Discover Education

Research

Student self-reflection as a tool for managing GenAI use in large class assessment

Celeste Combrinck[1] · **Nelé Loubser**[1]

Received: 12 December 2024 / Accepted: 12 March 2025
Published online: 26 March 2025

Abstract
Written assignments for large classes pose a far more significant challenge in the age of the GenAI revolution. Suggestions such as oral exams and formative assessments are not always feasible with many students in a class. Therefore, we conducted a study in South Africa and involved 280 Honors students to explore the usefulness of Turnitin's AI detector in conjunction with student self-reflection. Using a Mixed Methods Research (MMR) approach, we analysed data generated from the Turnitin AI reports, our grading rubrics, and qualitative student self-reflection. The findings show that incorporating self-reflection into assessments supports ethical GenAI use and improves the transparency lecturers need for decision-making. A declaration form allowed the students to be upfront about using Generative Artificial Intelligence tools. We found that students who can reflect on their learning relied less on generated content. However, students with high AI detected scores (> 20%) did not adequately reflect on how the tools supported their learning and could not give credible explanations of use. We contribute to the body of knowledge by providing students and academics with examples of responsibly handling AI-detected scores in large-class settings. We present a guided self-reflection and declaration with an AI detector to support students and help lecturers make decisions when grading. We also present a decision tree that lecturers and graders can use when evaluating AI use in assessments.

Keywords Student self-reflection · Turnitin AI detection tool · Large class assessment · Mixed methods research (MMR) · Generative Artificial Intelligence (GenAI or GAI) · Higher education

1 Introduction

The release of ChatGPT, a Large Language Model (LLM) with an accompanying chatbot, introduced new and significant opportunities and challenges for assessment in Higher Education (HE) [1–3]. Research on integrating Generative AI (GenAI) tools in teaching, learning, assessment and policy is still emerging [4, 5]. However, comprehensive guidelines on assessing the originality of students' work in the context of GenAI are limited [6, 7]. Developing GenAI detectors like Turnitin's AI report [8] can assist educators in identifying AI-generated text, but interpreting these reports can be challenging. Reviews of policies on academic misconduct have also revealed several grey areas, especially regarding language and writing support [6, 9]. Universities tend to leave the responsibility of deciding when GenAI use is plagiarism with the lecturers, as found by Alqahtani and Wafula [10]. Assessing the ethical use of GenAI is a complex task, necessitating tools that can accurately differentiate between accepted and unethical uses.

✉ Celeste Combrinck, celeste.combrinck@up.ac.za; Nelé Loubser, nele.loubser@up.ac.za | [1]Science, Mathematics and Technology Education (SMTE), Faculty of Education, University of Pretoria, Pretoria, South Africa.

Turnitin's AI score detector is designed to help educators identify AI-generated content. However, as Boud and Soler [11] and Turnitin [12] point out, the effectiveness of such tools depends on the lecturers' ability to interpret the results accurately and contextually. False positives can undermine trust in these technologies, but it should be noted that Turnitin reports a false positive rate of 1% or less for documents with 20% or more AI-generated content [13, 14]. Detecting AI, knowing how to assist students in understanding the tool's fair use and developing fair grading practices is difficult, as stated by Hayes, Jandrić [15]. Embedding self-reflection into assessments is an innovative approach to support ethical AI use, as suggested in our study [16, 17]. Self-reflections encourage students to report and critically evaluate their AI use, promoting a deeper understanding of ethical considerations. Self-reflection can be a powerful tool in fostering academic integrity, as it engages students in a dialogue about their learning processes and the moral implications of their actions [18, 19]. Rubrics that include criteria for ethical AI use can further support this approach, which is our aim with the current study. Rubrics can guide students in understanding the expectations and standards for responsible AI use while providing graders with more precise guidelines on assessing GenAI use. Lecturers can help students navigate the complexities of integrating AI into their academic work by clearly articulating the criteria for ethical use. We developed an AI declaration form and a student self-reflection in response to the need for lecturer interpretation tools. The declaration form and self-reflection practice could support students' ethical and responsible integration of GenAI into assessments while providing educators with supplementary tools to evaluate the appropriate use.

1.1 Research questions

The overarching goal of the current study was to promote students' ethical use of GenAI in assessment and provide lecturers with supplementary decision-making tools in large classes. Therefore, our study was guided by the following research questions:

1. How do student reflections on Generative AI (GenAI) usage correlate with their Turnitin AI scores, and what does this reveal about their ethical engagement with AI tools?
2. How do student reflections, Turnitin AI detected scores, and the declaration guide grading decisions?

2 Background

The impact of recent GenAI advancements on HE assessment has been well established [20, 21]. The changing technological landscape necessitates exploring the latest tools to support collaboration between students and lecturers while promoting inquiry [22]. Generative AI (GenAI), which includes technologies such as GPT-4o, has demonstrated significant potential to enhance educational processes while creating concerns about the potential challenges [23]. The need for research on the educational implications of using GenAI to learn and to replace learning, the impact on human development and the need for strategies and guidance becomes increasingly evident, as reported by Bukar, Sayeed [24]. The current study focuses on assessment tools in the era of GenAI for large classes, which are incredibly challenging in educational massification [25]. Current suggestions for dealing with GenAI involve more personalised assessments and individualised feedback [26], such as oral exams and increased formative assessment [27]. However, when class sizes include over 500 students, this becomes too time-consuming and expensive. Therefore, self-reflection and declaration of use is presented as alternatives to be included in the assessment process. Technology could play a role in dealing with massification by employing GenAI to provide feedback, but that discussion is beyond the scope of the current paper.

2.1 AI and assessment in higher education

GenAI can facilitate personalised learning experiences [28], automate administrative and assessment tasks [29], and support creative and critical thinking for students in the class and outside [30, 31]. However, the benefits are countered by concerns about academic dishonesty and the pedagogical implications of using GenAI for learning and assessment [32]. The primary challenge is ensuring that students use GenAI to enhance learning and to support lecturers dealing with GenAI during assessments [33]. Detecting AI-generated content copied from a primary LLM is relatively easy [34]. When spinning has been applied by putting AI content through "humanising" AI, such as Quilbot, the detection becomes more difficult. Turnitin created an AI detection tool we used in our study [8]. AI detection tools are expensive, and universities need to learn how to use them well to make the cost worthwhile and, at the same time, support student learning in this new technological area [35]. Studies by Baron [34] and Halaweh and El Refae [36] found that AI detection is likely to become more complex and that running text through AI spinners multiple times can result in 0% being detected. Future assessments will have to occur with GenAI, and the AI will be part of understanding how much learning has occurred.

Additionally, we suspect that the role of written assessment will have to be re-evaluated. AI detection tools adapt each time the users find ways to beat the detector, resulting in a continuous cycle of adapting and counter-adapting. Luo [6] notes that distinguishing between students' intellectual contributions and AI-generated work is challenging. The current study offers tools that can be combined with human insight and AI detectors. Assessment practices will have to evolve to accommodate writing and learning in conjunction with AI [37, 38], and we suspect that this may change the assessment landscape considerably.

2.2 Distinguishing between plagiarism, originality and authenticity

Many higher education policies have noted plagiarism and concerns about academic integrity when AI is used [7, 37]. Luo [6] notes that several HE policies overlook students' efforts and contributions when working with GenAI, while Ardito [39] and Fredheim and Pamment [40] note that LLMs allow users to conduct extensive, iterative conversations with complex prompts. Therefore, we concur with Luo [6] that plagiarism, originality and authenticity should be re-examined in the context of AI. The Turnitin [41] plagiarism spectrum identifies ten types of plagiarism based on research and findings from secondary and higher education institutions. On this spectrum, submitting text from a source verbatim, using significant portions of work from a single source, retaining the essential content of a source, and changing keywords or phrases are considered some of the most severe types of plagiarism. Similarly, if a student deliberately uses GenAI to complete an assessment, it may be considered plagiarism and academic dishonesty [7].

