

This work is licensed
under a Creative
Commons Attribution
4.0 International
License.

AI-assisted writing feedback in EFL: Tracking student performance and reflections

Junko Otoshi

Okayama University, Japan
otoshi-j@okayama-u.ac.jp

Naomi Fujishima

Kawasaki Medical School, Japan
nfujishi@med.kawasaki-m.ac.jp

This study explores how AI-assisted writing feedback supports English as a Foreign Language (EFL) learners' writing development and feedback literacy in a Japanese university. Twenty-one first-year students completed nine Write & Improve (W&I) tasks, progressing from descriptive to argumentative essays. Each task was revised following W&I feedback, and students were asked to write a short reflective log. An explanatory mixed-methods approach was adopted, combining quantitative analyses of writing performance and linguistic features with qualitative coding of students' reflections. Findings showed measurable gains in both higher- and lower-level groups, with particularly notable improvement among lower-level students in complexity, fluency, and sophistication. While the higher group consistently outperformed the lower group in Term 1, this gap narrowed in Term 2. Reflection logs provided important insights into feedback literacy: students initially valued W&I for surface-level corrections but later expressed frustration as tasks became more complex and scores plateaued. This tendency was especially evident among lower-level students, reflecting both the affordances and limitations of automated feedback. The study concludes that AI-assisted tools can foster writing development and emerging feedback literacy, but sustained progress requires teacher mediation. Integrating Automated Writing Evaluation (AWE) into ecological feedback environments – combining AI, teacher, and student reflection – offers a promising approach for sustainable L2 writing instruction.

Keywords: EFL writing, AI-assisted tools, feedback literacy, ecological environments, Write & Improve

Introduction

Automated writing evaluation (AWE) tools such as Cambridge's Write & Improve (W&I) provide scalable feedback on grammar, vocabulary, and text organization. Although they are less powerful than generative AI, these tools remain useful by delivering immediate, individualized feedback that supports students' long-term grammatical accuracy and self-revision skills (Link et al., 2022).

Despite their potential, questions remain regarding how learners interpret and use AWE feedback. Much of the research has emphasized the accuracy of automated scoring or the technological affordances of AWE, while less attention has been paid to learners' agentic engagement – their beliefs, strategies, and emotional responses (Kasneci et al., 2023; Yang & Zhang, 2023).

The concept of feedback literacy, defined as the capacity to interpret and apply feedback for improvement (Sutton, 2012), has become increasingly central in sustainable learning, especially with the rise of generative AI. In L2 writing, feedback literacy entails not only revising text but also fostering social and emotional engagement (Rad & Mirzaei, 2024). Wang and Lee (2021) suggest that learners' self-initiated regulation of their writing processes through resource-seeking, meaning negotiation, and strategic adaptation leads to better writing gains.

Reflection logs are one pedagogical means to encourage this process. By recording learners' interpretations and emotional reactions, logs help students monitor their writing and give teachers insight into how AI feedback is received. However, without teacher mediation or self-regulation support, English as a Foreign Language (EFL) learners may misinterpret or inconsistently act on AWE feedback (Fan, 2023; Link et al., 2022). This calls for pedagogical designs that integrate AWE, teacher, and self-feedback within broader ecological feedback environments.

The present study explores the use of W&I in a Japanese university academic writing course. It investigates: (1) how students' writing developed over time with repeated use of W&I, (2) what differences emerged between higher- and lower-level writers, and (3) how reflection logs revealed students' engagement with AI feedback and evolving feedback literacy.

Methods

Context and participants

The study was conducted in a required academic writing course at a Japanese university in spring 2024. The eight-week course met weekly for 50 minutes. Due to limited class time, students who majored in law, economics, and education were required to practice independently outside class. Twenty-five students enrolled in the course; four were excluded due to incomplete work – defined as completing fewer than seven of the nine required writing tasks – leaving 21 participants for analysis. Their English proficiency ranged from A2 to B1 on the CEFR (Common European Framework of Reference for Languages). Ethical approval and informed consent were obtained.

Students completed nine W&I tasks – five in Term 1 and four in Term 2 – progressing from descriptive to argumentative genres. Each submission was revised at least once after W&I feedback. Students then wrote short (about 30 words) reflection logs

in English on the feedback and their learning experience. Logs were collected via Moodle. The initial coding was conducted by the first author, who carefully reviewed each reflection to ensure accurate representation of student voices. Subsequently, the first author consulted with the second author to review and refine the thematic structure and category definitions. This collaborative process helped strengthen coding validity by ensuring interpretive consistency.

Writing tasks

Table 1. Writing tasks for the two terms

Term	Task number	Essay type	Word count
1	1-2	Descriptive	150-300
1	3-5	Narrative	150-300
2	6-7	Comparison-contrast	330-350+
2	8-9	Argumentative	330-350+

Data analysis

An explanatory mixed-methods design integrated quantitative and qualitative data.

