

The Eurasia Proceedings of Educational and Social Sciences (EPESS), 2026

Volume 53, Pages 12-17

ICETI 2026: International Conference on Education in Technology and Innovation

Modes of Engagement with AI-Generated Feedback: A Case Study of Three Pre-Service EFL Teachers

Raigul Zheldibayeva

Zhetysu University named after Ilyas Zhansugurov

Abstract: Generative AI (GenAI) feedback tools are increasingly used to support formative writing instruction, yet most research reports feedback effects in aggregate, offering little insight into how individual learners actually engage with AI-generated feedback. This exploratory case study examines three pre-service EFL teachers' engagement with feedback from CyberScholar, a rubric-aligned AI feedback platform, during revision of a short reflective essay on AI in English language education. Drawing on a larger pilot dataset, three contrasting cases were purposively selected to illustrate distinct modes of engagement: regenerative rewriting, in which the participant produced an almost entirely new text in a markedly more formal register; minimal, selective engagement, in which only a narrowly targeted mechanical issue was addressed while conceptual feedback was left unaddressed; and partial, uneven engagement, in which some rubric criteria were revised while others remained unchanged. Comparing participants' draft-to-draft changes, the AI-generated feedback they received, and the resulting rating changes, the study develops a preliminary typology of engagement modes rather than a claim about writing-quality improvement. Findings suggest that the same feedback instrument can prompt qualitatively different revision behaviors depending on how learners interpret and act on it, with implications for how AI feedback tools are designed and how teacher educators might scaffold students' critical use of such tools. The study contributes a fine-grained, learner-centered perspective that complements aggregate analyses of AI feedback effectiveness.

Keywords: AI-generated feedback, Case study, EFL teacher education, Feedback engagement, Reflective writing

Introduction

Generative AI (GenAI) feedback tools are increasingly embedded in writing instruction, promising immediate, criterion-referenced responses that can supplement or extend teacher feedback (Escalante et al., 2023; Farrokhnia et al., 2026). Much of the empirical literature on these tools evaluates their impact in aggregate: across a sample of learners, do AI-generated ratings increase, and do learners report the feedback as useful (Banihashem et al., 2024; Lo et al., 2026)? This aggregate approach is valuable for establishing broad patterns, but it can obscure a more basic question: what do individual learners actually do when they receive AI-generated feedback? Two learners who receive similar feedback and produce similarly improved ratings may have arrived there through very different processes one through careful, targeted editing, another through wholesale regeneration of the text, and a third through selective compliance with only the most explicit directives.

Research on AI-generated feedback in language education has grown rapidly, generally comparing AI and human feedback on dimensions such as accuracy, usefulness, and impact on subsequent drafts (Banihashem et al., 2024; Escalante et al., 2023). Several studies suggest that AI feedback is well suited to visible, mechanical features of text—grammar, vocabulary, organization—while human feedback remains stronger for higher-order, content-focused concerns (Lo et al., 2026; Wahyuni et al., 2026). Farrokhnia et al. (2026) found that even carefully engineered AI feedback did not automatically translate into stronger revision outcomes, concluding

- This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 Unported License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

© 2026 Published by ISRES Publishing: www.isres.org

that feedback quality alone is insufficient to predict revision behavior; how learners interpret and apply feedback matters as much as the feedback itself. This point—that the same feedback can be taken up differently by different learners—has received comparatively little direct attention in the AI-feedback literature, which tends to report outcomes in aggregate rather than examining individual trajectories.

Case study approaches are well suited to this kind of question. Yin (2018) argues that case studies are appropriate when a “how” or “why” question is being asked about a contemporary phenomenon within a real-world context, particularly when the boundaries between phenomenon and context are not clear-cut—as is the case with a single revision episode embedded in an authentic classroom activity. This exploratory case study addresses this gap by examining, in depth, how three pre-service EFL teachers engaged with feedback from CyberScholar during revision of a short reflective essay on the topic “Using AI in English Language Education: My Experience and Expectations.” The three cases were purposively selected from a larger pilot dataset (Zheldibayeva et al., 2025) to represent contrasting patterns of engagement rather than a representative sample. Rather than asking whether AI feedback improved writing quality—a question that would require independent, validated assessment beyond the scope of this exploratory study—this paper asks: how did each participant interpret and act on AI-generated feedback, and what does the resulting pattern of textual and rating change reveal about the range of ways learners engage with such feedback?