In their review of HE policies on GenAI, Moorhouse, Yeo [7] reported that highly ranked universities view copying and pasting from GenAI, spinning or humanising AI-generated text to avoid detection and not documenting GenAI use in assessment as forms of plagiarism. Many higher education institutions require students to acknowledge and report their GenAI use in assessment [42]. Ardito [39] and Luo [6] agree that reporting GenAI use might need to be more practical, as work sections have been co-created with AI. Furthermore, looking at guidelines on how instructors should respond when they suspect GenAI plagiarism, it seems that instructors are given the discretion to decide if a violation has occurred [7]. However, it is unclear what guidelines instructors should use to determine whether students are guilty of plagiarism with AI [43]. In light of the conversational and collaborative abilities of LLMs and AI chatbots, it is challenging to distinguish between the originality and authenticity of a student's work [44]. Luo [6] notes that there needs to be more understanding of what constitutes originality of a student's work in HE policies. The author highlights that originality is often assessed based on a declaration of authorship and originality score [6]. In this paper, we define authenticity as the student has completed the work on their own (it is their work mainly with sources only being used for scaffolding understanding) and as originality as reflecting a student's understanding of the sources, materials and content in question.

The definitions adopted in the current study are shown below:

- Authenticity—the text was written by the author in their own words, and ideas from other sources have appropriate citations.
- Originality—the ideas presented in the work are new, novel and original to the author.
- Humanising AI**—**Work written by generative artificial intelligence is rephrased in multiple ways by GenAI to make it sound more "human" and natural.

2.3 The need for ethical use guidelines and educational policy

Integrating ethical considerations into AI usage is crucial for maintaining the integrity of academic work [45]. Floridi and Cowls [46] argue that the ethical use of AI in education should be guided by principles such as transparency, accountability, and fairness. These principles can help develop policies that delineate acceptable uses of GenAI, prevent misuse, and allow students to use the tool correctly [47]. Recent studies emphasise the role of institutional guidelines in promoting ethical AI use; for example, Chan [48] offers a comprehensive framework that includes the dimensions of pedagogical, governance, and operational aspects to guide educators in judging the use of AI in student work. Yusuf, Pervin [49] found that clear, well-communicated policies significantly reduce instances of academic dishonesty related to AI. They advocate for including AI use guidelines in academic integrity policies, which can provide students with a framework for ethical AI use.

On the other hand, McDonald, Johri [50] point out that extensive revisions of pedagogical policies can burden academic staff and that such changes require careful thought. The responsible use of Generative AI (GenAI) in higher education requires a balanced approach that fosters innovation while maintaining academic integrity. Recent studies underscore the importance of clear policies, practical assessment tools, and reflective practices in achieving this balance [51–53].

3 Methods

Mixed Methods Research (MMR) is a relevant methodology for this study as it explores multiple dimensions inherent in the motivation and the result of using AI in assessment practices [54–56]. Our study investigated the usefulness of a GenAI declaration and reflection in combination with the Turnitin AI score in honours assignments. The current study investigated how reliable and valid the reflections are for gauging the fair and responsible use of Generative Artificial Intelligence. Students had to agree to our declaration, and when they used GenAI, they had to reflect on how learning took place and provide their prompts and outputs. The reflection was graded as part of their assignments, creating qualitative data. The AI score provided by the Turnitin platform was quantitively analysed.

3.1 Sample

Two hundred eighty honours students in the Faculty of Education at the University of Pretoria in South Africa participated in our study. The students are enrolled in distance education modules. We drew a stratified, random sample of participants' assignments for the qualitative analysis, sampling 25 texts (assignments). Stratification for the qualitative sampling was done on the variable *Turnitin AI scores* (coded as above 20% and below 20%), and we aimed to have half the sample comprised of each. The Turnitin report highlights where AI-generated texts could be found in the writing and the percentage detected.

3.2 Point of departure

The current paper is theoretical, with data to support our claims. We designed the study and the analysis to be inductive and contribute to the development of theory within the arena of AI in assessment.

3.3 Instruments

Our data collection utilised three instruments: the student declaration, the self-reflection item and the Turnitin AI score.

3.4 Student declaration

Table 1 shows the declaration participants must complete before submitting the assessment. The declaration was presented as a form they had to click through before the assignment submission link would be available.

Our university made guidelines available to students outlining the proposed and acceptable use of GenAI. Furthermore, only lecturers and graders can see the Turnitin AI score, and we wanted participants to be fully aware. The form also explicitly indicates to participants the expectations by agreeing.

3.4.1 Prompt and reflection items

Table 2 shows the two items included and graded in the student assignment, along with their rubric and examples of poor, good and excellent responses.

For the reflection section, participants had to write a brief paragraph explaining how they used AI in their assignment, ideally highlighting how learning was supported and enhanced. Our rubric required the allocation of 3 out of 3 marks only if the reflection showed personal growth and rich references to one's learning, and graders used the examples as a guideline to make judgements.

Table 1 Generative artificial intelligence declaration of use

Statement	Response options
1. Have you used any artificial intelligence/chatbots in your assignment? (Please note this is allowed if you rework and reword as required)	1. I have used AI/Chatbots or will be using them 2. I have not used any artificial intelligence
2. I avoided copying and pasting text directly from Generative AI (GenAI)	1. Agree 0. Disagree
3. If I used GenAI, I only did so to scaffold my learning	1. Agree 0. Disagree
4. If/When using Generative artificial intelligence, I made sure it aligns with university policies	1. Agree 0. Disagree
5. If I used any GenAI tools, I declared and explained it in the assignment	1. Agree 0. Disagree
6. I acknowledge that graders can see my AI score, even though it won't show on my side	1. Agree 0. Disagree

3.5 Administration and data collection

In the declaration form, participants gave consent for their assessment answers to be used for research purposes. Participation was voluntary; 32% of students declined consent, and their data was removed from the analysis. The data analysis and interpretation are illustrated in Table 3.

3.6 Data analysis

A quantitative data analysis of the markings done on the student reflections was conducted. The statistical data analysis was used to gauge the relationship between the Turnitin AI Detection Score and grader perceptions of participants' fair use (graded item). The Gamma Correlation Coefficient was deemed appropriate for calculating the relationship between AI scores and reflection marks due to the graded reflection's categorical nature and the AI score's limited variability and non-normality [57, 58]. From a mixed-methods perspective, the Gamma coefficient complements the qualitative analysis by quantifying patterns observed in students' reflective responses. Integrating statistical correlation measures with qualitative coding allowed a more nuanced interpretation of how the reflection correlates with the AI-detected score. This approach strengthens the study's findings by ensuring robust quantitative associations support insights drawn from qualitative reflections. The Statistical Package for the Social Sciences (SPSS) was used for the quantitative analysis [59], and Winsteps software for the reliability and validity calculations [60]. To interpret the Gamma correlation, we consulted Cohen's recommendations and used these guidelines [61, 62]:

< 0.10 = Very weak or no association
± 0.10 to ± 0.30 = Weak association
± 0.30 to ± 0.50 = Moderate association
± 0.50 to ± 1.00 = Strong or very strong association

We conducted a reflexive thematic (TA) content analysis to evaluate participants' understanding of GenAI in their textual reflections [63, 64]. During our qualitative analysis, we focused on assigning codes to the reflections and prompts students used regarding how ethical the use was, the level of thinking in prompt design and exploring participants' reasons for relying on GenAI to complete the assessment. The coding and themes were done in Excel, our preferred way to handle our QUAL data. The quantitative and qualitative data were integrated so that findings aligned with the concurrent mixed methods approach. The first step of our MMR analysis was to prepare the QUAL and QUANT analyses separately [65]. Next, we compared and contrasted our findings [66], synthesised the data, addressed the divergent findings, and concluded the analysis by creating a convergence table [67].

