Generative AI and College Students: Use and Perceptions." *Teaching of Psychology*. doi:10.1177/00986283241280350.

Grove, Jack. 2024. "Academics Despair as ChatGPT-Written Essays Swamp Marking Season." *Times Higher Education (*THE), 2024 https://www.timeshighereducation.com/news/academics-despair-chatgpt-written-essays-swamp-marking-season.

HESA. 2023. "Who's Studying in HE? | Higher Education Statistics Agency." 2023. https://www.hesa.ac.uk/news/19-01-2023/sb265-higher-education-student-statistics/numbers.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT Is Bullshit." *Ethics and Information Technology* 26 (2): 38. doi:10.1007/s10676-024-09775-5.

Higher Education Statistics Agency. 2024. "HE Student Data | HESA." 2024. https://www.hesa.ac.uk/data-and-analysis/students.

Lépine, Aurélia, Carole Treibich, and Ben D'Exelle. 2020. "Nothing but the Truth: Consistency and Efficiency of the List Experiment Method for the Measurement of Sensitive Health Behaviours." *Social Science & Medicine (1982)* 266 (December): 113326. doi:10.1016/j.socscimed.2020.113326.

Locquiao, Jed, and Bob Ives. 2024. "Replication to First-Year Undergraduate Students' Knowledge of Academic Misconduct." *Discover Education* 3 (1): 99. doi:10.1007/s44217-024-00190-y.

Luo, Jiahui. 2024. "A Critical Review of GenAI Policies in Higher Education Assessment: A Call to Reconsider the "Originality" of Students' Work." *Assessment & Evaluation in Higher Education* 49 (5): 1–14. doi:10.1080/02602938.2024.2309963.

Mollick, Ethan. 2023. "The Homework Apocalypse." July 15, 2023. https://www.oneusefulthing.org/p/the-homework-apocalypse.

Naughton, John. 2024. "AI Cheating Is Overwhelming the Education System – but Teachers Shouldn't Despair." *The Observer*, August 24, 2024, sec. Opinion. https://www.theguardian.com/commentisfree/article/2024/aug/24/ai-cheating-chat-gpt-openai-writing-essays-school-university.

Newstead, Stephen E., Arlene Franklyn-Stokes, and Penny Armstead. 1996. "Individual Differences in Student Cheating." *Journal of Educational Psychology* 88 (2): 229–241. doi:10.1037/0022-0663.88.2.229.

Newton, Philip, and Maira Xiromeriti. 2024. "ChatGPT Performance on Multiple Choice Question Examinations in Higher Education. A Pragmatic Scoping Review." *Assessment & Evaluation in Higher Education* 49 (6): 781–798. doi:10.1080/02602938.2023.2299059.

Newton, P. M. 2016. "Academic Integrity: A Quantitative Study of Confidence and Understanding in Students at the Start of Their Higher Education." *Assessment & Evaluation in Higher Education* 41 (3): 482–497. doi:10.1080/02602938.2015.1024199.

Newton, P. M. 2023. "Design, Run, and Interpret Survey-Based Research in the Fields of Academic Integrity and Misconduct." In *Handbook of Academic Integrity*, edited by Sarah Elaine Eaton. 2nd ed., 1–18. Singapore: Springer Nature. doi:10.1007/978-981-287-079-7_182-1.

Newton, P. M., and S. Jones. 2024. "Education and Training Assessment and Artificial Intelligence. A Pragmatic Guide for Educators." *British Journal of Biomedical Science* 81 (February): 14049. doi:10.3389/bjbs.2024.14049.

Newton, P. M., Christopher J. Summers, Uzman Zaheer, Maira Xiromeriti, Jemima R. Stokes, Jaskaran Singh Bhangu, Elis G. Roome, et al. 2025. "Can ChatGPT-4o Really Pass Medical Science Exams? A Pragmatic Analysis Using Novel Questions." *Medical Science Educator* 35 (2): 721–729. doi:10.1007/s40670-025-02293-z.

Nguyen, Hung Manh, and Daisaku Goto. 2024. "Unmasking Academic Cheating Behavior in the Artificial Intelligence Era: Evidence from Vietnamese Undergraduates." *Education and Information Technologies*. doi:10.1007/s10639-024-12495-4.

Oliver, Sebastian, Ben Marder, Antonia Erz, and Jan Kietzmann. 2022. "Fitted: The Impact of Academics' Attire on Students' Evaluations and Intentions." *Assessment & Evaluation in Higher Education* 47 (3): 390–410. doi:10.1080/02602938.2021.1921105.

Palan, Stefan, and Christian Schitter. 2018. "Prolific.Ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17 (March): 22–27. doi:10.1016/j.jbef.2017.12.004.

Perkins, Mike, Jasper Roe, Darius Postma, James McGaughran, and Don Hickerson. 2024. "Detection of GPT-4 Generated Text in Higher Education: Combining Academic Judgement and Software to Identify Generative AI Tool Misuse." *Journal of Academic Ethics* 22 (1): 89–113. doi:10.1007/s10805-023-09492-6.

Prolific.com. 2022. 'Prolific's Attention and Comprehension Check Policy'. 2022. https://researcher-help.prolific.com/en/article/fb63bb.

Smerdon, David. 2024. "AI in Essay-Based Assessment: Student Adoption, Usage, and Performance." *Computers and Education: Artificial Intelligence* 7: 100288. doi:10.1016/j.caeai.2024.100288.

Sorenson, Benjamin, and Kenneth Hanson. 2024. "Identifying Generative Artificial Intelligence Chatbot Use on Multiple-Choice, General Chemistry Exams Using Rasch Analysis." *Journal of Chemical Education* 101 (8): 3216–3223. doi:10.1021/acs.jchemed.4c00165.

Stojanov, Ana, Annegret Hannawa, and Lee Adam. 2025. "Communication Following the SACCIA Framework May Weaken the Relationship Between the Dark Triad and Academic Misconduct." *Communication Reports* 38 (1): 1–12. doi:10.1080/08934215.2024.2424535.

Tight, Malcolm. 2024. "Challenging Cheating in Higher Education: A Review of Research and Practice." *Assessment & Evaluation in Higher Education* 49 (7): 911–923. doi:10.1080/02602938.2023.2300104.

Turnitin. 2024. 'AI Paraphrasing Detection: Strengthening the Integrity of Academic Writing'. July 16, 2024. https://www.turnitin.com/blog/ai-paraphrasing-detection-strengthening-the-integrity-of-academic-writing.

Wang, Chaoran, Zixi Li, and Curtis Bonk. 2024. "Understanding Self-Directed Learning in AI-Assisted Writing: A Mixed Methods Study of Postsecondary Learners." *Computers and Education: Artificial Intelligence* 6 (June): 100247. doi:10.1016/j.caeai.2024.100247.

Weber-Wulff, Debora, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. "Testing of Detection Tools for AI-Generated Text." *arXiv* 10.48550/arXiv.2306.15666.

Wiese, Juliane V. 2023. "Karma and Honest Behavior: An Experimental Study." *Journal of Behavioral and Experimental Economics* 104 (June): 102018. doi:10.1016/j.socec.2023.102018.