Computers & Education 257 (2027) 105767

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

GenAI-supported internal feedback intervention for writing improvement: The roles of evaluative judgment and student-AI interaction in a cluster-randomized experimental study

Shijun (Cindy) Chen [a,*], Christian D. Schunn [c], Jie Cao [d], Seyyed Kazem Banihashem [e], Dragan Gašević [a,b,f], Yizhou Fan [g]

[a] Faculty of Education, The University of Hong Kong, Hong Kong, China
[b] School of Computing and Data Science, The University of Hong Kong, Hong Kong, China
[c] Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, USA
[d] School of Education, University of North Carolina at Chapel Hill, North Carolina, USA
[e] School of Health Professions Education, Maastricht University, Maastricht, the Netherlands
[f] Centre for Learning Analytics, Faculty of Information Technology, Monash University, Melbourne, Australia
[g] Graduate School of Education, Peking University, Beijing, China

ARTICLE INFO

Keywords:
Artificial intelligence
Human-computer interface
Pedagogical issues
Teaching and learning strategies

ABSTRACT

Generative artificial intelligence (GenAI) enables scalable feedback provision. However, its effectiveness depends not only on the feedback information provided but also on how students internalize that information. While GenAI has commonly been examined as a source of external feedback, its role in supporting internal feedback remains underexplored. This 14-week cluster-randomized experimental study examined the effects of GenAI-supported internal feedback (i.e., students' seeking, evaluating, and comparing GenAI outputs and feedback to regulate their own work) on STEM undergraduates' writing performance in comparison with business-as-usual teacher feedback. The study involved 368 students (experimental: $n = 169$; control: $n = 199$) at a university in China. Data included pre- and post-test writing scores, comparison sheets, and student-AI interaction logs. Students in the GenAI-supported internal feedback condition achieved higher adjusted post-test writing performance than those in the control group. Within the experimental group, evaluative judgment (a cognitive indicator) and prompt complexity (a behavioral indicator of interaction quality) were more strongly associated with better writing performance, whereas the contribution of interaction quantity was weaker. While both quantity and quality of student-AI interactions increased over time, evaluative judgment showed limited development across the intervention. The findings suggest that GenAI can support internal feedback processes within a student-centered learning framework, while behavioral engagement with GenAI may develop without corresponding growth in evaluative judgment. Our study underscores the importance of explicitly scaffolding evaluative judgment to support the responsible and effective educational use of GenAI.

* Corresponding author.Faculty of Education, The University of Hong Kong, Pokfulam, Hong Kong, China.
E-mail address: shijunchen@connect.hku.hk (S. Chen).

https://doi.org/10.1016/j.compedu.2026.105767
Received 15 May 2026; Received in revised form 8 September 2026; Accepted 10 October 2026
Available online 17 October 2026
0360-1315/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Feedback can be viewed as external information provided by various agents, including teachers, textbooks, peers (Hattie & Timperley, 2007), and more recently, generative artificial intelligence (GenAI). This perspective is considered product-oriented, focusing on information transmission from one agent to another (e.g., teacher to students). Over the last decade, however, feedback research has shifted toward a process-oriented, student-centered conceptualization that emphasizes students' proactive engagement (Boud & Molloy, 2013; Winstone & Carless, 2020; Winstone et al., 2026). In parallel, GenAI has shown considerable promise in providing timely, personalized feedback at scale (e.g., Banihashem et al., 2024; Cao & Zhou, 2025; Debets et al., 2025). However, positioning GenAI primarily as a scalable source of external feedback risks reducing feedback to information transmission, overlooking the active role of learners in making feedback meaningful (Chen & Carless, 2026; Corbin et al., 2025; Darvishi et al., 2023).

From a learner-centered perspective, the effectiveness of GenAI feedback depends on students' active engagement with it, including seeking feedback, evaluating its quality, and using it to improve their work (Bearman et al., 2024; Winstone et al., 2026; Zhan & Yan, 2025). Such active engagement emphasizes *internal feedback*, which refers to how learners interpret, compare, and generate meaning from external information, such as feedback, exemplars, and assessment criteria (Lipnevich & Smith, 2022; Nicol, 2021; Nicol & Kushwah, 2024). Despite its centrality to learning, existing research on GenAI has largely focused on its role as an external feedback source or as a supplement to teacher and peer feedback (e.g., Noroozi et al., 2025). Comparatively little attention has been paid to how GenAI may support internal feedback generation and, critically, how such active processes translate into learning and improved achievement. Emerging qualitative evidence (e.g., Chen & To, 2026; Tam, 2025) suggests that tools such as ChatGPT or DeepSeek can stimulate internal feedback processes and support self-regulation, yet their impact on measurable learner outcomes remains insufficiently understood.

In particular, limited research has examined how GenAI can be integrated into pedagogically structured activities that support internal feedback generation processes and writing improvement. To address this gap, the present study evaluates a GenAI-supported internal feedback intervention (GenAI-IFB) that combines AI-generated feedback with structured evaluative and interactive activities, including exemplar comparison, AI feedback evaluation, and iterative interaction with GenAI. Rather than isolating the effects of AI-generated feedback alone, the study investigates the effectiveness of this integrated pedagogical approach in comparison with business-as-usual teacher feedback. Building upon Lipnevich and Smith's (2022) model, we conceptualize GenAI-IFB as comprising two interrelated dimensions of active student engagement. The cognitive dimension is performance-domain specific and captures students' evaluative judgment (EJ), referring to how effectively students evaluate products and external AI feedback on those products (Nicol, 2021; Tai et al., 2018). The behavioral dimension is GenAI-specific, encompassing both the quantity and quality of student-AI interactions (e.g., Zhan & Yan, 2025).

This study makes three contributions. First, it extends current understanding of internal feedback theory by unfolding how internal feedback operates in human-AI interaction. Second, it provides empirical evidence on how GenAI might support internal feedback within an environment intentionally designed for this purpose, and how students' engagement in such internal feedback processes is associated with their writing development over time. Third, by integrating cognitive and behavioral dimensions, the study offers a more fine-grained account of learners' agentic engagement with GenAI feedback, with implications for both GenAI-supported feedback theory and the ethical and responsible design of AI-supported learning environments.

2. Literature review

2.1. External feedback vs. internal feedback on writing

Teacher feedback, as a source of external feedback (Nicol, 2021), has long been regarded as a promising approach to enhancing students' writing performance (Cui et al., 2022). It can support writing by providing individual guidance on linguistic accuracy (Hyland & Hyland, 2006) and facilitating students' revision processes (Sheen et al., 2009). When it is timely, specific, and focused on meaning-level aspects, teacher feedback has been shown to significantly enhance learners' writing performance (e.g., Peltzer et al., 2024). However, when decontextualized or delivered in unprincipled ways, it may impede students' learning (Mao & Lee, 2025). These include providing unclear or overly critical feedback, overemphasizing language-related aspects, and misalignment with students' expectations, as synthesized in Mao and Lee's (2025) review. These teacher feedback practices are especially evident in large, mixed-ability classrooms with time and resource limitations (Chen & Carless, 2026; Ho et al., 2020). Further, teacher feedback becomes particularly ineffective when students adopt passive responses by not proactively reading, interpreting, evaluating, or using it (Carless & Winstone, 2023; Winstone et al., 2017).

Self-feedback has been proposed as an alternative to address challenges faced by teacher feedback (e.g., Panadero et al., 2019). While self-feedback and internal feedback are sometimes used interchangeably (Lipnevich & Smith, 2022; Narciss et al., 2022), we use the term *internal feedback* to emphasize the learner's active generation of evaluative information through comparison, monitoring, reflection, and judgment (Butler & Winne, 1995; Carless, 2022; Nicol, 2021; Yan & Brown, 2017). The concept of internal feedback highlights feedback as an internal process of meaning-making rather than simply self-generated comments. In this sense, we conceptualize internal feedback as the underlying cognitive process through which effective self-feedback is generated. We therefore adopt the term *internal feedback* in this study. However, because much of the existing literature uses the term self-feedback, we review that literature below to inform internal feedback and our research.

Self-feedback can increase student motivation, active participation, and writing autonomy (Yu et al., 2020), although its

effectiveness on learning outcomes varies across contexts. For example, self-feedback produced greater improvements than GenAI feedback for students' English-Chinese translations (Cao & Zhong, 2023), similar improvements for concept learning (Narciss et al., 2022), and smaller improvements in collaborative writing (Zou et al., 2023). Further self-feedback showed smaller improvements in essay quality than from peer feedback (Heil & Ifenthaler, 2026). These mixed findings suggest a need to better understand the conditions under which internal feedback processes can effectively support learning. Prior research has typically compared self-feedback with external feedback sources, overlooking how structured external support may strengthen internal feedback. In this regard, recent advances in GenAI offer new possibilities.

2.2. *GenAI-supported external feedback: needs for GenAI-supported internal feedback*

Recent research has increasingly examined the role of GenAI as a source of external feedback. GenAI systems can provide timely, scalable, and detailed feedback (Banihashem et al., 2024; Debets et al., 2025), supporting students in writing revisions and learning improvement. At the cognitive level, GenAI has been shown to empower students in peer feedback provision (Noroozi et al., 2025), facilitate cognitive engagement strategies, such as extracting key information (Zhan & Yan, 2025), and enhance students' self-monitoring during internal feedback generation (Tam, 2025). At the behavioral level, it can promote students' feedback uptake (Lu et al., 2025) and self-regulated learning behaviors, including reflective thinking and iterative revisions (Yan et al., 2026). Despite these merits, existing research has primarily focused on examining how GenAI supports external feedback sources (e.g., teacher feedback or peer feedback), or has investigated it as an independent external feedback source, while research focusing on GenAI-supported internal feedback remains scarce. This gap is particularly important given that internal feedback is central to learning and self-regulation (Butler & Winne, 1995) and may be influenced when mediated through interactions with GenAI.

Similar to teacher and peer feedback, the effectiveness of AI-generated feedback ultimately depends on how learners interpret and use it (Winstone et al., 2026). While GenAI can rapidly provide abundant feedback, the feedback can be of uneven quality, and this information does not automatically translate into learning outcomes (see Pozdniakov et al., 2026). To address quality issues, learners must proactively process and evaluate the feedback. Furthermore, to influence learning even when feedback quality is high, students must internalize it. As Yan and Carless (2022) noted, "feedback has to become internal to have an impact on student learning because all information, no matter where it is from, has to be interpreted and filtered by the student's own cognitive system before being used" (p. 1118). Yet, the effectiveness of internal feedback itself remains constrained in several ways. Learners with limited knowledge of assessment criteria or insufficient task expertise often struggle to generate meaningful internal feedback (Narciss, 2017; Panadero et al., 2019). Without appropriate support, they often demonstrate weak monitoring, set unrealistic goals, or fail to follow up through their own plans (Chou & Zou, 2020). These challenges are further shaped by differences in language proficiency, with higher-proficiency learners showing greater capacity to apply feedback for self-correction (Ko, 2022). In response to these limitations, GenAI offers new possibilities for augmenting internal feedback in a hybrid human-AI model (Banihashem et al., 2026).

We propose a model for the effective use of GenAI to support internal feedback through the GenAI-IFB intervention (see Fig. 1). This model is framed within Lipnevich and Smith's (2022) general feedback processing model, involving cognitive and behavioral elements. While their model was developed primarily for human feedback contexts, we adapted it to account for the interactive nature of AI feedback. Unlike traditional human feedback, in which students typically receive feedback from teachers or peers, GenAI requires learners to actively seek feedback through prompting and iterative interaction. Accordingly, in the GenAI-IFB model, behavioral processing captures learners' observable interactions with GenAI (e.g., prompting, requesting clarification, and refining prompts), whereas cognitive processing captures the internal mental processes involved in interpreting, evaluating, and comparing AI-generated outputs. Although these behavioral interactions necessarily involve cognition, we distinguish the two dimensions by focusing on learners' observable actions versus their mental processing of AI outputs.

