Education and Information Technologies
https://doi.org/10.1007/s10639-025-13870-5

Transforming written assessment design to embrace AI: what needs to be changed to encourage higher-order critical thinking

Huahui Zhao[1] · **Thi Ngoc Yen Dang**[1]

Received: 9 July 2025 / Accepted: 8 December 2025

Abstract
Uncritical use of generative AI (GenAI) responses is a major concern among educators as it can hinder knowledge development, creativity, critical thinking and academic misconduct. To mitigate these repercussions, current discussions predominantly focus on changing assessment methods or policing students' use of GenAI, which greatly shapes students' GenAI use, for better or for worse. Few studies have examined how different assessment designs impact the way students use GenAI for coursework and the quality of GenAI-assisted writing. This study uncovered the relationships between assessment design and critical thinking in students' writing through analysing lecturer feedback on 51 postgraduate ChatGPT-assisted students' assignments across fourteen modules and assessment information related to the assignments. Results revealed that word limits, genres, information about organisational structures, and cognitive domains required by assessments significantly determined students' critical thinking performance in their disciplinary writing. Based on the results, we suggested (a) setting word limits based on task complexity rather than module credits, (b) designing integrated tasks with varied assessment methods to encourage critical thinking and knowledge development, (c) providing an appropriate amount of structural information to create space for critical thinking and (d) explicitly signalising cognitive domains required by assessments to address GenAI's impact on writing. We further encourage educators to critically reflect on the existing assessment guidance and practices to design assessments that cultivate critical AI users in an AI-empowered world.

Keywords Assessment design · Critical thinking · Lecturer feedback · ChatGPT · AI

Extended author information available on the last page of the article

Published online: 21 January 2026

1 Introduction

Uncritical use of AI responses can result in many negative consequences, including over-reliance on AI to substitute human creativity (Lo et al., 2024), metacognitive laziness (Fan et al., 2025), biased output with discrimination (Kasneci et al., 2023) and academic misconduct like plagiarism (Laflamme & Bruneault, 2025) . In the long run, these repercussions will undermine human knowledge and their distinctive intelligence and even hinder societal progress when plausible-sounding misinformation is increasingly used to justify viewpoints and provide solutions.

To address the detrimental effects, educators are recommended to redesign assessments to encourage critical GenAI use and retain unique human intelligence (e.g., originality and creativity) (e.g., Beckingham et al., 2024; Chan & Colloton, 2024). A main suggestion is to reduce written assessments or even replace them with other assessment methods, such as exams and presentations (Chiu, 2024; Kofinas et al., 2025).

Nevertheless, writing is a fundamental skill for developing rhetorical knowledge and critical thinking (CT) (Barrett & Pack, 2023). It has unique roles in capturing learners' thinking and eliciting evidence to measure learning outcomes at a depth that other methods cannot. Furthermore, written tasks are unavoidable activities in students' future workplace. Rather than eliminating written assessments to comprise their authenticity to mitigate unethical or ineffective GenAI use, educators should explore how assessment design influences GenAI use and redesign written assessments accordingly to cultivate students' CT, deepen subject knowledge, and enhance their competence in real-world problem solving.

No studies have yet explored how CT skills in GenAI-assisted writing vary across different assessment designs or what elements cause these differences. Undoubtedly, students' use of GenAI for writing shapes their writing processes and CT when writing. Assessments must be revised to address these changes. To redesign assessments that foster and evaluate CT, we need to understand the facets of assessment design that bring about the variations in CT skills in GenAI-assisted writing.

Although studies used surveys or interviews to elicit students' and staff's viewpoints about the impact of GenAI on writing, none of them exclusively focused on CT. Perception data are not necessarily congruent with practice and lack depth in revealing the actual impact on writing. To counteract these limitations, this study qualitatively analysed lecturer feedback on ChatGPT-assisted writing and related assessment design and quantitatively uncovered the elements of assessment design that led to significant differences in criticality in ChatGPT-assisted coursework. The results provided evidence-based implications for redesigning assessments to enhance CT and retain this distinctive human intelligence in the GenAI-enabled world.

2 Critical thinking in academic writing

The importance of CT for knowledge development and life has been unanimously stated by researchers and educators (Elbow, 2007; Facione, 1990; Martín-Raugh et al., 2023; Paul & Elder, 2014a). Its importance for critical GenAI use has been high-

lighted (Cordero et al., 2025; Sallam, 2023). Despite writing being both a process of performing critical thinking (CT) and a product that communicates its outcomes (Bean & Melzer, 2021), few studies have examined the manifestation of CT in writing. Research on the impact of GenAI on CT in its assistive writing is also few and far between.

CT requires writers to raise vital questions, gather and assess relevant information, think, recognise and assess open-mindedly about alternative viewpoints, evaluate them against criteria and standards, and reach well-reasoned conclusions and solutions (Bean & Melzer, 2021; Paul & Elder, 2014b). Disciplinary lecturers set specific expectations for CT in their students' disciplinary writing. Zhao et al. (2024) conducted corpus analyses of lecturer feedback on 230 students' disciplinary writing in Education and identified nine categories of CT that determine the quality of writing. Promoting CT in writing requires effective assessment design. Kurfiss (1988, cited in Bean & Melzer, 2021, p. 22) suggested that to encourage CT in assessment, asking students "to develop a 'best solution' to an ill-structured problem and justify the proposed solution with appropriate reasons and evidence".

3 Impact of GenAI on writing quality and critical thinking

Most existing studies on the impact of GenAI on academic writing have employed interviews to elicit students' perceptions. For instance, Kim et al. (2025) carried out in-depth interviews with twenty undergraduate and postgraduate students at a Chinese university after they used ChatGPT4 for writing IELTS essays. Students in their study highlighted ChatGPT as a non-judgmental proofreader that improved their overall writing quality. However, they reported mixed findings regarding its impact on CT: While ChatGPT helped them develop topic knowledge, analyse and identify gaps in existing studies, provide different perspectives, and engage in discussions, it was criticised for lacking higher-order thinking in Bloom's taxonomy (i.e., analysis, evaluation and creation) and human awareness (e.g., unique writing styles and learning environment), due to limitations such as fabricated information (i.e., hallucination) and a lack of contextual understanding.

Survey data from 28 students in Lin and Chang (2020) at a Canadian university reported that questioning and prompting with a locally designed chatbot helped them with their writing and made them more critical and reflective. In addition, after students used the chatbot to assist their essay outlines and peer review, they found that grades from the essay outlines without the chatbot were significantly lower than those using the chatbot. Studies also report that active engagement with automatic feedback from AI tools improves language accuracy or readability, although not necessarily writing quality (Barrett & Pack, 2023; Jiang, 2025). GenAI is also reported to enhance inclusivity by supporting disabled students (Coughlan & Iniesto, 2025).

