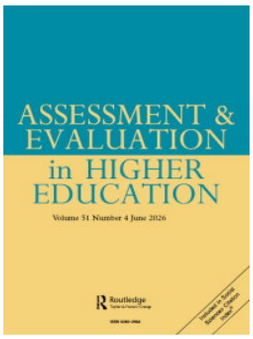

Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

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

Artificial intelligence and feedback in university education: effectiveness and student perceptions

Valentina Grion, Beatrice Doria, Daniele Agostini & Giorgia Slaviero

To cite this article: Valentina Grion, Beatrice Doria, Daniele Agostini & Giorgia Slaviero (08 Jul 2026): Artificial intelligence and feedback in university education: effectiveness and student perceptions, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2026.2697962

To link to this article: https://doi.org/10.1080/02602938.2026.2697962

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2026.2697962

Routledge
Taylor & Francis Group

Artificial intelligence and feedback in university education: effectiveness and student perceptions

Valentina Grion[a], Beatrice Doria[a], Daniele Agostini[b] and Giorgia Slaviero[c]

[a]Department of Human, Education and Sport Sciences, Pegaso Telematic University, Naples, Italy; [b]Department of Psychology and Cognitive Science, University of Trento, Trento, Italy; [c]Department of Philosophy, Sociology, Education and Applied Psychology (FisPPA), University of Padua, Padua, Italy

ABSTRACT

The integration of generative artificial intelligence (AI) into Higher Education has intensified debates about the role of technology in formative assessment. This study examines the effectiveness and practical comparability of AI-generated feedback in a project-based university course, comparing two large language models (GPT-o4-mini and DeepSeek R1) with feedback provided by an expert human teacher. Adopting a quasi-experimental design, 47 student groups ($N = 238$) were randomly assigned to one of three feedback conditions. Changes in project performance were analysed using non-parametric tests, robust models, and non-inferiority and equivalence analyses. Students' perceptions were also assessed through a validated questionnaire ($N = 200$). Results showed significant improvement in project performance from pre- to post-feedback across all conditions ($rrb = 0.77$), with no significant differences between feedback sources. Equivalence analyses indicated practical comparability between GPT-o4-mini and teacher feedback, while DeepSeek R1 demonstrated non-inferiority. Students' perceptions of mastery, emotions, and satisfaction were similarly high across conditions. Findings suggest that feedback effectiveness depends less on its source than on the pedagogical architecture in which it is embedded. When supported by strong assessment literacy and explicit criteria, AI-generated feedback can function as a credible component of formative assessment in higher education.

KEYWORDS

Formative assessment; AI-generated feedback; assessment literacy; higher education

Introduction

Assessment feedback is central to student learning in higher education. High-quality feedback clarifies expectations, justifies grades, identifies strengths and weaknesses, and supports self-regulation (Lipnevich and Smith 2022; Panadero, Jonsson, and Botella 2017; Yan and Carless 2022). Conversely, delayed or misaligned feedback may undermine engagement and achievement (Middleton et al. 2023). Timeliness is especially

CONTACT Beatrice Doria beatrice.doria@unipegaso.it Department of Human, Education and Sport Sciences, Pegaso Telematic University, Naples, Italy.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02602938.2026.2697962.

critical in iterative tasks such as project-based learning, where late feedback may lose its regulatory value (Wible et al. 2001). Recent perspectives conceptualise feedback as a dynamic process shaped by contextual factors, feedback agents, and students' cognitive, motivational, and emotional characteristics (Panadero 2023).

Despite its importance, providing timely, personalised, and high-quality feedback remains challenging. Growing enrolments and increasingly complex tasks limit instructors' capacity to deliver detailed formative input (Henderson, Ryan, and Phillips 2019). Students frequently report dissatisfaction and call for feedback that is constructive, actionable, and improvement-oriented (Henderson, Ryan, and Phillips 2019; Mulliner and Tucker 2017). These pressures have intensified interest in technological solutions capable of scaling feedback while reducing workload (Dai et al. 2024).

AI has emerged as a promising yet debated response. Early implementations relied on Learning Analytics systems that automate feedback or trigger interventions based on behavioural and performance data (Chatti et al. 2012; Pardo et al. 2018). While enhancing scalability and timeliness, such systems are constrained by rule-based logic, predefined indicators, and the need for teacher mediation to ensure quality (Dai et al. 2024), particularly in open-ended and complex tasks such as essays or project proposals (Beckman et al. 2021).

Recent advances in generative AI, especially Large Language Models (LLMs) based on Generative Pre-trained Transformer architectures, have reshaped this debate. Unlike rule-based systems, LLMs generate coherent, context-sensitive feedback on open-ended student work (Dai et al. 2024; Yan et al. 2024). Empirical studies highlight their potential to support formative feedback, rubric-based assessment, and iterative revision across contexts (Agostini and Picasso 2024; MacNeil et al. 2022; Pankiewicz and Baker 2023).

Against this background, the present study examines the educational effectiveness and practical comparability of AI-generated feedback in higher education. Rather than offering a comprehensive review of AI-supported systems (Ba et al. 2025; Shi and Aryadoust 2024), it addresses a focused question: to what extent can contemporary AI-generated feedback effectively support project-based learning and be considered comparable to feedback provided by an experienced teacher?

Adopting an empirical comparative design, the study analyses changes in project performance over time, compares outcomes across feedback sources, explores attendance status as a moderating factor, and investigates students' perceptions of AI-generated versus human feedback. By integrating objective performance indicators with subjective experiences, it seeks to clarify whether, and under what conditions, AI-based feedback can function as a credible and educationally meaningful component of formative assessment in higher education.

Literature review

From learning analytics–supported feedback to generative AI: scalability and tensions

Learning Analytics (LA) has emerged as an interdisciplinary field aimed at generating actionable insights to improve teaching and learning processes (Chatti et al. 2012).

In assessment contexts, LA-supported systems have been adopted to address challenges of scalability and timeliness in feedback provision, particularly in large-enrolment higher education courses (Dai et al. 2024). Research shows that LA approaches can support personalised and real-time feedback by leveraging behavioural and performance data from learning management systems (Arthars et al. 2019; Pardo et al. 2018). Systems such as OnTask enable instructors to automate feedback messages based on predefined indicators (Lim et al. 2021; Pardo et al. 2018), while predictive analytics models can trigger targeted interventions to address disengagement or underperformance, with reported benefits for motivation and outcomes (Arthars et al. 2019; Azcona, Hsiao, and Smeaton 2019).

Despite these advantages, LA-supported feedback remains limited. Although such systems may reduce workload, they continue to rely heavily on teachers for interpretation, evaluation, and the formulation of pedagogically meaningful feedback (Dai et al. 2024). Moreover, rule-based logics and predefined indicators constrain their capacity to address complex, open-ended, and creative tasks, such as essays or project-based assignments (Beckman et al. 2021; Dai et al. 2024). Consequently, LA systems are often more effective for monitoring engagement or delivering generalised performance feedback than for supporting nuanced formative assessment grounded in qualitative judgement. Overall, LA represents a valuable yet partial response to feedback challenges, improving scalability but struggling with adaptability, pedagogical depth, and alignment with student-centred conceptions of assessment.