Building on this framework, we conceptualize the cognitive dimension of GenAI-IFB as students' capacity to make evaluative judgments on GenAI feedback, AI-generated outputs (Bearman et al., 2024) or exemplars (Bearman & Ajjawi, 2023), and the numerical values students assigned to their own performance and the quality of others (Alemdag & Narciss, 2025; Nelson, 2018; Panadero et al., 2019), as this reflects the process of internal feedback generation (Nicol, 2021; Sadler, 2010; Tai et al., 2018). In particular, evaluating and comparing exemplars against criteria enables learners to identify quality differences, calibrate standards, and generate internal

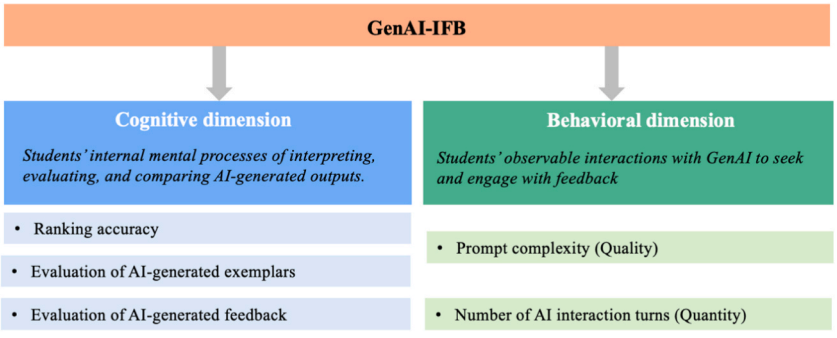

Fig. 1. GenAI-IFB model and its components.

feedback (Carless & Boud, 2018; Nicol, 2021). Empirical evidence in AI-mediated contexts supports this view. For example, Tam's (2025) qualitative study demonstrated that students actively generate internal feedback by evaluating whether to accept, adapt, or reject AI-generated suggestions. Similarly, drawing upon Nelson's (2018) framework, Xie et al. (2025) categorized students' evaluative judgment on ChatGPT feedback into hard (i.e., statistical accuracy), soft (i.e., ability to assess the significance, relevance, and value of GenAI outputs), and dynamic types (i.e., ongoing monitoring, critiquing, and adjusting).

Then we conceptualize the behavioral dimension of GenAI-IFB as both quantity and quality of students' interactions with GenAI to seek and engage with feedback. In this sense, behavioral engagement comprises a set of observable actions, including generating effective prompts, engaging in dialogue with GenAI, and eliciting GenAI feedback with interactive turns (Zhan & Yan, 2025).

2.3. Research gaps and the current study

Despite growing interest in GenAI feedback, little research has examined how pedagogically designed GenAI interventions can support students' internal feedback generation and learning outcomes. In particular, there is limited empirical evidence examining whether GenAI-IFB can match or exceed the effectiveness of established feedback approaches, such as teacher feedback. Addressing this gap is essential for determining whether GenAI can move beyond information provision to support meaningful learning processes. Accordingly, this study compares a GenAI-supported internal feedback intervention (GenAI-IFB) with business-as-usual teacher feedback (TFB). The GenAI-IFB intervention integrates AI-generated feedback with structured evaluative and interactive activities, whereas TFB represents a well-established and pedagogically grounded form of external feedback. This comparison provides a meaningful benchmark for evaluating whether a pedagogically designed intervention that supports internal feedback generation can achieve writing outcomes comparable to or better than those associated with teacher feedback.

RQ1. What is the effect of GenAI-supported internal feedback intervention (GenAI-IFB) on students' writing performance compared with the business-as-usual teacher feedback (TFB)?

Drawing on theories of internal feedback (e.g., Nicol, 2021) and the pedagogical perspectives emphasizing learner engagement with feedback, we hypothesize that students participating in the GenAI-IFB intervention, which integrates AI-generated feedback with structured activities, obtain higher post-test scores than students in the TFB condition, in which students are normally not required to evaluate teacher feedback.

H1. Students participating in the GenAI-IFB condition will obtain significantly higher post-test writing scores than those in the TFB condition.

Beyond overall effectiveness, the processes through which the GenAI-supported internal feedback intervention influences learning remain underexplored. Prior research has identified key aspects of engagement in GenAI contexts: behavioral engagement, reflected in students' interaction patterns with GenAI when seeking GenAI feedback (e.g., number of interaction turns, prompting strategy), and cognitive engagement, reflected in students' evaluative judgment of GenAI outputs. However, these dimensions have often been examined in isolation. For example, students' interaction patterns (e.g., multiple turns, refining, iterative inquiring) with GenAI are shown to be positively related to academic writing outcomes (Kim et al., 2025; Nguyen et al., 2024), while research on students' evaluative judgments of GenAI feedback has largely been qualitative (Xie et al., 2025; Chen & To, 2026) or conceptual (Bearman et al., 2024), offering limited evidence linking it to measurable learning gains.

Treating behavioral and cognitive engagement as separate processes is problematic because GenAI-supported internal feedback inherently requires students to both engage behaviorally with AI systems and critically evaluate the quality and usefulness of their outputs (Bai & Wang, 2025; Bearman et al., 2024; Xie et al., 2025; Zhan et al., 2025). Learning, therefore, depends not merely on the presence of interaction or evaluation, but on how these processes are coordinated during feedback. Addressing this gap enables a more precise understanding of how different forms of engagement with GenAI translate into meaningful learning gains, thereby informing both theoretical understanding and pedagogical feedback design.

Accordingly, the current study aimed to examine how behavioral and cognitive dimensions of GenAI-IFB jointly relate to students' writing improvement. Behavioral engagement is operationalized as both student-AI interaction quantity (i.e., number of interaction turns) and interaction quality (i.e., prompt complexity), as prior research suggests that sustained interaction reflects effortful engagement, while more sophisticated prompting indicates deeper strategic use of GenAI for feedback generation (e.g., Kim et al., 2025; Nguyen et al., 2024). Cognitive engagement is operationalized as students' evaluative judgment ability in assessing GenAI outputs, given that the capacity to critically evaluate feedback is central to internal feedback processing and learning (Carless, 2022; Nicol, 2021).

RQ2. To what extent does the effectiveness of GenAI-IFB intervention depend on students' interactions with GenAI?

H2a. Higher *prompt complexity* is associated with greater writing improvement.

H2b. Greater *interaction quantity* is associated with greater writing improvement.

RQ3. To what extent does the effectiveness of GenAI-supported internal feedback intervention depend upon students' evaluative judgment ability?

H3. Student evaluative judgment ability is associated with greater writing improvement.

Understanding the role of GenAI-supported internal feedback in learning requires distinguishing between performance outcomes

and the processes that generate them. While writing performance is often used as an indicator of learning outcome, measurable task performance and learning are not equivalent constructs (Soderstrom & Bjork, 2015). A comprehensive account, therefore, requires examining not only whether GenAI-supported internal feedback intervention improves writing performance, but also how it influences the key cognitive and behavioral processes that underpin learning. In the present study, these processes are operationalized as students' evaluative judgment (cognitive engagement) and AI interaction (behavioral engagement). However, it remains unclear whether these capacities develop through repeated engagement with GenAI-supported internal feedback or require more explicit instructional support. Examining their development over time is therefore important for understanding whether GenAI-IFB intervention promotes transferable learning beyond immediate performance (Yan et al., 2025). Because prior research provides limited evidence on the development of AI interaction and evaluative judgment in a GenAI-supported internal feedback context, RQ4 was formulated as an exploratory research question.

RQ4. To what extent do AI interaction and evaluative judgment improve with practice?

3. Methodology

3.1. Context and participants

The 14-week-long experiment was conducted in a compulsory English course at a normal university (teacher education) in China. Ethical approval has been obtained from the first author's institution. Informed consent was secured from both the teacher and student participants, with assurance that all data would be used confidentially. The course named *Critical Thinking and Academic Writing* met weekly for 90-min sessions and was divided into six sections that met at different times each week. The course was delivered in the second semester across multiple class sections, each enrolling between 56 and 72 students. All student participants were instructed by the same course instructor who had over 10 years of experience teaching English writing and 15 years of experience teaching English.

Participants were 368 undergraduate students (Age: 18-22; 251 females, 117 males) in their second semester of the second year of STEM-related education majors (e.g., science, artificial intelligence, medicine, engineering). They learned English as a second language and had passed the Chinese CET (College English Test) Band-4, which aligns with an English proficiency level ranging from B1 to B2 based on the Common European Framework of Reference for Languages (CEFR).

3.2. Experimental design and procedures

The study adopted a cluster-randomized experimental design. To minimize disruption to regular classroom teaching, the intervention was implemented at the class section level rather than through individual student assignment. Accordingly, three class sections experienced the GenAI-supported internal feedback intervention (GenAI-IFB; 3 sections, $n = 169$), whereas three class sections received the teacher feedback condition (TFB; 3 sections, $n = 199$). The experimental condition was designed to explicitly scaffold students' internal feedback generation through structured activities. In contrast, the TFB condition reflected typical classroom

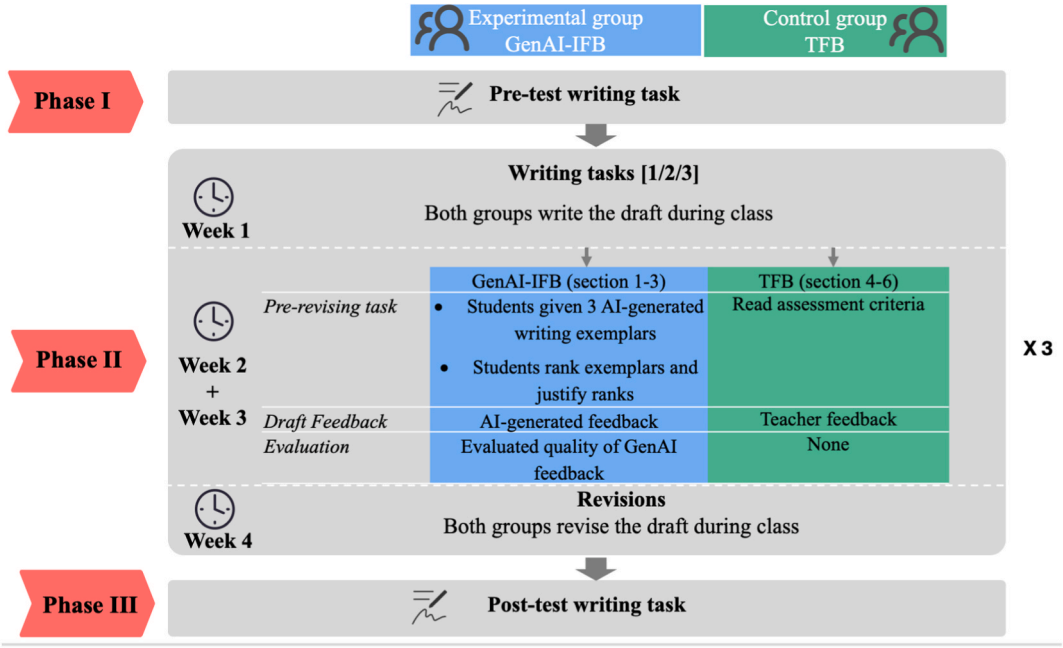

Fig. 2. In-class feedback and revising processes across conditions that were implemented across different sections of the course.

practice, in which students engaged in feedback uptake without a required evaluation step. This design preserved ecological validity while enabling a contrast between explicit self-generated feedback with GenAI support and naturally occurring external feedback uptake.

The intervention spanned three writing tasks across the semester, each lasting 4 weeks. For each task, students produced an initial draft and a subsequent revision, resulting in six pieces of writing per student (three drafts and three revised essays). The experiment involved three phases. In Phase I, students completed a pre-test writing task to verify that participants in each condition were initially equivalent. The task's assessment criteria (see Appendix A) were adapted from the CET-Band 6 rubric, and these criteria were applied consistently across all writing tasks throughout the study. The course teacher explained these criteria in class before the pre-test writing task was submitted to ensure that all students had a shared understanding of the criteria.

In Phase II, students completed three writing tasks, with key stages of drafting, feedback engagement, and revision conducted in class to ensure consistency in task conditions and independent student work. The three writing tasks included two argumentative tasks and one chart description task, with comparable task requirements and similar amounts of instructional information provided in the prompts. The required length of all tasks was 150-250 words.

In week 1 (Phase II), students wrote a first draft in class for 30 min, under exam-like conditions. No AI tools, online sources, or textbook references were permitted. Drafts were submitted immediately after completion. The instructor was present to guide task engagement and ensure independent work. In weeks 2 and 3, the two groups experienced different conditions under teacher monitoring during scheduled class sections. (see Fig. 2).