Educators are more concerned than students about the use of AI in instruction. Teachers in K-12 schools in Lee et al. (2024) asserted that GenAI's lack of human awareness significantly affected their perceived usefulness and intention to use AI in their teaching. Merine and Purkayastha (2022) reiterated other researchers' worries about incorporating GenAI into education, including impairing students' CT capabil-

ity and explanatory skills. Likewise, Li et al. (2023) expressed their concerns about students exploiting ChatGPT to achieve a good grade with minimum effort, thereby losing their opportunities to develop CT and idea-synthesising skills. Kitamura (2023) analogised invalid responses and plagiarism related to ChatGPT use as thorns alongside the benefits of using ChatGPT as roses.

The likely underperformance of CT in ChatGPT-assisted writing can be explained by Borji (2023) test of ChatGPT's capability of dealing with varied aspects. He reported that ChatGPT responses often include hallucination and bias due to limitations in training data and its inability to differentiate the nature of data (e.g., between factual and fictional information). However, its sophisticated language may seduce students into adopting these invalid responses without critically evaluating their quality and relevance to their writing. He also noted that ChatGPT underperforms in reasoning which requires familiarity with real-world knowledge. It does not fully understand the social world in the training data as well as the connections between concepts and entities. Consequently, it cannot perform well critical thinking tasks such as making predictions about events (i.e., temporal reasoning), understanding human behaviours and mental processes (i.e., psychological reasoning) and judging things that are generally accepted as true, correct or appropriate that require life experience and observation (i.e., commonsense reasoning). However, these limitations were identified based on ChatGPT's responses to a few prompts, lacking systematic examination and rigour in educational contexts. As ChatGPT advances, its performance may improve, and its responses become more sophisticated and harder to detect hallucinations. This makes CT even more crucial for judging the relevance and value of GenAI's responses in academic work.

To address this gap, survey data alone is insufficient. A systematic examination of CT in writing across assessment design is essential to reveal how assessment can be designed to mitigate the detrimental consequences of GenAI use and enhance its positive impacts on CT and knowledge development. This study aimed to reveal which assessment elements impede critical thinking in ChatGPT-assisted writing and provide implications for assessment design to promote higher-order CT in postgraduates' disciplinary writing.

4 Factors of assessment design related to assessment performance

Multiple factors are suggested to be taken into consideration when designing assessments and presented in assessment information. Assessment design needs to consider purposes, contexts, learner outcomes, tasks, feedback process and interactions with learners and colleagues (Bearman et al., 2016) . Bean and Melzer (2021) suggested that a good assessment task should include a disciplinary problem for students to address, introduce the role or purpose of their writing, specify the audience, genre and implied discourse community, include interactive components with various stages, and provide evaluation criteria. The UK Quality Code for Higher Education (Quality Assurance Agency, 2024) suggests that assessment information should include learning outcomes, the purpose of assessment, provision of learning support (e.g., feedback) and academic integrity. Other essential information includes task

descriptions and structural guidance to break complex tasks into manageable sections (Swales & Feak, 2012) and word limits, depending on the module's credit-bearing (e.g., 20-credit modules require 200 h of learning) (Quality Assurance Agency, 2021). This commonly translates into longer outputs for modules with more credits in assessment practice. Zhao (2024) stressed accessible assessment rubrics as part of the information for students to promote a shared understanding of assignment quality between academics and students.

While these sets of information are commonly observed in assessment information across settings, no studies have systematically researched how they might shape writing and CT differently. When redesigning assessments to address repercussions of GenAI use, it is essential to scrutinise whether and how these facets need to be revised to evaluate and develop critical thinking and knowledge development.

The assessment information provided by the modules in this study commonly included assessment purposes related to the modules' learning outcomes, assessment support, assessment rubrics, academic integrity and submission guidance (e.g., submission platform, deadlines, and submission format). They vary in terms of task complexity, the amount of information about organisational structure and the length of outputs. Therefore, due to limited studies on assessment design and its impact on CT, we synthesised various studies from theoretical perspectives and practical suggestions across disciplines to understand the potential impact of these variations on CT. Unsurprisingly, little evidence is related to GenAI-assisted writing.

4.1 Task complexity: cognitive demands

Task complexity is determined by cognitive demands and the number of required elements (Skehan, 1998). Deane (2011) suggests that different writing tasks require different cognitive strategies. Since students have limited cognitive resources when performing a task, increasing task complexity forces trade-offs between different aspects of performance (Skehan, 2009). Robinson (2001) suggested a triadic componential framework of cognitive complexity of assessment tasks that affected assessment outputs, consisting of the number of reasoning steps and linguistic structures, interactional demands (i.e., the extent to which students need to engage in negotiation of meaning) and task conditions including modality (written vs. spoken). More complex tasks are believed to promote higher-order thinking by requiring problem-solving, self-regulation and creativity (Maranna et al., 2025), although the required CT skills need to be within the Zone of Proximal Development (i.e., the zone between learners' independent problem-solving capacity and solving with help) (Vygotsky, 1978). Open-ended tasks can lead to deeper cognitive engagement than closed-ended tasks (Raz et al., 2024). No studies have examined how the requirements of cognitive domains in assessments relate to CT manifested in the writing product. In this study, we examined how task complexity in terms of the required CT skills and genres of tasks relates to CT manifested in ChatGPT-assisted writing.

4.2 The provision of organisational structure

The provision of organisational **structure** may also affect students' engagement with assessments and their outputs. Swell's cognitive load theory (CLT) suggests that structured instruction - such as step-by-step worked examples - helps reduce the unnecessary mental effort to process tasks (i.e., extraneous cognitive load) and enhance germane load to enhance knowledge acquisition (e.g., design tasks that requiring higher-order thinking and problem solving to develop cognitive schemas) (Paas et al., 2003; Sweller, 1988, 2020). This is supported by scaffolding, evolving from the ZPD in socio-cultural theory (Vygotsky, 1978). For example, students are observed to struggle with structuring positioning in their disciplinary writing (Wingate, 2012). Providing them with information on how to structure assignments can reduce their extraneous cognitive load, allowing them to focus more on the content. This study will examine whether providing students with structural information facilitates or hinders postgraduate students' CT in ChatGPT-assisted disciplinary writing.

4.3 Output length

The cognitive load theory indicates that excessive length can dilute focus or cause cognitive overload (Sweller, 1988). Although longer papers allow for comprehensive literature reviews, methodological details, and nuanced argumentation, Bean and Melzer (2021) indicated that they may lead to content inflation. Students tend to use long quotations or lengthy paraphrases to fill their pages without considering their relevance and importance. In contrast, a short paper requires students to condense their thinking into a small amount of writing, often producing more learning than a traditional research paper (Bean & Melzer, 2021).