Method

Context and Data Source

The three cases examined in this study are drawn from a larger pilot activity in which pre-service EFL teachers at a university in Kazakhstan used CyberScholar, a GenAI-based feedback platform, to draft and revise a short reflective essay (approximately 250 words) on the topic “Using AI in English Language Education: My Experience and Expectations.” CyberScholar generated criterion-based feedback and a star rating (1–4) for each of five rubric criteria: Academic Writing Style and Format, Reflection and Critical Thinking, Language Accuracy and Vocabulary, Structure and Coherence, and Content Relevance and Depth. Participants received this feedback after their initial submission, were invited to revise, and received a second round of feedback and ratings on their revised draft. The larger pilot and its aggregate findings are reported elsewhere (Zheldibayeva et al., 2025); the present study draws on the same underlying dataset but asks a different question, using a subset of cases for close qualitative analysis rather than cross-case aggregation.

Case Selection

Three cases were purposively selected from the pilot dataset to represent maximally contrasting patterns of engagement with AI-generated feedback, following a theoretical (rather than representative) sampling logic (Yin, 2018). Selection was based on an initial review of all available complete cases (first draft, first-round feedback and ratings, revised draft, second-round feedback and ratings), comparing the direction and consistency of rating change across the five criteria and the nature of the textual change between drafts. Participant A showed a uniform, substantial rating increase across all five criteria accompanied by an almost complete rewriting of the essay. Participant B showed a single-criterion rating increase with the other four criteria unchanged, accompanied by minimal, targeted textual edits. Participant C showed a mixed pattern—rating increases on some criteria, no change on others—accompanied by moderate, uneven revision. To protect participant anonymity, pseudonymous labels (Participant A, B, C) are used throughout; no names, institutional affiliations, or other potentially identifying information are disclosed.

Analytic Approach

For each case, the initial draft, the first-round AI feedback, the revised draft, and the second-round AI feedback and ratings were compared. Analysis proceeded in three steps: (1) identifying what feedback was given for each criterion in round one; (2) identifying what changed—and did not change—in the revised draft relative to that feedback; and (3) examining whether and how the round-two rating and commentary reflected those changes. This within-case analysis was followed by a cross-case comparison to identify a preliminary typology of engagement modes. As in the broader pilot, ratings were generated by CyberScholar itself; they are treated here as part of the phenomenon under study rather than as an independent or validated measure of writing quality. Table 1 presents the rating changes and predominant engagement patterns across the three cases.

Table 1. Rating change and engagement pattern by case

Participant	Round 1 ratings*	Round 2 ratings*	Predominant engagement pattern
A	3, 2, 2, 3, 3	4, 4, 4, 4, 4	Regenerative rewriting
B	3, 3, 3, 3, 3	4, 3, 3, 3, 3	Minimal, selective engagement
C	2, 2, 2, 2, 2	3, 2, 3, 3, 2	Partial, uneven engagement

Note. Ratings are listed in the order: Academic Writing Style and Format, Reflection and Critical Thinking, Language Accuracy and Vocabulary, Structure and Coherence, Content Relevance and Depth.