Against this backdrop, advances in generative AI – particularly Large Language Models (LLMs) based on Generative Pre-trained Transformer architectures – have renewed interest in the automation and enhancement of feedback (Dai et al. 2024; Yan et al. 2024). Unlike indicator-driven systems, LLMs generate context-sensitive and linguistically rich feedback in response to open-ended student work. Empirical studies document applications of GPT-based models in programming tasks (MacNeil et al. 2022; Pankiewicz and Baker 2023), online discussions (Lim et al. 2021), and the production of feedback components such as praise, justification, and improvement-oriented suggestions (Hirunyasiri et al. 2023). LLMs appear particularly suited to linguistically complex and revision-oriented tasks, including essays and project-based assignments (Agostini and Picasso 2024; Dai et al. 2024).

Some evidence suggests that AI-generated feedback may outperform human feedback on dimensions such as readability and structural clarity, as well as in articulating strengths and areas for improvement (Dai et al. 2023). In project-based learning contexts, GPT-4 has been shown to provide detailed feeding-forward information supporting iterative revision (Dai et al. 2024). These features position generative AI as a potentially transformative tool for formative assessment, addressing issues of timeliness, workload, and consistency.

However, concerns extend beyond effectiveness to issues of validity, transparency, trust, and ethical responsibility (Siddiq and Murchan 2024). The alignment of AI-generated feedback with theoretically grounded models, such as Hattie and Timperley (2007) levels of feedback, remains inconsistent (Dai et al. 2023), and studies report only moderate agreement with human expert judgement in evaluative accuracy and feedback polarity (Dai et al. 2024; Leighton and Bustos Gómez 2018).

Poorly calibrated feedback may undermine trust, engagement, and learning outcomes (Leighton and Bustos Gómez 2018).

Generative AI thus represents not a definitive solution, but a qualitatively different paradigm that reshapes the conditions of feedback production and use. While it offers unprecedented scalability for open-ended tasks, its integration into assessment requires careful attention to validation, instructional mediation, and alignment with feedback literacy, student agency, and self-regulated learning (Carless and Boud 2018; Dai et al. 2024; Grion and Serbati 2019).

The growing use of AI in educational contexts has attracted increasing scholarly attention, particularly regarding its potential to support feedback processes and assessment practices (Doria et al. 2025a; Zawacki-Richter et al. 2019). While feedback is widely recognised as a key driver of learning (Hattie and Timperley 2007), its effectiveness depends not only on the quality of the information provided but also on affective, motivational, and relational dimensions that shape how learners interpret and act upon feedback (Carless 2015; Chan and Hu 2023; Lipnevich and Smith 2018).

Recent developments in generative AI have opened new possibilities for educational feedback, including the generation of forms of internal feedback that may stimulate students' reflective and self-regulatory processes (Nicol 2021). At the same time, these developments raise important questions regarding the educational effectiveness of AI-generated feedback and students' perceptions of such tools when compared with feedback provided by human experts. Against this background, the present study investigates the extent to which AI-generated feedback can support project-based learning in higher education and whether it can be considered comparable to expert human feedback in terms of both learning outcomes and students' perceptions.

Methods

Aim of the study

The overall aim of the present study is to examine the effectiveness of AI-generated feedback in supporting students' project-based learning in higher education, by comparing it with feedback provided by an expert human teacher. The study adopts a dual focus on learning outcomes and students' perceptions of feedback, with the specific goal of investigating whether AI-based feedback can be considered practically comparable to human expert feedback within assessment processes.

Accordingly, the study addresses the following research questions:

RQ1. To what extent does feedback, irrespective of its source, lead to changes in students' project performance over time (PRE–POST)?

RQ2. To what extent do students' learning outcomes differ as a function of feedback source (expert human teacher vs AI-based systems)?

RQ3. To what extent does students' attendance status moderate the relationship between feedback source and learning outcomes?

RQ4. To what extent can AI-generated feedback be considered comparable to expert human feedback in terms of both learning outcomes and students' perceptions?

Study context and design

The study was conducted within the undergraduate course Assessment and Learning, offered to third-year students enrolled in the single-cycle Master's degree programme in Primary Teacher Education at the University of Padua. The course involved a total of 238 students, of whom 146 were attending and 92 were non-attending. Students were initially organised into 49 working groups (28 attending groups and 21 non-attending groups). However, two groups did not submit the project within the required timeframe and were therefore excluded from the study. Consequently, the final sample consisted of 47 working groups. The course was designed to foster the development of professional assessment competence by introducing core principles of formative assessment, assessment for learning, and summative assessment. The participant group consisted largely of female students (94%), with males accounting for 6% of the sample. Most participants were between 20 and 23 years old (66,95%), while smaller shares fell within the 24–26 (10,73%) and 27–30 (9,87%) age ranges, or were over 31 (12,45%). Overall, the age and gender distribution aligns with the typical demographic profile of students enrolled in Primary Teacher Education programmes. Information regarding participants' prior experience with ChatGPT or other generative AI tools was not collected.

As part of the course activities, students were involved in an experimental task focused on project-based learning and formative feedback (see Figure 1). Projects were evaluated using an analytic rubric specifically developed for the activity. The rubric generated a total score ranging from 0 to 30 points and was used both for feedback generation and for the assessment of project quality. The rubric assessed key dimensions of instructional design, including the alignment between learning objectives, learning activities, assessment strategies, and the overall pedagogical coherence of the project. The complete rubric is provided in Online Appendix A[1]. An example of the project assignment provided to students is included in Online Appendix B to facilitate the interpretation of the feedback process and assessment outcomes. For the purposes of the study, students were randomly assigned to one of three experimental feedback conditions: AI-generated feedback provided by GPT-o4-mini, feedback provided by an expert human teacher, and AI-generated

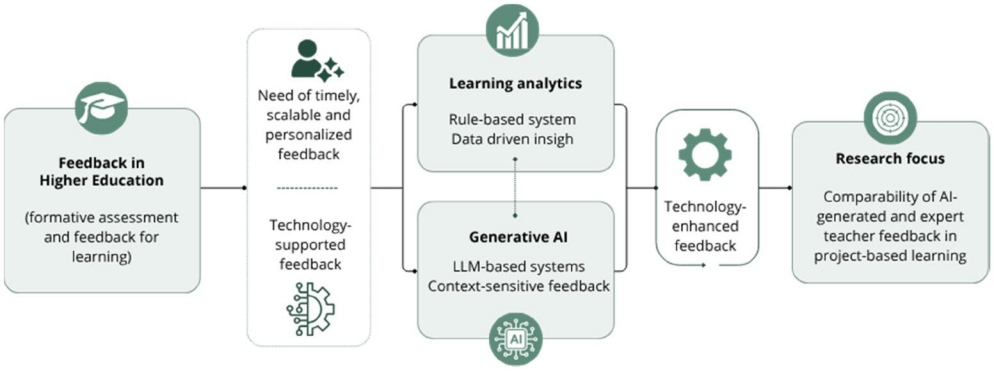

Figure 1. Conceptual framework guiding the study.

feedback provided by DeepSeek R1. The expert human teacher condition was implemented by two instructors working collaboratively throughout the study. The same instructors jointly provided the feedback in the human-feedback condition and evaluated all project submissions. During the evaluation process, they remained blind to both the assigned feedback condition and the project stage (PRE vs. POST). Random assignment was performed prior to the implementation of the activity in order to ensure initial comparability across conditions.