Anderson et al. (2016) argues that deep learning via writing depends less on the amount of writing assigned in a course than on the design of the writing assignments themselves. Although no empirical results have been reported about the impact of length on CT, it is clear that students change their writing strategies to meet word counts, which could affect the variety and focus of CT and their output quality (Evans & Harrington, 2024). This study aimed to substantiate whether and how word counts related to the manifestation of CT in ChatGPT-assisted disciplinary writing.

To address the aforementioned knowledge gaps on the relationships between assessment design and CT in ChatGPT-assisted writing, we asked the following five research questions:

RQ1. What is the distribution of positive and negative lecturer feedback on critical thinking skills in ChatGPT-assisted writing?
RQ2. How is teacher feedback on critical thinking skills associated with the length of written coursework?
RQ3. How is teacher feedback on critical thinking skills associated with cognitive domains required by written coursework?
RQ4. How is teacher feedback on critical thinking skills associated with different genres of written coursework?

RQ5. How is teacher feedback on critical thinking skills associated with the provision of organisational structures of written coursework?

RQ1 provided an overview of teacher feedback on critical thinking on the whole dataset whilst RQ 2–5 examined how the feedback provision on each critical thinking differed across different elements of assessment design, including cognitive domains required by the assessments (RQ2), genres (RQ3), the provision of organisational structures for the written coursework (RQ4) and word limits (RQ5). Mixed research methods were adopted to answer each question.

5 Methodology: research context, data background and participants

After obtaining ethical approval from our university ethics committee, volunteer students uploaded their marked ChatGPT-assisted written coursework to a Microsoft Form saved in a password-protected OneDrive folder hosted by our institution. They were assured that their work would be used solely for research purposes and would not be examined for academic integrity investigation. Each participant received £10 for their contributions. After receiving assignments, we informed lecturers of the project and sought their consent to analyse their feedback.

Participating students were from diverse backgrounds, including Chinese, Vietnamese, Indonesian, Japanese, Persian, Malayalam, Spanish, Hindi, British, and Arabic. The diversity of the student population is a very typical in an Education postgraduate programme in the U.K. They did not receive training in using ChatGPT for writing.

5.1 Data analysis of lecturer feedback on critical thinking

Lecturers used the existing assessment criteria and provided feedback on knowledge and understanding, argument, and academic presentation. The feedback was provided before this study in a naturalistic context without interference from the researchers; therefore, lecturers were not told whether ChatGPT was used when feedback was provided.

Overall, lecturers provided approximately 20,000 words of feedback on various aspects of these 51 assignments. The mean coursework mark was 60.61 ($SD = 9.30$), with most scripts falling into the 50 s and 60 s bands (Table 1).

1. **Manually reviewed** each feedback file to record all feedback terms lecturers use to evaluate CT, regardless of their frequency, starting from the nine CT categories developed in Zhao et al. (2024). Wildcards were used to record terms with the same root (e.g., implement* as a wildcard to cover implements, implemented, implementing and implementation). The categories embody:

    1. **Explaining**: explain writing by defining terms or concepts, providing rationales and details, identifying gaps, comparing and contrasting different resources, and summarising resources and own work.

Table 1 Construction of lecturer feedback corpus

Grade	Number of Scripts	Total Tokens of Feedback
Fail (below 50 s)	5	2,767
Pass (50s)	17	6,881
Merit (60s)	21	7,540
Distinction (70s and above)	8	3,233
Total	**51**	**20,421**

We followed similar steps to those in our previous project that explored lecturer feedback on CT in student-only writing from the same cohort of tutors, using the same assessment criteria (Zhao et al., 2024) :

2. **Analysing**: explore sources, analyse tasks, data and materials, and interpret phenomena, analysis, and findings with references and enquiry questions.
3. **Evidencing**: substantiate writing with readings, data, examples, and personal viewpoints.
4. **Contextualising**: provide contexts to set the scene, identify problems, strengthen analysis, present implications, locate discussion, perform reflection, provide overviews and set out aims.
5. **Evaluating**: assess various resources for writing, reflect on existing knowledge to inform knowledge and practice, measure with tools/scales, and weigh information.
6. **Establishing relevance**: create links with literature, personal experience, and context, and stay focused on relevant issues.
7. **Synthesising**: integrate different lines of literature and different resources and perspectives.
8. **Extrapolating**: focus on the implication of knowledge across domains or in practical contexts.
9. **Positioning**: articulate and justify different viewpoints, present authorial suggestions and hypotheses, challenge assumptions, make predictions and problematise existing issues.

Two new categories were created from unfitted relevant feedback terms to the nine existing ones: **balancing** (i.e., balancing different views and pros and cons) and **applying** (i.e., designing and implementing a procedure to address real-world issues).

2. **Created the feedback corpus** of the 51 feedback files in AntConc 4.3.1, a widely used concordancing software (Anthony, 2023). A concordance is a list of all occurrences of a search word in the corpus displayed together with the words in its immediate context (Sinclair, 1991) (Figure 1).

35-4.docx	in the writing' something which is especially important in a	reflective	piece. There seems to be little in the way

Fig. 1 An example of a concordance related to the search word 'reflective'

3. **Used the 'KWIC (key word in context)'** function in AntConc 4.3.1 to generate relevant concordances which show how a word was used in a feedback instance to:

    a. **Refine sub-categories to the eleven CT skills**. For instance, concordance lines revealed the term *differ** falling into multiple categories, including explaining, synthesising, and balancing. Therefore, EXdiffer*, SYNdiffer* and BALdiffer* were created to record their multiple associations.
    b. **Distinguish the nature of feedback**. We classified feedback into positive, negative, mixed (of positive and negative feedback), overlapping, or irrelevant feedback, after copying all concordances in Excel files (one file for each term). For instance, in the concordance line: "*References should not only support your claims but also add value to your discussion*", '*references*' and '*support*' were feedback terms for the same CT skill: evidencing. It was only counted once in '*referencing*' and labelled as overlapping feedback in '*support*'. Irrelevant feedback refers to instances where feedback terms were used to describe other aspects instead of CT skills. For example, in the concordance line in Fig. 1 above, reflective was used to refer to the writing as a reflective piece rather than a CT skill, so it was categorised as irrelevant feedback.

The two authors independently coded ten of the feedback files, following the steps above. A full agreement was achieved mainly because of the co-construction of coding schemes in the previous project and the rigorous steps established to code the data. After counting the frequency of each type of feedback related to each term, the results were recorded in an Excel file (one worksheet per CT skill) and imported to SPSS version 28.0.1.1 for statistical analyses.

5.2 Data construction regarding different elements of assessment design

Assessment design was explained to students using a faculty-wide template, consisting of key information about assignment titles, assignment descriptions, rationale for design, word limits, draft and feedback provision, learning outcomes assessed, assignment guidance, assessment criteria, and academic integrity. NVivo 14 was used to analyse commonalities and variations in assessment design. The results were recorded in an MS Excel, which revealed word limits, genres, and cognitive domains required by the assessment, and provision of organisational structures as the most striking differences.

