Comparative analysis of peer group and AI-generated feedback in peer assessment: Insights into feedback quality and student perceptions in higher education

Xin Li, Yue Zhang, Tingyu Liu
Intelligent Learning and Assessment Jiangsu Provincial Industrial Technology Engineering Center, Jiangsu Normal University, Xuzhou, China

The rapid integration of generative artificial intelligence (AI) into peer assessment is reshaping feedback practices in higher education. However, empirical evidence directly comparing human and AI-generated feedback remains limited. This mixed methods study compared feedback quality and student perceptions between 13 undergraduate peer groups comprising 39 students and ChatGPT-4o in a smart learning environment design course. Quantitative analyses indicated no significant difference during the first assessment stage; however, in the second stage, ChatGPT-4o provided feedback that was significantly higher in constructiveness, accuracy and concreteness. Thematic analysis further revealed that students perceived group feedback as contextually rich but sometimes subjective, whereas ChatGPT-4o feedback was viewed as more objective and well structured, though occasionally lacking contextual sensitivity. Overall, students demonstrated high levels of acceptance towards both feedback sources, but high-acceptance groups made more effective use of AI feedback during revision, resulting in higher accuracy scores. These findings highlight the complementary roles of human and AI feedback in strengthening the formative value of peer assessment and provide implications for the design of AI-augmented feedback systems in higher education.

Implications for practice or policy:

- Clear rubrics and explicit guidance for AI use are essential to improve the quality, consistency,and transparency of peer assessment.
- Combining peer group feedback with AI-generated feedback can enhance formative assessment by integrating contextualised human insights with structured and objective AI support.
- AI-generated feedback should be used as a supplementary tool rather than a standalone assessor, with ongoing human oversight to address potential inaccuracies and ensure pedagogical quality.

Keywords: peer assessment, group assessment, AI-generated assessment, feedback quality, higher education

Introduction

In recent years, a growing body of research has highlighted the pivotal role of assessment in higher education, not only as a mechanism for measuring learning outcomes but also as a powerful driver of learning processes (Cheong et al., 2023; Ma & Bui, 2022). Among various assessment forms, peer assessment has been widely recognised for fostering critical thinking, collaborative learning and self-regulated learning (Butler & Winne, 1995; Freeman & McKenzie, 2002; Hwang & Chang, 2021; Topping, 2017). However, persistent challenges – such as biased grading, insufficient feedback provision (Cho & Cho, 2011) and lack of recognition or acceptance by high-achieving students (Li & Gao, 2016; Ramon-Casas et al., 2019) – often undermine its pedagogical potential (Panadero, 2016). These limitations highlight the need for more reliable and engaging feedback mechanisms that can sustain the educational value of peer assessment.

To address these challenges, researchers have explored two complementary approaches: group-based peer feedback and technology-enhanced assessment. Group-based peer feedback allows students to

engage in collective evaluation and negotiation, thereby promoting deeper reflection and improving the reliability of evaluative judgements through shared reasoning (Zhang et al., 2023). In parallel, technology-enhanced assessment – particularly those powered by generative artificial intelligence (AI) – has shown growing potential in automating and enriching feedback delivery (Guo & Wang, 2024). Recent studies have suggested that generative AI tools such as ChatGPT can produce feedback that is coherent, specific and sometimes comparable to that of human tutors (Banihashem et al., 2024; Dai et al., 2023). These emerging capabilities position generative AI as a potentially valuable partner in assessment processes, calling for empirical scrutiny of how AI-generated feedback might complement human evaluative processes within peer assessment contexts.

Studies have primarily compared AI-generated feedback with teacher or individual peer feedback (Banihashem et al., 2024; Guo et al., 2024; Lu et al., 2024). However, limited attention has been paid to feedback jointly generated through peer-group discussion, which can be conceptualised as a collaborative activity involving socially shared and co-regulation (Er et al., 2021). From this perspective, comparing group and AI feedback is theoretically significant: group feedback reflects socially shared regulation processes (Hadwin et al., 2018), while AI-generated feedback represents a data-driven, algorithmic form of external regulation. Examining their relative effectiveness can thus deepen understanding of how human and artificial agents co-construct evaluative knowledge in higher education.

To address this gap, this study conducted a comparative analysis of feedback produced by student groups and ChatGPT-4o within a peer assessment activity embedded in a smart learning environment design course. Specifically, it investigates differences in feedback quality, students' perceptions of feedback usefulness and how students acted upon feedback during revision. By integrating perspectives from collaborative learning and feedback uptake research, this study advances theoretical understanding of human–AI interaction in formative assessment and provides pedagogical insights for designing hybrid peer assessment models that leverage both human and AI feedback strengths.

Literature review

Feedback in assessment

Feedback has long been recognised as a central mechanism in assessment and learning. In educational research, feedback is commonly defined as information provided regarding aspects of a learner's performance or understanding that can be used to reduce the gap between current and desired performance (Hattie & Timperley, 2007). Rather than serving merely as corrective information, feedback increasingly functions as a mechanism that guides learners' reflection, judgement and improvement throughout the learning process. Scholarship has therefore conceptualised feedback as an interactive and dialogic process that supports learners in interpreting information, making judgements about quality and taking action to improve their work (Boud & Molloy, 2013; Nicol, 2021). Furthermore, Carless and Boud (2018) have emphasised that feedback is effective only when learners are able to interpret and use it appropriately, highlighting the importance of considering how feedback is understood and applied in learning contexts.

Feedback can take different forms and serve multiple functions. Hattie and Timperley (2007) distinguished four levels – task-level feedback, process-level feedback, self-regulation feedback and self-level feedback. They argued that feedback directed at the task, the processes underlying task performance and self-regulation can facilitate learning, whereas self-level feedback is generally less effective. Feedback may also vary in source, including teacher feedback, peer feedback, automated feedback and self-generated feedback (Nicol, 2021; Shute, 2008), each contributing distinct perspectives and informational value to the learning process.

Within collaborative learning contexts, feedback often emerges through peer or group interactions, where evaluative judgements are not simply transmitted but co-constructed through dialogue, discussion and negotiation. Understanding feedback in such contexts therefore requires attention not only to the

information provided but also to the social processes through which feedback is generated, interpreted and enacted.