Table 2 Generative artificial intelligence use declaration

Item	Rubric	Examples
Q.1 Prompt used: Please paste your AI prompts and the outputs received and indicate for which aspect(s) of the assignment a specific prompt was used	Award 0 if the prompts are absent Award 1 mark if the prompts were included	Poor prompt: What are research question? Good prompt: Factors to consider in developing a qualitative research question Excellent prompt: You are a qualitative researcher in an education Faculty in South Africa. What are some effective qualitative research questions focused on bullying among young children? Please provide me with examples and explanations as to why these questions could support a case study design
Q.2 Reflection: Write a narrative wherein you reflect on how your research methodology skills have evolved since the beginning of the module. How have your experiences during the module contributed (or not) to your personal growth? If you used Generative Artificial Intelligence (GenAI) for your assignment, use this section to discuss how you used the tool(s) and ensured that learning took place while using GenAI	Award 1 mark if the reflection is basic in nature and lacks a personal reflection Award 2 marks if the reflection shows some insight and personal growth (this could be cognitive, social, emotional or behavioural) Award 3 marks if the student connects the content to their own experiences. A type of growth (cognitive, socio-emotional, spiritual) is discussed personally. There is evidence of personal growth and identifying areas for improvement in the reflection. The student provides rich descriptions of their own experiences	Poor reflection: I used ChatGPT to understand, rephrase and repurpose what I learned Good reflection but lacks depth: I used Generative AI (GenAI) to help me clarify how I would apply phenomenology in my own study, and how this is related to other ways to do research such as qualitative design. I now have a better grasp of the way forward and I used the bot to help clarify the statistical interpretations for the quantitative sections of the assignments, which are honestly quite difficult Excellent reflection: At the beginning of the module, my understanding of research methodologies was superficial, primarily focused on the broad distinctions between qualitative and quantitative approaches because that is all I learned in undergrad courses. Now, I have developed a more nuanced understanding of when and how to apply different research methodologies. For instance, I have gained a deeper appreciation for the intricacies of mixed-methods research, which I previously viewed as merely a combination of qualitative and quantitative techniques. I now understand the importance of methodological alignment and how integrating both approaches can yield richer, more comprehensive insights that neither method alone could provide. I used to view myself as a quantitative person, I like numbers. But now I think the world is more complex. I used GenAI to get feedback on my research design to see how my work aligns with the assignment guidelines and rubric. The responses helped me to see that my design spoke to something other than the research questions and that there were areas of my understanding with gaps. The bot then suggested that I use a different design. I thought about it and decided to modify what ChatGPT suggested so that the design aligns more with my social values of inclusion and participant-focused research but at the same time that, I addressed the gaps pointed out to me by the AI

Table 3 Instruments, data type, sample size and analysis

Source (instrument)	Data type	N	Analysis
AI Declaration	Declaration score out of 10	280	Checked the AI declaration for admittance of use (score used to filter data)
Assignment and the Turnitin AI detector	Turnitin AI detected score out of 100%	280	Students completed the assignment as part of their coursework, and Turnitin provided an AI-detected score and report
Reflection question, rubric and memorandum	Score out of 4 for reflection, marked as part of the assignment	Quant: 280 Qual: 25	Quantitative analysis of the mark assigned which was correlated with the Turnitin AI score Qualitative analysis of the reflections to find themes related to the use of GenAI

3.7 Methodological integrity (norms)

The trustworthiness of the qualitative data analysis was enabled by using two coders; both authors coded separately and then reached a consensus to produce the final themes and write-ups. The reliability and validity of the quantitative instruments were checked by looking at the overall reliability coefficient of the assignment (all items), which was 0.863, indicating that the underlying construct of the research methodology was consistently measured, including the reflection item. We also ran a Rasch Partial Credit Model [68] and found that the reflection item did not misfit the measurement model (Outfit MNSQ = 1.42), which we consider further evidence that the item aligns with the construct being measured (construct validity) [69]. We checked the normality of the distribution and other assumptions of correlations and chose the Gamma as the best suited for the analysis. To mix the results, we used triangulation to help us ensure that the interpretation of the combined QUAL and QUANT was credible and supported by our data.

3.8 Ethical considerations

Permission was requested from the Dean of our Education Faculty. He agreed to allow the study if proper ethical procedures were followed. Further permission was granted by the Faculty of Education Ethics Committee and our University's Survey Committee. Participants were asked for consent to use their data for the research purposes described in this paper. On the consent form, we clearly explained that participation in the study is voluntary and that if they prefer not to share their assessments for research purposes, we will not use the answers and results. There were no adverse consequences if a student declined to participate in the study.

3.9 AI use declaration

During this research and writing up our findings, we used ChatGPT 4.o [70] to ask for suggested headings for the literature review based on our research questions, talked to ChatGPT about our findings and what it could mean, and solicited advice from the bot on how to present our findings best [71]. We also used ChatGPT 4.o with Canva to receive feedback on our writing and suggestions on improving this article. We were careful not to use anything from the Chad (ChatGPT) verbatim, and when we disagreed with it, we chose our own insight as superior.

4 Results

4.1 Qualitative findings

Table 4 summarises the GenAI use in participants' assignments, showing the split between those with detected and undetected AI scores.

AI scores were deemed non-problematic when the percentage was between 0 and 20%. This range was guided by Turnitin's assertion that there is a higher possibility of false positives below this range [14]. While there are guidelines for acceptable similarity percentages, we could not find similar guidelines for GenAI-detected reports. Therefore, assignments that had AI scores above 20% required further investigation. Participant reflections were used to assess whether their reported and detected use was problematic and to what degree. Six of the 17 participants with high AI scores used GenAI to create their reflections. Next, participants' reflections were coded and analysed to understand better how the participants used GenAI to support them in completing their assignments. The themes derived from our codes are presented in the next section.