We collected 51 assignments from fourteen modules across four postgraduate programmes related to Education (i.e., Teaching English to Speakers of Other Languages, Education, Childhood Studies, and Digital Education) at a major university in the U.K. The assignments were typically with a length of 2,000 to 6,000 words from modules with 10 to 30 credits, with varied genres.

The cognitive domains were classified using verbs that described the actions students were required to perform to complete their assignments. These verbs aligned with the key search terms related to critical thinking (CT) categories found in lec-

turer feedback, likely because both the assessment documents and feedback were produced by the same group of lecturers. Based on content and thematic analyses of assessment information from the fourteen modules where the 51 assignments originated, three groups of cognitive domains surfaced:

- **Category 1**: Creating new materials – justifying (based on readings) – evaluating (based on feedback or reflection): five modules and 39 assignments belonged to this category ($N=39$).
- **Category 2**: Identifying a concept/problem – evidencing (through readings) – analysing/criticising the problem: four modules and five assignments belonged to this category ($N=5$).
- **Category 3**: Identifying a problem in a local context – analysing (based on readings) – providing solutions: five modules and seven assignments belonged to this category ($N=7$).

We distinguished genres into four categories based on CT skills identified from the analysis of lecturer feedback, to enhance the comparability of required CT skills and CT skills manifested in ChatGPT-assisted writing. This resulted in four genres:

Genre 1. Portfolio essays with various tasks of different natures, including designing materials (main CT skills including extrapolating and applying), providing rationales (main CT skills including analysing and explaining) and reflecting on the design (main CT skills including evaluating) ($N=38$).
Genre 2. Argumentative essays based on a specific topic (main CT skills including explaining and synthesising) ($N=4$).
Genre 3. Essays following oral presentation (main CT skills: evaluating via reflecting on activities such as micro-teaching sessions or applying procedures introduced in the oral presentation) ($N=5$).
Genre 4. Portfolio essays include a collection of mini tasks (main CT skill: analysing materials) ($N=4$).

Three categories were observed regarding the provision of organisational structures (i.e., how each section of assignments can be organised):

- **Assessments with detailed instructions**: These assessments provide comprehensive, step-by-step guidance for each part of an assignment. In most cases, they include an approximate word count for each section ($N=5$).
- **Assessments with brief instructions**: These assessments offer a general outline of what should be included in the assignment without detailed steps or main points in each section ($N=41$).
- **Assessment with no instructions**: These assessments do not provide any information about the structure of the assignments ($N=5$).

We adopted Kruskal-Wallis H tests to investigate how CT differed across these different elements of assessment design. There were big differences in sample sizes across groups related to each element. Following the suggestion regarding a small sample

size project (i.e., 51 assignments in this study) from statisticians, including Gibbons and Chakraborti (2020), a small sample size >=4 was used as an acceptable minimum sample size for Kruskal-Wallis H tests. We referred to adjusted significant values for paired comparison results to identify differences between groups in lecturer feedback on each CT skill across cognitive domains, genres, and provision of organisational structures. We employed Mann-Whitney tests to examine the impact of word limits on CT, since there were two groups. For both types of tests, we applied Cohen's (1988) guideline to measure the effect sizes: 0.2 = small effect, between 0.5 = medium effect and between 0.8 = large effect.

6 Results

The results were reported in the order of the research questions from both quantitative and qualitative perspectives.

6.1 Lecturer feedback on critical thinking in ChatGPT-assisted writing (RQ1)

In total, lecturers provided 990 comments on CT (SD = 0.24–4.04.24.04) on the 51 assignments, with 94 mixed feedback instances. The comments were unevenly distributed across the eleven CT skills: explaining, evidencing, establishing relevance, positioning, extrapolating, contextualising, evaluating, analysing, applying, synthesising, and balancing in descending order (Fig. 2). None of the skills received over 30% of the total number of comments, although over 65% of the feedback instances tackled explaining, evidencing, and establishing relevance.

Lecturers provided 175 positive comments on ChatGPT-assisted writing, unevenly distributed across CT skills (SD = 0.24–4.04), accounting for 17.7% of the overall feedback instances. Figure 2 shows the total number and percentage for each skill.

Figure 3 shows that *explaining (EXP)* was the most frequently praised skill in students' ChatGPT-assisted writing. *Evidencing (EVI)*, *evaluating (EVA)*, *contextual-*

35-4.docx	in the writing' something which is especially important in a	reflective	piece. There seems to be little in the way

Fig. 2 Lecturer feedback on critical thinking in ChatGPT-assisted writing

35-4.docx	in the writing' something which is especially important in a	reflective	piece. There seems to be little in the way

Fig. 3 Distribution of positive feedback on CT skills in ChatGPT-assisted writing

35-4.docx	in the writing' something which is especially important in a	reflective	piece. There seems to be little in the way

Fig. 4 Distribution of negative feedback on CT in ChatGPT-assisted writing

ising (CONT), and *establishing relevance (RELE)* received over 10% of the positive feedback. A few positive comments were provided on *positioning (POSI)* and *analysing (ANAL)* and even fewer positive comments on *extrapolating (EXTRA)*, *synthesising (SYN)*, and *balancing (BALA)*.

Lecturers provided 721 negative comments with different distributions across CT skills (SD = 0.24–4.04), accounting for 72.8% of the total number of lecturer comments.

Figure 4 shows that 67% of negative feedback focused on *explaining*, *evidencing*, and *establishing relevance*, in descending order. Additionally, 8.32% of the feed-

back instances addressed issues on *extrapolating* theories to practice or suggestions to other contexts (SD = 1.35). Lecturers provided 25 negative comments on *applying* (SD = 0.86), encouraging students to make careful decisions when implementing a procedure in their practice. Furthermore, 5.55% of negative feedback explicitly required students to *contextualise* their writing within its theoretical and practical contexts. Lecturers provided thirty instances of negative feedback on *positionality* (SD = 0.94) and thirty on analysing (*SD = 0.78).*

6.2 Word limits and critical thinking in ChatGPT-assisted writing

The data consists of one essay each requiring 1200, 2000, and 2500 words. Nine assignments require 3000 words, and two require 4,000 words. There are 37 of 6000 words. To meet the minimum sample size of a non-parametric test, we merged them into two groups: one with a length ≥ 3000 words ($n = 12$) and the other with a length ≥ 4000 words ($n = 39$). We carried out a Mann-Whitney test to uncover the impact of word limits on positive and negative lecturer feedback on CT in ChatGPT-assisted writing.