The 47 working groups, each consisting of four to five students, were distributed across the feedback conditions as follows: DeepSeek R1 ($n = 16$ groups), expert human teacher ($n = 16$ groups), and GPT-o4-mini ($n = 15$ groups). These working groups represented the unit of analysis for the experimental phase. Feedback was delivered consistently within each group according to the assigned condition and learning outcomes were aggregated and analysed at the group level to preserve the independence of observations and avoid pseudo-replication.

The study was structured into five sequential phases encompassing instructional design, rubric co-construction, experimental feedback provision, revision, and outcome measurement (see Figure 2).

Prompt design and construction

To generate AI-based formative feedback, the models were provided with the full set of instructional materials used during the course, including lecture slides, assigned readings, and reference texts. In addition, a dedicated document was uploaded, containing a detailed description of the assignment, the pedagogical framework underpinning the task, and the assessment rubric co-constructed with students during the course.

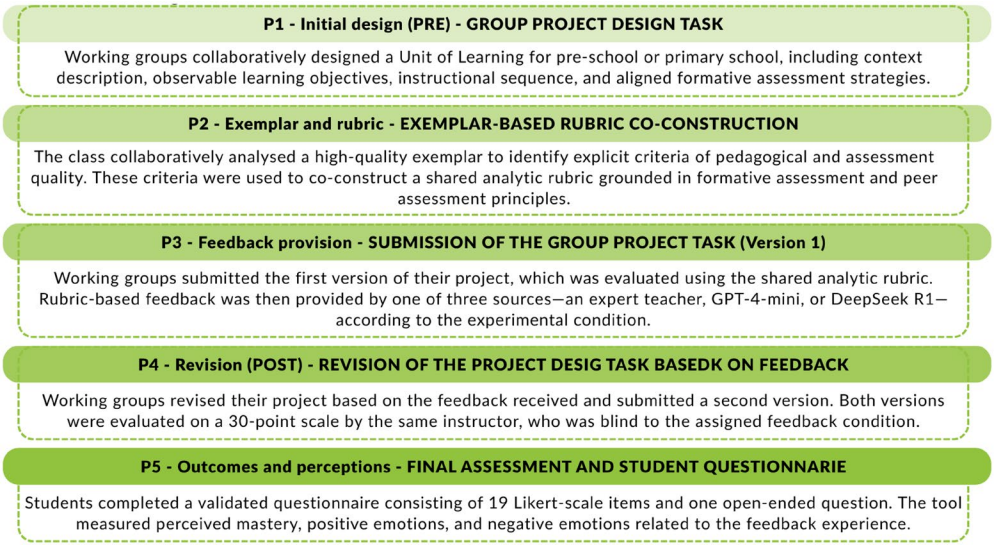

Figure 2. Instructional design.

The system-level instructions specified that the AI should act as a university professor with expertise in didactics, educational design, assessment, and feedback, and that it should provide objective, comprehensive, and well-justified formative feedback. The models were explicitly instructed to base their evaluation on the assignment specifications, the course materials, and the assessment rubric included in the provided documentation.

All materials were made accessible through a Retrieval-Augmented Generation (RAG) process, allowing the models to retrieve and reference relevant course content when generating feedback. Feedback was requested in narrative form and was expected to balance evaluative comments with actionable suggestions for improvement.

This design aimed to ensure that AI-generated feedback was aligned with the instructional context, the theoretical foundations of the course, and the assessment criteria shared with students. An example of the feedback prompt is provided in Online Appendix C.

Data analysis

Quantitative analyses were conducted using R (R Core Team) through reproducible scripts. Descriptive statistics – including mean, median, standard deviation, and range – were computed for the total project score and each rubric dimension at both pre-feedback (PRE) and post-feedback (POST).

Given the discrete and bounded nature of rubric-based scores and the observed ceiling effect at POST, the analytical strategy prioritised non-parametric and robust approaches. PRE-POST changes were examined using Wilcoxon signed-rank tests, with effect sizes estimated *via* rank-biserial correlation (*rrb*).

To examine differences in learning outcomes across feedback sources (GPT-o4-mini, expert human teacher, DeepSeek R1), raw gain scores (POST–PRE) were initially computed. However, raw gains showed strong dependency on baseline performance, limiting interpretability. They were therefore retained for descriptive purposes only; inferential analyses used more robust indicators of change, including baseline-adjusted difference scores (DIFF_ADJ) and normalised gain indices (GAIN_NORM).

Between-group comparisons were conducted using Kruskal–Wallis tests on post-test scores and baseline-adjusted change indices. The comparability of baseline (PRE) performance across conditions was preliminarily verified using a Kruskal–Wallis test. Pairwise post-hoc comparisons (Dwass–Steel–Critchlow–Fligner, DSCF) were planned only in the event of a statistically significant omnibus test; otherwise, pairwise contrasts were summarised through Hodges–Lehmann difference estimates with 95% confidence intervals, privileging estimation over significance testing. Effect sizes were quantified using epsilon squared (ε^2).

To investigate the role of attendance status (attending vs non-attending) and its interaction with feedback source, robust linear models were estimated using HC3 heteroskedasticity-consistent standard errors.

As a robustness check, a one-way Welch ANOVA with Games–Howell post-hoc comparisons was conducted on raw gain scores. Given the aforementioned limitations, these analyses were treated as sensitivity analyses and were not used to inform the main conclusions.

Finally, because the absence of statistically significant differences does not imply practical equivalence, non-inferiority and equivalence analyses were performed. These were based on Welch-adjusted 90% confidence intervals and a pre-specified equivalence margin of ±1 point on the 30-point grading scale to assess the practical comparability of AI-generated and expert human feedback.

Analysis of students' perceptions questionnaire

In parallel with the assessment of project performance, students' perceptions of the feedback received were collected using a validated questionnaire consisting of 19 closed-ended items and one open-ended item (Baydas Onlu, Abdusselam, and Yilmaz 2022).

The internal consistency of the three questionnaire scales was assessed using Cronbach's alpha. Reliability coefficients indicated satisfactory internal consistency for all dimensions: $\alpha = 0.81$ for Perceived mastery, $\alpha = 0.85$ for Positive emotions, and $\alpha = 0.73$ for Negative emotions. These values are consistent with the original validation (Baydas Onlu, Abdusselam, and Yilmaz 2022). The three composite scales were retained for subsequent analyses.

To examine potential differences in students' perceptions as a function of feedback source, composite scores (mean of the items per dimension) were compared across the three experimental conditions using Kruskal–Wallis tests for independent samples. Post-hoc pairwise comparisons (Dwass–Steel–Critchlow–Fligner, DSCF) were planned only in the event of a statistically significant omnibus test. Effect sizes were estimated using ε^2.

Results

Descriptive statistics

Descriptive statistics for overall project performance at pre- and post-feedback are reported below. Considering the full sample of project groups ($N = 47$), results indicate a marked increase in performance from PRE to POST, accompanied by a substantial reduction in score variability at post-test. Specifically, the mean total project score increased from $M = 23.81$ (SD = 4.63) at PRE to $M = 27.70$ (SD = 0.95) at POST, with the post-test median reaching 28, close to the upper bound of the 30-point grading scale. This compression is consistent with a ceiling effect.