In week 2 (Phase II) of the GenAI-IFB condition, the instructor used a GenAI model (DeepSeek R1) to generate three writing exemplars for each task at high, medium, and low quality levels on the assessment criteria (Appendix B). To ensure the validity of the quality categories, the AI-generated exemplars were independently reviewed by two experienced university writing instructors and the first author, who also had years of university English teaching experience. Using the CET-6 writing rubric as the evaluation framework, the reviewers confirmed that the exemplars appropriately represented high-, medium-, and low-quality performance before they were used in the intervention. Students received these exemplars in class and were required to fill a GenAI outputs comparison sheet (see the upper part of Fig. 3), which guided their engagement through a structured, two-stage internal feedback generation process. Students first ranked the exemplars and their own writing from high to low quality (e.g., B/C/my writing/A). Second, they justified their rankings by referring explicitly to the content, organization, and language in relation to the assessment criteria (e.g., "I ranked essay B first because it demonstrated stronger coherence in content").

In week 2 of the TFB condition (Phase II), students were asked to carefully read the assessment criteria during class.

In week 3, students in the GenAI-IFB condition engaged with an AI tool (DeepSeek R1) to generate feedback on their own drafts. In this stage, students were instructed to prompt the AI with targeted questions informed by issues identified during their prior evaluations (e.g., How can I improve the coherence of paragraph 2 in my writing?). To ensure that students examined the AI feedback and to document their reactions to it, students were asked to give a summary of their evaluation of the received feedback (see the lower part of

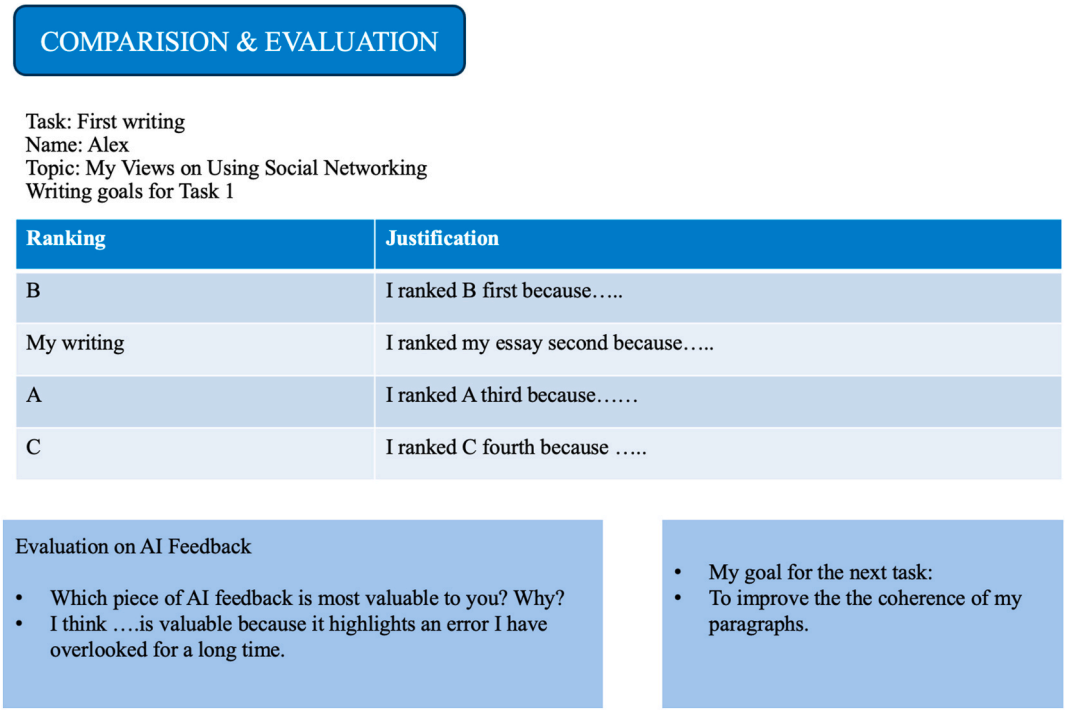

Fig. 3. GenAI outputs comparison and AI feedback evaluation sheet in GenAI-IFB condition.

Fig. 3).

In week 3 of the TFB condition, students received written feedback from their instructor, which was guided by the same rubrics and best practices in effective teacher feedback (Nicol & Macfarlane-Dick, 2006). Teacher feedback typically included: general comments on content development, evaluation of organization and coherence, sentence-level language suggestions, and prioritized revision focus. Students were asked to carefully read the received feedback, but they were not asked to evaluate the teacher feedback in a naturalistic setting.

In week 4 (Phase II), students in both groups revised their drafts in class with teacher monitoring. Students in the GenAI-IFB group revised the draft using the feedback generated through the GenAI-supported internal feedback process, whereas the TFB group revised their draft based on explicit feedback provided by the teacher. Note that the conditions were not meant to be equivalent during Phase II in the type of information received or required reflection. Rather, classroom time was kept constant and a specific AI-based approach was contrasted against typical practice (i.e., teacher-only feedback), where teacher feedback is provided under naturalistic conditions rather than being standardized or controlled for experimental equivalence.

In Phase III, two weeks after finishing the third writing task, students completed the post-test, which was a new argumentative writing task assessed using the same criteria as the pre-test. The instructor considered the post-test more demanding for this student cohort. This judgment was based on students' prior curricular exposure: they had greater familiarity with the general content area addressed in the pre-test, whereas the post-test addressed a less familiar topic in the regular curriculum.

3.3. Measurements

The dataset consisted of pre- and post-writing test scores for all students; six writing samples per student with three initial drafts and three revised versions; GenAI outputs comparison and AI feedback evaluation sheets; and student-AI interaction log data. This dataset was used to capture students' writing performance, evaluative judgment skills, and GenAI interactions. Interrater reliability was assessed in terms of Cohen's κ by recording a random sample of 100 comparison sheets and 100 student-AI interaction logs.

Table 1
Evaluative judgment and interaction complexity coding schemes, including levels, definitions, and examples.

Construct	Subdimensions	Score	Definitions	Examples
Evaluative Judgment: Exemplar Evaluation	Global description	1	Surface description or evaluation of the text without reference to specific features, explicit criteria, or support reasoning	*The essay is good.*
	Feature evaluation without evidence	2	Listing isolated features as reasoning for a ranking	*I ranked essay A first because it has rich vocabulary, clear organization, diverse argument, and few language errors.*
	Feature evaluation with evidence	3	Listing isolated features with specific examples for a ranking	*There are many language errors, and the expression is relatively simplistic. Numerous grammatical errors are present, such as 'They think tomorrow is another day, can do later'.*
	Integrated quality judgment	4	Synthesizing features into an overall judgment of text quality or effectiveness	*I ranked essay A first because it demonstrates varied sentence structures, employs advanced vocabulary, and makes frequent use of cohesive devices, which makes the text flow more smoothly.*
	Criteria-based comparison	5	Explicitly weighing criteria, comparing two qualities or essays, and clearly stating which is more important for the ranking	*I ranked my essay second because it is on topic, clearly expresses ideas, and is coherent. Compared with essay C, it uses less advanced vocabulary and sentence structures but contains fewer errors.*
Evaluative Judgment: GenAI Feedback Evaluation	General intuition	1	Evaluating GenAI as generally useful without specific justification	*AI feedback is helpful.*
	Task-referenced	2	Evaluating GenAI feedback by aligning its usefulness with the criteria of the current task	*AI feedback helps me work on coherence in this writing*
	Learning-referenced	3	Evaluating GenAI feedback by aligning its usefulness with long-term English writing improvement	*AI feedback pointed out a subject-verb agreement error, which is a problem I often overlook, and this helped improve my grammatical accuracy.*
Evaluative Judgment: Ranking Accuracy	Wrong sequence	0	Ranked the low exemplars as the highest quality	/
	Misidentified sequence	1	Failed to identify the top exemplar	/
	Misplaced sequence	2	Correctly identified the high one exemplar, but swapped the other two	/
	Correct sequence	3	Fully correct ordering	/
GenAI Interaction: Prompt Complexity	Simple feedback seeking	1	Prompts request general feedback without specifying evaluation criteria, writing dimensions, or intended revisions	*Comment on my writing*
	Specific feedback seeking	2	Prompts request targeted feedback on a particular aspect of writing, such as organization, coherence, or language accuracy	*Please explain how I can improve my paragraph organization*
	Critical dialogue	3	Prompts reflect critical, dialogic engagement with GenAI feedback	*Is there another way to improve this paragraph without changing the main idea?*

3.3.1. GenAI interaction

For students in the GenAI-IFB condition, their GenAI interactions in feedback seeking were measured in terms of two indicators: prompt complexity and interaction quantity (see Table 1).

Prompt complexity (H2a) captures the qualitative sophistication of each student's behavioral engagement with GenAI. Prompts were coded on a three-level ordinal scale: simple feedback seeking; specific feedback seeking; and critical dialogue (see Table 1). The coding scheme draws on literature on prompt literacy and AI literacy (Federiakin et al., 2024; Kim et al., 2025). For each task, students' highest prompt complexity level was retained, and scores were then averaged across the three tasks. Because students with more interaction turns may have more opportunities to produce higher-level prompts, we examined the robustness of our findings using two alternative measures : (1) mean prompt complexity, calculated by averaging levels of all prompts within each task, and (2) the proportion of higher-level prompts, defined as the proportion of prompts coded as Levels 2 or 3 within each task. For both measures, task-level scores were then averaged across the three tasks.

Interaction quantity (H2b) reflects the quantity of behavioral engagement and was operationalized as the total number of GenAI-student dialogue turns across the three tasks. Following typical practice in research on human-AI dialogue (e.g., Jiang et al., 2023), one turn was defined as one message generated by either the student or GenAI. Prompt complexity and interaction quantity were moderately correlated with the number of interaction turns (Spearman's $\rho = 0.57$, $p < .001$), while variance inflation factors were low (VIFs = 1.3 for prompt complexity and 1.2 for interaction turns), indicating limited multicollinearity.

3.3.2. Evaluative judgment

Evaluative judgment (H3) in the GenAI-IFB condition was assessed in relation to students' cognitive engagement with GenAI. Drawing upon prior conceptualization of evaluative judgment (e.g., Nelson, 2018; Sadler, 2010), three indicators were used: (a) evaluation of AI exemplars, (b) ranking accuracy, and (c) evaluation of AI feedback. These indicators capture complementary aspects of students' ability to interpret, compare, and evaluate quality. Correlations among these indicators in the current study were small (Pearson's $r = 0.15$-0.25, Spearman's $\rho = 0.14$-0.25, all $ps < 0.01$). The significant positive associations suggest some commonality among the three indicators as components of EJ, whereas their small magnitude indicates that they capture relatively distinct aspects of this broader capacity. For example, ranking accuracy represents a hard form of evaluative judgment, evaluating AI-generated exemplars reflects a soft form, and evaluating AI-generated feedback represents a dynamic form involving ongoing monitoring, critiquing, and adjusting (e.g., Nelson, 2018; Xie et al., 2025). We therefore retained the three indicators separately rather than combining them into a single composite measure. Table 1 presents the coding schemes with definitions and examples.

First, exemplar evaluation reflects the depth of students' evaluation of AI-generated exemplars. It was coded using a five-level hierarchical scheme: global description; feature evaluation without evidence; feature evaluation with evidence; integrated quality judgment; or criteria-based comparison (see Table 1). This coding scheme was informed by prior literature (e.g., Sadler, 2010) and refined inductively through analysis of data.

Second, ranking accuracy captures students' ability to discriminate between levels of quality. Students were asked to rank three AI-generated exemplars, and accuracy was scored on a four-point scale (0-3), ranging from incorrect ordering to fully correct sequencing.

Third, GenAI feedback evaluation reflects the depth of focus. It was coded using a three-level framework: general intuition, task-referenced, and learning-referenced (see Table 1). The coding scheme was developed deductively based on Sadler's (2010) theory of capability in complex appraisal, and inductively through iterative coding and constant comparison across students and tasks. Again, the highest level was assigned to reflect students' maximum demonstrated evaluative capacity.

Interrater reliability was calculated based on independent coding by the first author and the third author on 20% of the dataset. Cohen's κ for all variables was satisfactory: $\kappa = 0.79$ for maximum prompt complexity, $\kappa = 0.76$ for individual prompt-level coding; $\kappa = 0.83$ for exemplar evaluation, $\kappa = 0.89$ for GenAI feedback evaluation. Unlike the other evaluative judgment indicators, ranking accuracy was based on an objective comparison between students' rankings and predetermined reference ordering of the validated exemplars. Following the establishment of satisfactory agreement, the third author coded the remaining data using the agreed coding scheme.