4.1.1 Theme 1: AI as a tool for creativity and active engagement

Five of the reflections had aspects of creativity, application and engagement. Participants said they used the GenAI as a tutor and for enhanced learning. One participant also used the GenAI tool to get formative feedback on their assignment by prompting: *Does my answer cover all expectations for the question? Does my answer make sense?*

Table 4 Breakdown of GenAI scores included in qualitative analysis

Participant no.	GenAI score (%)	AI generated content (%)	GenAI paraphrasing (%)	AI use detected in reflection
Participant 1	0	0	0	None
Participant 2	0	0	0	None
Participant 3	0	0	0	None
Participant 4	0	0	0	None
Participant 5	0	0	0	None
Participant 6	0	0	0	None
Participant 7	0	0	0	None
Participant 8	0	0	0	None
Participant 9	0	0	0	None
Participant 10	< 20	< 20	< 20	None
Participant 11	< 20	< 20	< 20	None
Participant 13	21	21	0	None
Participant 15	26	0	26	None
Participant 16	37	0	37	None
Participant 17	41	0	41	None
Participant 20	70	70	0	None
Participant 21	70	70	0	None
Participant 24	85	85	0	None
Participant 14	22	22	0	Yes, Generated
Participant 12	20	0	20	Yes, Generated
Participant 19	67	50	17	Yes, Generated
Participant 22	72	72	0	Yes, Generated
Participant 23	78	78	0	Yes, Generated
Participant 18	57	20	37	Yes, Rephrased
Participant 25	100	16	84	Yes, Rephrased

(Participant 7). These prompts align with our institutional guidelines for acceptable use, our teaching goal for students to utilise AI as a cognitive aid, and the general definition of mind tools [72]. Generative AI (GenAI) can be an inspiration source, helping students explore and expand their ideas, perspectives and approaches. One participant highlighted that the AI helped them improve their writing skills by providing critical feedback and said that:

> It helped me to customise my work in collaboration with other learners. It also improved my learner engagement by allowing me to be out there and academically learn how to receive critiques. It further developed my writing skills, and I write like a pro (Participant 10).

Another student said it assisted them with more profound and independent work by helping them break down their work into more understandable sections. Learning is supported and developed when students engage dynamically with AI. The participants were aware and able to use the tool effectively, as shown by Participant 1, who said in their reflection that:

> To ensure I was learning, I carefully checked the AI's outputs against academic sources. I used AI suggestions as a starting point for deeper exploration, which kept me actively involved in the research process.

Overall, the results support the idea that integrating GenAI into written assignments improves students' academic writing and can reduce the time spent retrieving information [73]. Implications of the quick and convenient answers students can receive from AI tools may lead to diminished critical thinking abilities. Reduced critical thinking would happen if students do not independently research the topic and cannot thoroughly evaluate the AI outputs [73].

4.1.2 Theme 2: AI for promoting skills and supporting knowledge construction

In seven reflections, participants reported using AI to supply examples, assist with understanding, and help them apply what they had learned. For example, the assignment required students to interpret statistical outputs, which we know from experience can be challenging. Participants with low AI scores said they used LLMs to help them understand concepts, such as statistical significance. They also asked the bot to check their interpretations. For example, participant 13 said in their reflection that:

> For the question about quantitative data interpretation, due to not understanding the value $p < 0.05$, I used AI to interpret. I interpreted the inferential statistics so that it draws the reader's attention to the significant findings with values.

The participant who could avoid misusing AI asked the bots for information and inputs but drew their own conclusions and reworked materials accordingly. For example, participant 2 said in their reflection that:

> I used GenAI as a study tool to remind me of the different types of research methodologies, why they are used and what influences them. This gave me the framework to answer questions but since I didn't input specific questions or details about the given scenarios, it forced me to apply the framework myself and put in effort to understand.

This theme reflects the potential of GenAI to foster development and encourage self-improvement when it supports students in growing their insights. Conversely, without the necessary prior knowledge to identify incorrect information, students risk becoming overly reliant on AI tools that have been found to give incorrect answers to content questions [74, 75].

4.1.3 Theme 3: AI for donkey work (mindless uses)

Three of the reflections drew attention to AI's uses for formatting, rephrasing and technical assistance. While participants were strictly prohibited from using GenAI for language editing, the assessment guidelines were not explicit about allowances for technical editing. Participants who avoided high AI scores could employ the LLM machine for mundane tasks like formatting, as Participant 6 notes:

> I used GenAI (ChatGPT) for various reasons, including assignment layout, literature review, generating of information, summarising of text extracted from a textbook and for reference. I also used Microsoft Office and Grammarly to check spelling and grammatical errors. I used the information generated by AI to develop ideas and could relate my responses according to the assignment guidelines. I avoided copying and pasting information generated by AI as I know the consequences of that conduct are dire. I have also improved my reading and creative writing skills following the multiple sources I consulted for data on this assessment.

While not as exciting as knowledge construction or creativity, the use of GenAI for mundane tasks is encouraged in our modules as it frees up time for thinking and engagement. These results support previous studies that have outlined how GenAI tools can make the research process more efficient, improve academic writing, structure answers and improve students' self-confidence in large classes [76–78]. Students should be encouraged to continue engaging deeply with learning materials to develop their own ideas and thought processes [73].

4.1.4 Theme 4: lack of reflection, denial and parroting

The reflections of participants with high AI scores were typically absent. Participants with AI scores above 20% also tended to deny using generative artificial intelligence or parroted what was said during lectures in their reflections. Some participants declared that they did not use GenAI in their assignments or did not respond to the reflection question on GenAI usage, but they had Turnitin AI scores above 20%. Generated content was detected for three assignments out of the randomly sampled 25, and the score was between 70 and 85%. GenAI paraphrasing was detected for another remaining three, with scores ranging from 26 to 41%. Since paraphrasing can include spinning, we speculate that the participants did not view this as unethical use. Only three of the 17 participants with high AI scores included their prompts. The prompts used sections from the assignment in verbatim. Below, we show two of the three prompts:

Write a detailed overview of the key challenges faced by teachers in South Africa, including factors such as workload, resources, and working conditions (Participant 18).
What are some effective qualitative research questions focused on bullying among young children? (Participant 22).

Although these prompts align with our institutional guidelines, how outputs were directly taken from GenAI platforms violates the rules. Concerns about the overreliance on GenAI to complete academic tasks include reduced decision-making, critical thinking and analytical thinking abilities [73]. Without the proper evaluation of GenAI outputs, students are at risk of LLMs being found to present misleading and incorrect information as factual. Li and Little [79] and Albus and Seufert [74] point out that students with decreased content knowledge are particularly at risk of becoming over-reliant on AI as they cannot distinguish between facts and AI hallucination.

4.2 Quantitative findings

First, we show the descriptive statistics in this section and the inferential correlations we calculated. Table 5 summarises the total sample obtained (those who agreed to be included in the study) regarding their AI scores.

Our achieved sample contained 51 participants (18%) who had AI-detected scores above 20%. We calculated the sampling power for the total sample in g*power [80, 81]. We concluded that a sample size of 280 achieved a power of 99.6%, far beyond the usual requirements of 80%. The Turnitin AI report shows a breakdown of AI-generated and AI-paraphrased scores, which make up the overall AI score. We show this split along with the descriptive statistics in Table 6.

Of the 51 participants with AI scores above 20%, the average AI percentage was 46% (almost half of their content was AI-sourced). There is a split between AI-generated content, where the average percentage was 24% and the AI paraphrased score, where the average was 22%. The AI scores could go as high as 100%.

Table 5 Sample displayed by AI score above or below 20%

AI score		Frequency	Percent	Valid percent
Valid	Below 20%	211	75.4	80.5
	Above 20%	51	18.2	19.5
	Total	262	93.6	100.0
Missing	System	18	6.4	
Total		280	100.0	

Table 6 AI scores breakdown into generated and paraphrased with descriptive statistics

	AI score	AI-generated score	AI paraphrased score
Valid N—AI score present	51	51	51
Mean	46%	24%	22%
Median	41%	17%	21%
Std. deviation	19%	26%	22%
Range	80%	85%	84%
Minimum	20%	0%	0%
Maximum	100%	85%	84%

Table 7 Correlation between reflection grades and AI scores

		Value	Asymptotic standard error	Approximate T[b]	Approximate significance
Ordinal by ordinal	Gamma	– 0.364	0.226	– 1.554	0.120
N of valid cases		27			

[a]Not assuming the null hypothesis

[b]Using the asymptotic standard error assuming the null hypothesis

Table 7 shows the Gamma correlation coefficient between the reflection grade (out of 3) and the AI scores (between 20 and 100).