Group assessment

Within the broader framework of feedback in assessment, peer assessment has long been regarded as a critical pedagogical approach within computer-supported collaborative learning, fostering reflective thinking, mutual understanding and accountability among learners (Tenenbaum et al., 2020; Topping, 2017). Within this paradigm, group assessment – a process in which one group evaluates the work of another – extends the pedagogical value of traditional peer assessment by embedding evaluation within collective reasoning and negotiated judgement (Topping, 1998; Zhang et al., 2021; Zhang et al., 2023). Such inter-group interactions promote shared knowledge construction, enhance evaluative judgement and cultivate metacognitive awareness at the group level (Chen et al., 2021; Tan & Chen, 2022). From a theoretical standpoint, group assessment aligns with the socially shared regulation of learning framework, which views collaborative regulation as an emergent, co-constructed process rather than an aggregation of individual self-regulation (Hadwin et al., 2018). During the feedback generation process, groups collectively plan, monitor and evaluate their understanding, engaging in joint regulation of the evaluative task.

Empirical studies have shown that group-generated feedback tends to emphasise conceptual clarity and content relevance more strongly than individual feedback (Kessler, 2009; Memari Hanjani, 2016). Moreover, collaborative discussions within groups mitigate issues of linguistic insecurity and self-doubt, particularly among second-language learners (Alshuraidah & Storch, 2019). Despite these benefits, challenges persist: the quality of group-generated feedback often depends on group dynamics, equity in participation and the group's regulatory capacity (Kirschner & Erkens, 2013). Uneven participation, limited evaluative expertise or insufficient coordination may weaken the depth and reliability of feedback. Those social and regulatory dynamics are not merely contextual variables but integral determinants of how effectively feedback is produced and utilised within groups (Yukawa, 2006).

Nevertheless, while research has demonstrated that group assessment enhances interaction and reflective engagement, relatively few studies have examined how group-generated feedback is subsequently interpreted and enacted by learners. As Butler and Winne (1995) and Nicol (2021) have emphasised, feedback effectiveness depends not merely on its quality but on how learners interpret, internalise and act upon it. Therefore, examining both feedback production and feedback uptake provides a more comprehensive perspective for understanding the educational value of group assessment.

AI-generated assessment

Recent advancements in large language models have transformed how feedback can be generated and delivered. ChatGPT, powered by generative pre-trained transformer architecture, enables context-sensitive dialogue and personalised feedback that surpasses the limitations of static automated scoring systems (Barrot, 2024; Preiksaitis & Rose, 2023; Sol & Heng, 2024). Its linguistic fluency and adaptability enable it to produce detailed, structured and constructive feedback in near real time, expanding the accessibility of formative assessment (Guo & Wang, 2024; Su et al., 2023). Empirical findings have suggested that AI-generated feedback can enhance learners' understanding by offering elaborated explanations, scaffolding cognitive engagement, and sustaining learning motivation through supportive tone (Dai et al., 2023; Lu et al., 2024). However, scholars caution that ChatGPT's performance is not without limitations. The model may generate factually incorrect or contextually irrelevant responses (Thorp, 2023), lack socio-emotional sensitivity and interpersonal attunement typical of human feedback (Allagui, 2023), and exhibit biases in tone or depth (Mohamed, 2024). Such limitations may reduce learners' trust in AI-generated feedback and influence their willingness to adopt it, particularly in learning environments where affective and relational dimensions play an important role (Carless & Boud, 2018).

The release of ChatGPT-4o in May 2024 marked a significant leap in multimodal comprehension and real-time responsiveness (Kleinman, 2024; OpenAI, 2024). Its improved contextual reasoning and linguistic

coherence suggest the potential to provide feedback comparable to or even exceeding human-generated responses, especially in terms of structural clarity and descriptive depth (Wiggers, 2024). Despite these advancements, empirical studies directly comparing feedback produced by ChatGPT-4o and human groups remain scarce. Studies have primarily contrasted teacher feedback with AI feedback (Guo et al., 2024; Lu et al., 2024) or individual peer feedback with AI feedback (Banihashem et al., 2024), overlooking the collective and socially regulated nature of group assessment.

Theoretically, comparing AI-generated feedback with group-generated feedback provides a valuable lens for exploring two distinct forms of regulation: data-driven algorithmic regulation versus socially shared human regulation. While ChatGPT relies on probabilistic language modelling to evaluate and suggest improvements, human groups rely on negotiation, shared meaning-making and the co-construction of evaluative standards. This distinction reflects a fundamental shift in the source and nature of regulation – from socially constructed to computationally generated (Urzúa et al., 2025). Investigating these differences not only advances understanding of how feedback functions across human and machine systems but also informs the design of hybrid assessment models that integrate human judgement with AI support to enhance feedback quality, fairness and uptake (Shvets et al., 2025).

Research questions

Building on the preceding review, this study sought to compare feedback generated by student groups and ChatGPT-4o in the context of peer assessment. Specifically, the study examined differences in feedback quality, explored students' perceptions of feedback from both sources and investigated how learners adopt and apply the feedback during the revision process. The following research questions (RQs) guided the study:

- RQ1: What are the differences in feedback quality between those generated by groups and those generated by ChatGPT-4o?
- RQ2: What are the students' perceptions of the feedback produced by groups compared to that generated by ChatGPT-4o?
- RQ3: How do students adopt and apply feedback from both groups and ChatGPT-4o in the process of improving their learning outcomes?

Methods

Participants and learning context

The study was conducted in an undergraduate course titled Smart Learning Environment Design at a university in China. The course focuses on the conceptualisation and design of smart learning environments across diverse settings, including kindergartens, schools, homes and libraries, with the aim of developing students' abilities to design innovative learning environments supported by digital technologies. The study involved 39 third-year undergraduates (comprising 13 males and 26 females aged between 20 and 21) majoring in educational technology. Participants were randomly assigned to 12 groups, each consisting of three or four members.

The study followed established ethical guidelines for educational research and was approved by the Institutional Ethics Review Board of the university prior to data collection. Students were informed of the study's purpose and procedures before participation, and informed consent was obtained from all participants. Participation was voluntary and had no influence on course grades. All data were anonymised prior to analysis and used solely for research purposes.

Materials

Group tasks and assessment criteria

This study followed a two-phase design. In Phase 1, groups designed a smart library; in Phase 2, they designed a smart home learning environment. For each phase, groups submitted a comprehensive design report (approximately 1,000 words). These tasks were selected to align with the course objectives and to encourage collaborative problem-solving and creative design.

Designs were evaluated across five dimensions: innovation, practicality, technical feasibility, user-friendliness and environmental adaptability. Each dimension was assessed using a 5-level rubric ranging from *very poor* (1–5) to *excellent* (21–25).

To ensure the validity of the rubric, two external senior scholars in educational technology reviewed and refined the assessment criteria. The initial rubric was generated using a large language model and subsequently revised by the experts to ensure content validity and clarity of evaluation standards. These experts were not involved in teaching, scoring or data analysis.