Descriptive statistics stratified by feedback source are presented in Table 1. At PRE, baseline performance varied descriptively across experimental conditions, with greater variability in the expert human teacher condition, although these differences were not statistically significant (Kruskal–Wallis $H(2) = 1.24$, $p = 0.538$; see RQ2). At POST, however, mean scores were uniformly high across all feedback sources, with low dispersion and values clustered near the maximum score. Despite initial differences, all groups converged towards comparable post-test performance.

Substantial baseline imbalances emerged in some subgroups (especially the expert human teacher condition); given small subgroup sizes and the strong dependency observed between baseline performance and subsequent gains, statistics should be interpreted with caution. Their role is primarily contextual, supporting the rationale for the baseline-adjusted and robust analyses presented in the subsequent sections.

Descriptive statistics for raw change scores (POST – PRE) in Table 1. As shown, raw gains displayed marked variability across conditions, especially in the expert human teacher group. As noted above, these indices are reported for descriptive purposes only.

RQ1. Effect of feedback on students' project performance over time

To address RQ1, PRE-POST changes were analysed across the full sample of project groups ($N = 47$), using Wilcoxon signed-rank tests.

Results indicated a statistically significant improvement in overall project performance from PRE to POST (see Table 1), $W = 1081$, $p < 0.001$, with a large effect size (rank-biserial correlation, $rrb = 0.77$). On average, total project scores increased by approximately 3.9 points on the 30-point grading scale. Despite the ceiling effect noted above, the magnitude and consistency of the PRE–POST change provide robust evidence of improved project performance following feedback across the sample.

RQ2. Effect of feedback source on students' project performance

To address RQ2, differences in project performance as a function of feedback source were examined by comparing post-feedback scores across the three experimental conditions (expert human teacher, GPT-o4-mini, DeepSeek R1). Post-test mean scores were uniformly high across all conditions, with limited variability (Table 2).

A Kruskal–Wallis test revealed no statistically significant differences in post-feedback project performance across feedback sources, H(2) = 1.91, $p = 0.384$, with a negligible effect size ($\varepsilon^2 = 0.042$; Table 4). Baseline (PRE) scores were also comparable across conditions, H(2) = 1.24, $p = 0.538$, $\varepsilon^2 = 0.027$, supporting the equivalence of the groups prior to the feedback intervention.

Notably, 91% of the groups achieved scores of 27 or higher on the 30-point scale at post-test, indicating a substantial ceiling effect. As a result, comparisons based solely on post-test scores may underestimate potential differences between conditions. To account for this limitation, gain scores (POST – PRE), which preserve variability in baseline performance, were also examined. Consistent with the post-test analysis, gain scores did not differ significantly across feedback conditions, H(2) = 0.74, $p = 0.690$, $\varepsilon^2 = 0.016$, indicating that the source of feedback did not significantly influence improvements in project performance.

Table 1. Descriptive statistics by feedback source (PRE and POST).

Source	n	PRE mean	PRE SD	POST mean	POST SD	DIFF Mean	DIFF SD
DeepSeek	16	24.94	3.73	28.00	0.37	3.06	3.94
Teacher	16	22.69	6.24	27.44	1.37	4.75	6.59
GPT-o4-mini	15	23.80	3.28	27.67	0.82	3.87	3.50
Total	47	23.81	4.63	27.70	0.95	3.89	4.84

Table 2. Kruskal–Wallis test comparing post-feedback project performance across feedback sources.

Outcome	Test	H	df	p	ε^2
TOT_POST	Kruskal–Wallis	1.91	2	.384	.04

Note. Kruskal–Wallis test results for post-feedback total project scores (TOT_POST) across feedback conditions. Effect size is expressed as epsilon squared (ε^2), indicating negligible between-group differences.

Consistent with the analytical strategy outlined above, no post-hoc tests were conducted, given the non-significant omnibus results. Pairwise contrasts were instead summarised through Hodges–Lehmann estimates of the difference in gains, with 95% confidence intervals: expert human teacher vs GPT-o4-mini, HL = 0.00, 95% CI [−4.50, 5.00]; expert human teacher vs DeepSeek R1, HL = 1.00, 95% CI [−4.00, 6.00]; GPT-o4-mini vs DeepSeek R1, HL = 1.00, 95% CI [−2.00, 4.00]. All intervals included zero, indicating no evidence of differences between feedback sources; at the same time, their width shows that modest differences cannot be ruled out, and similar post-test scores should not, by themselves, be read as demonstrating equivalence (see RQ4 and Limitations)

RQ3. Role of attendance status and interaction with feedback source

To address RQ3, the role of students' attendance status (attending vs non-attending) and its potential interaction with feedback source were examined using baseline-adjusted change scores (DIFF_ADJ) to account for the dependency observed between raw gains and baseline performance.

Adjusted learning gains were analysed using a robust linear model with HC3 heteroskedasticity-consistent standard errors. The model included feedback source (expert human teacher, GPT-o4-mini, DeepSeek R1), attendance status, and their interaction as predictors.

Attendance showed no significant main effect on DIFF_ADJ, $F(1, 41) = 1.52$, $p = .225$. Likewise, neither the main effect of feedback source, $F(2, 41) = 1.09$, $p = 0.345$, nor the interaction between feedback source and attendance status, $F(2, 41) = 0.97$, $p = 0.389$, reached statistical significance (Table 3).

RQ4. Practical equivalence and non-inferiority of AI-generated feedback

To address RQ4, non-inferiority and equivalence analyses were conducted to evaluate whether AI-generated feedback could be considered practically comparable to feedback provided by an expert human teacher, given that non-significant differences do not imply practical equivalence.

Equivalence and non-inferiority were assessed using Welch-adjusted 90% confidence intervals, with a pre-defined equivalence margin of ±1 point on the 30-point grading scale, reflecting a substantively meaningful difference in grading practice.

Results based on post-feedback project scores (TOT_POST) are reported in Table 4.

Table 3. Robust linear model (HC3) testing the effects of feedback source, attendance status, and their interaction on baseline-adjusted improvement (DIFF_ADJ).

Predictor	df	F	p
Feedback source	2	1.09	.345
Attendance status	1	1.52	.225
Feedback source × Attendance	2	0.97	.389

Note. DIFF_ADJ = baseline-adjusted change score. Robust linear model estimated with HC3 heteroskedasticity-consistent standard errors. Feedback source was treated as a three-level categorical factor (expert human teacher, GPT-o4-mini, DeepSeek R1); attendance status was coded as attending vs non-attending. These findings indicate that, once baseline performance was accounted for, students' attendance status neither influenced learning gains, nor moderated the effect of feedback source.

Table 4. Non-inferiority and equivalence tests comparing AI-based feedback with expert human feedback on post-test project performance (90% CI, equivalence margin = ±1 point).

Comparison	nAI	nHuman	Mean difference	90% CI	Non-inferior	Equivalent
GPT-o4-mini vs Teacher	15	16	0.23	[−0.46, 0.91]	Yes	Yes
DeepSeek R1 vs Teacher	16	16	0.56	[−0.05, 1.18]	Yes	No

Note. Equivalence margin set at ±1 point on the 30-point grading scale. Non-inferiority is supported when the lower bound of the confidence interval is above −1. Equivalence requires the entire confidence interval to fall within the equivalence margin.

Table 5. Non-inferiority and equivalence tests comparing AI-based feedback with expert human feedback on baseline-adjusted improvement (DIFF_ADJ).