The correlation coefficient shows a negative, moderate relationship. Higher AI scores are associated with a lower grade on the reflection. The correlation coefficient is not statistically significant ($p = 0.120$), which may be due to the non-parametric nature of the data. However, the detection of a negative, moderate relationship ($r = -0.364$) shows there is an association.

4.3 Concurrent MMR findings

Table 8 shows our comprehensive findings from combining the QUANT and QUAL data in the concurrent analysis.

We first analysed the quantitative data, including AI detection scores and reflection grades, establishing statistical relationships using the Gamma Correlation Coefficient. Separately, a reflexive thematic analysis was conducted on the self-reflections, categorising students' responses into themes based on their depth of engagement with AI tools. Thereafter, we conducted a triangulated synthesis by comparing individual students' AI scores with their corresponding qualitative reflections. We used a side-by-side joint display analysis to systematically integrate these findings, visually aligning qualitative themes with quantitative scores (see Table 8). The concurrent analysis highlights that while AI detection tools provide useful indicators, their effectiveness is enhanced when complemented with self-reflective assessments, offering a richer understanding of students' engagement with GenAI.

Our quantitative and qualitative analyses found that the self-reflections were associated with improved GenAI ethical and pedagogical use. However, here we acknowledge that students who are already equipped with the knowledge and skills to achieve well academically may also know how to use such tools responsibly and can reflect on their own learning. Those who struggle with the module may need more introspection skills and fall back on the LLMs to substitute for their lower academic abilities. Grading the self-reflections provides additional evidence of student intent and awareness of using Generative AI (GenAI), but it does not necessarily provide insights to make decisions beyond what the AI score already indicated.

5 Discussion

In this section, we answer questions about our results and build an argument for the usefulness of AI detectors and self-reflection when dealing with GenAI in student assessments. We also propose a decision tree to help lecturers and markers make fair and consequent decisions about students' GenAI use in written assessments.

5.1 How do the student self-reflections add value to the assessment and AI judgement process?

The findings show a moderate relationship between the quality of the student self-reflection process and the AI score detected. Our findings suggest that students with better academic performance and higher AI literacy require less help from LLMS, are more able to reflect metacognitively and are better prepared to write good prompts for responsible use. For the current study, we defined high academic performance and skills as mastery of theories and frameworks, engagement and contribution to discourse, academic writing skills and application of knowledge to practice. The importance of crafting and refining prompts for meaningful outputs has been well established [82]. Using GenAI in assessment requires students to design precise prompts to interact with the tools effectively and critically evaluate and refine the outputs [74, 82]. Therefore, students who struggle scholastically may find it difficult to reflect on the learning process. As a result, they seem prone to over-reliance on AI tools to complete both the assignment tasks and the reflection. This finding aligns with the work of Albus and Seufert, who note that a student's prior knowledge plays a key part in their success in working with AI tools. Our findings from the reflection confirm their notion that students with lower levels of prior knowledge may need more learning assistance but struggle to use AI tools effectively. The student self-reflections are a valuable indicator of what the higher AI-detected scores mean: Students with academic challenges need support for understanding content and broader insights into their own learning. At this point in their learning, they may not yet be able to self-reflect. The self-reflections also helped us interpret the AI-detected scores as they gave insight into what students thought when they used the GenAI. These findings further highlight the importance of teaching students to design prompts rather than input the assignment instructions [83]. Students who denied using LLMs for assistance could be flagged and asked for

Table 8 Concurrent mixed methods research findings

No.	Source	Findings
First finding	Finding	More comprehensive self-reflections tend to be associated with lower AI scores and more responsible use of the tool(s)
	Quantitative result	Moderate negative correlation (– 0.364) between AI scores & reflection grades
	Qualitative finding	Theme 1: Creativity and Active Engagement Theme 2: Skill Development and Knowledge Construction Theme 3: Technical Support ('Donkey Work')
	Integrated interpretation	Students with AI scores below 20% wrote self-reflection, which indicated responsible use of GenAI and using the tools as a scaffold to enlarge and improve their learning. The higher the reflection score, the lower the accompanying AI score detected. These students also used the AI for technical editing and other less cognitively intensive tasks
Second finding	Finding	Students with AI-generated content scores above 20% exhibited a paradoxical trend: they were more likely to deny using AI while also engaging in higher rates of problematic practices, such as direct copying and pasting from artificially generated text. These finding underscores potential discrepancies in self-reporting and highlights challenges in AI literacy among students
	Quantitative result	Moderate negative correlation (– 0.364) between AI scores & reflection grades
	Qualitative finding	Theme 4: Lack of Reflection, denial and parroting
	Integrated interpretation	Students who need help to use AI responsibly can also not reflect comprehensively on their learning trajectory. They tend to deny using AI at all (until confronted with the Turnitin AI report) and use GenAI to write their AI reflections, but they may require more support and further interventions

further explanation. We here note that once presented with the evidence of the Turnitin AI report, most students admitted using the LLMs to generate content beyond what was allowable. These instances further highlight the importance of promoting AI literacy in higher education institutions, including evaluating AI-generated content, ethical consideration of using AI in assessments and AI policies [74, 84].

5.2 How does the AI detector fit into the decision-making process?

Students will continue looking for new ways to fool the AI detectors, which is possible with repeated spinning [85]. One study found that Turnitin's AI detectors showed the most significant drop in accuracy compared to other models when applying adversarial techniques to deceive the detector [85, 86]. While Turnitin and other AI detectors continue to adapt to attacks, there are also several online networks dedicated to exposing the vulnerabilities of the detectors [86]. The Turnitin AI detector helped us identify students who used AI more extensively than allowable in our module (18% detected as having more than 20% generated text), and as students did not deny this when presented with the evidence, we concluded that the scores were accurate and helped to quality assure the assessment process. To ensure that assessments remain fair and inclusive, we propose combining Turnitin AI detection reports with additional aids like student reflections to guide decisions regarding suspected misuse of GenAI.

5.3 To what extent were student declarations helpful in the grading process?

The AI declaration helped when high AI scores were found to see if students acknowledged that they would use generative artificial intelligence in their assignments. Those who denied using GenAI and did not declare they intended to use the machines did not comply with the assignment and student guidelines for GenAI use. While it is not intended to be punitive with this declaration, several factors may have contributed to non-compliance. A study by Gonsalves [87] found that students experience fear regarding grade penalties and accusations of academic misconduct, and they feel intimidated by AI declarations. We suspect confusion about what constitutes AI use might have impacted several student declarations and self-reflections in this study, especially where paraphrased content was detected. Some students may have used paraphrasing tools like Quilbot and did not declare this. Another possibility is that students copied their work into ChatGPT to paraphrase and did not consider this AI use. Further research is needed into the reasons for non-compliance with declaration requests. Therefore, it remains important that clear expectations regarding integrating GenAI into teaching, learning, and assessment are communicated to students at the beginning of the module. Lecturers and assessors are encouraged to be transparent with expectations and how the student use of GenAI is evaluated, as well as how lecturers use GenAI to ensure that student–lecturer trust remains intact [88].