To ensure scoring reliability, one faculty member and two teaching assistants independently evaluated all submissions. The final teacher score was calculated as the average of the three ratings. The intra-class correlation coefficient was 0.882, indicating high inter-rater reliability. The two teaching assistants also reviewed group scores to ensure that evaluations were generally consistent with the rubric.

Feedback instrument

To evaluate feedback quality, a feedback rating instrument was developed. It measured five dimensions: effectiveness, constructiveness, accuracy, completeness and concreteness (Banihashem et al., 2024; Guo & Wang, 2024; Steiss et al., 2024):

- effectiveness: the extent to which the feedback was perceived as helpful in improving designs
- constructiveness: whether the feedback provided actionable suggestions for refinement
- accuracy: whether the feedback correctly identified strengths and weaknesses
- completeness: whether the feedback covered all key aspects of the design
- concreteness: whether the feedback contained specific examples and explanations.

Students rated each dimension on a 5-point Likert scale (1 = *very low*, 5 = *very high*). They were also required to justify their scores in writing, fostering reflection and transparency.

Preparation for ChatGPT-4o feedback

ChatGPT-4o was used to generate feedback on group designs. This multimodal model processes text, images and video, and outputs in corresponding formats (OpenAI, 2024).

To ensure consistency in prompt design, we adopted the CO-STAR framework, a prompt engineering approach commonly used in AI-assisted instructional design (Kurniawan et al., 2025). Each prompt consisted of six elements:

- context: the role and expertise of the evaluator ("as a university faculty member specialising in smart learning environment design")
- objective: the intended purpose of the feedback (evaluating the student-designed program using the given rubric)
- style: the expected writing style (academic and professional)
- tone: the desired tone of the feedback (objective and neutral)
- audience: the target recipients of the feedback (student groups)
- response: the expected content of the output (rating, detailed rationale, and suggestions for improvement).

All prompts were pilot-tested internally by the research team prior to the study to ensure clarity, alignment with the rubric and consistency across design submissions.

Procedure

The experiment included two rounds of activities, each lasting 3 weeks (see Figure 1). The procedures were identical across rounds, differing only in the design tasks.

In Week 1, groups collaboratively developed a smart learning environment design. In Week 2, each design was evaluated by another group and by ChatGPT-4o using established grading criteria. Groups discussed the criteria before assigning scores, providing justifications and offering improvement suggestions. Each group member then evaluated the two types of feedback, which were anonymised as Feedback 1 (group feedback) and Feedback 2 (ChatGPT-4o feedback).

In Week 3, groups revised and resubmitted their designs based on the received feedback. The same procedure was repeated in the second round.

After completing both rounds, six groups were randomly selected to participate in 15-minute focus group interviews (one interview per round). The interviews were audio-recorded and transcribed to explore participants' perceptions of the two feedback types, consisting of their strengths, weaknesses, overall preference and expectations for feedback.

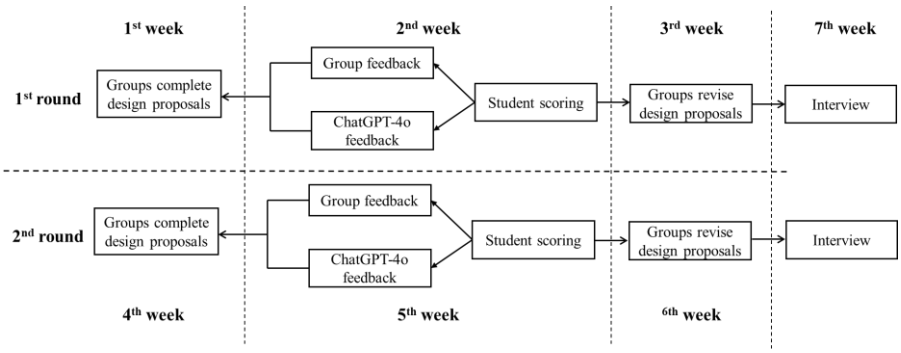

Figure 1. Experimental procedure of the study

Data collection and analysis

To address RQ1, students' ratings of the two feedback types were compared across five dimensions. Descriptive statistics were calculated, followed by Mann–Whitney U tests to examine statistically significant differences between group feedback and ChatGPT-4o feedback.

To address RQ2, interview data were analysed using thematic analysis following Braun and Clarke's (2006) framework. Transcripts were iteratively reviewed to generate initial codes representing students' perceptions of Feedback 1 and Feedback 2, which were subsequently refined into overarching themes. To enhance analytical credibility, the coding scheme was reviewed and revised collaboratively by two of us authors. Both then independently applied the finalised scheme, achieving an inter-coder agreement of 82.6%. Discrepancies were resolved through discussion until consensus was reached.

To address RQ3, students' revision behaviours were examined using the "Track Changes" function in Microsoft Word. Revisions were categorised into four patterns: revisions based solely on group feedback, revisions based solely on ChatGPT-4o feedback, revisions informed by both feedback sources and revisions unrelated to either feedback source. Feedback acceptance was operationalised as the number of revisions derived from each feedback source. Based on acceptance rates, groups were classified into high-acceptance and low-acceptance categories, and their revision patterns were compared to investigate differences in feedback adoption and utilisation.

Results

Differences in feedback quality between group-generated and ChatGPT-4o feedback

Table 1 presents a statistical comparison of group feedback and ChatGPT-4o feedback across five dimensions. In the first round, the average scores for both feedback types were comparable across all dimensions. Specifically, group feedback had a slight advantage in most areas, while ChatGPT-4o feedback was marginally better in concreteness. However, in the second round, ChatGPT-4o feedback consistently achieved higher average scores across all dimensions, indicating a change in the relative performance of the two assessment methods from the first round.

Table 1
Results of statistical analysis of the two rounds of assessment

Dimension	Round 1				Round 2			
	Group feedback		ChatGPT-4o feedback		Group feedback		ChatGPT-4o feedback	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Effectiveness	3.79	0.732	3.69	0.766	3.77	0.667	4.05	0.647
Constructiveness	3.77	0.842	3.67	0.772	3.64	0.537	4.10	0.598
Accuracy	3.90	0.680	3.54	0.854	3.74	0.637	4.13	0.615
Completeness	3.72	0.686	3.51	0.914	3.77	0.667	3.92	0.664
Concreteness	3.54	0.854	3.74	0.818	3.59	0.715	4.00	0.688

Subsequently, the Mann–Whitney U test was employed to analye the feedback from the two rounds across the five dimensions, with results presented in Tables 2 and 3. In the first round, no significant differences were detected between group and ChatGPT-4o feedback for any dimension, indicating that students perceived the quality of feedback from their peers and ChatGPT-4o to be comparable. However, in the second round, differences emerged in constructiveness ($p < .001$), accuracy ($p = .010$) and concreteness ($p = .021$), with ChatGPT-4o feedback receiving higher average score. This trend suggests an enhanced quality of assessment provided by ChatGPT-4o in the second round.