Comparison	nAI	nHuman	Mean difference	90% CI	Non-inferior	Equivalent
GPT-o4-mini vs Teacher	15	16	0.26	[−0.42, 0.94]	Yes	Yes
DeepSeek R1 vs Teacher	16	16	0.62	[0.02, 1.23]	Yes	No

The comparison between GPT-o4-mini and the expert human teacher yielded a small mean difference of 0.23 points (AI – Teacher), with the entire 90% confidence interval falling within the equivalence margin [−0.46, 0.91]. In contrast, the comparison between DeepSeek R1 and the expert human teacher yielded a small positive mean difference of 0.56 points (AI – Teacher). The lower bound remained above −1, supporting non-inferiority; however, the upper bound slightly exceeded the equivalence margin, preventing conclusive equivalence. This result suggests practical comparability but with greater uncertainty relative to the GPT-o4-mini condition.

To verify the robustness of these findings, the same comparisons were also examined using a one-way Welch ANOVA on post-test scores, followed by Games–Howell post-hoc tests. Results converged with the non-parametric analyses (RQ2), strengthening confidence in the stability of the findings.

The same pattern held when using DIFF_ADJ (Table 5): GPT-o4-mini met both non-inferiority and equivalence criteria, while DeepSeek R1 met non-inferiority only.

Overall, these findings indicate that AI-generated feedback – particularly when provided by GPT-o4-mini – can be considered practically equivalent to expert human feedback in supporting students' project performance, while feedback generated by DeepSeek R1 demonstrates robust non-inferiority relative to the human condition. These conclusions hold within the bounds of the rubric-based assessment used in this study: given the compression of post-test scores near the scale maximum, equivalence results based on post-test scores should be interpreted with caution, whereas the analyses based on baseline-adjusted improvement are less affected by this constraint.

Students' perceptions of feedback sources

Students were not informed about the source of the feedback they received. Throughout the intervention, participants remained blind to the assigned feedback

Table 6. Descriptive statistics for questionnaire dimensions by feedback condition.

Dimension	Feedback condition	N	Mean	Median	SD	Min	Max
Perceived mastery (P1–P8)	DeepSeek R1	81	4.22	4.25	0.44	2.88	5
	Teacher	63	4.14	4.13	0.48	3	5
	GPT-o4-mini	56	4.17	4.13	0.54	2.88	5
Positive emotions (EP1–EP6)	DeepSeek R1	81	4.21	4.33	0.59	2.5	5
	Teacher	63	3.99	4.17	0.65	2.33	5
	GPT-o4-mini	56	4.2	4.33	0.6	2.67	5
Negative emotions (EN1–EN5)	DeepSeek R1	81	1.22	1.2	0.35	1	3.6
	Teacher	63	1.31	1.2	0.3	1	2.2
	GPT-o4-mini	56	1.39	1.2	0.52	1	3

condition and therefore did not know whether the feedback had been generated by GPT-o4-mini, DeepSeek R1, or provided by the expert human teacher. Students' perceptions were analysed across the three feedback conditions (GPT-o4-mini, expert human teacher, and DeepSeek R1). The final sample consisted of 200 respondents, predominantly female (95.5%) with 3.5% male and 1% preferring not to disclose gender. Participants' age was mainly concentrated in the 20–23 age range (65.5%), followed by smaller proportions aged 24–26 (9.5%), 27–30 (11.5%), and over 31 (13.5%).

Prior to inferential analyses, the internal consistency of the three questionnaire scales was assessed. Cronbach's alpha coefficients indicated satisfactory reliability for perceived mastery ($\alpha = 0.81$), positive emotions ($\alpha = 0.85$) and negative emotions ($\alpha = 0.73$), consistent with the original validation. All three scales were therefore retained for subsequent analyses.

Descriptive statistics for the three composite indices, perceived mastery (P1–P8), positive emotions (EP1–EP6), and negative emotions (EN1–EN5), are reported in Table 6 by feedback condition.

Overall, mean scores reflect favourable perceptions across all groups. Perceived mastery was high and comparable ($M = 4.14$–4.22), indicating that feedback supported project design competence. Positive emotions were similarly elevated ($M = 3.99$–4.21), suggesting that feedback was experienced as motivating. Negative emotions were consistently low ($M = 1.22$–1.39), indicating minimal discomfort or disengagement regardless of feedback source.

To examine differences in perceptions by feedback source, Kruskal–Wallis tests were conducted on the three composite indices. No significant differences emerged for perceived mastery, $\chi^2(2, N = 200) = 0.89$, $p = 0.642$, $\varepsilon^2 = 0.004$, indicating comparable perceptions of competence development across conditions. Likewise, positive emotions showed no significant differences, $\chi^2(2, N = 200) = 4.76$, $p = 0.092$, $\varepsilon^2 = 0.024$, although descriptively higher means were observed in the AI conditions compared to the teacher condition.

For negative emotions, the Kruskal–Wallis test did not reach statistical significance, $\chi^2(2, N = 200) = 5.23$, $p = 0.073$, $\varepsilon^2 = .026$. In accordance with the analytical strategy described in the Methods, no post-hoc comparisons were conducted. Descriptively, negative emotions were slightly lower in the AI conditions than in the teacher condition.

Overall, students' perceptions were largely comparable across conditions, both cognitively (perceived mastery) and affectively (positive and negative emotions),

indicating that AI-generated feedback was experienced as an acceptable and supportive form of formative input comparable to expert teacher feedback.

Overall satisfaction was very high (98%), with no meaningful differences between sources (DeepSeek R1: 97.5%; teacher: 94%; GPT-o4-mini: 100%). Due to the near-ceiling distribution, satisfaction was analysed descriptively only, further confirming the high acceptability of AI-based feedback in this context.

Discussion

In recent years, assessment literature in higher education has emphasised that feedback is effective not for its mere presence, but for its capacity to function as usable information within processes of review, regulation, and student agency (Doria et al. 2025a; 2025b; Grion et al. 2021; Nicol 2021). From this perspective, feedback becomes transformative only when embedded in frameworks that make criteria and quality standards explicit, thereby supporting feedback literacy and self-regulation (Carless and Boud 2018; Grion and Serbati 2019; Hattie and Timperley 2007; Nicol 2021; Panadero, Jonsson, and Botella 2017; Yan and Carless 2022). The integration of generative AI in feedback processes should therefore be understood not as a replacement of the teacher, but as a reconfiguration of how feedback is produced, circulated, and interpreted, with implications for pedagogical quality and evaluative responsibility (Chan and Hu 2023; Lipnevich and Smith 2018).

The findings show a marked improvement in project performance from pre- to post-test across all conditions, irrespective of feedback source. Although the design does not allow the isolation of specific instructional components (e.g. rubrics or exemplars), the results suggest that feedback effectiveness depends more on the pedagogical configuration in which it is embedded than on the source itself.

Moreover, the absence of significant differences between conditions, together with evidence of non-inferiority and practical equivalence relative to teacher feedback, shifts the focus from a reductive 'AI vs. teacher' debate to a more relevant question: under what pedagogical conditions can AI-generated feedback function as an effective component of assessment practice. This interpretation aligns with research highlighting the transformative potential of generative technologies (Agostini 2024; Agostini and Picasso 2024; Dai et al. 2024; Doria et al. 2025a), provided they are integrated within theoretically grounded and pedagogically mediated assessment architectures (Dai et al. 2024; Yan et al. 2024). In this sense, AI should be understood not as an autonomous solution to feedback challenges, but as a support for teachers with developed assessment literacy and as an amplifier of well-designed pedagogical structures.