5.4 How are our findings related to the research questions in the current paper?

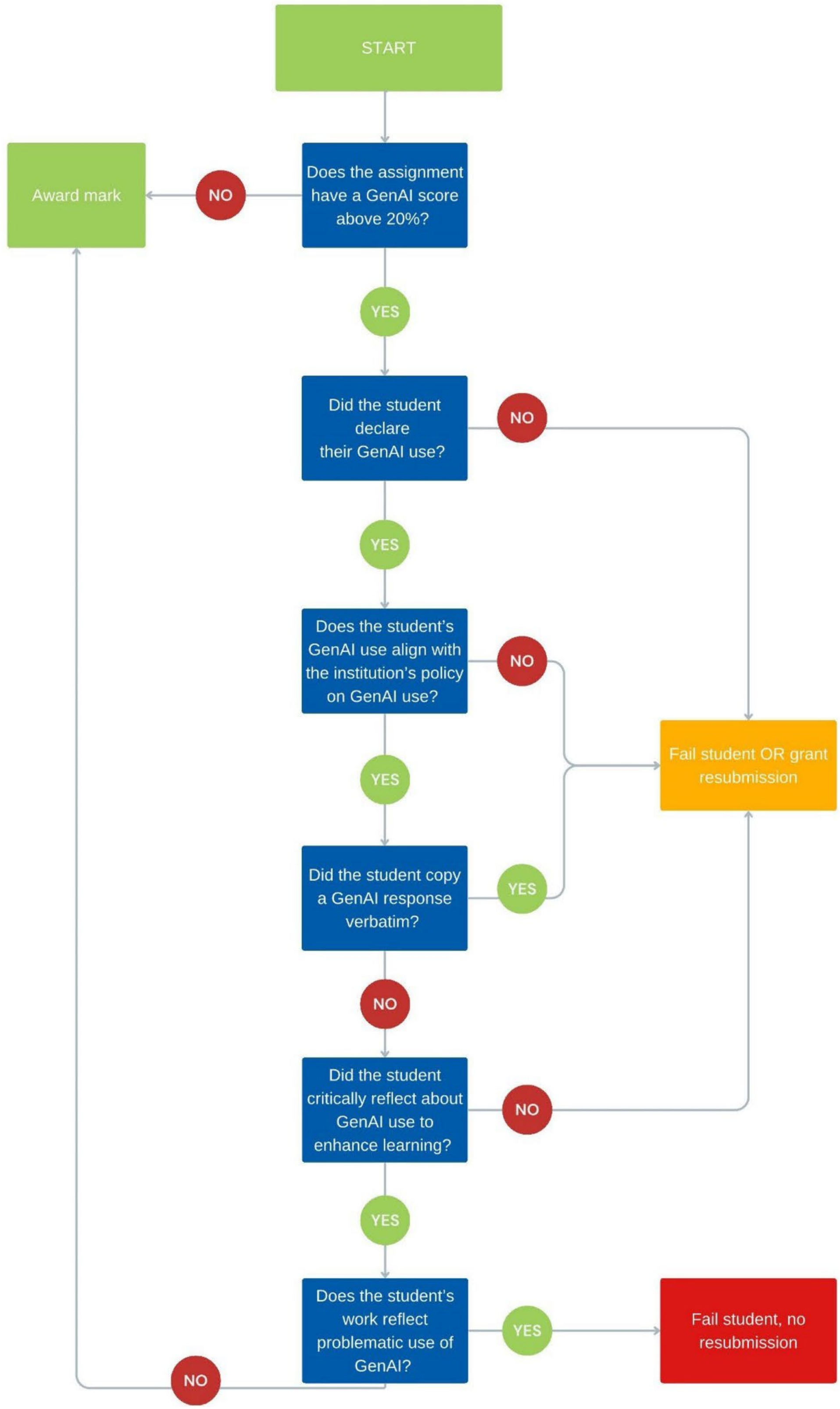
Our first research question was about the correlation between student self-reflections and the Turnitin AI detection scores. Here, we found what we expected: a moderate negative relationship. The better students were at reflecting on their own learning, the lower their corresponding AI score was to a moderate degree. The self-reflection item can be included for students and graders as an additional indicator of metacognition and the ability to use AI ethically. The second research question asked about the combination which could be used to judge students' ethical use of AI and transparency in assessments. Here, we found that the AI declaration questions, the self-reflections, and the Turnitin AI scores used in combination worked best and helped us decide if a student had failed the assessment due to misuse of GenAI. We present these tools for use by other researchers and academics to be adjusted to your specific purposes and context.

5.5 Contribution of the study

Based on the findings of this study, we designed the decision tree captured in Fig. 1 to assess GenAI use in written assessments submitted in Turnitin. The purpose of the decision tree is primarily to train our graders to assign marks fairly and only to fail students where Generative AI (GenAI) is problematic [35]. The decision tree shows how graders could easily use the results from student self-reflection, Turnitin AI score, and student-declared use of GenAI to make fair and accurate decisions.

The tree starts with reviewing the detected AI score report from Turnitin. While our institution has guidelines on the allowable threshold for similarity, we did not have guidelines for acceptable Generative AI (GenAI) scores at the time

Fig. 1 Decision tree for assessing GenAI use in Turnitin



of this publication. Next, the student declaration is reviewed to see if they reported their use. If the student did not use GenAI responses verbatim in their assessment, their reflection on reported use was assessed to establish to what extent the use supported student learning. This process allowed the assessor to conclude. If the assessor establishes that GenAI use was aligned with the assessment requirements and institutional guidelines, the use was deemed appropriate, and a grade was awarded for the assessment.

If GenAI use has been detected but not declared, the use does not align with institutional guidelines; the student failed the assessment or would be granted a resubmission in line with the institutional assessment guidelines. If a student's AI use was deemed inappropriate, the student failed the assignment without the opportunity to resubmit. It is important to note that the decision tree will be piloted in the module's next examination cycle. We acknowledge that further research is needed to determine the effectiveness of the decision tree in enhancing fair and consistent marking. Therefore, we do not propose using the decision tree as is but rather adapt it to suit the educator's context, assessment guidelines and institutional guidelines on GenAI and assessment.

5.6 Significance of findings for international educators

We recommend combining quantitative measures, such as AI detectors, with qualitative reflections to promote GenAI's fair and ethical use. Combining these measures will also support students using GenAI, as the judgements made can be clarified with assessment rubrics for the reflections. The decision tree and declaration can be adapted to various international educational systems. We also recommend emphasising reflective learning to foster critical thinking, self-regulation and awareness of accountability for students. Our use of AI detection tools sees it as being employed in conjunction with human insight and educational values across different learning spheres. Our final recommendation is that technological changes and innovations be aligned with ethical and pedagogical priorities using resources like those offered in our paper.

Our recommendations are based on the empirical data presented in the current paper. The recommendations are given so lecturers and researchers can make informed decisions about assessment practices and GenAI.

1. Use AI detectors with care, but they do have a role to play. When students submit written work, especially for content-heavy subjects, measures such as AI detectors help quality assure the writing. Writing with AI should be a goal, but until such practices have been established, the emphasis should be on students writing in their own words.
2. Train graders and lecturers on how to deal with generated text. Students will use GenAI in their written assignments; therefore, graders and lecturers must be equipped to respond. Here, the decision tree (Fig. 1) can be helpful.
3. Carefully consider the role and necessity of writing in assessment. The role of writing in assessment may be changing, and if software can write for us, this may not be the most accurate way to assess student understanding. Asking students to make presentations, complete tests and conduct practical assessments may be the way forward.
4. Use reflective practices to help students in their GenAI use: As found in this paper, self-reflection can help students use GenAI more responsibly and increase their reasoning skills about their own learning. Therefore, we recommend including self-reflection tasks in written assignments to help students and lecturers decide about GenAI use.