Table 2
Results of the Mann–Whitney U test for the 1st round of assessment

Dimension	Assessment type	*N*	Average rank	Sum of ranks	Mann–Whitney U	*Z*	*p*
Effectiveness	Groups	39	40.82	1592.00	709.000	-0.578	0.563
	ChatGPT-4o	39	38.18	1489.00			
Constructiveness	Groups	39	40.78	1590.50	710.500	-0.544	0.586
	ChatGPT-4o	39	38.22	1490.50			
Accuracy	Groups	39	43.23	1686.00	615.000	-1.624	0.104
	ChatGPT-4o	39	35.77	1395.00			
Completeness	Groups	39	41.17	1605.50	695.500	-0.717	0.473
	ChatGPT-4o	39	37.83	1475.50			
Concreteness	Groups	39	36.42	1420.50	640.500	-1.294	0.196
	ChatGPT-4o	39	42.58	1660.50			

Table 3
Results of the Mann–Whitney U test for the 2nd round of assessment

Dimension	Assessment source	*N*	Average rank	Sum of ranks	Mann–Whitney U	*Z*	*p*
Effectiveness	Groups	39	35.44	1382.00	602.000	-1.796	0.073
	ChatGPT-4o	39	43.56	1699.00			
Constructiveness	Groups	39	32.17	1254.00	474.500	-3.304	<0.001
	ChatGPT-4o	39	46.83	1826.00			
Accuracy	Groups	39	33.78	1317.50	537.500	-2.567	0.010
	ChatGPT-4o	39	45.22	1763.50			
Completeness	Groups	39	37.76	1472.50	692.500	-0.783	0.433
	ChatGPT-4o	39	41.24	1608.50			
Concreteness	Groups	39	34.15	1332.00	552.000	-2.315	0.021
	ChatGPT-4o	39	44.85	1749.00			

Students' perceptions of group-generated and ChatGPT-4o feedback

Thematic analysis of interview data revealed distinct perceptions of group and ChatGPT-4o feedback, as shown in Tables 4 and 5.

For group feedback, students valued its comprehensiveness, noting that it often addressed multiple aspects of their design proposals, including ideas, expression, logic and formatting. For example, a student from Group 6 remarked that their peers' feedback was thorough, covering technological features, practicality, strengths, costs, user needs and formatting. Group feedback was also appreciated for its specificity, as it provided detailed and precise evaluations. A student from Group 7 emphasised that their peers accurately identified both strengths and weaknesses, including minor issues. However, students also recognised its subjective nature. Some expressed concerns that personal opinions, emotions, experiences or even considerations related to final grades could bias peer feedback. As one student from Group 7 observed, certain comments and scores appeared less objective and potentially inaccurate due to these subjective factors.

For ChatGPT-4o feedback, students praised its constructiveness, highlighting its actionable and improvement-oriented suggestions. For instance, a student from Group 1 mentioned that ChatGPT-4o offered detailed recommendations, such as the need for a comprehensive user manual. Its specificity was also valued, as it provided targeted comments that identified both strengths and shortcomings. A student from Group 4 noted that ChatGPT-4o captured key aspects of their design and offered concrete suggestions for refinement. Moreover, ChatGPT-4o's objectivity was a major advantage: students appreciated that it adhered strictly to the assessment criteria and delivered consistent, unbiased evaluations. A student from Group 8 described it as "objective and acceptable". Nevertheless, some students perceived ChatGPT-4o's feedback as somewhat mechanistic or template-driven. A student from Group 5 described it as rigid and "robotic", suggesting that it sometimes lacked depth and personalisation.

Table 4
Students' perceptions of group feedback

Themes	Definition	Examples
Completeness	The feedback covers all aspects of the work, including ideas, expression, logic and formatting.	Feedback 1 was comprehensive, mentioning the technological features, practicality, strengths, costs, user needs and formatting of our proposal (Group 6-1).
Specificity	The feedback is clear, detailed, and targeted, accurately reflecting the true level of the work.	Feedback 1 accurately identified the strengths and weaknesses of our proposal, including some detailed issues (Group 7-2).

Themes	Definition	Examples
Subjectivity	The feedback is influenced by personal emotions, opinions, experiences and biases.	Feedback 1 may have been influenced by personal factors or final grades, making some comments and scores seem less objective and accurate (Group 7-2).

Note. In each identifier, the first number indicates the group and the second indicates the student within that group; for example, Group 6-1 refers to Student 1 in Group 6.

Table 5
Students' perceptions of ChatGPT-4o feedback

Themes	Definition	Examples
Constructiveness	The feedback provides targeted suggestions and ideas for improvement based on the strengths and weaknesses of the work.	Feedback 2 included detailed suggestions for improvement, such as providing a comprehensive user manual (Group 1-2).
Specificity	The feedback is clear, detailed and targeted, accurately reflecting the true level of the work.	Feedback 2 accurately captured the highlights and shortcomings of the proposal and offered detailed suggestions for improvement (Group 4-3).
Objectivity	The feedback accurately reflects the quality and level of the work based on the scoring criteria.	Feedback 2 was scored strictly according to the scoring criteria, and I found it quite objective and acceptable (Group 8-1).
Templatisation	The feedback follows a fixed format and phrasing, appearing mechanical and lacking a personal touch.	The phrasing in Feedback 2 seemed somewhat rigid, like it was generated by a robot, lacking flexibility (Group 5-3).

Note. In each identifier, the first number indicates the group and the second indicates the student within that group; for example, Group 6-1 refers to Student 2 in Group 1.

Adoption and application of group-generated and ChatGPT-4o feedback in learning improvements

Table 6 summarises the extent to which groups adopted group feedback and ChatGPT-4o feedback across the two experimental rounds. Across both rounds, a total of 127 revisions were made. In the first round, 29 revisions were based on group feedback, compared to 19 from ChatGPT-4o. This trend reversed in the second round, where 28 revisions stemmed from ChatGPT-4o feedback and 27 from group feedback. Overall, group feedback was applied slightly more often than ChatGPT-4o feedback, with implementation rates of 44.09% and 37%, respectively. Additionally, 3.15% of revisions integrated both feedback types. In total, 107 out of 127 revisions were directly adopted from one or both feedback sources (group feedback only, ChatGPT-4o feedback only, or combined feedback), corresponding to an overall acceptance rate of 85.09%. Notably, some groups did not rely directly on either feedback source, instead opting for independent group discussions or alternative methods, which accounted for 15.75% of revisions (20 total). This tendency was more evident in the second round, where 16 out of 72 revisions were independent, reflecting an increased inclination toward self-directed revision in the later stage of the experiment.