Feedback effectiveness as a systemic outcome

The findings provide clear evidence that feedback plays a significant role in supporting project development. Across the full sample, performance improved from pre- to post-feedback (Wilcoxon: $W = 1081$, $p < 0.001$, $rrb = 0.77$), with an average increase of 3.9 points on a 30-point scale. This improvement occurred across all

conditions, irrespective of feedback source, and was accompanied by reduced score variability (PRE SD = 4.63; POST SD = 0.95), with scores converging towards the upper limit (POST median = 28). This convergence indicates substantial overall improvement, with most groups meeting the well-defined standards operationalised in the rubric. However, because reduced variability is an expected consequence of measurement near the scale maximum, the low post-test dispersion should not be read as independent evidence of homogeneous final quality (see Limitations).

Baseline variability and deviations from normality at PRE should be interpreted in light of the decision to preserve naturally formed working groups to maintain ecological validity. Initial heterogeneity likely reflects genuine differences in prior competence, while small subgroup sizes and discrete rubric scores make departures from normality unsurprising in classroom-based research. Baseline non-normality should therefore be considered an expected feature rather than a methodological flaw.

Crucially, feedback source did not produce significant differences in outcomes. Post-feedback scores did not differ across conditions (Kruskal–Wallis $H(2) = 1.91$, $p = 0.384$, $\varepsilon^2 \approx 0$), and equivalence analyses confirmed the practical comparability of AI-generated and teacher feedback. GPT-o4-mini met both non-inferiority and equivalence criteria relative to the teacher, while DeepSeek R1 met non-inferiority. These patterns were consistent for both post-test scores and baseline-adjusted gains (DIFF_ADJ), indicating that improvements cannot be attributed to feedback source (= 0.384, $\varepsilon^2 = 0.042$), and the same held for gain scores, $H(2) = 0.74$, $p = .690$. However, the interpretive value of post-test comparisons is limited by the ceiling effect: similar post-test scores cannot, by themselves, demonstrate that the feedback sources were equally effective. The convergence of gain-score comparisons, baseline-adjusted analyses, and non-inferiority/equivalence tests on adjusted improvement supports the comparability of feedback sources, while modest differences cannot be conclusively ruled out given the sample size.

Results were robust across contextual variables: attendance did not affect learning gains nor interact with feedback source, and students' perceptions and satisfaction were comparable across conditions. Overall, performance improvements appear stable across sources, contexts, and subjective experiences, supporting a view of feedback as a systemic and relational process rather than a function of isolated variables (Panadero 2023).

AI as an effective support for teachers with assessment literacy

In this paper, assessment literacy refers to the knowledge, skills, and dispositions required to design, implement, and interpret assessment processes. Specifically, teacher assessment literacy denotes educators' capacity to construct sound assessment tasks and criteria, make dependable judgements about student work, and use assessment to support learning (Pastore and Andrade 2019; Xu and Brown 2016). Student assessment literacy refers to students' understanding of assessment purposes, criteria, and standards, and to their capacity to use this understanding – including the feedback they receive – to improve their own work (Carless and Boud 2018; Smith et al. 2013).

These findings should not be read as implying that AI-generated and teacher feedback are the same kind of feedback, or that they are interchangeable in general. Rather, they indicate that, under the conditions of this study—in which both were

anchored to the same explicit assessment criteria within a structured instructional design—the two sources produced comparable effects on project performance. Consistent with systemic perspectives on feedback, effectiveness appears less dependent on the source itself than on the instructional conditions shaping how feedback is produced, interpreted, and used (Carless and Boud 2018; Nicol 2021; Panadero 2023).

In this study, the architecture relied on the instructor's assessment literacy. AI-generated feedback was anchored to a shared analytic rubric, co-constructed with students through exemplar analysis, which made the assessment criteria explicit and shared (the complete rubric is provided in Appendix A). Research on assessment for learning shows that transparency, comparison, and engagement with exemplars enhance students' capacity to interpret and use feedback effectively (Grion & Serbati, 2019; Hattie and Timperley 2007; Nicol 2021).

Crucially, the exemplar embodied the teacher's evaluative expectations regarding high-quality performance in that specific context, providing the AI with an explicit interpretative anchor often tacit in human assessment. In this sense, the exemplar functioned as a calibration device through which teacher assessment literacy shaped the production of AI-generated feedback.

In short, AI-generated and teacher feedback produced comparable results when grounded in the same assessment criteria. This suggests that generative AI is best understood neither as an autonomous solution nor as a threat to teachers, but as a support for educators with strong assessment literacy: deliberately aligning the AI with context-specific quality standards—here, through the rubric and the exemplar—is what made its feedback educationally meaningful. Under such conditions, AI can enhance the scalability and timeliness of feedback while its interpretative and ethical dimensions remain pedagogically grounded.

At the same time, educators should remain aware of potential risks associated with AI-supported feedback, including students' over-reliance on automated suggestions and unequal access to AI tools across educational contexts. These concerns reinforce the importance of maintaining teacher oversight and fostering students' critical engagement with feedback rather than encouraging passive acceptance of AI-generated responses (Miao et al., 2021; Zengin and Korkut 2026).

Student assessment literacy as a possible complementary condition: a direction for future research

Alongside teacher assessment literacy, students' assessment literacy may represent a complementary condition for the effective use of AI-generated feedback. Feedback influences learning not merely through its provision, but through students' capacity to interpret and act upon it within iterative processes (Carless and Boud 2018; Nicol 2021; Panadero 2023). In this study, students engaged actively with the feedback received, revising their projects accordingly. It should be noted, however, that student assessment literacy was not directly measured: the considerations developed in this section are therefore conceptual, and should be read as hypotheses rather than as empirically supported explanations of the results. The present results support this view. Perceived mastery was consistently high across conditions ($M \approx 4.14$–4.22), indicating that feedback – regardless of source – was experienced as useful for improving project design competence. Positive emotions were similarly elevated ($M \approx 3.99$–4.21), while negative emotions remained low ($M \approx 1.22$–1.39), and overall

satisfaction was extremely high (≈98%). These findings suggest that students were able to engage productively with feedback information irrespective of whether it originated from AI systems or an expert teacher.

Such engagement likely reflects the course's pedagogical design, including rubric co-construction, exemplar analysis, and iterative revision, practices known to strengthen evaluative judgement and feedback literacy (Carless and Boud 2018; Grion & Serbati, 2019; Hattie and Timperley 2007; Nicol 2021). The comparable effectiveness of AI and teacher feedback may thus depend not only on properties of the feedback source, but also on students' preparedness to use feedback meaningfully—a hypothesis that the present design cannot test directly. Examining whether, and how, student assessment literacy moderates the productive use of AI-generated feedback is an important direction for future research.