A significant difference was found in **positive feedback on *analysing*** (U = 176, Z = −2.158, $p = .031$), assignments with a length ≥ 3000 words ($n = 12$) received more positive feedback on analysing (Md = 0, mean rank = 30.83) than those with a length ≥ 4000 words (Md = 0, mean rank = 24.51). The effect size of $r = .30$ reveals a medium effect on the strength of positive feedback on *analysing*.

A significant difference was also observed in terms of **negative feedback on *extrapolating*** (U = 327.5, z = 2.196, $p = .028$). Longer assignments received more negative feedback (Md = 1.00, mean rank = 28.40) than shorter ones (Md = 0.00, mean rank = 18.21). The effect size $r = .31$ revealed a medium effect of word limits on negative feedback on extrapolating in assignments.

A significant difference was observed in terms of **positive feedback on *synthesising*** (U = 195, z = −2.575, $p = .010$). Shorter assignments received more positive feedback (Md = 0, mean rank = 29.25) than longer assignments (Md = 0, mean rank = 25). An effect size of $r = .36$ revealed a medium effect on positive feedback on synthesising.

A significant difference was observed in terms of **positive feedback on *evaluating*** (U = 313, z = 2.013, $p = .044$). Shorter assignments received less positive feedback on evaluating (Md = 0, mean rank = 19.42) than longer assignments (Md = 0, mean rank = 28.03). An effect size of $r = .28$ revealed a medium effect of word limits on positive feedback on evaluating.

6.3 Cognitive domains in assessment information and ChatGPT-assisted writing

We carried out Kruskal-Wallis Tests to uncover any significant differences in lecturer feedback on CT in ChatGPT-assisted writing across the three categories.

A significant difference was observed in **negative feedback on *extrapolating*** [x^2 (2, $N = 51$) = 6.495, $p = .039$]. Pairwise comparisons revealed that significant differences only existed between Categories 1 and 2 ($x^2 = 16.397$, adjusted $p = .042$). Assignments in Category 1 received more negative feedback (Md = 1, average rank = 28.40) than those in Category 2 (Md = 0.00, average rank = 12.00). The effect size of $\varepsilon^2 = 0.13$ sug-

gests a small effect. Extrapolating requires students to apply knowledge to practice. Assessments in Category 1 require writers to extrapolate their knowledge from readings to design new materials. In contrast, assessments in Category 2 require students to evidence their analysis with readings rather than designing new materials, making them less demanding in extrapolating than those in Category 1 and thus potentially receiving less negative feedback.

Another significant difference was observed in **positive feedback on *evaluating*** [x^2 (2, N=51)=7.205, p=.027]. Pairwise comparisons revealed significant differences only existed between Categories 1 and 3 (x^2=13.128, adjusted p=.041). The former received more positive feedback (Md=1, average rank=28.63) than the latter (Md=0.00, average rank=15.50). The effect size of ε^2=0.14 suggests a small effect. The difference is understandable as evaluating is required by assessments in Category 1, whereas students are not required to evaluate their solutions by assessments in Category 3.

6.4 Assessment genres and critical thinking in ChatGPT-assisted writing

The Kruskal-Wallis H test results revealed a significant difference in lecturers' **positive feedback on *analysing*** across genres [X^2 (3, N=51)=15.25, p=.002]. The adjusted significant value showed that significant differences only exist between assignments with Genre 1 and Genre 3 (X^2=13.59, adjusted p=.008). The median rank for Genre 1 (Md=0.00) is lower than that for Genre 3 (Md=1.00). This suggests that although both genres require applying theories to design procedures or materials, either in written (Genre 1) or verbal (Genre 3) forms, students performed better at *analysing* in Genre 3 than in Genre 1. This suggests the facilitative role of oral presentation in activating analysing with a medium effect size (ε^2=0.30).

The statistical results also showed a significant difference in lecturers' **positive feedback on *evidencing*** across genres [X^2(3, N=51)=9.69, p=.021]. The adjusted significant value revealed significant differences between Genre 1 and Genre 2 (X^2=19.474, adjusted p=.011). The median rank of positive feedback for Genre 1 (Md=0.00) is lower than Genre 2 (Md=1.50). This suggests that argumentative essays encourage more supporting with evidence than portfolio assessments requiring multiple cognitive steps, although the effect size of ε^2=0.19 suggests a small effect.

6.5 Provision of organisational structures

We carried out the Kruskal-Wallis H test to uncover whether no instructions, brief instructions, or detailed instructions about assignment structure was associated with significant differences in positive and negative lecturer feedback on each CT skill.

The results revealed a significant difference in **positive feedback on *analysing*** across detailed instruction, brief instruction and no instruction about structure [X^2(2, N=51)=10.81, p=.004]. The adjusted significant value showed significant differences between assignments with detailed instructions and no instructions (X^2=15.70, adjusted p=.015), and between brief instructions and no instructions (X^2=13.26, adjusted p=.005). No significant difference was observed between detailed and brief

instructions. Assessments without instructions have the highest median rank of positive feedback on *analysing* (Md = 1.00, with an average rank = 38.20), followed by those with brief instructions (Md = 0.00, with an average rank = 24.94) and detailed instructions (Md = 22.50, with an average rank = 22.50). The results indicated that providing instructions about organisational structures might stop students from critically engaging with assessment, with a medium effect size of $\varepsilon^2 = 0.22$.

7 Discussions and implications for Language assessment in the AI-enabled era

This study scrutinised the impact of assessment design on CT in disciplinary writing. It conducted (a) corpus analyses of lecturer feedback in 51 students' assignments, (b) content and thematic analyses of the 14 assessment briefs guiding the 51 assignments, and (c) statistical comparison analyses to reveal their relationships. The results provided significant implications for redesigning assessments to enhance postgraduates' knowledge development and CT during their assessment preparation and in their ChatGPT-assisted writing outputs.

7.1 Setting word limits based on task complexity to enhance critical thinking

The effect sizes of comparison analyses indicated word limits as the most significant factor impacting CT in ChatGPT-assisted writing. Shorter assignments encourage more concise and focused analysis as students need to be more selective and precise in their arguments to meet word limits. The result corroborates Bean and Melzer (2021) that shorter papers can facilitate deeper engagement with CT. It also substantiates the cognitive load theory that longer assignments could overwhelm students, leading to less effective analysis. Increased length might lead students to spread the analysis too thin across multiple points, thereby diluting its quality.

Likewise, shorter assignments encourage students to be more concise and clearer in their synthesis of information. Limited word counts require students to focus on the most relevant points and report their relationships in a coherent and well-integrated synthesis. This leads to more effective and impactful synthesis and hence, more positive feedback on synthesising. Bean and Melzer (2021) stated about content inflation in terms of long quotes from references in long papers. Longer ChatGPT-assisted assignments might prompt students to include a large chunk of ChatGPT's responses to their writing without carefully synthesising across resources. This led to increasing negative feedback on synthesising. Longer assignments also compel students to cover more ground, leading to overextension or less accurate extrapolation, resulting in more negative feedback on extrapolating than shorter ones. Nevertheless, when *evaluating* is required, longer assignments provide more space for students to explore more complex concepts, describe a more thorough reflection, and elaborate on their evaluations.