Table 6
Revision behaviours matched with group and ChatGPT-4o feedback

Reference source	Round 1	Round 2	Total
Revision based on group feedback only	29	27	56 (44.09%)
Revision based on ChatGPT-4o feedback only	19	28	47 (37.00%)
Revision based on both	3	1	4 (3.15%)
Extra revision	4	16	20 (15.75%)
Total	55	72	127 (100%)

To further examine feedback acceptance, we aggregated the total revisions per group across both rounds and categorised them into four types: group feedback only, ChatGPT-4o feedback only, combined feedback, extra revisions. The mean number of accepted revisions per group was 8.92, obtained by averaging the total accepted revisions across the 12 groups (N = 107). Groups above this threshold were classified as "high-acceptance", while those below were considered "low-acceptance" (Table 7).

Table 7
Grouping of high- and low-acceptance groups

Group	Group assessment	ChatGPT-4o assessment	Both	Neither	Acceptance group
1	7	0	0	1	Low acceptance
2	3	3	1	2	Low acceptance
3	2	11	0	0	High acceptance
4	3	4	0	0	Low acceptance
5	3	8	0	1	High acceptance
6	3	4	1	0	Low acceptance
7	5	1	0	0	Low acceptance
8	10	6	1	10	High acceptance
9	8	0	0	1	Low acceptance
10	4	4	0	1	Low acceptance
11	6	4	0	0	High acceptance
12	2	2	1	4	Low acceptance

A comparative analysis (Figure 2) revealed distinct patterns between high- and low-acceptance groups. High-acceptance groups incorporated more ChatGPT-4o feedback, averaging 7.25 accepted revisions, whereas low-acceptance groups relied more on peer feedback, averaging 4.38 revisions. Furthermore, differences in the perceived quality of feedback (Figure 3) highlighted that high-acceptance groups achieved significantly higher scores across all five dimensions, with accuracy showing the largest discrepancy.

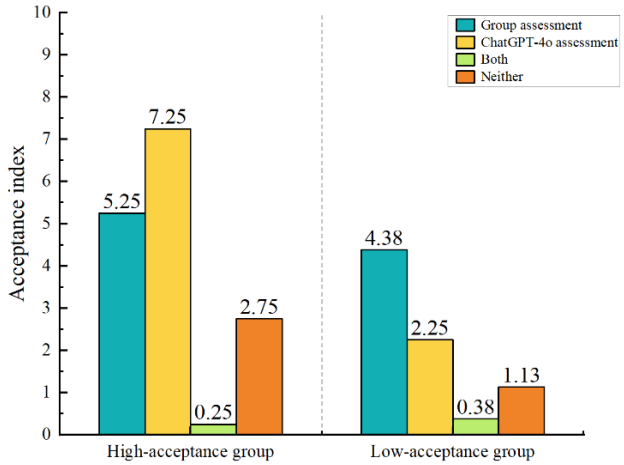

Figure 2. Differences in feedback uptake between high- and low-acceptance groups

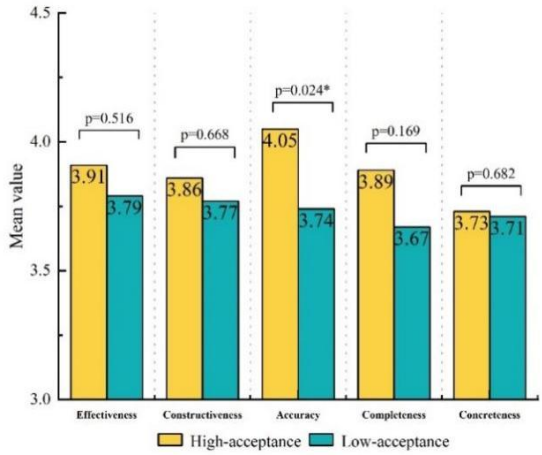

Figure 3. Differences in feedback quality between high- and low-acceptance groups

Discussion

Discussion on the results of RQ1

The first-round peer assessment findings indicated no statistically significant differences in quality between group feedback and ChatGPT-4o-generated feedback. This outcome may be attributed to the use of clearly defined evaluation criteria in both conditions, which promoted consistency in feedback quality. This finding is consistent with research emphasising the role of explicit criteria in improving feedback effectiveness (Roscoe et al., 2013). The slightly higher mean quality of group feedback across the five dimensions may reflect the collaborative nature of group discussions, where diverse perspectives are integrated to generate well-grounded arguments (Noroozi et al., 2011; Weinberger et al., 2007). In contrast, in the second-round peer assessment, ChatGPT-4o feedback received higher average ratings in constructiveness, accuracy, and concreteness than group feedback. This shift can be explained by task- and context-related factors. Compared with the initial submissions, the second-round designs were more refined and structured, enabling the model to generate more precise and actionable feedback. Studies have similarly indicated that the effectiveness of AI-generated feedback depends strongly on input clarity and completeness (De Winter et al., 2023; Jo, 2024). In addition, the model's consistent application of evaluation criteria may have provided an advantage over human groups, whose judgements are more susceptible to subjective interpretation and fatigue across repeated assessments.

Taken together, these findings suggest that although group assessments may initially benefit from collaborative reasoning processes, ChatGPT-4o demonstrates strong iterative assessment capabilities, achieving comparable or even superior feedback quality in later stages. This highlights the potential of generative AI as a reliable assessment partner that can complement – and in some contexts enhance – traditional peer assessment practices (Barrot, 2024; Farrokhnia et al., 2024; Guo & Wang, 2024; Su et al., 2023).

Discussion on the results of RQ2

Analysis of interview data revealed distinct strengths and weaknesses in both group and ChatGPT-4o feedback. Group feedback was perceived as comprehensive and context sensitive, largely due to the collaborative process through which members shared knowledge and integrated multiple perspectives related to content, language and structure. Such collaborative sense-making enables more nuanced evaluation and aligns with research showing that collective discussion supports clearer judgement and deeper understanding (Yukawa, 2006). Students' disciplinary backgrounds and interactive discussions further contributed to the specificity of group feedback, supporting findings that peer collaboration enhances critical thinking and evaluative precision (Zhang et al., 2023). However, participants also

highlighted the subjective nature of group feedback, as evaluations may be shaped by individual emotions and experiences, potentially affecting consistency and impartiality (Panadero, 2016).