3.3.3. Quantitative writing scores

Students' writing performance was assessed through the pre-test, the first drafts of the three writing tasks, and the post-test, all completed in class without AI assistance. All writing tasks were scored using the *iWrite 3.0* system, an automated writing evaluation system. The instructor administered the scoring using the assessment criteria. The platform used a 15-point scale as its default setting, aligning with the CET-Band 6 framework adopted in the study. Automated scoring provided a standardized scoring procedure across classes and helped reduce teacher scoring fatigue and potential bias, given the large volume of student submissions. This system has demonstrated high reliability and validity in essay scoring (Liu, 2022). To enhance scoring accuracy, the instructor conducted a post-hoc moderation of the *iWrite* scores for each task, reviewing them against the criteria and making adjustments where necessary; these revised scores were used in analyses. To evaluate the potential influence of post-hoc moderation and possible condition-related bias, we compared the original automated scores with final moderated scores and conducted an independent blind re-scoring of a subsample (see Appendix C). The proportion of adjusted essays did not differ significantly between conditions (Fisher's exact test, $p = .82$). Neither the mean signed adjustment ($t = -0.43$, $p = .67$), nor the absolute magnitude of adjustment ($t = -0.21$, $p = .84$) differed significantly between conditions. Agreement between the original automated scores and the final moderated scores was excellent (ICC = 0.95 for pre-test and 0.94 for post-test). The pre- and post-test essays of 64 students (32 per condition; approximately 20% of the analytical sample) were independently re-scored by a rater blind to condition membership. Agreement between the blind and final moderated scores was high (ICC = 0.96 for pre-test and 0.98 for post-test). The discrepancy between blind and final

moderated post-test scores did not differ between the GenAI-IFB ($M = 0.14$, $SD = 0.44$) and TFB conditions ($M = 0.12$, $SD = 0.30$), $t = 0.23$, $p = .82$.

3.4. Data analysis

Statistical analyses were conducted using *R 4.4.1*.

3.4.1. Analysis of writing pre- and post-test scores (RQ1, H1)

Students in the GenAI-IFB condition ($n = 141$; 83.4%) and in the TFB condition ($n = 174$; 87%) were included if they completed the pre- and post-tests and all three writing tasks. The sample size was determined by the number of students enrolled in the course sections and task completion rather than through a priori power analysis, reflecting the practical constraints of conducting a classroom-based intervention. To evaluate the adequacy of the achieved sample, a sensitivity power analysis indicated that the study had 80% power to detect an effect size of approximately Cohen's $d = 0.32$ or larger ($\alpha = .05$, two-tailed).

A complete-case (listwise deletion) approach was adopted to ensure consistency across repeated measures and comparability of within-student outcomes (Newman, 2014). Given the relatively high and comparable completion rates across conditions, missing data were minimal, and thus no imputation procedures were applied. Baseline attrition analyses indicated that participants who withdrew did not differ significantly from those who completed the study in pretest writing performance, $t(68.7) = -1.1$, $p = .29$. Attrition was also unrelated to instructional condition, $\chi^2(1) = 0.7$, $p = .41$, or class membership, $\chi^2(5) = 7.2$, $p = .21$. There was no significant difference between completers and dropouts in either the GenAI-IFB group ($p = .27$) or the TFB group ($p = .50$). These results suggest that attrition was unlikely to introduce systematic bias.

To address RQ1, which examines differences in writing performance between conditions, a two-step analysis was conducted. Because students were nested in six intact class sections (3 in GenAI-IFB, 3 in TFB) per week, mixed-effects models with class section specified as a random intercept were used to account for the hierarchical data structure. First, a mixed-effects Analysis of Variance (ANOVA) was used to test baseline equivalence by condition in pre-test writing scores. Second, a mixed-effects Analysis of Covariance (ANCOVA) was run on post-test writing scores as a function of condition, controlling for pre-test scores and again including a random intercept for sections. Model assumptions were checked by examining residual plots. The homogeneity of regression slopes assumption was examined by testing the Condition × Pre-test interaction. Adjusted marginal means and 95% confidence intervals were used to interpret condition differences. The estimated between-section variance was negligible (Variance = 0.00), indicating little detectable between-section variability after controlling for students' pre-test performance. However, because only six class sections were included, this estimate should be interpreted with appropriate caution. Cohen's *d* was calculated for post hoc comparisons based on adjusted means to estimate the standardized magnitude of between-group differences (Lakens, 2013). Model-level explanatory power was evaluated using R^2-type indices where applicable.

3.4.2. Testing the relationship of GenAI feedback interactions and evaluative judgments with improvements in writing scores (RQ2, RQ3)

Analyses were conducted within the GenAI-IFB group to examine whether AI interaction quality (RQ2) and evaluative judgment (RQ3) predicted learning outcomes.

A series of mixed-effects ANCOVA regression models were fitted with post-test writing score as the dependent variable, pretest score as a covariate, and a random intercept for class section to account for section effects. Three models were specified. Model 1 examined the effects of AI interaction variables, including prompt complexity (H2a) and interaction quantity (H2b). Model 2 examined the effects of evaluative judgment variables, including exemplar evaluation, ranking accuracy, and AI feedback evaluation (H3). Model 3 integrated both sets of predictors to assess their unique contributions when considered simultaneously. This approach enabled examination of the contribution of each predictor set and whether their effects were reduced in the integrated model. To assess the robustness of the RQ2 findings to the operationalization of prompt complexity, Models 1 and 3 were re-estimated using mean prompt complexity and proportion of higher-level prompts as alternatives to the maximum-based measures, while retaining the same predictors and controls.

To support interpretation of potential non-linear patterns, prompt complexity and evaluative judgment variables were additionally categorized into low, medium, and high levels using tertile-based binning for visualization and post-hoc comparison. Adjusted post-test means were estimated using estimated marginal means. Post-hoc pairwise comparisons were conducted using Tukey-adjusted comparisons to control for multiple testing. Cohen's *d* was calculated to estimate the effect size.

Before the regression analyses, potential multicollinearity among the predictors was assessed using variance inflation factors (VIFs), and assumptions concerning residual normality and homoscedasticity were examined. Standardized coefficients were reported to facilitate the interpretation of effect sizes. Multicollinearity was low across the primary models (VIFs = 1.20 to 1.98). Residuals approximated normality in Models 2 and 3 (Shapiro-Wilk tests: $W = 0.99$, $p = .45$; $W = 0.99$, $p = .65$), whereas Model 1 showed a slight deviation ($W = 0.98$, $p = .01$). Breusch-Pagan tests indicated homoscedastic residuals for Models 1 and 3 ($ps > 0.1$), whereas Model 2 showed a slight departure from homoscedasticity (BP = 12.2, $p = .03$). Diagnostic checks were also conducted for the sensitivity models (see Appendix D).

3.4.3. Analysis of GenAI interaction and evaluative judgment development over time (RQ4)

To address RQ4, the development of GenAI interaction and evaluative judgment across tasks was examined. The distributions of the variables were first inspected using histograms. Changes over time were analysed using repeated-measures ANOVA, with task specified as a within-subject factor. Where sphericity assumptions were violated, Greenhouse-Geisser corrections were applied. Mean values

with standard errors were plotted to visualise trends across tasks. As a sensitivity check, changes in prompt complexity were also examined using mean prompt complexity and the proportion of higher-level prompts. Effect sizes for the repeated measures ANOVA tests were reported using partial eta squared (η_p^2).

4. Findings

4.1. Comparison between writing scores in GenAI-IFB and TFB conditions (RQ1, H1)

As shown in Fig. 4, there were no significant differences in writing performance at pre-test (GenAI-IFB: $M = 9.70$, $SD = 1.44$; TFB: $M = 9.82$, $SD = 1.65$, $d = -0.08$, $p > .48$, 95% CI [−0.30, 0.14]). At post-test, students in the GenAI-IFB condition achieved higher writing scores than those in the TFB condition (GenAI-IFB: $M = 10.2$, $SD = 1.69$; TFB: $M = 9.44$, $SD = 1.92$). After controlling for pre-test performance, this between-condition difference remained significant ($\beta = 0.87$, SE = 0.19, $t(312) = 4.46$, $p < .001$, 95% CI [0.50, 1.24]), corresponding to a moderate effect size (Cohen's $d = 0.52$, 95% CI [0.15, 0.86]).

Students in the GenAI-IFB condition showed improvement in writing performance compared with students in the TFB condition, in which they were not required to evaluate teacher feedback explicitly (see Fig. 4). Because the pre- and post-test involved different writing tasks, and their relative difficulty was not independently calibrated, raw within-condition score changes should be interpreted cautiously. Therefore, H1 was evaluated primarily through the adjusted post-test comparison, which accounted for baseline writing performance. Students in the GenAI-IFB condition achieved significantly higher adjusted post-test scores than those in the TFB condition, a pattern consistent with H1.

4.2. Relationships between AI interactions, evaluative judgments, and writing performance (RQ2, RQ3)

Table 2 presents the results of regression models testing different factors shaping writing improvements in the GenAI-IFB condition: Model 1 examines the GenAI interaction (RQ2) variables alone, Model 2 examines the evaluative judgment variables (RQ3) alone, and Model 3 includes all predictors simultaneously to test the robustness of findings. These models accounted for substantial amounts of variance in writing outcomes.

4.2.1. AI interactions and writing improvement (RQ2)

Across Models 1 and 3, prompt complexity was consistently associated with higher posttest scores after controlling for pretest performance, and it was the largest predictor of gains. Therefore, H2a was confirmed: students who generated more complex prompts when interacting with GenAI tended to demonstrate greater writing improvement. Fig. 5 presents adjusted post-test writing scores across low, medium, and high levels of prompt complexity, controlling for pretest, section, and evaluative judgment predictors.

Pairwise comparisons based on estimated marginal means with Tukey adjustment indicate that the high-level group significantly outperformed both low-level groups ($MD = -1.26$, $SE = 0.21$, $t(131) = -6.07$, $p < .001$, $d = 1.64$) and medium groups ($MD = -0.95$, $SE = 0.18$, $t(131) = -5.19$, $p < .001$, $d = 1.23$). The difference between the low and medium groups was not statistically significant ($MD = -0.31$, $SE = 0.18$, $t(131) = -1.77$, $p = .19$). These findings suggest that the benefits of prompt complexity emerge primarily at higher levels.

In the primary analyses, interaction quantity was not significantly associated with post-test writing performance in either Model 1 or 3. However, sensitivity analyses using alternative operationalizations of prompt complexity revealed a more nuanced pattern (see Appendix D). Prompt complexity remained significant when operationalized as either mean prompt level (Model 1: $\beta = 0.34$, $p < .001$; Model 3: $\beta = 0.36$, $p < .001$) or the proportion of high-level prompts (Model 1: $\beta = 0.29$, $p < .001$; Model 3: $\beta = 0.32$, $p < .001$), providing robust support for H2a. Interaction quantity became significant in the corresponding Model 1 analyses using mean prompt complexity ($\beta = 0.22$, $p < .01$) and the proportion of higher-level prompts ($\beta = 0.26$, $p < .001$). Still, it remained nonsignificant in the integrated Model 3 specifications. Accordingly, evidence for H2b was conditional rather than consistently supported: the quality of

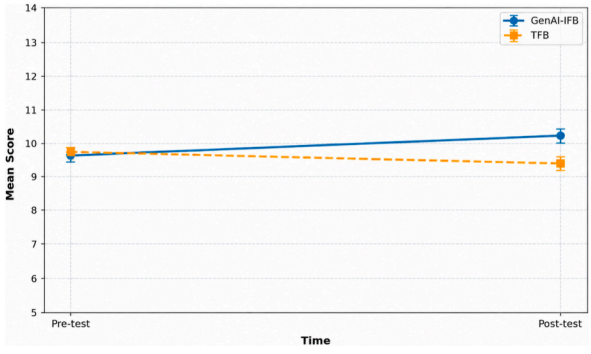

Fig. 4. Mean (with SE bars) pre- and post-test writing scores within each condition. The y-axis reflects the observed score range (5-14); the full writing scale ranges from 0 to 15.

Table 2
Regression models predicting posttest writing scores from interaction quality and evaluative judgment.