The results suggest that the length of assignments should be decided by the nature of tasks and required CT skills in the module assessment rather than by their credits.

In general, shorter assignments facilitate more concise and effective delivery of CT than longer ones.

7.2 Explicitly signalling a wide range of cognitive domains to encourage critical thinking

Cognitive domains have a small effect on CT regarding extrapolating and evaluating. Requiring students to justify their material design encourages extrapolating, although negative feedback indicated that students need to make more efforts to develop this CT skill while working with ChatGPT. The explicit requirement of evaluating materials in assessments promotes evaluating. With the assistance of ChatGPT, students performed this CT skill well. This was confirmed by reflection notes in students' assignments about how useful ChatGPT feedback was in helping them evaluate their lesson and material design. One student wrote:

> Overall, the feedback from ChatGPT provided valuable insights and practical suggestions to enhance my lesson plan. By critically reflecting on these suggestions and comparing them with existing knowledge, teacher inputs and literature, I was able to validate their relevance and identify areas for improvement. This process emphasised the importance of integrating various sources of feedback to create a well-rounded and effective lesson design for Year 5 Malaysian primary school learners.

The results of the impact of cognitive domains on CT corroborate the role of genres in shaping CT as discussed above. They suggested the importance for assessment designers to think carefully about what cognitive skills would be required to complete these assessment tasks and how they could evidence students' learning outcomes. In addition, explicit assessment vocabulary related to CT should be used in assessment information to set clear cognitive pathways to encourage CT. This aligns with Andrade's (2000) suggestion about using CT-related wordings in assessment rubrics to direct students towards critical tasks and Zhao et al.'s (2025) advice about incorporating vocabulary related to writing proficiency in instruction and assessment to guide students towards effective language communication.

7.3 Use varied genres to encourage critical thinking

The effect size showed genres as the second most influential assessment element for performing CT. The results showed that a genre combining written and oral presentation (i.e., integrated tasks) encourages deep engagement with analysis. This might be because oral presentation requires students' active participation and engagement with content, developing a deeper understanding of the material and consequently, better analytical skills when writing the subsequent essays. The positive effect might also be attributed to immediate feedback from peers and instructors during oral presentations. This real-time feedback can help students refine their understanding and improve their analytical skills, leading to higher-quality analysis. Oral presentations typically have a clear structure and focus. This clarity can help students organise

their thoughts more effectively when writing their essays, resulting in more coherent and well-analysed content. Reflecting on the oral presentation and applying the procedures introduced can also enhance CT, leading to more insightful and analytical essays.

The results also showed that students performed better in evidencing in essay-based assessments than in portfolios requiring material design, rationale and reflections. This finding is unsurprising, as designing materials, providing rationales, and reflecting on design emphasise creativity, justification and evaluation, which may not always require extensive evidence. In contrast, argumentative essays inherently demand robust evidence to support claims.

The results indicated that assessment methods combining written and oral presentation could be a promising approach to designing GenAI-allowed assessments. This dovetails with the CRADEL suggestion about using a range of different sequential tasks to evidence outcomes in varied ways to address the relationship between assessment and GenAI (Bearman et al., 2023). However, if an assessment aims to assess students' understanding of relevant readings or concepts, an essay-based assessment works better than assessments requiring *applying* as the main CT skill. Nevertheless, strategies need to be adopted to ensure students' independent work on such essays as GenAI excels at writing essays based on evidence from their training data with good structures.

The results resonate with the implication above about the purpose of assessment, such as required CT to demonstrate learning outcomes, determining assessment design, echoing the decisive role of the purposes of assessments in assessment design (Bean & Melzer, 2021; Quality Assurance Agency, 2024).

7.4 Provide an appropriate amount of structural information to encourage critical thinking

The amount of information about organisational structure was the third influential element shaping CT in ChatGPT-assisted disciplinary writing. The results suggested that assignments without instructions allow students to approach the task in their own way, encouraging creativity and independent thinking. This freedom can lead to more innovative and thorough analysis with ChatGPT. Conversely, assignments with detailed or brief instructions provide a framework that might limit the students' ability to explore different analytical approaches, potentially leading to more uniform and less creative responses. In addition, when students are given no instructions, they are forced to engage more deeply with the tasks, conversing with ChatGPT, and relying on their CT skills to determine the best way to approach the assignment, enhancing their analytical abilities.

In contrast, detailed instructions can lead to a more mechanical approach and reduce the need for independent CT, where students follow the steps without fully engaging with their coursework. Brief instructions offer some flexibility but still provide a structure that might limit deeper exploration. Both detailed and brief instructions provide some level of guidance, which can result in similar levels of analysis, explaining why there are no significant differences in feedback between these two types of instructions.

The results were corroborated by students' chat histories collected in this study. Students copied and pasted the detailed information about how to structure the assignments to their ChatGPT prompts, either all at once or part by part, followed by requests for ChatGPT to produce the outlines or drafts. This could lead to academic misconduct if no CT is performed to evaluate ChatGPT responses. This also causes students' dilemma on how to use ChatGPT. On one hand, they worry about potential collisions with another student's work if both use the structure information to elicit assistance in outlines or drafts. On the other hand, they hesitate to use their own organisational structure, worrying that their analysis might not meet their tutors' expectations.

No instructions about the structure of an assignment encourage significantly more analytic thinking than brief or detailed instructions. This suggests removing explicit guidance about structuring writing products to motivate students to analyse the requirements (with ChatGPT) and complete the assignment creatively. However, as the cognitive load theory highlights, the provision of clear instructions needs to cater to individual needs (Sweller, 2024), confirmed by discussions about fading scaffolding to align with human cognitive architecture (Paas et al., 2003). Therefore, the amount of organisational structure information needs to be decided by the targeted students. Postgraduate students might need more such information in their first few assignments, but less for later assignments to develop their CT skills.

Traditionally, specifics are regarded as a necessity of assignment briefs to offer explicit instructions on formatting, structure, length/word count and other related issues (The University of Reading, 2024). This needs to be reconsidered in the GenAI-enabled era, as they are a double-edged sword. We recommend replacing specifics in assessment information with class time for students to explore and co-construct the organisational structure with the assistance of GenAI. The skills can be transferrable across modules and beyond graduation to make assessment more sustainable (Boud & Soler, 2016). Additionally, to avoid copying and pasting instructional information, which causes plagiarism, such ChatGPT use should be included in the unacceptable use of AI. As Perkins (2023) suggests, the eligible use of AI tools should be made clear to the students to reduce occurrences of unethical use of GenAI and academic misconduct. The conversation should be prioritised, as Barrett and Pack (2023) reported that 89.7% of teachers had never provided training to students on the acceptable usage of GenAI. However, knowledge of AI has been found to largely influence users' perceptions and experience, as shown in.