ChatGPT-4o feedback, in contrast, was perceived as constructive, specific and objective. Its cross-disciplinary knowledge base enables the generation of clear and actionable suggestions, consistent with evidence that AI systems can provide detailed diagnostic and solution-oriented feedback (Finnie-Ansley et al., 2022; Ouyang et al., 2024). Moreover, the model's systematic adherence to evaluation criteria contributes to greater consistency and reduced emotional bias (Fokides & Peristeraki, 2025; Kooli & Yusuf, 2025). Despite these strengths, students noted a tendency towards formulaic expression, reflecting the model's reliance on standardised linguistic and evaluative patterns. Such templated responses may limit personalisation and reduce sensitivity to nuanced aspects of student work, a limitation also identified in research (AlGhamdi, 2024).

Overall, the findings suggest a complementary relationship between group and AI feedback. While group assessment provides contextualised and socially grounded insights, ChatGPT-4o contributes consistency and structured guidance. Integrating both sources may therefore produce more balanced and effective feedback outcomes, supporting emerging evidence on the synergistic potential of human–AI collaborative assessment (Banihashem et al., 2024).

Discussion on the results of RQ3

The findings indicated a high level of student acceptance of group and ChatGPT-4o feedback, with an overall acceptance rate of 85.09% during the revision process. This suggests that both feedback sources were perceived as reliable and actionable. Interview evidence further showed that group feedback was valued for its comprehensiveness and specificity, whereas ChatGPT-4o feedback was perceived as constructive, specific, and objective (Tables 4 and 5). Importantly, acceptance extended beyond surface-level satisfaction and shaped students' revision strategies: group feedback often supported immediate textual revisions, while ChatGPT-4o feedback encouraged reconsideration of broader design coherence and logic. This indicates that feedback acceptance involves an interpretive process in which students decide not only whether to adopt feedback but also how to integrate it into their learning processes.

Approximately 15.75% of revisions were classified as extra revisions, suggesting that feedback also functioned as a catalyst for additional improvements beyond explicit suggestions. Similarly, research has shown that general and forward-looking feedback can stimulate further revisions and reflective engagement (Lu et al., 2024). Rather than responding mechanically to corrective comments, students engaged in reflective dialogue with feedback, evaluating its relevance and generating new revisions. This process reflects metacognitive regulation, whereby learners monitor, evaluate and adapt their strategies during revision (AlGhamdi, 2024).

A notable shift was observed across rounds: group feedback was more readily accepted in the first round, whereas acceptance of ChatGPT-4o feedback increased substantially in the second round. Interview data suggest that this change resulted not only from improved feedback quality but also from students' growing familiarity with the model's evaluative logic. Repeated exposure enabled students to better interpret AI-generated suggestions, fostering trust and supporting the development of self-regulatory revision routines. This finding aligns with studies indicating that sustained interaction enhances learners' openness to AI-supported feedback (Noroozi et al., 2016; Valero Haro et al., 2024).

Overall, group feedback showed slightly higher acceptance, likely because its suggestions were closely aligned with students' intentions and easier to implement. In contrast, ChatGPT-4o feedback, while often more constructive, sometimes required greater cognitive effort to interpret due to its professional and formulaic expression. From a self-regulation perspective, acceptance appears to depend not only on perceived quality but also on cognitive accessibility: feedback that can be readily translated into action is more likely to be adopted, whereas more complex feedback demands additional reflective effort (Grimes & Warschauer, 2010).

Accuracy also emerged as a critical factor influencing feedback uptake (Figure 3). Accurate and contextually relevant feedback enhanced students' confidence in self-evaluation and strengthened trust in the feedback source, supporting autonomous learning behaviours. At the same time, acceptance was shaped by task conditions, including the initial quality of group work, which influenced perceived needs for revision (Kerman et al., 2024; Noroozi et al., 2016). Taken together, these findings suggest that feedback acceptance should be understood as a multidimensional process shaped by feedback quality, cognitive accessibility, self-regulatory engagement and learners' reflective practices. This perspective highlights the value of hybrid feedback systems that integrate the contextual sensitivity of peer feedback with the consistency and reliability of AI-generated feedback.

Implications, limitations and conclusions

By comparing feedback quality, students' perceptions and revision behaviours across group and ChatGPT-4o feedback, this study provides pedagogical and analytical insights for improving peer assessment in higher education.

Implications

First, the findings highlight the central role of clear evaluation criteria in ensuring feedback quality. Providing explicit rubrics and targeted training can enhance consistency, fairness and effectiveness in peer assessment processes. Second, group and ChatGPT-4o feedback demonstrated complementary strengths and were broadly accepted by students, suggesting that integrating human and AI feedback can optimise assessment outcomes. While group feedback offers contextualised and discussion-based insights, ChatGPT-4o contributes structured, constructive and objective guidance. Third, educators should remain attentive to potential risks associated with AI-generated feedback, particularly hallucinations and pedagogically unsound suggestions. Human oversight remains essential to verify AI outputs and provide corrective guidance. The observed improvement in ChatGPT-4o feedback across rounds further indicates that prompt design, task clarity and iterative interaction may influence AI feedback quality, highlighting the importance of structured instructional design when integrating AI into assessment practices.

From a practical perspective, these findings support positioning ChatGPT-4o as a supplementary feedback source rather than a standalone assessment tool. A hybrid feedback approach may enhance learning outcomes by combining AI consistency with human contextual sensitivity, while maintaining instructor supervision to ensure educational reliability.

Limitations

Several limitations should be acknowledged. First, the study involved a relatively small sample and short intervention period, and participants had limited prior experience with ChatGPT-4o. Future studies should include longer-term implementations and multiple experimental rounds to examine developmental trends in AI-supported feedback. Second, participants were primarily drawn from the educational technology field, limiting generalisability across disciplines. Broader and more diverse samples are needed to validate the findings in varied academic contexts. Third, the study focused mainly on textual assessment and did not fully explore ChatGPT-4o's multimodal evaluation capabilities. Future research could investigate prompt designs supporting multimodal feedback. In addition, the study did not directly examine how feedback influenced measurable improvements in revised work, leaving aspects of the feedback–learning relationship under-explored. Finally, the validity of AI-generated feedback was not systematically evaluated. Future research should incorporate validation mechanisms, such as expert review or fact-checking protocols, to ensure that AI feedback is not only fluent but also accurate and pedagogically sound.

Conclusions

This study provides a comprehensive comparison of group and ChatGPT-4o feedback within peer assessment contexts. Results show comparable feedback quality in the first round, whereas ChatGPT-4o demonstrated advantages in constructiveness, accuracy and specificity in the second round. Interview findings indicated that students perceived group feedback as comprehensive yet subjective, while ChatGPT-4o feedback was viewed as constructive and objective but sometimes formulaic. Both feedback sources achieved high acceptance rates, though preferences varied across learner groups.