Predictor	Model 1		Model 2		Model 3	
	β	*SE*	β	*SE*	β	*SE*
AI Interaction						
Prompt Complexity	0.45***	0.08			0.38***	0.05
Interaction Quantity	0.13	0.08			0.04	0.05
Evaluative Judgment						
Exemplar evaluation			0.35***	0.06	0.30***	0.05
Ranking accuracy			0.30***	0.06	0.26***	0.05
AI Feedback evaluation			0.22***	0.06	0.19***	0.04
Controls						
Pretest	0.64***	0.07	0.66***	0.05	0.74***	0.04
Section 2	0.14†	0.07	0.00	0.06	0.02	0.04
Section 3	0.11	0.08	0.00	0.06	0.01	0.05
Model statistics						
N	141		141		141	
R^2	0.52		0.68		0.82	
Adjusted R^2	0.51		0.66		0.81	
F(df)	29.58 (5135)		46.40 (6134)		75.08 (8132)	

Note. †$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$. *SE* = standard errors.

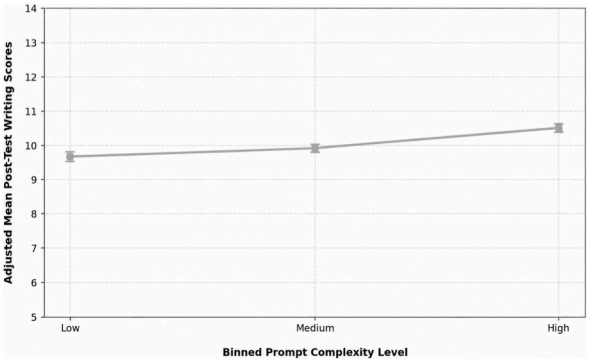

Fig. 5. Adjusted mean post-test writing scores (with SE bars) as a function of prompt complexity, with controls for pre-test, section, and evaluative judgment predictors. Writing score range: 5-14 (full scale: 0-15).

student-AI interaction showed a robust association with writing performance, whereas the contribution of interaction quantity varied depending on how prompt complexity was operationalized and the predictors included in the model.

4.2.2. Evaluative judgment and writing improvement (RQ3)

Across Models 2 and 3, all three evaluative judgment components (exemplar evaluation, ranking accuracy, and GenAI feedback evaluation quality) significantly predicted improved writing outcomes (RQ3). Fig. 6 shows adjusted mean post-test scores across low,

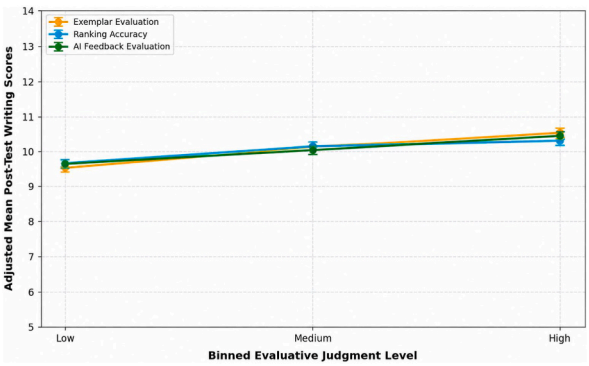

Fig. 6. Adjusted mean post-test writing scores (with SE bars) as a function of evaluative judgment scores, with controls for pre-test, section, and prompt complexity. Writing score range: 5-14 (full scale: 0-15).

medium, and high evaluative judgment levels, enabling clearer interpretation of potential non-linear patterns. These findings indicate that students with stronger evaluative judgment across multiple dimensions demonstrated greater gains in writing proficiency, particularly in separating low vs. medium evaluative judgment. Note that the post-test means at the low evaluative judgment levels were similar to those at pre-test, suggesting that students with low evaluative judgment essentially made no measurable progress in writing. These results also provide predictive validity evidence for the use of these measures of evaluative judgment.

4.3. Development of AI interactions and evaluative judgment over time (RQ4)

Histogram analyses revealed that students spanned a broad range of values on the scales, but with different distribution patterns (Fig. 8). Prompt complexity and the number of AI interaction turns were positively skewed, with many cases at the low end of the scales. That is, many students made limited use of the AI. By contrast, exemplar evaluation and AI feedback evaluation were normally distributed with means and medians around the middle of the scales. Ranking accuracy was negatively skewed, with students clustering at perfect exemplar rankings. This pattern suggests a potential ceiling effect for ranking accuracy, leaving relatively limited scope for further improvement on this indicator compared with other EJ measures.

Prompt complexity increased progressively across the tasks, $F(1.90, 265.70) = 25.85$, $p < .001$, $\eta_p^2 = 0.156$ (Fig. 7). This developmental pattern was robust to alternative operationalizations of prompt complexity: significant increases were also observed using mean prompt level, $F(1.84, 256.92) = 26.30$, $p < .001$, $\eta_p^2 = 0.159$, and the proportion of higher-level prompts, $F(1.87, 261.32) = 24.59$, $p < .001$, $\eta_p^2 = 0.150$, with all pairwise task comparisons remaining significant ($ps \leq 0.015$). Interaction quantity also showed a significant but more modest increase across tasks, $F(1.87, 261.51) = 6.49$, $p = .002$, $\eta_p^2 = 0.044$.

In contrast, evaluative judgment measures did not show comparable changes over time. No significant changes were observed for exemplar evaluation, $F(2, 280) = 0.12$, $p = .883$, $\eta_p^2 = 0.001$. AI feedback evaluation, $F(2, 280) = 0.96$, $p = .384$, $\eta_p^2 = 0.007$, or ranking accuracy, $F(2, 280) = 2.51$, $p = .083$, $\eta_p^2 = .018$. Thus, addressing RQ4, student-AI interaction became progressively more sophisticated and somewhat more extensive with practice, whereas their evaluative judgment remained relatively stable across the three tasks.

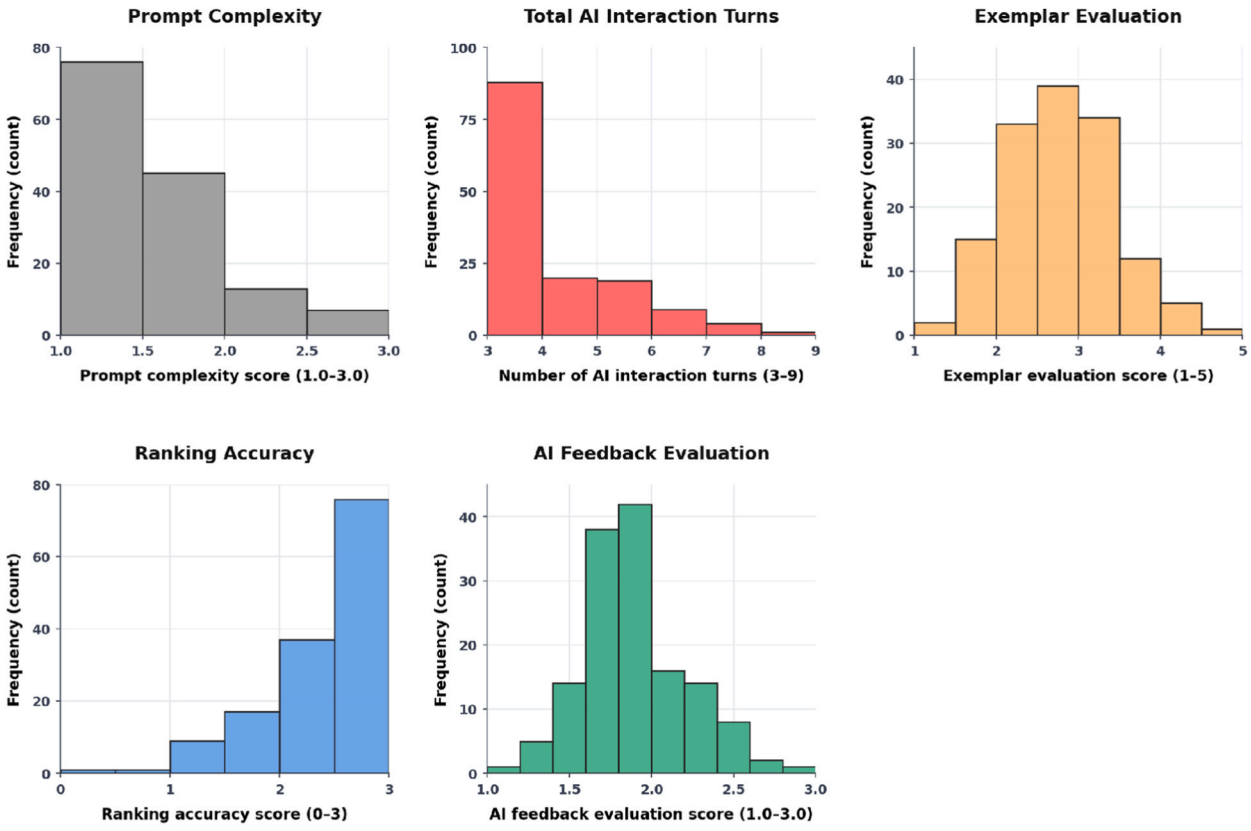

Fig. 7. Histograms of prompt complexity, AI interaction, and Evaluative judgment measures.

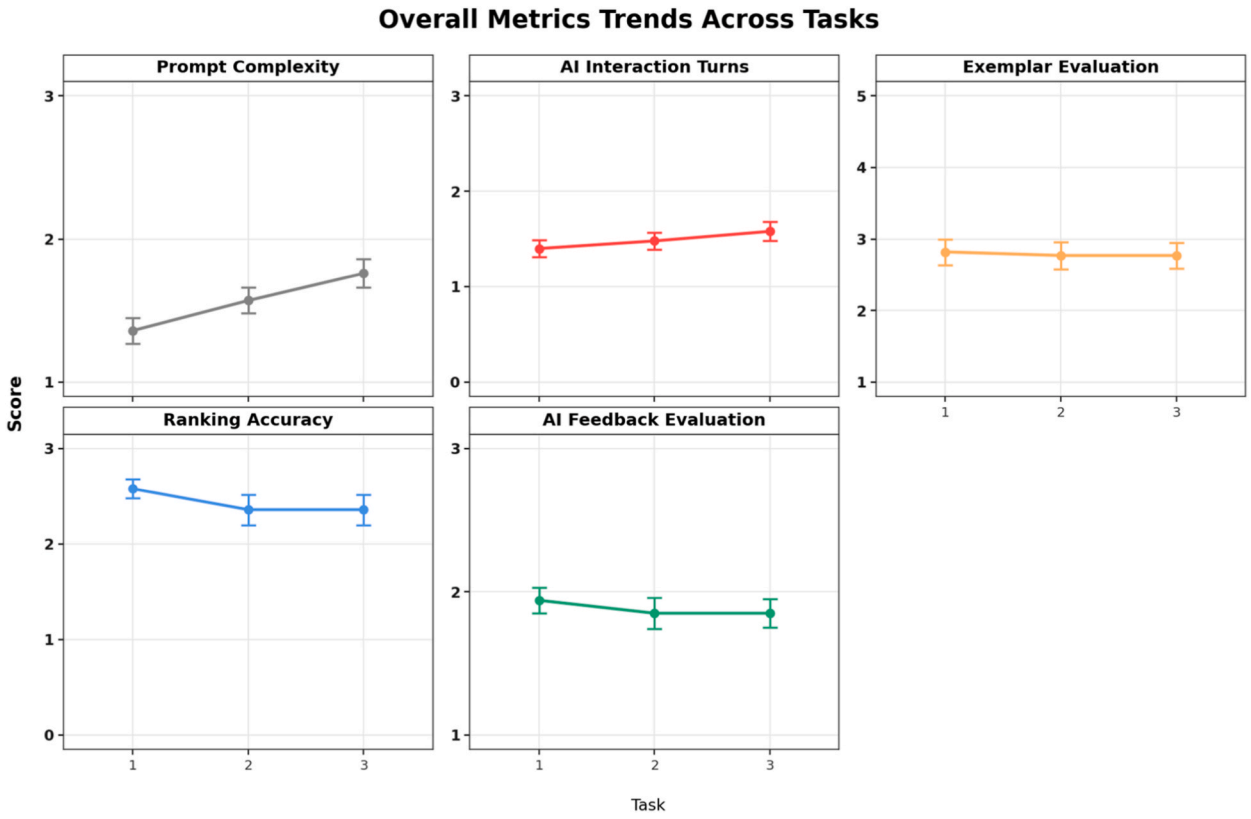

Fig. 8. Mean levels (and SE bars) across practice tasks in prompt complexity, AI interaction, and evaluative judgment measures.