Limitations

Several limitations should be acknowledged. First, post-test scores clustered near the maximum of the 30-point rubric (91% of groups scored 27 or above; SD = 0.95), indicating a ceiling effect. This compression limits the sensitivity of post-test comparisons between feedback sources: similar post-test scores are compatible both with genuinely comparable effectiveness and with differences that the measure could not capture. Although gain-score and baseline-adjusted analyses converge with the post-test results, and equivalence analyses on adjusted improvement support comparability, modest differences between feedback sources cannot be conclusively ruled out. Future studies should employ assessment scales offering more headroom, or more demanding performance criteria. Second, the group-level sample ($N = 47$) limits statistical power, as reflected in the width of the confidence intervals for pairwise contrasts. Third, information on students' prior experience with generative AI tools was not collected, and differences in prior familiarity may have influenced engagement with AI-mediated feedback. Fourth, students' assessment literacy was not directly measured, so its contribution to the effective use of feedback remains hypothetical. Finally, the study was conducted within a single course and disciplinary context, and the transferability of the findings to other settings remains to be established.

Conclusion

Placed within the broader trajectory of technology-supported feedback in higher education, these findings suggest that generative AI should not be understood merely as a more advanced technical solution to persistent feedback challenges. Earlier Learning Analytics approaches have shown how technological systems can enhance scalability, timeliness, and personalisation, while also revealing limitations related to pedagogical quality, mediation, and alignment with formative assessment principles. In this respect, generative AI represents less a rupture than a continuation of these developments, reaffirming a central insight from assessment research: technological innovation alone does not guarantee pedagogical effectiveness.

What appears decisive is the pedagogical architecture within which feedback is embedded. When supported by strong assessment literacy, transparent criteria,

exemplar-based calibration, and iterative engagement, AI-generated feedback can function as a credible formative resource within the conditions examined in this study, without displacing the relational and interpretative dimensions central to assessment practice. Under such circumstances, generative AI may help address tensions related to scalability, timeliness, and equitable access while preserving the core principles of formative assessment.

For assessment scholarship, these findings invite a shift away from reductive 'AI versus teacher' framings towards a more productive inquiry into how feedback ecosystems can be designed so that technological tools – whether analytics-based or generative – strengthen rather than redefine the educational purposes of assessment. Future research should further examine the transferability of these findings across disciplinary contexts, longitudinal timeframes, and alternative prompt and feedback architectures.

Note

1. Online supplementary materials (Online Appendices A–C) are available at https://doi.org/10.5281/zenodo.20814177.

Authors' contributions

Grion, V.: Conceptualisation, methodology, project administration, resources & supervision, writing-review and editing. Doria, B.: Conceptualisation, data curation, methodology, project administration, resources, supervision, visualisation, investigation, writing–original draft preparation & writing–review and editing. Agostini, D.: Formal analysis, data curation, methodology, project administration, visualisation, resources and investigation, writing–original draft preparation & writing–review and editing. Slaviero, G.: Conceptualisation, methodology, investigation, resources, supervision, validation, visualisation, writing–review and editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Supplementary materials

Online Appendices A–C and the anonymised datasets supporting the findings of this study are available on Zenodo at https://doi.org/10.5281/zenodo.20814177. The supplementary materials include: (A) the assessment rubric used in the study; (B) an example of the project assignment provided to students; (C) the AI prompting protocol and custom instructions used to generate formative feedback; and (D) the anonymised datasets and analysis files underlying the reported results.

ORCID

Valentina Grion http://orcid.org/0000-0002-2051-1313
Beatrice Doria http://orcid.org/0000-0002-3894-9460
Daniele Agostini http://orcid.org/0000-0002-9919-5391
Giorgia Slaviero http://orcid.org/0009-0000-9312-4683

References

Agostini, D. 2024. "Large Language Model sono capaci di valutare i compiti scritti degli studenti? uno studio pilota in Università." Research Trends in Humanities Education & Philosophy 11: 38–60. https://doi.org/10.6093/2284-0184/10671.

Agostini, D., and F. Picasso. 2024. "Large Language Models for Sustainable Assessment and Feedback in Higher Education: Towards a Pedagogical and Technological Framework." *Intelligenza Artificiale* 18 (1): 121–138. https://doi.org/10.3233/IA-240033.

Arthars, N., M. Dollinger, L. Vigentini, D. Y.-T. Liu, E. Kondo, and D. M. King. 2019. "Empowering Teachers to Personalize Learning Support." In *Utilizing Learning Analytics to Support Study Success*, edited by D. Ifenthaler, D.-K. Mah, and J. Y.-K. Yau, 223–248. Charm, CH: Springer. https://doi.org/10.1007/978-3-319-64792-0_13.

Azcona, D., I.-H. Hsiao, and A. F. Smeaton. 2019. "Detecting Students-at-Risk in Computer Programming Classes with Learning Analytics from Students' Digital Footprints." *User Modeling and User-Adapted Interaction* 29 (4): 759–788. https://doi.org/10.1007/s11257-019-09234-7.

Ba, S., L. Yang, Z. Yan, C. K. Looi, and D. Gašević. 2025. "Unraveling the Mechanisms and Effectiveness of AI-Assisted Feedback in Education: A Systematic Literature Review." *Computers and Education Open* 9: 100284. https://doi.org/10.1016/j.caeo.2025.100284.

Baydas Onlu, O., M. S. Abdusselam, and R. M. Yilmaz. 2022. "Students' Perception of Instructional Feedback Scale: Validity and Reliability Study." *Contemporary Educational Technology* 14 (3): Ep368. https://doi.org/10.30935/cedtech/11811.

Beckman, K., T. Apps, S. Bennett, B. Dalgarno, G. Kennedy, and L. Lockyer. 2021. "Self-Regulation in Open-Ended Online Assignment Tasks: The Importance of Initial Task Interpretation and Goal Setting." *Studies in Higher Education* 46 (4): 821–835. https://doi.org/10.1080/03075079.2019.1654450.

Carless, D. 2015. "Exploring Learning-Oriented Assessment Processes." *Higher Education* 69 (6): 963–976. https://doi.org/10.1007/s10734-014-9816-z.

Carless, D., and D. Boud. 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.

Chan, C. K. Y, and W. Hu. 2023. "Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education." *International Journal of Educational Technology in Higher Education* 20(43):1–18. https://doi.org/10.1186/s41239-023-00411-8.

Chatti, M. A., A. L. Dyckhoff, U. Schroeder, and H. Thüs. 2012. "A Reference Model for Learning Analytics." *International Journal of Technology Enhanced Learning* 4 (5/6): 318–331. https://doi.org/10.1504/IJTEL.2012.0518.

Dai, J., X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang. 2023. Safe RLHF: Safe Reinforcement Learning from Human Feedback (arXiv:2310.12773). arXiv. https://doi.org/10.48550/arXiv.2310.12773.

Dai, W., Y.-S. Tsai, J. Lin, A. Aldino, H. Jin, T. Li, D. Gašević, and G. Chen. 2024. "Assessing the Proficiency of Large Language Models in Automatic Feedback Generation: An Evaluation Study." *Computers and Education: Artificial Intelligence* 7: 100299. https://doi.org/10.1016/j.caeai.2024.100299.

Doria, B., L. C. Foschi, G. Slaviero, C. Zaggia, and V. Grion. 2025a. "Feedback e Feedback Automatizzato: Pratiche e Percezioni Dei Docenti Universitari." *Research Trends in Humanities Education & Philosophy* 12 (1): 25–42. https://doi.org/10.6093/2284-0184/11562.

Doria, B., V. Grion, F. Picasso, and L. Li. 2025b. "Faculty Assessment Development in Higher Education: The CRAFT Framework Emerging from a Systematic Literature Review." *Assessment & Evaluation in Higher Education* 51 (1): 16–40. https://doi.org/10.1080/02602938.2025.2584121.