Results and Discussion

Participant A: Regenerative Rewriting

Participant A's initial draft was short, colloquial in register, and organized as three brief paragraphs describing a period of AI use, its help with speaking practice and feedback, and a hope that AI would eventually assist with pronunciation and fluency. CyberScholar's first-round feedback rated the draft 3 (Academic Writing Style and Format), 2 (Reflection and Critical Thinking), 2 (Language Accuracy and Vocabulary), 3 (Structure and Coherence), and 3 (Content Relevance and Depth). Feedback across criteria converged on similar points: the tone was inconsistent with academic register, at least one sentence was grammatically unclear in describing the participant's stage of study, and—most consistently—the reflection remained at the level of personal benefit without connecting to broader pedagogical implications or theoretical perspectives, and without addressing the participant's own future teaching practice. The revised draft, submitted under a new heading indicating a reflective analysis, bore little structural resemblance to the original. Where the initial draft had been organized around chronological personal experience, the revision opened with a more formal framing and introduced content not present in the original at all: an explicit discussion of pedagogical potential, a named risk (“an over-dependence on AI for immediate answers could hinder the development of independent problem-solving and critical thinking skills”), and a two-part conclusion addressing both the participant's continued learning and prospective teaching practice—directly responding to the round-one critique that the essay lacked a connection to teaching. Vocabulary shifted markedly, introducing terms such as “scaffolding,” “differentiated learning materials,” and “formative assessment” that did not appear in the original draft or, notably, in CyberScholar's feedback itself.

Round-two ratings rose to 4 across all five criteria. The accompanying commentary specifically praised the essay's “nuanced discussion,” its “critical reflection on potential benefits and limitations,” and its integration of “theoretical perspectives” — the precise gaps identified in round one. The scale and uniformity of this change, however, is at least as notable as its direction: rather than incrementally addressing each piece of feedback in turn, Participant A appears to have substantially regenerated the entire text in a new register, addressing multiple pieces of feedback simultaneously through what reads as a single global revision rather than a series of local edits.

Participant B: Minimal, Selective Engagement

Participant B's initial draft discussed AI use for language learning generally and for preparation toward an important English proficiency examination specifically, expressed a nuanced view acknowledging both AI's convenience and its limitations in providing genuine emotional connection, and closed with a caution against overreliance. First-round ratings were uniform: 3 across all five criteria. Feedback was correspondingly consistent in tone—each criterion's commentary described the essay as competent but incomplete, most often citing two specific, recurring issues: a tendency to refer to the AI system using a personal pronoun rather than the grammatically correct impersonal form, and length slightly exceeding the 250-word limit. Several criteria also noted, more substantively, that the essay's reflection remained under-theorized and insufficiently connected to teaching practice. The revised submission addressed the two most explicit, mechanical issues directly: the pronoun was corrected to the grammatically appropriate impersonal form throughout, and the essay was trimmed toward the word limit. Beyond these two changes, the revised text was substantively similar to the original—the same examples, the same paragraph structure, and the same level of theoretical engagement. The more substantive feedback repeated across all five criteria in round one, to connect personal experience with pedagogical implications and to engage theoretical perspectives, went largely unaddressed in the revision.

Round-two ratings reflected this pattern closely: Academic Writing Style and Format rose from 3 to 4, the only criterion whose specific, named issues (pronoun use, word count) had been directly and mechanically fixable. The remaining four criteria—Reflection and Critical Thinking, Language Accuracy and Vocabulary, Structure and Coherence, and Content Relevance and Depth—remained at 3, with round-two commentary repeating, in some cases almost verbatim, the same substantive critiques offered in round one. Participant B's engagement with feedback in this case appears highly selective: responsive to feedback that named a specific, low-effort correction, and largely unresponsive to feedback requesting deeper conceptual revision.

Participant C: Partial, Uneven Engagement

Participant C's initial draft focused on AI use for learning a different target language than the one specified by the assignment, discussed AI as a time-saving and accessible alternative to formal courses, and closed with a call for AI literacy education in schools and workplaces. First-round ratings were low and uniform: 2 across all five criteria. Feedback identified overlapping problems across criteria: an informal, conversational tone throughout, a text formatted as a single block without paragraph breaks, length exceeding the word limit, and—as with the other two cases—limited connection between the participant's personal experience and the essay's required topic (AI in English language education specifically). The revised draft showed clear engagement with some, but not all, of this feedback. Paragraph breaks were introduced, bringing structure to what had been a single block of text, and several informal phrases were reworded in a more formal register. However, the essay's substantive focus remained unchanged: it continued to center on the other target language rather than the assigned topic of English language education, and the participant's own future teaching practice was still not addressed.