Rather than advocating the uncritical adoption of AI as an assessment tool, the findings highlight a balanced perspective: generative AI can enhance peer assessment through consistency, structured guidance and scalability, but risks such as hallucination necessitate sustained human oversight. Future research should further explore long-term human–AI collaborative feedback models and develop safeguards that ensure the validity, reliability and educational integrity of AI-supported assessment across diverse learning environments.

Author contributions

Xin Li: Conceptualisation, Methodology, Supervision, Writing – review and editing; **Yue Zhang:** Data curation, Formal analysis, Investigation, Writing – original draft; **Tingyu Liu:** Resources, Software, Validation, Writing – review and editing.

Acknowledgements

This paper was supported by the National Natural Science Foundation of China (grant number 62507025), Ministry of Education Youth Program for Humanities and Social Sciences Research (grant number 25YJCZH138) and 2025 Jiangsu Provincial Social Science Fund Youth Project (grant number 25JYC007).

References

AlGhamdi, R. (2024). Exploring the impact of ChatGPT-generated feedback on technical writing skills of computing students: A blinded study. *Education and Information Technologies*, *29*(14), 18901–18926. https://doi.org/10.1007/s10639-024-12594-2

Allagui, B. (2023). Chatbot feedback on students' writing: Typology of comments and effectiveness. In O. Gervasi, B. Murgante, A. M. A. C. Rocha, C. Garau, F. Scorza, Y. Karaca, & C. M. Torre (Eds.), *Lecture notes in computer science: Vol. 14112*. *Computational science and its applications – ICCSA 2023 workshops* (pp. 377–384). Springer. https://doi.org/10.1007/978-3-031-37129-5_31

Alshuraidah, A., & Storch, N. (2019). Investigating a collaborative approach to peer feedback. *ELT Journal*, *73*(2), 166–174. https://doi.org/10.1093/elt/ccy057

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*, Article 23. https://doi.org/10.1186/s41239-024-00455-4

Barrot, J. S. (2024). ChatGPT as a language learning tool: An emerging technology report. *Technology, Knowledge and Learning*, *29*(2), 1151–1156. https://doi.org/10.1007/s10758-023-09711-4

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

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

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/10.3102/00346543065003245

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

Chen, W., Tan, J. S., & Pi, Z. (2021). The spiral model of collaborative knowledge improvement: An exploratory study of a networked collaborative classroom. *International Journal of Computer-Supported Collaborative Learning*, *16*(1), 7–35. https://doi.org/10.1007/s11412-021-09338-6

Cheong, C. M., Luo, N., Zhu, X., Lu, Q., & Wei, W. (2023). Self-assessment complements peer assessment for undergraduate students in an academic writing task. *Assessment & Evaluation in Higher Education*, *48*(1), 135–148. https://doi.org/10.1080/02602938.2022.2069225

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, *39*, 629–643. https://doi.org/10.1007/s11251-010-9146-1

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y. S., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

De Winter, J. C., Dodou, D., & Stienen, A. H. (2023). ChatGPT in education: Empowering educators through methods for recognition and assessment. *Informatics*, *10*(4), Article 87. https://doi.org/10.3390/informatics10040087

Er, E., Dimitriadis, Y., & Gašević, D. (2021). A collaborative learning approach to dialogic peer feedback: A theoretical framework. *Assessment & Evaluation in Higher Education*, *46*(4), 586–600. https://doi.org/10.1080/02602938.2020.1786497

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2022). The robots are coming: Exploring the implications of OpenAI Codex on introductory programming. In J. Sheard & P. Denny (Eds.), *Proceedings of the 24th Australasian Computing Education Conference* (pp. 10–19). Association for Computing Machinery. https://doi.org/10.1145/3511861.3511863

Fokides, E., & Peristeraki, E. (2025). Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek. *Education and Information Technologies*, *30*(2), 2577–2621. https://doi.org/10.1007/s10639-024-12912-8

Freeman, M., & McKenzie, J. (2002). SPARK, a confidential web-based template for self and peer assessment of student teamwork: Benefits of evaluating across different subjects. *British Journal of Educational Technology*, *33*(5), 551–569. https://doi.org/10.1111/1467-8535.00291

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment*, *8*(6). https://ejournals.bc.edu/index.php/jtla/article/view/1625

Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education*, *63*, Article 100962. https://doi.org/10.1016/j.iheduc.2024.100962

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hadwin, A., Järvelä, S., & Miller, M. (2018). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 83–106). Routledge. https://doi.org/10.4324/9781315697048-6

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

Hwang, G. J., & Chang, S. C. (2021). Facilitating knowledge construction in mobile learning contexts: A bi-directional peer-assessment approach. *British Journal of Educational Technology*, *52*(1), 337–357. https://doi.org/10.1111/bjet.13001

Jo, H. (2024). From concerns to benefits: A comprehensive study of ChatGPT usage in education. *International Journal of Educational Technology in Higher Education*, *21*, Article 35. https://doi.org/10.1186/s41239-024-00471-4

Kerman, N. T., Noroozi, O., Banihashem, S. K., Karami, M., & Biemans, H. J. (2024). Online peer feedback patterns of success and failure in argumentative essay writing. *Interactive Learning Environments*, *32*(2), 614–626. https://doi.org/10.1080/10494820.2022.2093914

Kessler, G. (2009). Student-initiated attention to form in wiki-based collaborative writing. *Language Learning & Technology*, *13*(1), 79–95. https://doi.org/10.64152/10125/44169

Kirschner, P. A., & Erkens, G. (2013). Toward a framework for CSCL research. *Educational Psychologist*, *48*(1), 1–8. https://doi.org/10.1080/00461520.2012.750227

Kleinman, Z. (2024, May 14). OpenAI's new model GPT-4o can teach maths and flirts - but still glitches. *BBC News*. https://www.bbc.com/news/articles/cv2xx1xe2evo

Kooli, C., & Yusuf, N. (2025). Transforming educational assessment: Insights Into the Use of ChatGPT and large language models in grading. *International Journal of Human-Computer Interaction*, *41*(5), 3388–3399. https://doi.org/10.1080/10447318.2024.2338330

Kurniawan, C., Pujiarti, H., Taufiqurrahman, F., Yunikawati, N. A., & Nordin, R. M. (2025). AI-assisted differentiated instruction (DI) for optimizing more personal learning paths. In H. Hanafi (Chair), *Proceedings of the 2025 11th International Conference on Education and Technology* (pp. 412–416). IEEE. https://doi.org/10.1109/icet67257.2025.11290951