5. Discussion

5.1. Interpretation of findings

The greater improvements in writing performance for students engaging in the GenAI-IFB intervention than business-as-usual TFB condition (RQ1, H1) extend prior research by demonstrating that GenAI's effectiveness lies not simply in its capacity to generate feedback, but in how it is integrated into students' internal feedback processes (e.g., Chen & Carless, 2026; Chen & To, 2026; Tam, 2025; Zhan & Yan, 2025). The different learning trajectories observed across conditions indicate that feedback design may be as important as the feedback source itself. Our findings support an agentic view of feedback, in which students iteratively generate, seek, evaluate, and apply feedback to regulate their learning (Nicol, 2021; Panadero et al., 2024; Sadler, 2010; Winstone et al., 2017). Within this process, GenAI serves as an on-demand text generator and responsive feedback partner that supports continuous monitoring and reflection. These findings also align with earlier conclusions that well-scaffolded internal feedback can enhance performance (e.g., Narciss et al., 2022).

At the behavioral level, the findings highlight the importance of student-AI interaction quality (RQ2). Prompt complexity, reflecting students' ability to articulate, refine, and critically engage with queries, was consistently associated with writing performance (H2a). In contrast, interaction quantity showed less consistent associations (H2b): although it was positively associated with writing performance in the sensitivity versions of Model 1, this association was not significant when EJ was included in the integrated models. This suggests that the quality of student-AI interaction provides a more robust indicator than the sheer number of interaction turns. In line with prior research, more sophisticated prompting may help students elicit more relevant and targeted feedback (e.g., Kim et al., 2025; Nguyen et al., 2024), thereby facilitating feedback uptake. This interpretation aligns with perspectives emphasizing that productive feedback engagement involves evaluation, comparison, and refinement rather than interaction frequency alone (cf. Panadero et al., 2019). Findings for RQ2 suggest that behavioral engagement in GenAI-IFB should be considered in terms of both its qualitative sophistication and quantitative extent, with the former showing a more consistent association with writing performance. Relatedly, prompt complexity increased progressively across tasks, while interaction quantity showed a more modest increase (RQ4), suggesting that students' behavioral engagement became more sophisticated with practice.

At the cognitive level, evaluative judgment emerged as a significant predictor of writing performance (RQ3, H3), with contributions from students' ability to evaluate both essays and AI-generated outputs. This finding reinforces the central role of evaluative judgment in feedback processes, supporting the view that the ability to judge quality underpins effective feedback use (Carless & Boud, 2018; Sadler, 2010). Extending prior research, our study provides empirical evidence that EJ is an important learner capacity

associated with greater benefit from GenAI-supported internal feedback. The findings suggest that effective engagement with AI-generated outputs requires not only general AI literacy, but also domain-specific evaluative knowledge.

Overall, the composite measures of EJ showed limited development over the GenAI-IFB intervention (RQ4). However, this finding should be interpreted cautiously because the three EJ indicators exhibited different developmental patterns. In particular, ranking accuracy showed a pronounced ceiling effect, with many students correctly ordering the exemplars from the outset, leaving limited room for further improvement. By contrast, exemplar and AI feedback evaluation provided greater opportunity for developmental change, yet neither showed significant improvement across the three writing tasks. This finding differs from prior studies suggesting that students' EJ can be supported through ChatGPT feedback (Xie et al., 2025), dialogic exemplar use (Chong, 2021), and rubric-supported evaluative activities (Gyamfi et al., 2022). Our findings suggest that structured GenAI-supported feedback activities, even with reflection prompts, may not be sufficient to promote measurable development in higher-order evaluative judgment over three writing tasks

Although EJ did not show significant improvement, students who demonstrated stronger EJ capacity during the intervention tended to benefit more from the GenAI-supported internal feedback activities (RQ3, H3). This suggests that EJ may function as an enabling process through which students benefit from GenAI-IFB, rather than a learning outcome that develops rapidly through relatively short-term AI interactions.

Another possible explanation for limited EJ development concerns how GenAI shapes students' cognitive engagement. Interaction with GenAI may facilitate either cognitive onloading (e.g., students actively engage in evaluating, comparing, and refining feedback) or cognitive offloading, whereby evaluative judgment is deferred to the system (Risko & Gilbert, 2016). In contexts where students rely heavily on AI-generated responses, opportunities for developing evaluative judgment may be reduced, particularly when students lack the capacity to critically interrogate the feedback provided.

A further, related explanation for the limited development of exemplar and AI feedback evaluation (RQ4) concerns students' tendency to trust AI-generated outputs, particularly when they have a weaker understanding of quality (Bearman et al., 2024). Positive initial experiences with GenAI may then have further reinforced this trust (Brummernhenrich et al., 2025), leading students (particularly those with weaker evaluative capacities) to accept AI outputs uncritically despite potential biases, inaccuracies, and hallucinations (Jia et al., 2024; Lindsay et al., 2025; Ruwe & Kuklick, 2026). Such reliance may then have reduced students' need to construct and refine their understanding of quality, thereby constraining EJ development. To address this, engaging students in comparisons across multiple feedback sources, such as teacher input, peer perspectives, and textbooks may help calibrate their evaluative judgments and promote higher-order reasoning (Alfieri et al., 2013; Nicol, 2021).

5.2. Contributions, implications, and future research

5.2.1. Theoretical contributions

This study advances internal feedback theory by extending Lipnevich and Smith's (2022) framework into AI-mediated contexts. Specifically, it conceptualizes GenAI-supported internal feedback as comprising two interrelated dimensions: a cognitive set of dimensions, reflected in students' evaluative judgment of exemplars and AI-generated feedback (RQ3), and a behavioral dimension, reflected in the quantity and quality of their interactions with GenAI (RQ2). The extended framework positions AI not merely as an external feedback provider but as a resource through which learners generate, evaluate, and regulate their own internal feedback.

The GenAI-supported internal feedback intervention was associated with greater writing improvement than business-as-usual teacher feedback conditions (RQ1). This finding should, however, be interpreted cautiously given the confounded intervention design. The observed differences may not be attributable solely to the feedback internalization process, but may also reflect cognitive offloading, whereby AI undertakes part of the evaluative processing, potentially reducing opportunities for students to independently compare, question, and refine their own judgments. This highlights a theoretical tension between GenAI-supported internal feedback generation and reliance on AI-mediated evaluation.

Beyond extending internal feedback theory, our findings contribute to theories of evaluative judgment (e.g., Sadler, 2010; Tai et al., 2018). Although evaluative judgment predicted writing performance, it showed limited development over time despite students' repeated comparisons, evaluations, and justifications. This suggests that engaging in evaluative practices does not necessarily translate into measurable improvements in evaluative capacity, indicating that such capacity may develop more slowly than observable behavioral engagement in AI-mediated learning.

The study also extends current understanding of the self-feedback process by highlighting the role that student-AI interaction can play in this process. Students actively elicited AI feedback through their prompts, with greater prompt complexity consistently associated with better writing performance. This suggests that student-AI interaction can provide additional information that learners can draw on to understand the quality of writing and inform subsequent improvement.

Importantly, our study advances understandings of student-AI interaction by distinguishing between interaction quality and interaction quantity. Our findings suggest that the two aspects may be relevant, but that the quality of student-AI interaction is the stronger predictor of writing performance than the quantity alone.

Our study further indicates that the cognitive and behavioral dimensions of GenAI-supported internal feedback may follow different trajectories, refining our proposed framework by demonstrating that behavioral engagement with AI can improve without corresponding growth in evaluative judgment capacities.

5.2.2. Implications for practice

The findings carry important practical implications for both pedagogical practice and the design of AI feedback systems. From a pedagogical perspective, teachers can move beyond encouraging frequent GenAI interaction and instead scaffold students in crafting clear, specific, and contextually grounded prompts aligned with their learning needs and goals, as RQ2 showed that prompt complexity was more strongly associated with writing performance than interaction quantity. From a design perspective, GenAI systems could be developed to better support internal feedback processes. For example, AI systems could prompt learners to refine their queries, ask clarification questions before generating responses, and surface uncertainty to encourage critical evaluation (Chang et al., 2023). Such features may help shift learners from passive recipients of AI-generated feedback to active agents in internal feedback generation.

The findings also point towards evaluative judgment as an important learner capacity associated with better writing performance within GenAI-IFB (RQ3). Meanwhile, its limited development over time (RQ4) suggests that repeated participation in evaluative activities is insufficient to foster this capacity. As Bearman et al. (2024) noted, "we should not assume that because the act of using GenAI exercises evaluative judgment, students are developing good evaluative judgment" (p.897). Developing EJ may therefore require intentional instructional design. One promising approach is to integrate GenAI with other feedback approaches, such as peer feedback, which has been widely recognized for fostering higher-order thinking, facilitating meaning negotiation, and supporting the development of evaluative judgment (Carless & Boud, 2018; Chen et al., 2027; Mu & Schunn, 2025; Nicol et al., 2014; Tai et al., 2018).

In terms of practical significance of writing gains, the adjusted between-condition difference in post-test writing performance was moderate in magnitude ($d = 0.52$). Given the absence of established interpretive benchmarks for this outcome measure, the practical significance for the GenAI-IFB should be interpreted cautiously.

Beyond these implications, the limited development of evaluative judgment observed in RQ4 highlights the importance of responsible GenAI use and the "care-full craft of feedback" (Winstone et al., 2026, p. 1). Although issues such as AI hallucinations, misinformation, fairness, and feedback immediacy were not directly examined, they warrant attention when students are expected to generate internal feedback through sustained interaction with GenAI. Our findings suggest that structured evaluative and interactive activities alone may be insufficient to develop students' evaluative judgment. This suggests that students should be explicitly scaffolded in identifying AI hallucinations, misinformation, or culturally biased content against domain-specific criteria, standards, and credible sources. Without such scaffolding, there is a risk that learners may over-rely on AI as an authoritative source (Fan et al., 2025), while AI systems themselves may face the challenge of 'AI echo-chamber', i.e., LLMs trained on their own output will spiral into regurgitating more and more of this biased content (Lindsay et al., 2025, p. 4). Designing feedback landscapes that encourage comparison across AI and non-AI feedback from credible sources may help mitigate these risks and prompt responsible AI use.

Fairness is another important consideration in the design and implementation of GenAI-supported feedback. AI feedback systems should not amplify existing individual differences but aim to reduce disparities and support learning across diverse student cohorts (Winstone et al., 2026). To this end, AI systems can be designed to externalize and scaffold evaluative processes. This may involve embedding structured comparison tasks (exemplars in feedback processes; Chen & To, 2026) rather than providing standalone feedback, adapting feedback in response to learners' evaluative capacity, and guiding prompt construction rather than leaving it entirely to students. The immediacy of AI-generated feedback also warrants consideration. Although instant feedback can enhance efficiency, it may reduce opportunities for students to engage in evaluative processing. Designing mechanisms that deliberately slow down the feedback loop may create space for deeper cognitive engagement. These considerations highlight the value of moving beyond outcome-based fairness metrics (cf. Viberg et al., 2024) toward capability-sensitive approaches that consider whether learners are equally positioned to benefit from GenAI feedback.

5.2.3. Limitations and future research

Several limitations should be acknowledged. First, we did not examine the affective dimension, despite its critical role in internal feedback models (Lipnevich & Smith, 2022; Yan & Brown, 2017). Future research should therefore attend more closely to the relational and emotional aspects of GenAI-supported feedback, including how students experience, trust, resist, or depend on AI.

Second, the GenAI-IFB intervention involved multiple components (e.g., ranking exemplars, justifying rankings, requesting feedback, and evaluating feedback), and it was not possible to extract the more specific contributions from these components or determine whether similar learning outcomes could have been achieved through comparable evaluative activities embedded within teacher feedback. Likewise, because the study did not include a GenAI-only condition without structured evaluative activities, the findings should be interpreted as the effect of an integrated instructional intervention rather than the isolated effect of AI feedback. In addition, differences in genre, topic familiarity, and task difficulty should be considered when interpreting the writing outcomes. Future studies should use tasks that are more closely matched in genre and topic familiarity and, when possible, independently calibrated for difficulty.

Third, although prompt complexity and evaluative judgment were positively associated with writing performance, these findings are inherently correlational. Since these results involved multiple regression with several critical controls and temporal ordering among variables, many plausible confounds are addressed. Experimental studies are needed to more directly test their causal role (e.g., experimentally manipulating prompt complexity or implementing training interventions targeting evaluative judgment). In addition, students generally engaged in a relatively small number of AI interaction turns during the intervention. The restricted range of interaction quantity may have limited statistical power to detect its association with writing performance and may constrain the generalizability of the findings to contexts involving more sustained student-AI interaction. Future research involving greater variation in interaction quantity could more clearly examine how the extent of student-AI interaction relates to learning outcomes.