Round-two ratings reflected this partial engagement: Academic Writing Style and Format and Structure and Coherence both rose from 2 to 3, criteria whose feedback had focused on formatting, tone, and paragraphing—exactly the changes the participant made. Reflection and Critical Thinking and Content Relevance and Depth remained at 2; both had been criticized in round one specifically for the essay's weak connection to English language education, which the revision did not address. Language Accuracy and Vocabulary rose from 2 to 3, corresponding to the surface-level wording changes described above. Participant C's pattern thus sits between the other two cases: more responsive than Participant B, engaging with several distinct pieces of feedback rather than just one, but—like Participant B—leaving the most conceptually demanding feedback unaddressed.

Cross-Case Discussion

Read together, these three cases suggest that revision in response to AI feedback is not a single behavior but a family of qualitatively different behaviors, even when the feedback instrument, task, and rubric are held constant. Participant A's regenerative rewriting addressed nearly every point of round-one feedback simultaneously, producing a text that reads less like an edited version of the original than a new composition responding to the feedback's overall thrust. Participant B engaged with feedback selectively and literally, correcting the two most explicitly named, mechanically simple issues while leaving substantive critique unaddressed across all five criteria. Participant C's engagement was broader than Participant B's but still partial, responding to feedback about form and tone while leaving feedback about content and topic relevance unaddressed across two revision rounds.

A pattern worth highlighting is that in all three cases, the feedback that went unaddressed was consistently the same type: feedback asking participants to connect personal experience with pedagogical theory, or to more explicitly address their own future teaching practice. This was among the most frequently repeated critiques across all three participants' round-one feedback, and it was the critique most consistently absent from participants' revisions except in Participant A's case. This raises a question about the interaction between feedback specificity and uptake: feedback naming a concrete, low-effort correction (a pronoun, a word count) was reliably acted upon; feedback requesting a more abstract cognitive shift (deeper theoretical engagement) was inconsistently acted upon, and where it was addressed—as in Participant A's case—the response involved regenerating substantial portions of the text rather than incrementally developing the existing argument. This last observation has a specific implication for how AI-assisted revision is understood and taught. If the most demanding feedback is addressed primarily through wholesale regeneration rather than incremental development, then a straightforward “revision occurred, and ratings improved” narrative may obscure an important distinction: between revision as a process of developing and refining one's own thinking, and revision

as a process of producing a new text that satisfies stated criteria. Both are legitimate responses to feedback, but they are pedagogically different, and pre-service teachers—who will themselves need to model and evaluate revision practices for their future students—may benefit from explicit discussion of this distinction rather than treating rating increases alone as a sufficient indicator of engagement.

Conclusion

This case study examined how three pre-service EFL teachers engaged with AI-generated feedback during revision of a reflective essay, identifying three distinct patterns—regenerative rewriting, minimal selective engagement, and partial uneven engagement—that were not distinguishable from aggregate rating data alone. Across all three cases, feedback requesting a concrete, mechanically simple correction was consistently acted upon, while feedback requesting deeper theoretical or pedagogical engagement was inconsistently addressed, and was addressed most fully only through substantial text regeneration rather than incremental development. These findings suggest that examining individual revision trajectories alongside aggregate outcome measures can reveal meaningful variation that a purely aggregate analysis would miss, with implications for how AI feedback tools are designed and for how teacher educators might scaffold pre-service teachers' critical engagement with AI-generated feedback, both as learners and as future evaluators of their own students' writing.

This study has several limitations that should be considered when interpreting these findings. The three cases were purposively selected to illustrate contrast rather than to represent the full pilot sample. The analysis relies on textual comparison and the AI system's own ratings and commentary rather than participant interviews, so engagement patterns are inferred rather than confirmed by participants themselves. Ratings were generated by CyberScholar, the same system whose feedback is under study, and are treated here as part of the phenomenon under investigation rather than as independent, validated indicators of writing quality; this study makes no claim that any participant's writing quality improved or declined in an externally verifiable sense. Finally, the analysis was conducted by a single researcher without a second coder, and the small number of cases and single-institution context limit generalizability; the typology proposed here should be treated as a preliminary, theory-generating contribution.