- Quantitative: W&I’s CEFR scores were converted to a 6-point scale. Students were divided into higher ($n = 9$) and lower ($n = 12$) groups using a median split. Mann–Whitney U tests compared group scores for each task, while Wilcoxon signed-rank tests assessed development in linguistic features (complexity, accuracy, fluency, logicity, sophistication) using the CEFR-based Writing Level Analyzer (Uchida & Negishi, 2025).
- Qualitative: 260 reflection logs were thematically coded using a deductive approach (Saldaña, 2021). Eight focused categories emerged, grouped into three dimensions: AI assessment (Accuracy, Trust), personal experience (Emotional Reaction, Perceived Learning, Recognition of Progress, Effort Level), and AI feedback utility (Clarity, Helpfulness). To explore changes across groups and terms, Fisher’s exact tests were applied.

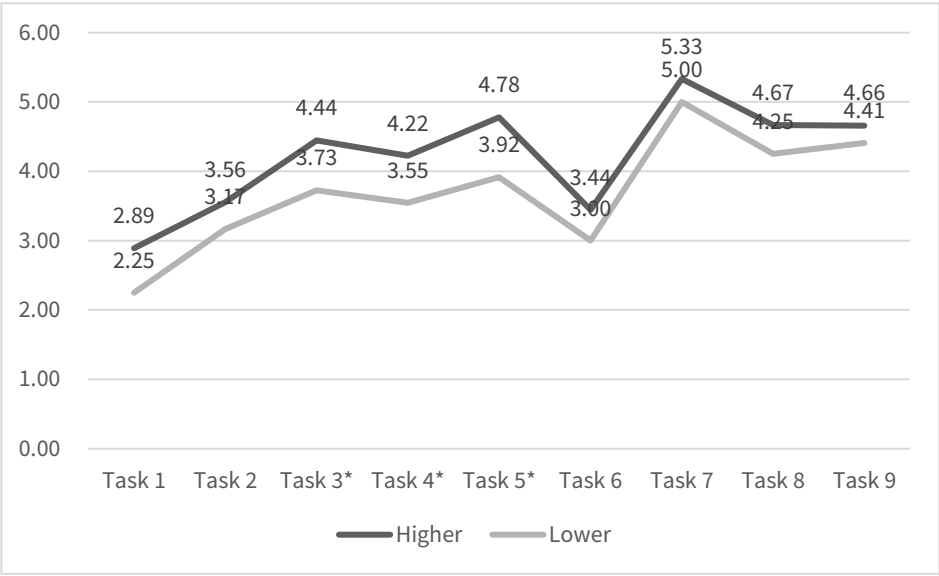
Results

Writing development

As shown in Figure 1, both groups improved across tasks, with the higher group consistently outperforming the lower group. Significant group differences were found in Term 1 (Tasks 3–5, $U = 19-21$, $p < .05$), but not in Term 2, suggesting the performance gap narrowed over time.

A notable turning point occurred in Task 6 (comparison–contrast, 4-paragraph essay), when average scores dropped, and reflection logs showed frustration. After the instructor provided a model essay, scores for Task 7 rose sharply – the highest across all tasks – indicating the value of teacher scaffolding alongside AI feedback.

The model essay provided by the instructor made key genre expectations for comparison–contrast writing explicit, thus giving students effective guidance.



Note. Asterisk marks for Tasks 3–5 show statistical differences between groups by Mann–Whitney U tests.

Figure 1. Transition of average writing task scores between groups

Linguistic features

Comparisons between Task 5 (narrative) at the end of Term 1 and Task 9 (argumentative) at the end of Term 2 were conducted to examine changes in linguistic features using Wilcoxon signed-rank tests, which assess within-group change over time. In this test, Z represents the standardized test statistic. The results are as follows:

- Higher group: significant improvement in complexity ($Z = -2.06, p = .039$) and fluency ($Z = -2.16, p = .031$).
- Lower group: significant improvement in complexity ($Z = -2.97, p = .003$), fluency ($Z = -2.33, p = .020$), and sophistication ($Z = -2.60, p = .009$).

Feedback literacy

Fisher’s exact test was utilized to examine statistically significant changes. Table 2 summarizes coded categories across groups and terms. Only the lower group showed a significant change in two categories: Emotional Reaction and Helpfulness.

Table 2. Frequency of student responses between groups and terms

Group	Total counts	Accuracy	Trust	Emotional reaction	Perceived learning	Recognition of progress	Effort level	Clarity	Helpfulness
LT1	83	3 (3.6%)	5 (6.0%)	10 (12.0%)	19 (22.8%)	3 (3.6%)	18 (21.6%)	9 (10.8%)	16 (19.2%)
LT2	53	0 (0%)	6 (11.3%)	16 (30.1%)	16 (30.1%)	0 (0%)	7 (13.2%)	5 (9.4%)	3 (5.6%)
HT1	55	2 (3.6%)	3 (5.4%)	4 (7.2%)	11 (20%)	1 (1.8%)	14 (25.4%)	10 (18.1%)	10 (18.1%)
HT2	43	0 (0%)	2 (4.6%)	9 (20.9%)	13 (30.2%)	3 (6.9%)	7 (16.2%)	4 (9.3%)	5 (11.6%)

Note. LT1 refers to the lower group in Term 1, LT2 to the lower group in Term 2, HT1 to the higher group in Term 1, and HT2 to the higher group in Term 2. The counts indicate the frequency of coded responses for each focused category by group and term. Significant differences between Term 1 and Term 2 within groups were assessed using Fisher’s Exact test.