Fourth, the findings should not be interpreted as evidence for the unique superiority of GenAI feedback over other feedback sources. Iterative and reflective engagement with non-AI feedback, including teachers and peers, may be just as useful. The aim of this

study was not to establish the unique value of GenAI feedback, but to examine a particular design for supporting internal feedback through GenAI. Finally, as the study was conducted within a single context, the generalizability of the findings may be limited. Future studies should extend this line of inquiry across disciplines, cultural contexts, and educational levels to examine the robustness and transferability of these findings.

Several additional directions for future research emerge from this study. Additional comparison conditions (e.g., teacher feedback with structured evaluative and interactive activities or GenAI feedback without such activities) could help isolate the contribution of different instructional elements. Given the evaluative judgment was positively associated with writing performance but showed limited development over time , future studies should investigate how GenAI-supported feedback can be designed to actively foster this capacity. It is also worthwhile to explore how process-oriented GenAI feedback designs, such as adaptive prompting and structured comparison tasks, can enhance the effectiveness of internal feedback generation across diverse learner populations. While this study focused on short-term writing improvement, future research should examine broader and longer-term learning achievement, including feedback literacy, student feedback agency, and other lifelong learning capacities. Understanding how GenAI supports these lifelong capacities is critical for evaluating its role in education beyond immediate performance gains (Yan et al., 2025).

6. Conclusion

This study compared GenAI-supported internal feedback intervention with teacher feedback as a form of business-as-usual external feedback over 14 weeks. The findings indicate that GenAI-IFB supported student learning outcomes, particularly when students engaged in key behavioral and cognitive ways. Specifically, prompt complexity, as a behavioral indicator, and evaluative judgment as a cognitive indicator, were significant positive predictors of writing performance. Although students' interactions with GenAI became more complex over time, their evaluative judgment did not demonstrate corresponding development. Our study addresses an important research gap by providing empirical evidence on how GenAI supports internal feedback and by extending internal feedback theory within an AI-mediated context.

CRediT authorship contribution statement

Shijun (Cindy) Chen: Writing – original draft, Conceptualization, Formal analysis, Visualization, Project administration, Investigation, Data curation, Resources, Methodology, Validation, Writing – review & editing. **Christian D. Schunn:** Writing – review & editing, Validation, Methodology, Conceptualization. **Jie Cao:** Writing – review & editing, Visualization, Validation, Methodology. **Seyyed Kazem Banihashem:** Writing – review & editing, Validation. **Dragan Gašević:** Writing – review & editing, Validation. **Yizhou Fan:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no competing interests.

Appendix A. Assessment Criteria

The CET-6 writing section adopts a holistic impression scoring approach. The full score for the essay is 15 points, divided into five score bands.

- 14-point band (13–15 points)
- 11-point band (10–12 points)
- 8-point band (7–9 points)
- 5-point band (4–6 points)
- 2-point band (1–3 points)

Scoring Criteria

- **14-point band:** The essay fully addresses the topic. Ideas are expressed clearly; the writing is fluent and coherent, with basically no language errors, except for a few minor mistakes.
- **11-point band:** The essay addresses the topic. Ideas are expressed clearly and the writing is coherent, but there are a small number of language errors.
- **8-point band:** The essay basically addresses the topic. In some places, ideas are not expressed clearly; the writing is only barely coherent, and there are quite a number of language errors, some of which are serious.
- **5-point band:** The essay basically addresses the topic. Ideas are expressed unclearly, coherence is poor, and there are many serious language errors.
- **2-point band:** The organization is unclear, ideas are confused, and the language is fragmented, or most sentences contain errors, the majority of which are serious.

Appendix B. Prompts and Essay Exemplar

The following are the scoring criteria for the College English Test Band 6 (CET-6) writing section:

14-point band: The essay fully addresses the task. Ideas are expressed clearly, and the writing is fluent and coherent. There are basically no language errors, with only a few minor mistakes.

11-point band: The essay fully addresses the task. Ideas are expressed clearly and the writing is coherent, but there are a small number of language errors.

8-point band: The essay basically addresses the task. In some places, ideas are not expressed clearly. The writing is barely coherent, and there are quite a number of language errors, some of which are serious.

5-point band: The essay basically addresses the task. Ideas are unclear, coherence is weak, and there are many serious language errors.

2-point band: The essay lacks clear organization and logical flow. Ideas are disorganized, language is fragmented, or most sentences contain errors, the majority of which are serious.

This is the first English writing task:

Title: My Views on Using Social Networking.

Directions:

Social networking, such as WeChat, plays an important role in our lives. However, some people find it annoying. Write an essay to express your views on using social networking. You should write at least 150 words but no more than 250 words.

The target writers are second-year university students.

Based on the CET-6 writing criteria, please generate:

one essay at the 14-point level,

one essay at the 11-point level, and

one essay at the 5-point level.

14-Point Band Sample Essay (189 words)

Social networking platforms like WeChat have become integral to modern communication, yet their omnipresence sparks polarized opinions. As a digital native, I maintain that rational utilization of these tools can yield substantial benefits while mitigating potential drawbacks.

Undeniably, social media bridges geographical divides effectively. Through instant messaging and video calls, families separated by continents can maintain emotional connections. My own experience confirms this: weekly video chats with my studying-abroad cousin through WeChat have preserved our bond. Moreover, professional networking opportunities abound on these platforms. University students frequently access internship information through official accounts, while alumni groups facilitate career guidance exchange.

However, prudent management remains crucial. The “always-online” culture may breed anxiety, as evidenced by my friend's compulsive checking of WeChat notifications during lectures. A 2022 Pew Research study reveals that 43% of young adults experience digital fatigue from excessive social media use. To address this, I consciously disable non-essential notifications and allocate specific time slots for social media engagement.

Ultimately, social networking tools mirror kitchen knives - their value depends on proper handling. By maintaining digital discipline and prioritizing real-world interactions, we can harness these platforms' advantages without falling prey to their addictive traps.

11-Point Band Sample Essay (176 words)

Nowadays social networking like WeChat is very popular, but some people think it's annoying. In my view, we should use it in proper way to get benefits and avoid problems.

First, social media makes communication easier. We can chat with friends far away and share our life by photos and videos. For example, when my classmates go to other cities, we still keep contact through WeChat. Second, it's useful for study. Many teachers send learning materials through WeChat groups. Students can discuss homework and ask questions online.

But there's some bad points too. Some people spend too much time watch short videos or play mini games. This waste time and not good for study. Also, privacy problem is worrying. If we post too much personal information, maybe someone bad will use it.

So how to solve this? We need control our time. Set specific time for check messages instead of always looking at phone. Don't believe strange information. Use real name and phone number carefully.

In conclusion, social networking is helpful if we use it correctly. We should be smart users instead of let it control our life.

5-Point Band Sample Essay (153 words)

Social networking is very popular today. Many people use WeChat every day. But some people don't like it. I think social networking have good and bad.

First, good things. We can talk to friends easy. Send message fast. Make new friends online. See many news and information. Help our life convenient. For example, pay money by WeChat, no need carry cash.

But have problems too. Some people always play phone, ignore real friends. My roommate always look at phone when we eat together. Not good. Another problem is fake news. Many people share wrong information, make others confused. Sometimes receive too much advertisement, feel annoying.

Some people say social media bad for health. Because use phone long time, eyes hurt. Also, some students sleep late because chat online.

I think we should use less time on social media. Go out with friends real life. Don't believe all information online. Government should manage bad content.

In a word, social networking is important but need use careful. Hope everyone can use it right way.

Appendix C. Post-hoc Score Adjustments by Condition

Table C1
Direction and Magnitude of Post-hoc Score Adjustments by Condition

Condition	Total essays, *n*	Adjusted, *n* (%)	Upward, *n*	Downward, *n*	Mean signed adjustment (SD)	Mean absolute adjustment (SD)
GenAI-IFB	141	9 (6.4%)	1	8	−0.43 (0.48)	0.59 (0.23)
TFB	174	10 (5.7%)	3	7	−0.33 (0.58)	0.61 (0.20)

Note. Signed adjustment = moderate score – original *iWrite* score; negative values indicate downward adjustment.

Table C2
Agreement and Score Discrepancies in the Blind Re-scoring Subsample

Assessment	Overall ICC	GenAI-IFB discrepancy, *M* (*SD*)	TFB discrepancy, *M* (*SD*)	*t*	*p*
Pre-test	0.96	−0.18 (0.43)	−0.03 (0.42)	−1.42	0.161
Post-test	0.98	0.14 (0.44)	0.12 (0.30)	0.23	0.817

Appendix D

Table D1
Sensitivity Analyses of RQ2 Using Alternative Operationalizations of Prompt Complexity

Predictor	Mean complexity Model 1 β (SE)	Mean complexity Model 3 β (SE)	Proportion high-level Model 1 β (SE)	Proportion high-level Model 3 β (SE)
AI interaction				
Prompt complexity	0.34*** (0.08)	0.36*** (0.05)	0.29*** (0.08)	0.32*** (0.05)
Interaction quantity	0.22** (0.08)	0.06 (0.05)	0.26*** (0.08)	0.09† (0.05)
Evaluative judgment				
Exemplar evaluation		0.32*** (0.05)		0.31*** (0.05)
Ranking accuracy		0.27*** (0.05)		0.28*** (0.05)
AI feedback evaluation		0.20*** (0.04)		0.21*** (0.04)
Pretest	0.63*** (0.07)	0.74*** (0.04)	0.62*** (0.07)	0.74*** (0.04)
Section 2	0.30* (0.15)	0.03 (0.09)	0.31* (0.15)	0.04 (0.09)
Section 3	0.32† (0.17)	0.05 (0.11)	0.35* (0.17)	0.07 (0.11)
N	141	141	141	141
R^2	0.48	0.82	0.469	0.81
Adjusted R^2	0.46	0.81	0.449	0.80
F(df)	25.28 (5, 135)***	73.33 (8, 132)***	23.86 (5, 135)***	68.86 (8, 132)***

Note. †$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$. SE = Standard Error.

Table D2
Diagnostic Checks for the RQ2 Sensitivity Models

Sensitivity model	Maximum VIF	Shapiro–Wilk *W*	*p*	Breusch–Pagan χ^2(df)	*p*
Mean complexity – Model 1	1.80	0.98	0.10	5.22 (5)	0.39
Mean complexity – Model 3	1.86	0.99	0.71	14.86 (8)	0.06
Proportion high-level – Model 1	1.65	0.99	0.15	5.59 (5)	0.35
Proportion high-level – Model 3	1.72	0.99	0.32	17.23 (8)	0.03

Appendix E. Sample writing task

Task 1: Social networking, such as WeChat, plays an important role in our lives. However, some people find it annoying. Write an essay to express your views on using social networking. You should write at least 150 words but no more than 250 words.

Data availability

The data that has been used is confidential.

References

Alemdag, E., & Narciss, S. (2025). Promoting formative self-assessment through peer assessment: Peer work quality matters for writing performance and internal feedback generation. *International Journal of Educational Technology in Higher Education, 22*(1), Article 22. https://doi.org/10.1186/s41239-025-00522-4

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist, 48*(2), 87–113. https://doi.org/10.1080/00461520.2013.775712

Bai, Y., & Wang, S. (2025). Impact of generative AI interaction and output quality on university students' learning outcomes: A technology-mediated and motivation-driven approach. *Scientific Reports, 15*(1), Article 24054. https://doi.org/10.1038/s41598-025-08697-6

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Banihashem, S. K., Noroozi, O., Khosravi, H., Schunn, C. D., & Drachsler, H. (2026). Pedagogical framework for hybrid intelligent feedback. *Innovations in Education & Teaching International, 63*(2), 554–570. https://doi.org/10.1080/14703297.2025.2499174

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box : Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology, 54*(5), 1160–1173. https://doi.org/10.1111/bjet.13337

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgment for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education, 49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Brummernhenrich, B., Paulus, C. L., & Jucks, R. (2025). Applying social cognition to feedback chatbots: Enhancing trustworthiness through politeness. *British Journal of Educational Technology, 56*(6), 2321–2340.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.2307/1170684

Cao, S., & Zhong, L. (2023). Exploring the effectiveness of ChatGPT-based feedback compared with teacher feedback and self-feedback: Evidence from Chinese to English translation. https://doi.org/10.48550/arxiv.2309.01645.