8 Conclusions

This study highlights how assessment design elements—such as word limits, genre conventions, cognitive demands, and structural guidance—significantly influence students' capacity for critical thinking (CT) in written coursework, as evidenced by lecturer feedback, as revealed from lecturer feedback. It reveals that every decision made about assessment design can significantly affect opportunities to develop students' CT and knowledge development.

In a GenAI-enabled world, only the tasks that requires a wide spectrum of lower and higher-order CT can foster meaningful intelligent partnerships between students and GenAI. Written assessments should also design in a way that promotes the development of disciplinary knowledge, critical positionality and creativity of applying knowledge to solve discipline-specific problems in local contexts, effectively preparing graduates for their postgraduate workplace empowered by GenAI.

To achieve these benefits, educators must critically evaluate how each design element supports CT, informed by close analyses of student work. Long-standing assessment policies and guidelines need to be scrutinised in the context where new technologies such as GenAI penetrating the education sector. There is an urgent need for targeted training in critical GenAI use to address the wide existence of unpreparedness of GenAI use among students and academics (Barrett & Pack, 2023; Gernal et al., 2024).

This study focused on one discipline that relies heavily on written assessment. Future studies can explore other disciplines where different AI tools, such as machine learning in business-related education (e.g., Al Jaghoub et al., 2024; Gilani et al., 2023; Gilani et al., 2024). Further studies could also investigate the impact of implementing the recommendations outlined here, either by replicating the research design or by gathering perspectives from students and lecturers.

This research offers practical suggestions for reforming assignment design by considering these factors for critical thinking in GenAI-assisted written coursework. The less radical changes are more feasible considering the heavy workload of academics and less risky regarding the unclear picture of the impact of GenAI on writing. Future work can build on this study by exploring alternative assessment methods, such as analysing materials, summaries or abstracts of articles, and writing a dialogue between characters with different viewpoints (Bean & Melzer, 2021), progressive feedback to refine CT and develop self-regulation (Nicol & and Macfarlane-Dick, 2006), and gamified assessment (Yasin et al., 2022; Yasin et al., 2022b). Ultimately, researching the impact of GenAI on assessment performance to develop our understanding of its affordance for learning and skill development and fostering transparent and meaningful dialogue between students and staff remains paramount (Gilbert & Maguire, 2014; Zhao, 2024).

Acknowledgements We would like to express our gratitude to the participating lecturers and students for their contribution to this project, and to Katie Gathercole for her feedback on the early version of this manuscript.

Author contributions Huahui Zhao: writing – original draft, conceptualisation, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualisation. Thi Ngoc Yen Dang: writing – review and editing, data curation, funding acquisition, investigation, methodology

Funding This work was funded by the British Council (grant numbers GES_2312). However, the views, findings, conclusions, or recommendations presented in this article are solely those of the authors and do not necessarily represent those of the British Council or its affiliated partners.

Data availability The datasets generated and analysed during the current study are not publicly available to protect participant anonymity and confidentiality, as required by the ethics approval and the conditions of informed consent.

Declarations

Competing interests The authors declare that they have no competing interests.

References

Al Jaghoub, J., Suleiman, A., Takshe, A. A., Moussa, S., Gilani, S. A. M., Sheikh, S., & Tantry, A. (2024). The Role of Innovation in Waste Management for Enterprises: A Critical Review of the Worldwide Literature. In R. El Khoury (Ed.), *Technology-Driven Business Innovation: Unleashing the Digital Advantage, Volume 1* (pp. 453–464). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-51997-0_38

Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, *57*(5), 13–18.

Anthony, L. (2023). *AntConc (Version 4.2.2)*. In Waseda University. https://www.laurenceanthony.net/software

Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education*, *20*(1), 59. https://doi.org/10.1186/s41239-023-00427-0

Bean, J. C., & Melzer, D. (2021). *Engaging Ideas: The Professor's Guide to Integrating Writing, Critical Thinking, and Active Learning in the Classroom*. Wiley.

Bearman, M., Ajjawi, R., Boud, D., Tai, J., & Dawson, P. (2023). CRADLE-Suggestes: Assessment and GenAI.

Bearman, M., Dawson, P., Boud, D., Bennett, S., Hall, M. and Molloy, E. (2016). Support for assessment practice: developing the Assessment Design Decisions Framework. Teaching in higher education. 21(5), pp.545–556.

Beckingham, S., Lawrence, J., Powell, S., & Hartley, P. (2024). *Using generative AI effectively in higher education: Sustainable and ethical practices for learning, teaching and assessment*. Routledge. https://doi.org/10.4324/9781003482918

Borji, A. (2023) A Categorical Archive of ChatGPT Failures. https://doi.org/10.48550/arXiv.2302.03494}

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment and Evaluation in Higher Education*, *41*(3), 400–413. https://doi.org/10.1080/02602938.2015.1018133

Chan, C., & Colloton, T. (2024). *Generative AI in Higher Education: The ChatGPT Effect*. Routledge.

Chiu, T. K. F. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence*, *6*, 100197. https://doi.org/10.1016/j.caeai.2023.100197

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second edition. ed.). L. Erlbaum Associates.

Cordero, J., Torres-Zambrano, J., & Cordero-Castillo, A. (2025). Integration of generative artificial intelligence in higher education: Best practices. *Education Sciences*, *15*(1), 32. https://doi.org/10.3390/educsci15010032

Coughlan, T., & Iniesto, F. (2025). What should I know? Analysing behaviour and feedback from student use of a virtual assistant to share information about disabilities. *The Internet and Higher Education*, *66*, 101002. https://doi.org/10.1016/j.iheduc.2025.101002

Deane, P. (2011). Writing assessment and cognition. *ETS Research Report Series*, *2011*(1), i–60. https://doi.org/10.1002/j.2333-8504.2011.tb02250.x. https://doi.org/https://doi.

Elbow, P. (2007). Reconsiderations: Voice in writing Again–Embracing contraries. *College English*, *70*(2), 168–188. https://doi.org/10.58680/ce20076342

Evans, C. and Harrington, P. (2024). Applying Scalability to Meet Word Count Requirements in Written Assessments. *College teaching*, pp.1–8.

Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. Research Findings and Recommendations.*. https://eric.ed.gov/?id=ED315423

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Gernal, L., Tantry, A., Gilani, S. A. M., & Peel, R. (2024). The Impact of Online Learning and Soft Skills on College Student Satisfaction and Course Feedback. In R. El Khoury (Ed.), *Technology-Driven Business Innovation: Unleashing the Digital Advantage, Volume 1* (pp. 515–528). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-51997-0_44

Gibbons, J. D and Chakraborti, S. (2020). Nonparametric Statistical Inference [Online]. CRC Press.