Recommendations

Three practical implications follow from these findings. First, developers of AI feedback tools might consider how feedback is framed for criteria that require conceptual or theoretical development rather than mechanical correction: feedback naming a concrete, low-effort fix was consistently acted upon in this study, while feedback requesting deeper engagement was not, suggesting that such feedback may need to be paired with more explicit guidance on how to act on it. Second, teacher educators working with pre-service teachers on AI-assisted writing tasks might explicitly discuss the distinction between incremental revision and wholesale regeneration, since both were observed in this study but carry different implications for what students actually practice and learn during revision. Third, given that rating increases alone did not reliably indicate the depth or nature of a student's engagement with feedback, instructors evaluating AI-assisted revision tasks may wish to review draft-to-draft textual change directly, rather than relying on automated ratings as a proxy for engagement or improvement. Future research with larger samples, participant interviews, and multiple coders could test whether the typology proposed here holds across a broader range of learners, tasks, and AI feedback platforms.

Scientific Ethics Declaration

Prior to data collection, the pilot study from which these cases are drawn received institutional permission from the participating university and was reviewed by the relevant Scientific Ethics Committee, which raised no objection to the study. Students were informed of the study's purpose and procedures, and informed consent was obtained before participation. For the present case study, all data were fully anonymized prior to analysis: participants are identified only by non-identifying case labels (Participant A, B, C), and no names, institutional affiliations, or other potentially identifying details are disclosed. Because case study reporting inherently involves a level of individual detail that can risk indirect identification even without names, direct quotations from participants' essays have been lightly paraphrased and specific biographical details not essential to the analysis have been generalized; all substantive claims about content, feedback, and rating change remain accurate to the source data.

* The authors declare that the scientific ethical and legal responsibility of this article published in EPESS Journal belongs to the authors.

Conflict of Interest

* The author declares that there is no conflict of interest.

Funding

* This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under the project “Artificial intelligence as an assistant in acquiring academic English by non-language major students: Randomized study” (AP25794537).

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Education in Technology and Innovation (www.iceti.net) held in İstanbul/Türkiye on September 17-18, 2026

References

- 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>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, Article 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Farrokhnia, M., Latifi, S., Papadopoulos, P. M., Hogenkamp, L., Gijlers, H., Khosravi, H., & Noroozi, O. (2026). Generative AI offers more, but students revise less: Comparing the effects of teacher and AI feedback on student essay revisions. *International Journal of Educational Technology in Higher Education*, 23, Article 6. <https://doi.org/10.1186/s41239-026-00579-9>
- Lo, J., Wong, C., Ng, A., Wong, P., Cheung, D., & Lai, P. (2026). Stretching AI's reach: Assessing an AI-driven feedback system for extended academic writing. *Computers and Education: Artificial Intelligence*, 10, Article 100511. <https://doi.org/10.1016/j.caeai.2025.100511>
- Wahyuni, S., Dewi, D. S., Wulandari, R. S., Thohiriyah, T., Kurniadi, D., Muawanah, F. H., Chairunnisa, N. Z., & Ardhi, M. A. (2026). Evaluating students' perceptions, writing outcomes, and challenges in using AI-based feedback systems for English academic writing. *Multidisciplinary Reviews*, 9(8), e2026380. <https://doi.org/10.31893/multirev.2026380>
- Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.
- Zheldibayeva, R., Nascimento, A. K. de O., Castro, V., Kalantzis, M., & Cope, B. (2025). The impact of AI-driven tools on student writing development: A case study. *Online Journal of Communication and Media Technologies*, 15(3), e202526. <https://doi.org/10.30935/ojcm/16738>

Author(s) Information

Raigul Zheldibayeva

Zhetysu University named after Ilyas Zhansugurov
Zhansugurov St. 187 A, Taldykorgan, Kazakhstan
ORCID iD: <https://orcid.org/0000-0002-8757-0748>

*Contact e-mail: r.zheldibayeva@zu.edu.kz

To cite this article:

Zheldibayeva, R. (2026). Modes of engagement with AI-generated feedback: A case study of three pre-service EFL teachers. *The Eurasia Proceedings of Educational and Social Sciences (EPESS)*, 53, 12-17.