LI, L., & Gao, F. (2016). The effect of peer assessment on project performance of students at different learning levels. *Assessment & Evaluation in Higher Education*, *41*(6), 885–900. https://doi.org/10.1080/02602938.2015.1048185

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education*, *49*(5), 616–633. https://doi.org/10.1080/02602938.2024.2301722

XMa, M., & Bui, G. (2022). Implementing continuous assessment in an academic English writing course: An exploratory study. *Assessing Writing*, *53*, Article 100629. https://doi.org/10.1016/j.asw.2022.100629

Memari Hanjani, A. (2016). Collaborative revision in L2 writing: Learners' reflections. *ELT Journal*, *70*(3), 296–307. https://doi.org/10.1093/elt/ccv053

Mohamed, A. M. (2024). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: Perceptions of EFL faculty members. *Education and Information Technologies*, *29*(3), 3195–3217. https://doi.org/10.1007/s10639-023-11917-z

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

Noroozi, O., Biemans, H. J. A., Busstra, M. C., Mulder, M., & Chizari, M. (2011). Differences in learning processes between successful and less successful students in computer-supported collaborative learning in the field of human nutrition and health. *Computers in Human Behavior*, *27*(1), 309–318. https://doi.org/10.1016/j.chb.2010.08.009

Noroozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *The Internet and Higher Education*, *31*, 20–31. https://doi.org/10.1016/j.iheduc.2016.05.002

OpenAI. (2024, May 13). *Hello GPT-4o*. https://openai.com/index/hello-gpt-4o/

Ouyang, F., Guo, M., Zhang, N., Bai, X., & Jiao, P. (2024). Comparing the effects of instructor manual feedback and ChatGPT intelligent feedback on collaborative programming in China's higher education. *IEEE Transactions on Learning Technologies*, *17*, 2173–2185. https://doi.org/10.1109/tlt.2024.3486749

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 247–266). Routledge.

Preiksaitis, C., & Rose, C. (2023). Opportunities, challenges, and future directions of generative artificial intelligence in medical education: Scoping review. *JMIR Medical Education*, *9*, Article e48785. https://doi.org/10.2196/48785

Ramon-Casas, M., Nuño, N., Pons, F., & Cunillera, T. (2019). The different impact of a structured peer-assessment task in relation to university undergraduates' initial writing skills. *Assessment & Evaluation in Higher Education*, *44*(5), 653–663. https://doi.org/10.1080/02602938.2018.1525337

Roscoe, R. D., Varner, L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology*, *8*(4), 362–381. https://doi.org/10.1504/IJLT.2013.059131

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Shvets, O., Murtazin, K., Meeter, M., & Piho, G. (2025). Experiment with ChatGPT: Methodology of first simulation. *Frontiers in Education*, *8*(10), Article 1624516. https://doi.org/10.3389/feduc.2025.1624516

Sol, K., & Heng, K. (2024). AI-powered chatbots as personalized academic writing assistants for non-native English speakers. In M. A. Peters & R. Heraud (Eds.), *Encyclopedia of educational innovation* (pp. 1–5). Springer. https://doi.org/10.1007/978-981-13-2262-4_313-1

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, *57*, Article 100752. https://doi.org/10.1016/j.asw.2023.100752

Tan, J. S., & Chen, W. (2022). Peer feedback to support collaborative knowledge improvement: What kind of feedback feed-forward? *Computers & Education*, *187*, Article 104467. https://doi.org/10.1016/j.compedu.2022.104467

Tenenbaum, H. R., Winstone, N. E., Leman, P. J., & Avery, R. E. (2020). How effective is peer interaction in facilitating learning? A meta-analysis. *Journal of Educational Psychology*, *112*(7), 1303–1319. https://doi.org/10.1037/edu0000436

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, *379*(6630), 313. https://doi.org/10.1126/science.adg7879

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.2307/1170598

Topping, K. (2017). Peer assessment: Learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology*, *1*(1), Article 7. https://doi.org/10.31532/InterdiscipEducPsychol.1.1.007

Urzúa, C. A. C., Ranjan, R., Saavedra, E. E. M., Badilla-Quintana, M. G., Lepe-Martínez, N., & Philominraj, A. (2025). Effects of AI-assisted feedback via generative chat on academic writing in higher education students: A systematic review of the literature. *Education Sciences*, *15*(10), Article 1396. https://doi.org/10.3390/educsci15101396

Valero Haro, A., Noroozi, O., Biemans, H. J., Mulder, M., & Banihashem, S. K. (2024). How does the type of online peer feedback influence feedback quality, argumentative essay writing quality, and domain-specific learning? *Interactive Learning Environments*, *32*(9), 5459–5478. https://doi.org/10.1080/10494820.2023.2215822

Weinberger, A., Stegmann, K., Fischer, F., & Mandl, H. (2007). Scripting argumentative knowledge construction in computer-supported learning environments. In F. Fischer, I. Kollar, H. Mandl, & J. M. Haake (Eds.), *Scripting computer-supported collaborative learning: Cognitive, computational and educational perspectives* (pp. 191–211). Springer. https://doi.org/10.1007/978-0-387-36949-5_12

Wiggers, K. (2024, May 13). OpenAI debuts GPT-4o 'omni' model now powering ChatGPT. *TechCrunch*. https://techcrunch.com/2024/05/13/openais-newest-model-is-gpt-4o/

Yukawa, J. (2006). Co-reflection in online learning: Collaborative critical thinking as narrative. *International Journal of Computer-Supported Collaborative Learning*, *1*, 203–228. https://doi.org/10.1007/s11412-006-8994-9

Zhang, S., Li, H., Wen, Y., Zhang, Y., Guo, T., & He, X. (2023). Exploration of a group assessment model to foster student teachers' critical thinking. *Thinking Skills and Creativity*, *47*, Article 101239. https://doi.org/10.1016/j.tsc.2023.101239

Zhang, Y., Pi, Z., Chen, L., Zhang, X., & Yang, J. (2021). Online peer assessment improves learners' creativity: not only learners' roles as an assessor or assessee, but also their behavioral sequence matter. *Thinking Skills and Creativity*, *42*, Article 100950. https://doi.org/10.1016/j.tsc.2021.100950

Corresponding author: Xin Li, lix2023@jsnu.edu.cn

Please cite as: Lin, X., Zhang, Y., & Liu, T. (2026). Comparative analysis of peer group and AI-generated feedback in peer assessment: Insights into feedback quality and student perceptions in higher education. *Australasian Journal of Educational Technology*, *42*(4), 129–146. https://doi.org/10.14742/ajet.11467