Cao, S., & Zhou, T. (2025). Exploring the efficacy of ChatGPT-based feedback compared with teacher feedback and self-feedback: Evidence from Chinese-English translation. *Sage Open, 15*(3), Article 21582440251369204.

Carless, D. (2022). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education, 23*(2), 143–153. https://doi.org/10.1177/1469787420945845

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Carless, D., & Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education, 28*(1), 150–163. https://doi.org/10.1080/13562517.2020.1782372

Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability, 15*(17), Article 12921. https://doi.org/10.3390/su151712921

Chen, S., Borchers, C., & Schunn, C. D. (2027). From revision to learning transfer: Self- and socially shared-regulated pathways in AI-assisted feedback. Learning and Instruction, 107, 102489. doi:https://doi.org/10.1016/j.learninstruc.2026.102489.

Chen, S. (Cindy), & Carless, D. (2026). Teacher feedback dilemmas and the use of GenAI: Challenges or opportunities? *Innovations in Education & Teaching International*, 1–14. https://doi.org/10.1080/14703297.2026.2648040

Chen, S. (Cindy), & To, J. (2026). Unravelling undergraduates' development of evaluative judgments through AI-supported internal feedback. *Assessment & Evaluation in Higher Education*, 1–18. https://doi.org/10.1080/02602938.2026.2711099

Chong, S. W. (2021). University students' perceptions towards using exemplars dialogically to develop evaluative judgment: The case of a high-stakes language test. *Asian-Pacific Journal of Second and Foreign Language Education, 6*(1), 1–27. https://doi.org/10.1186/s40862-021-00115-4

Chou, C. Y., & Zou, N. B. (2020). An analysis of internal and external feedback in self-regulated learning activities mediated by self-regulated learning tools and open learner models. *International Journal of Educational Technology in Higher Education, 17*(1), 55.

Corbin, T., Tai, J., & Flenady, G. (2025). Understanding the place and value of GenAI feedback: A recognition-based framework. *Assessment & Evaluation in Higher Education, 50*(5), 718–731. https://doi.org/10.1080/02602938.2025.2459641

Cui, Y., Schunn, C. D., & Gai, X. (2022). Peer feedback and teacher feedback: A comparative study of revision effectiveness in writing instruction for EFL learners. *Higher Education Research and Development, 41*(6), 1838–1854. https://doi.org/10.1080/07294360.2021.1969541

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2023). Impact of AI assistance on student agency. *Computers & Education.* , Article 104967. https://doi.org/10.1016/j.compedu.2023.104967

Debets, T., Banihashem, S. K., Joosten-Ten Brinke, D., Vos, T. E., de Buy Wenniger, G. M., & Camp, G. (2025). Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts. *Computers & Education.* , Article 105323. https://doi.org/10.1016/j.compedu.2025.105323

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

Federiakin, D., Molerov, D., Zlatkin-Troitschanskaia, O., & Maur, A. (2024). Prompt engineering as a new 21st century skill. *Frontiers in Education, 9*. https://doi.org/10.3389/feduc.2024.1366434

Gyamfi, G., Hanna, B. E., & Khosravi, H. (2022). The effects of rubrics on evaluative judgment: A randomised controlled experiment. *Assessment & Evaluation in Higher Education, 47*(1), 126–143. https://doi.org/10.1080/02602938.2021.1887081

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heil, J., & Ifenthaler, D. (2026). Effects of teacher, peer and self-feedback on student improvement in online assessment: The role of individuals' presumptions and feedback literacy. *Assessment & Evaluation in Higher Education, 51*(2), 281–300. https://doi.org/10.1080/02602938.2025.2530452

Ho, P. V. P., Phung, L. T. K., Oanh, T. T. T., & Giao, N. Q. (2020). Should peer E-Comments replace traditional peer comments? *International Journal of Instruction, 13*(1), 295–314. https://doi.org/10.29333/iji.2020.13120a

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., & Gehringer, E. (2024). On assessing the faithfulness of LLM-generated feedback on student assignments. *International Educational Data Mining Society*, 491–499. https://doi.org/10.5281/zenodo.12729868

Jiang, S., Vakulenko, S., & de Rijke, M. (2023). Weakly supervised turn-level engagingness evaluator for dialogues. In J. Gwizdka, & S. Y. Rieh (Eds.), *Proceedings of the 2023 conference on human information interaction and retrieval* (pp. 258–268). ACM. https://doi.org/10.1145/3576840.3578319.

Kim, J., Lee, S., Detrick, R., Wang, J., & Li, N. (2025). Students-generative AI interaction patterns and its impact on academic writing. *Journal of Computing in Higher Education*. https://doi.org/10.1007/s12528-025-09444-6

Ko, C. (2022). Online individualized corrective feedback on EFL learners' grammatical error correction. *Computer Assisted Language Learning, 37*, 1449–1477. https://doi.org/10.1080/09588221.2022.2118783

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lindsay, E. D., Zhang, M., Johri, A., & Bjerva, J. (2025). The responsible development of automated student feedback with generative AI. In *2025 IEEE global engineering education conference (EDUCON)* (pp. 1–10). IEEE.
Lipnevich, A. A., & Smith, J. K. (2022). Student – Feedback interaction model: Revised. *Studies In Educational Evaluation, 75*, Article 101208. https://doi.org/10.1016/j.stueduc.2022.101208
Liu, S. (2022). A research on the blended evaluation mode in college English writing course. *Journal of Language Teaching and Research, 13*(4), 763–771. https://doi.org/10.17507/jltr.1304.09
Lu, Q., Yao, Y., Xiao, L., Zhu, X., & Yin, H. (2025). Exploring the impact of a GenAI-supported feedback practice on student feedback literacy in L2 writing. *Computer Assisted Language Learning*, 1–30.
Mao, Z., & Lee, I. (2025). Every advantage has its disadvantage: Side effects of teacher feedback in L2 writing. *RELC Journal*. https://doi.org/10.1177/00336882251376701
Mu, H., & Schunn, C. D. (2025). The good, bad, and ugly of comment prompts: Effects on length and helpfulness of peer feedback. *International Journal of Educational Technology in Higher Education, 22*(1), Article 4. https://doi.org/10.1186/s41239-025-00502-8
Narciss, S. (2017). Conditions and effects of feedback viewed through the lens of the interactive tutoring feedback model. In S. M. Bridges, C. K. Y. Chan, D. Carless, & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 173–189). Singapore: Springer. https://doi.org/10.1007/978-981-10-3045-1_12.
Narciss, S., Prescher, C., Khalifah, L., & Körndle, H. (2022). Providing external feedback and prompting the generation of internal feedback fosters achievement, strategies and motivation in concept learning. *Learning and Instruction, 82*, Article 101658.
Nelson, R. (2018). Barriers to the cultivation of evaluative judgment: A critical and historical perspective. In D. Boud, R. Ajjawi, P. Dawson, & J. Tai (Eds.), *Developing evaluative judgment in higher education* (1st ed., Vol. 1, pp. 51–59). Routledge. https://doi.org/10.4324/9781315109251-6.
Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17*(4), 372–411.
Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education, 49*(5), 847–864. https://doi.org/10.1080/03075079.2024.2323593
Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314
Nicol, D., & Kushwah, L. (2024). Shifting feedback agency to students by having them write their own feedback comments. *Assessment & Evaluation in Higher Education, 49*(3), 419–439. https://doi.org/10.1080/02602938.2023.2265080
Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment and Evaluation in Higher Education, 39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518.
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
Noroozi, O., Haddadian, G., Gao, X., Schunn, C., Alqassab, M., & Banihashem, S. K. (2025). The value of GenAI for peer feedback provision: Student perceptions and impacts. *International Journal of Educational Technology in Higher Education, 22*(1), Article 61. https://doi.org/10.1186/s41239-025-00558-6
Panadero, E., Fernández, J., Pinedo, L., Sánchez, I., & García-Pérez, D. (2024). A self-feedback model (SEFEMO): Secondary and higher education students' self-assessment profiles. *Assessment in Education: Principles, Policy & Practice, 31*(3–4), 221–253. https://doi.org/10.1080/0969594X.2024.2367027
Panadero, E., Lipnevich, A., & Broadbent, J. (2019). Turning self-assessment into self-feedback. In *The impact of feedback in higher education: Improving assessment outcomes for learners* (pp. 147–163). Cham: Springer International Publishing.
Peltzer, K., Lira Lorca, A., Krause, U.-M., & Busse, V. (2024). Effects of formative feedback on argumentative writing in English and cross-linguistic transfer to German. *Learning and Instruction, 92*, Article 101935. https://doi.org/10.1016/j.learninstruc.2024.101935. Art. 101935.
Pozdniakov, S., Brazil, J., Banihashem, S. K., Noroozi, O., Gašević, D., Sadiq, S., & Khosravi, H. (2026). AI assistance in peer feedback provision: Pedagogically sound, but minimally adopted. *Computers & Education, 248*, Article 105591. https://doi.org/10.1016/j.compedu.2026.105591
Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002
Ruwe, T., & Kuklick, L. (2026). Quality counts? Examining the role of feedback provider and feedback quality on students' feedback perceptions. *British Journal of Educational Technology, 57*(1), 272–298.
Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535–550. https://doi.org/10.1080/02602930903541015
Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*(4), 556–569. https://doi.org/10.1016/j.system.2009.09.002
Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000
Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgment: Enabling students to make decisions about the quality of work. *Higher Education, 76*(3), 467–481.
Tam, A. C. F. (2025). Interacting with ChatGPT for internal feedback and factors affecting feedback quality. *Assessment & Evaluation in Higher Education, 50*(2), 219–235. https://doi.org/10.1080/02602938.2024.2374485
Viberg, O., Kizilcec, R. F., Wise, A. F., Jivet, I., & Nixon, N. (2024). Advancing equity and inclusion in educational practices with AI-powered educational decision support systems (AI-EDSS). *British Journal of Educational Technology, 55*(5), 1974–1981. https://doi.org/10.1111/bjet.13507
Winstone, N. E., & Carless, D. (2020). *Designing effective feedback processes in higher education : A learning-focused approach*. Routledge.
Winstone, N. E., Gravett, K., Bearman, M., Noble, C., Jensen, L. X., Jones, A., & Nicola-Richmond, K. (2026). The care-full craft of feedback in an age of generative AI. *Assessment & Evaluation in Higher Education*, 1–17. https://doi.org/10.1080/02602938.2026.2643333
Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538
Xie, X., Zhang, L. J., & Wilson, A. J. (2025). Comparing ChatGPT feedback and peer feedback in shaping students' evaluative judgment of statistical analysis: A case study. *Behavioral Sciences, 15*(7), 884. https://doi.org/10.3390/bs15070884
Yan, D.(A.), Tian, F., Li, H., Gao, Y., & Li, J. (2026). A GenAI-supported dialogic model to enhance the functions of peer feedback in EFL writing: A mixed-method interventional study. *System, 136*, Article 103895. https://doi.org/10.1016/j.system.2025.103895
Yan, L., Greiff, S., Lodge, J. M., & Gašević, D. (2025). Distinguishing performance gains from learning when using generative AI. *Nature Reviews Psychology, 4*(7), 435–436. https://doi.org/10.1038/s44159-025-00467-5
Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247–1262. https://doi.org/10.1080/02602938.2016.1260091
Yan, Z., & Carless, D. (2022). Self-assessment is about more than self: The enabling role of feedback literacy. *Assessment & Evaluation in Higher Education, 47*(7), 1116–1128. https://doi.org/10.1080/02602938.2021.2001431
Yu, S., Jiang, L., & Zhou, N. (2020). Investigating what feedback practices contribute to students' writing motivation and engagement in Chinese EFL context: A large-scale study. *Assessing Writing, 44*, Article 100451. https://doi.org/10.1016/j.asw.2020.100451
Zhan, Y., Boud, D., Dawson, P., & Yan, Z. (2025). Generative artificial intelligence as an enabler of student feedback engagement: a framework. *Higher Education Research and Development, 44*(5), 1289–1304. https://doi.org/10.1080/07294360.2025.2476513.
Zhan, Y., & Yan, Z. (2025). Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 1–14. https://doi.org/10.1080/02602938.2025.2471821
Zou, D., Xie, H., & Wang, F. L. (2023). Effects of technology enhanced peer, teacher and self-feedback on students' collaborative writing, critical thinking tendency and engagement in learning. *Journal of Computing in Higher Education, 35*(1), 166–185. https://doi.org/10.1007/s12528-022-09337-y