5.7 Limitations of the current study

The current study is limited to a sample of honours students at a public university in South Africa (University of Pretoria). The generalisability includes theoretical, but external generalisability may be limited [89]. The sample size was large enough for adequate data analysis (see power reported in the sample section). Two more limitations to keep in mind are the rapidly changing landscape of generative artificial intelligence and the limitations inherent in the Turnitin AI detectors.

Acknowledgements None.

Author contributions CC contributed to the literature review, the methods write-up, the quantitative data analysis, the mixed methods analysis and the discussion. NL contributed to the introduction, literature review, the tools (instruments), the qualitative data analysis, the discussion and the overall conceptualisation of the paper.

Funding None.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate The current study was conducted in line with the *Human Research Ethics Committees (RECs)* in South Africa and was fully approved by the *University of Pretoria's Ethics Committee* in the Faculty of Education, reference number EDU087/24. Participants freely gave informed consent for us to use their data for the research purposes described in this paper.

Competing interests The authors declare no competing interests.

References

1. Islam I, Islam MN. Exploring the opportunities and challenges of ChatGPT in academia. Discov Educ. 2024. https://doi.org/10.1007/s44217-024-00114-w.
2. Schlagwein D, Willcocks L. 'ChatGPT et al.': the ethics of using (generative) artificial intelligence in research and science. J Inf Technol. 2023;38(3):232–8.
3. Strzelecki A. Students' acceptance of ChatGPT in higher education: an extended unified theory of acceptance and use of technology. Innov High Educ. 2023. https://doi.org/10.1007/s10755-023-09686-1.
4. Batista J, Mesquita A, Carnaz G. Generative AI and higher education: trends, challenges, and future directions from a systematic literature review. Information. 2024;15(11):676.
5. Ogunleye B, et al. A systematic review of generative AI for teaching and learning practice. Educ Sci. 2024. https://doi.org/10.3390/educsci14060636.
6. Luo J. A critical review of GenAI policies in higher education assessment: a call to reconsider the "originality" of students' work. Assess Eval High Educ. 2024; 1–14.
7. Moorhouse BL, Yeo MA, Wan Y. Generative AI tools and assessment: guidelines of the world's top-ranking universities. Comput Educ Open. 2023;5: 100151.
8. Turnitin LLC. Turnitin [Computer software]. 2024.
9. Chan CKY. Is AI changing the rules of academic misconduct? An in-depth look at students' perceptions of 'AI-giarism'. arXiv preprint arXiv:2306.03358, 2023.
10. Alqahtani N, Wafula Z. Artificial intelligence integration: pedagogical strategies and policies at leading universities. Innov High Educ. 2024. https://doi.org/10.1007/s10755-024-09749-x.
11. Boud D, Soler R. Sustainable assessment revisited. Assess Eval High Educ. 2016;41(3):400–13.
12. Turnitin LLC. AI writing detection capabilities: frequently asked questions. 2024. https://www.turnitin.com/products/features/ai-writing-detection/. Accessed 18 Jun 2024.
13. Turnitin LLC. Understanding false positives within Turnitin's AI writing detection capabilities. 2023, YouTube: YouTube.
14. Turnitin LLC. AI writing detection in the classic report view. 2024. https://guides.turnitin.com/hc/en-us/articles/28457596598925-AI-writing-detection-in-the-classic-report-view. Accessed 20 Nov 2024.
15. Hayes S, Jandrić P, Green BJ. Towards a postdigital social contract for higher education in the age of artificial intelligence. Postdigit Sci Educ. 2024;6(2):467–85.
16. Fisk GD. AI or human? Finding and responding to artificial intelligence in student work. Teach Psychol. 2024. https://doi.org/10.1177/00986283241251855.
17. Ratnayake A, et al. All "wrapped" up in reflection: supporting metacognitive awareness to promote students' self-regulated learning. J Microbiol Biol Educ. 2024;25(1):e00103-e123.
18. Hamilton A, Wildermuth S. Student reflection. Multidisciplinary approaches to culminating student experiences, 2024: p. 45.
19. Nganga C, Jamison K. Learning to lead for equity and social justice through critical reflection and autobiography. Urban Rev. 2024;56(2):349–68.
20. Barros A, Prasad A, Śliwa M. Generative artificial intelligence and academia: implication for research, teaching and service. Manag Learn. 2023;54(5):597–604.
21. Davidson A. New technologies always scare us. Is A.I. any different? Freakonomics. 2023.
22. Dewey J, Pautz MC, Diede MK. How do we address faculty burnout? Start by exploring faculty motivation. Innov High Educ. 2023.
23. Nketsiah I, Imoro O, Barfi KA. Postgraduate students' perception of plagiarism, awareness, and use of Turnitin text-matching software. Acc Res. 2023. https://doi.org/10.1080/08989621.2023.2171790.
24. Bukar UA, et al. Decision-making framework for the utilization of generative artificial intelligence in education: a case study of ChatGPT. IEEE Access. 2024;12:95368–89.
25. Tight M. Mass higher education and massification. High Educ Pol. 2019;32(1):93–108.