Gilani, S. A. M., Copiaco, A., Gernal, L., Yasin, N., Nair, G., & Anwar, I. (2023). Savior or distraction for survival: Examining the applicability of machine learning for rural family farms in the united Arab Emirates. *Sustainability*, *15*(4), 3720. https://www.mdpi.com/2071-1050/15/4/3720

Gilani, S. A. M., Tantry, A., Askri, S., Gernal, L., Sergio, R., & Mataruna-Dos-Santos, L. J. (2024). *2024//). Adoption of Machine Learning by Rural Farms: A Systematic Review*. Computing and Informatics.

Gilbert, F., & Maguire, G. (2014). *Assignment brief design guidelines: developing academic communication to enhance the student experience in assessment*. https://assignmentbriefdesign.weebly.com/

Jiang, Y. (2025). Interaction and dialogue: Integration and application of artificial intelligence in blended mode writing feedback. *The Internet and Higher Education*, *64*, 100975. https://doi.org/10.1016/j.iheduc.2024.100975

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large Language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, J., Yu, S., Detrick, R., & Li, N. (2025). Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies*, *30*(1), 1265–1300. https://doi.org/10.1007/s10639-024-12878-7

Kitamura, F. C. (2023). ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology*, *307*(2), e230171. https://doi.org/10.1148/radiol.230171

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

Laflamme, A.S. and Bruneault, F. (2025). Redefining Academic Integrity in the Age of Generative Artificial Intelligence: The Essential Contribution of Artificial Intelligence Ethics. Journal of scholarly publishing. 56(2), pp.481–509.

Lee, S. S., Li, N., & Kim, J. (2024). Conceptual model for Mexican teachers' adoption of learning analytics systems: The integration of the information system success model and the technology acceptance model. *Education and Information Technologies*, *29*(11), 13387–13412. https://doi.org/10.1007/s10639-023-12371-7

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large Language models write reflectively. *Computers and Education: Artificial Intelligence*, *4*, 100140. https://doi.org/10.1016/j.caeai.2023.100140

Lin, M. P. C., & Chang, D. (2020). Enhancing Post-secondary Writers' Writing Skills with a Chatbot.

Lo, C. K., Hew, K. F., & Jong, M.S.-y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education, 219*, Article 105100. https://doi.org/10.1016/j.compedu.2024.105100

Maranna, S., Claassen, A., Joksimovic, S., Willison, J., Parange, N., & Costabile, M. (2025). Cognitive presence and self-regulated learning: Learning transfer in an online allied health course. *Journal of University Teaching and Learning Practice, 22*(1). https://doi.org/10.53761/h2gnev81

Martín-Raugh, M., Kell, H., Ling, G., Fishtein, D., & Yang, Z. (2023). Noncognitive skills and critical thinking predict undergraduate academic performance. *Assessment & Evaluation in Higher Education, 48*(3), 350–361. https://doi.org/10.1080/02602938.2022.2073964

Merine, R., & Purkayastha, S. (2022). 11–14 June 2022). Risks and Benefits of AI-generated Text Summarization for Expert Level Content in Graduate Health Informatics. 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI).

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Paul, R., & Elder, L. (2014a). *Critical thinking: Tools for taking charge of your learning and your life* (Third edition. ed). Pearson.

Paul, R., & Elder, L. (2014b). The Miniature Guide to Critical Thinking Concepts and Tools. In (Seventh Edition ed.).

Perkins, M. (2023). Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*

Quality Assurance Agency (2021). *Higher Education Credit Framework for England: Advice on Academic Credit Arrangements*.

Quality Assurance Agency (2024). *UK Quality Code for Higher Education*.

Raz, T., Reiter-Palmon, R., & Kenett, Y. N. (2024). Open and closed-ended problem solving in humans and AI: The influence of question asking complexity. *Thinking Skills and Creativity, 53*, Article 101598. https://doi.org/10.1016/j.tsc.2024.101598

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57. https://doi.org/10.1093/applin/22.1.27

Sallam, M. (2023). The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv*. https://doi.org/10.1101/2023.02.19.23286155

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating Complexity, Accuracy, Fluency, and lexis. *Applied Linguistics, 30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Swales, J.M. and Feak, C.B. (2012). Academic writing for graduate students : essential tasks and skills 3rd edition. Ann Arbor: The University of Michigan Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Sweller, J. (2024). Cognitive load theory and individual differences. *Learning and Individual Differences, 110*, Article 102423. https://doi.org/10.1016/j.lindif.2024.102423

The University of Reading (2024). Focus on: Assessment & Feedback - Assignment briefs. In U. o. Reading (Ed.).

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wingate, U. (2012). Argument!' helping students understand what essay writing is about. *Journal of English for Academic Purposes*, *11*(2), 145–154. https://doi.org/10.1016/j.jeap.2011.11.001

Yasin, N., Gilani, S. A. M., Contu, D., & Fayaz, M. J. (2022a). Simulation-based Learning in Business and Entrepreneurship in Higher Education: A Review of the Games Available. In D. Hyams-Ssekasi & N. Yasin (Eds.), *Technology and Entrepreneurship Education: Adopting Creative Digital Approaches to Learning and Teaching* (pp. 25–51). Springer International Publishing. https://doi.org/10.1007/978-3-030-84292-5_2

Yasin, N., Majid Gilani, S. A., & Nair, G. (2022b). Dump the paper quiz"—The PERI model for exploring gamification in student learning in the United Arab Emirates. *Industry and Higher Education, 36*(5), 623–637. https://doi.org/10.1177/09504222211055067

Zhao, H. (2024). Promoting accessibility of assessment criteria: Shifting from a product- to a process- and future-oriented approach. *Teaching in Higher Education*, *29*(5), 1283–1301. https://doi.org/10.1080/13562517.2022.2129964

Zhao, H., Dang, T. N. Y., & Finlayson, N. (2024). Operationalising critical thinking in postgraduate disciplinary writing: Insights from corpus and cluster analyses of lecturer feedback. *Assessment and Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2024.2406869

Zhao, H., Dang, T. N. Y., & Finlayson, N. (2025). Education lecturers' expectations about writing proficiency: Insights from corpus analysis of teacher feedback on academic writing. *Journal of Second Language Writing, 67*, Article 101173. https://doi.org/10.1016/j.jslw.2024.101173

Authors and Affiliations

Huahui Zhao[1] · Thi Ngoc Yen Dang[1]

✉ Huahui Zhao
h.zhao1@leeds.ac.uk

Thi Ngoc Yen Dang
T.N.Y.Dang@leeds.ac.uk

[1] School of Education, Hillary Place, University of Leeds, Leeds LS2 9JT, UK