Grion, V., A. Serbati, V. Doria, and D. Nicol. 2021. "Rethinking Assessment and Feedback Practices in Higher Education: A Review of Recent Literature." *Innovations in Education and Teaching International* 58 (4): 405–416.

Grion, V., and A. Serbati. 2019. *Valutazione Sostenibile e Feedback Nei Contesti Universitari*. Lecce, IT: PensaMultimedia.

Hattie, J., and H. Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77 (1): 81–112. https://doi.org/10.3102/003465430298487.

Henderson, M., T. Ryan, and M. Phillips. 2019. "The Challenges of Feedback in Higher Education." *Assessment & Evaluation in Higher Education* 44 (8): 1237–1252. https://doi.org/10.1080/02602938.2019.1599815.

Hirunyasiri, D., D. R. Thomas, J. Lin, K. R. Koedinger, and V. Aleven. 2023. Comparative Analysis of GPT-4 and Human Graders in Evaluating Praise Given to Students in Synthetic Dialogues (arXiv:2307.02018). arXiv. https://doi.org/10.48550/arXiv.2307.02018.

Leighton, J. P., and M. C. Bustos Gómez. 2018. "A Pedagogical Alliance for Trust, Wellbeing and the Identification of Errors for Learning and Formative Assessment." *Educational Psychology* 38 (3): 381–406. https://doi.org/10.1080/01443410.2017.1390073.

Li, C., and W. Xing. 2021. "Natural Language Generation Using Deep Learning to Support MOOC Learners." *International Journal of Artificial Intelligence in Education* 31 (2): 186–214. https://doi.org/10.1007/s40593-020-00235-x.

Lim, L.-A., S. Gentili, A. Pardo, V. Kovanović, A. Whitelock-Wainwright, D. Gašević, and S. Dawson. 2021. "What Changes, and for Whom? A Study of the Impact of Learning Analytics-Based Process Feedback in a Large Course." *Learning and Instruction ?*72: 101202. https://doi.org/10.1016/j.learninstruc.2019.04.003.

Lipnevich, A. A., and Smith, J. K., eds. 2018. *The Cambridge Handbook of Instructional Feedback*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781316832134.

Lipnevich, A. A., and J. K. Smith. 2022. "Student – Feedback Interaction Model: Revised." *Studies in Educational Evaluation* 75: 101208. https://doi.org/10.1016/j.stueduc.2022.101208.

MacNeil, S., A. Tran, D. Mogil, S. Bernstein, E. Ross, and Z. Huang. 2022. "Generating Diverse Code Explanations Using the GPT-3 Large Language Model." *Proceedings of the 2022 ACM Conference on International Computing Education Research – Volume 2, ICER '22* 2:37–39. https://doi.org/10.1145/3501709.3544280.

Miao, F., W. Holmes, R. Huang, and H. Zhang. 2021. *AI and Education: Guidance for Policy-Makers*. UNESCO. https://doi.org/10.54675/PCSP7350

Middleton, T., A. Ahmed Shafi, R. Millican, and S. Templeton. 2023. "Developing Effective Assessment Feedback: Academic Buoyancy and the Relational Dimensions of Feedback." *Teaching in Higher Education* 28 (1): 118–135. https://doi.org/10.1080/13562517.2020.1777397.

Mulliner, E., and M. Tucker. 2017. "Feedback on Feedback Practice: Perceptions of Students and Academics." *Assessment & Evaluation in Higher Education* 42 (2): 266–288. https://doi.org/10.1080/02602938.2015.1103365.

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.

Panadero, E. 2023. "Toward a Paradigm Shift in Feedback Research: Five Further Steps Influenced by Self-Regulated Learning Theory." *Educational Psychologist* 58 (3): 193–204. https://doi.org/10.1080/00461520.2023.2223642.

Panadero, E., A. Jonsson, and J. Botella. 2017. "Effects of Self-Assessment on Self-Regulated Learning and Self-Efficacy: Four Meta-Analyses." *Educational Research Review* 22: 74–98. https://doi.org/10.1016/j.edurev.2017.08.004.

Pankiewicz, M., and R. S. Baker. 2023. "Large Language Models (GPT) for Automating Feedback on Programming Assignments." arXiv arXiv:2307.00150. https://arxiv.org/abs/2307.00150

Pardo, A., K. Bartimote, S. Buckingham Shum, S. Dawson, J. Gao, D. Gašević, S. Leichtweis, et al. 2018. "OnTask: Delivering Data-Informed, Personalized Learning Support Actions." *Journal of Learning Analytics* 5 (3): 235–249. https://doi.org/10.18608/jla.2018.53.15.

Pastore, S., and H. L. Andrade. 2019. "Teacher Assessment Literacy: A Three-Dimensional Model." *Teaching and Teacher Education* 84: 128–138. https://doi.org/10.1016/j.tate.2019.05.003.

Shi, H., and V. Aryadoust. 2024. "A Systematic Review of AI-Based Automated Written Feedback Research." *ReCALL* 36 (2): 187–209. https://doi.org/10.1017/S0958344023000265.

Siddiq, F., and D. Murchan. 2024. "Towards a Code of Ethics for Using Technology-Enabled Data and Related Analytic Approaches in Educational Assessment." *Assessment in Education: Principles, Policy & Practice* 31 (5–6): 376–400. https://doi.org/10.1080/0969594X.2025.2453138.

Smith, C. D., K. Worsfold, L. Davies, R. Fisher, and R. McPhail. 2013. "Assessment Literacy and Student Learning: The Case for Explicitly Developing Students' Assessment Literacy." *Assessment & Evaluation in Higher Education* 38 (1): 44–60. https://doi.org/10.1080/02602938.2011.598636.

Wible, D., C.-H. Kuo, F. Chien, A. Liu, and N.-L. Tsao. 2001. "A Web-Based EFL Writing Environment: Integrating Information for Learners, Teachers, and Researchers." *Computers & Education* 37 (3–4): 297–315. https://doi.org/10.1016/S0360-1315(01)00056-2.

Xu, Y., and G. T. L. Brown. 2016. "Teacher Assessment Literacy in Practice: A Reconceptualization." *Teaching and Teacher Education* 58: 149–162. https://doi.org/10.1016/j.tate.2016.05.010.

Yan, L., V. Echeverria, Y. Jin, G. Fernandez-Nieto, L. Zhao, X. Li, R. Alfredo, Z. Swiecki, D. Gašević, and R. Martinez-Maldonado. 2024. "Evidence-Based Multimodal Learning Analytics for Feedback and Reflection in Collaborative Learning." *British Journal of Educational Technology* 55 (5): 1900–1925. https://doi.org/10.1111/bjet.13498.

Yan, Z., and D. Carless. 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.

Zawacki-Richter, O., V. I. Marín, M. Bond, and F. Gouverneur. 2019. "Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators?" *International Journal of Educational Technology in Higher Education* 16 (1): 39. https://doi.org/10.1186/s41239-019-0171-0.

Zengin, R, and N. Korkut. 2026. "From Applications to Implications: A Systematic Review of AI in K-12 English Language Teaching." *İstanbul 29 Mayıs Üniversitesi Eğitim Fakültesi Dergisi* 1 (1): 59–75. https://izlik.org/JA34GS37ZG.