26. Rasul T, et al. The role of ChatGPT in higher education: benefits, challenges, and future research directions. J Appl Learn Teach. 2023;6(1):41–56.
27. Pishchukhina O, Allen A. Supporting learning in large classes: online formative assessment and automated feedback. IEEE.
28. Muhic, M. The transformative potential of generative Ai in education. in INTED2024 Proceedings. 2024. Valencia, Spain: IATED.
29. Naseer F, et al. Automated assessment and feedback in higher education using generative AI. In Transforming education with generative AI: prompt engineering and synthetic content creation. 2024, IGI Global. p. 433–461.
30. Ruiz-Rojas LI, Salvador-Ullauri L, Acosta-Vargas P. Collaborative working and critical thinking: adoption of generative artificial intelligence tools in higher education. Sustainability. 2024;16(13):5367.
31. Albadarin Y, et al. A systematic literature review of empirical research on ChatGPT in education. Discov Educ. 2024. https://doi.org/10.1007/s44217-024-00138-2.
32. AlAli R, Wardat Y. Opportunities and challenges of integrating generative artificial intelligence in education. Int J Religion. 2024;5(7):784–93.
33. Koh E, Doroudi S. Learning, teaching, and assessment with generative artificial intelligence: towards a plateau of productivity. Learn Res Pract. 2023;9(2):109–16.
34. Baron P. Are AI detection and plagiarism similarity scores worthwhile in the age of ChatGPT and other generative AI? Scholarsh Teach Learn South. 2024;8(2):151–79.
35. Perkins M, et al. Detection of GPT-4 generated text in higher education: combining academic judgement and software to identify generative AI tool misuse. J Acad Ethics. 2024;22(1):89–113.
36. Halaweh M, El Refae G. Examining the accuracy of AI detection software tools in education. In 2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA). 2024. DUBROVNIK, Croatia: IEEE.
37. Eaton SE. Postplagiarism: transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. Int J Educ Integr. 2023;19(1):23.
38. Salinas-Navarro DE, et al. Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. Interact Technol Smart Educ. 2024. https://doi.org/10.1108/ITSE-12-2023-0236.
39. Ardito CG. Generative AI detection in higher education assessments. New Dir Teach Learn. 2024. https://doi.org/10.1002/tl.20624.
40. Fredheim R, Pamment J. Assessing the risks and opportunities posed by AI-enhanced influence operations on social media. Place Brand Public Diplomacy. 2024. https://doi.org/10.1057/s41254-023-00322-5.
41. Turnitin LLC. The plagiarism spectrum: instructor insights into the 10 types of plagiarism. 2016. https://www.turnitin.com/static/plagiarism-spectrum/. Accessed 25 Nov 2024.
42. Dai Y, et al. University policies on generative AI in Asia: promising practices, gaps, and future directions. J Asian Public Policy. 2024. https://doi.org/10.1080/17516234.2024.2379070.
43. Falebita OS, Kok PJ. Strategic goals for artificial intelligence integration among STEM academics and undergraduates in African higher education: a systematic review. Discov Educ. 2024. https://doi.org/10.1007/s44217-024-00252-1.
44. Uddin MM. Rejection or integration of AI in academia: determining the best choice through the opportunity cost theoretical formula. Discov Educ. 2024. https://doi.org/10.1007/s44217-024-00349-7.
45. Hmoud M, et al. Higher education students' task motivation in the generative artificial intelligence context: the case of ChatGPT. Information. 2024;15(1):33.
46. Floridi L, Cowls J. A unified framework of five principles for AI in society. Machine learning and the city: applications in architecture and urban design, 2022: p. 535–545.
47. Grimes M, et al. From scarcity to abundance: scholars and scholarship in an age of generative artificial intelligence. Acad Manag J. 2023;66(6):1617–24.
48. Chan CKY. A comprehensive AI policy education framework for university teaching and learning. Int J Educ Technol High Educ. 2023. https://doi.org/10.1186/s41239-023-00408-3.
49. Yusuf A, Pervin N, Román-González M. Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives. Int J Educ Technol High Educ. 2024. https://doi.org/10.1186/s41239-024-00453-6.
50. McDonald N, et al. Generative artificial intelligence in higher education: evidence from an analysis of institutional policies and guidelines. arXiv preprint arXiv:2402.01659, 2024.
51. Chaka C. Accuracy pecking order—how 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. J Appl Learn Teach. 2024. https://doi.org/10.37074/jalt.2024.7.1.33.
52. Dwivedi YK, et al. Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int J Inf Manage. 2021;57: 101994.
53. Kamalov F, SantandreuCalonge D, Gurrib I. New era of artificial intelligence in education: towards a sustainable multifaceted revolution. Sustainability. 2023;15(16):12451.
54. Corrigan JA, Onwuegbuzie AJ. Toward a meta-framework for conducting mixed methods representation analyses to optimize meta-inferences. Qual Rep. 2020;25(3):785–812.
55. Malina MA, Nørreklit HS, Selto FH. Lessons learned: advantages and disadvantages of mixed method research. Qual Res Account Manag. 2011;8(1):59–71.
56. Onwuegbuzie AJ, Poth C. Special issue mixed methods. Int J Qual Methods. 2015;14(2):122–5.
57. Field AP. Discovering statistics using IBM SPSS statistics. Thousand Oaks: Sage publications limited; 2024.
58. Goodman LA, et al. Measures of association for cross classifications. Cham: Springer; 1979.
59. IBM. IBM SPSS Statistics for Windows (Version 29.0). 2024, IBM Corp.
60. Linacre JM. Winsteps® (Version 5.4.0.0) 2023: Portland, Oregon: Winsteps.com.
61. Agresti A. Categorical data analysis, vol. 792. New Jersey: John; 2012.
62. Cohen J. Statistical power analysis for the behavioral sciences, vol. 2. Milton Park: Routledge; 1988.
63. Braun V, Clarke V. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. Couns Psychother Res. 2021;21(1):37–47.
64. Braun V, Clarke V. Thematic analysis: a practical guide. London: SAGE; 2022.

65. PlanoClark VL, Ivankova NV. Mixed methods research: a guide to the field, vol. 3. Thousand Oaks: Sage publications; 2017.
66. Selznick, B.S., *Always Almost There: Perspectives on Mixed Methods Research in Higher Education*. Innovative Higher Education, 2024.
67. Hands AS. Integrating quantitative and qualitative data in mixed methods research: an illustration. Can J Inf Libr Sci. 2022;45(1):1–20.
68. Masters GN. Partial credit model. In van der Linden WJ, editor. Handbook of item response theory. Chapman and Hall/CRC; 2021. p. 137–154.
69. Combrinck C. Is this a useful instrument? An introduction to Rasch models for evaluating tests and questionnaires. In Kramer S, et al., Editors. Online Readings in Research Methods (ORIM). Psychological Society of South Africa: Pretoria, South Africa; 2020, p. 127–181.
70. Open AI, L.L.C. ChatGPT-4.o (13 May 2024 version) [Large language model]. 2024; https://chat.openai.com.
71. Combrinck C. A tutorial for integrating generative AI in mixed methods data analysis. Discov Educ. 2024. https://doi.org/10.1007/s44217-024-00214-7.
72. Jonassen DH, Carr C, Hsiu-Ping Y. Computers as mindtools for engaging learners in critical thinking. Tech Trends Washington DC. 1998;43:24–32.
73. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn Environ. 2024. https://doi.org/10.1186/s40561-024-00316-7.
74. Albus P, Seufert T. Artificial intelligence in education: the importance of metacognitive. The Impact of Artificial Intelligence on Societies: Understanding Attitude Formation Towards AI. 2024. 109.
75. Watts FM, et al. Comparing student and generative artificial intelligence chatbot responses to organic chemistry writing-to-learn assignments. J Chem Educ. 2023;100(10):3806–17.
76. Duhaylungsod AV, Chavez JV. ChatGPT and other AI users: innovative and creative utilitarian value and mindset shift. J Namibian Stud : History Politics Culture, 2023. **33**.
77. Malik AR, et al. Exploring artificial intelligence in academic essay: higher education student's perspective. Int J Educ Res Open. 2023;5: 100296.
78. Pokkakillath S, Suleri J. ChatGPT and its impact on education. Res Hosp Manag. 2023;13(1):31–4.
79. Li MD, Little BP. Appropriate reliance on artificial intelligence in radiology education. J Am Coll Radiol. 2023;20(11):1126–30.
80. Faul F, et al. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods. 2009;41(4):1149–60.
81. Faul F, et al. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175–91.
82. Hutson J, Plate D. Human-AI collaboration for smart education: reframing applied learning to support metacognition. Faculty Scholarship, 2023. 480.
83. Hutson J, Plate D. Human-AI collaboration for smart education: reframing applied learning to support metacognition. IntechOpen, 2023.
84. Lelescu A, Kabiraj S, et al. Digital assessment in higher education: sustainable trends and emerging frontiers in the AI era. In: Grosseck G, et al., editors. Digital assessment in higher education: navigating and researching challenges and opportunities. Singapore: Springer Nature Singapore; 2024. p. 27–44.
85. Perkins M, et al. Simple techniques to bypass GenAI text detectors: implications for inclusive education. Int J Educ Technol High Educ. 2024. https://doi.org/10.1186/s41239-024-00487-w.
86. Ardito CG. Contra generative AI detection in higher education assessments. rXiv preprint, 2023. 2312.05241.
87. Gonsalves C. Addressing student non-compliance in AI use declarations: implications for academic integrity and assessment in higher education. Assess Eval High Educ. 2024. https://doi.org/10.1080/02602938.2024.2415654.
88. Luo J. How does GenAI affect trust in teacher-student relationships? Insights from students' assessment experiences. Teach Higher Educ. 2024. https://doi.org/10.1080/13562517.2024.2341005.
89. Turner SF, Cardinal LB, Burton RM. Research design for mixed methods: a triangulation-based framework and roadmap. Organ Res Methods. 2017;20(2):243–67.