- **Emotional reaction:** increased significantly from Term 1 to Term 2 ($p = .01$). Early comments reflected mixed emotions:

My CEFR level improved, so I was happy. (Term 1)

But later logs revealed frustration:

The CEFR level decreased. I was shocked by it, or I worked hard on the essay, but the level didn’t go up to C. (Term 2)

- **Helpfulness:** perceived usefulness declined significantly ($p = .04$). Early comments valued W&I for highlighting surface errors:

I learned that I still made a lot of spelling mistakes. (Term 1)

However, later comments noted diminished support as tasks grew more complex.

Although no other subcategories showed statistical differences, Perceived Learning was the largest category across groups and terms. Comments evolved from grammar/vocabulary concerns to rhetorical awareness. For example, one student in Term 1 wrote:

I tried to use not ordinary adjectives, but they may not fit for my essay. (Term 1)

By Term 2, reflections were more metacognitive:

I used specific examples in this writing. I think it was very difficult to summarize the content using that specific example. (Term 2)

This shift suggests emerging feedback literacy, as students moved from surface-level corrections toward broader rhetorical concerns, even while expressing frustration.

Discussion

Developmental gains and narrowing gaps

The findings indicate that repeated engagement with W&I supported measurable writing development in both groups. Lower-level students, in particular, made notable gains in complexity, fluency, and sophistication. These results echo previous

research showing that automated feedback can scaffold incremental improvement (Link et al., 2022). Repeated practice appeared to benefit less-skilled learners more, enabling them to close the gap with higher-level peers by the end of the course.

Feedback literacy and emotional trajectories

Initially, students valued W&I primarily for surface-level corrections, but lower-level students' frustration increased as tasks shifted toward argumentative genres. This tension reflects both the affordances of AWE – such as immediacy and individualized feedback – and its constraints, particularly its emphasis on form. Importantly, this frustration became visible during classroom interaction and through declining scores. In response to this issue, the instructor, the first author of this paper, introduced a model text aligned with the task demands. Following this intervention, students' scores improved, and their reflections indicated increased confidence in their writing. From an ecological perspective, this episode illustrates feedback as an adaptive process emerging from learners' affective responses, teacher noticing, and timely pedagogical mediation. The teacher's intervention is especially important for lower-level students.

Additionally, this episode sheds light on the importance of reflection logs: they served as diagnostic tools, signaling when students struggled emotionally and cognitively. By recognizing these cues, teachers can intervene with timely scaffolding.

Implications for ecological feedback environments

This study highlights the value of integrating AWE into an ecological feedback design in which AI, teacher, and student roles dynamically intersect. While AWE supported independent learning and attention to surface-level features, teacher scaffolding became particularly important at moments when task demands increased, and rhetorical control was required. Reflection logs functioned as a mediating tool between these aspects, making students' perceptions and difficulties visible to the instructor and enabling timely pedagogical adjustments. Together, these elements formed a feedback environment in which different roles of support complemented one another, allowing the strengths of AWE to be leveraged while its limitations were addressed through human mediation.

A limitation of this study is that increased rhetorical demands and longer length requirements co-occurred in the Term 2 tasks, making it difficult to isolate their individual effects. Both factors likely contributed to student frustration and the narrowing performance gap. Future studies should disentangle these task variables through a more controlled task design.

Conclusion

AI-assisted feedback fostered measurable gains in writing development and feedback literacy among Japanese university students, particularly lower-level learners. However, reliance on surface-level feedback led to frustration as tasks became more complex, underscoring the continuing importance of teacher mediation.

These findings support the design of ecological feedback environments that balance AI-driven independent practice with human support. While the present study focused on AWE tools that primarily target surface-level features, future research should examine whether similar learning trajectories emerge when generative AI systems capable of offering more rhetorically oriented feedback are incorporated alongside teacher and peer mediation, and whether such combinations better support the development of long-term feedback literacy.

References

- Fan, N. (2023). Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study. *SAGE Open*, 13(2).
<https://doi.org/10.1177/21582440231181296>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 35(4), 605–634.
<https://doi.org/10.1080/09588221.2020.1743323>
- Rad, H. S., & Mirzaei, A. (2024). Developing feedback literacy, scaffolded writing, and resilience through intervention on feedback processes for L2 writing students. *Studies in Educational Evaluation*, 81, 101354.
<https://doi.org/10.1016/j.stueduc.2024.101354>
- Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage Publications.
- Sutton, P. (2012). Conceptualizing feedback literacy: Knowing, being, and acting. *Innovations in Education and Teaching International*, 49(1), 31–40.
<https://doi.org/10.1080/14703297.2012.647781>
- Uchida, A., & Negishi, M. (2025). *CEFR-based writing level analyzer* (Version 2.1) [Computer software]. CEFR-J Project.
- Yang, L., & Zhang, L. J. (2023). Self-regulation and student engagement with feedback: The case of Chinese EFL student writers. *Journal of English for Academic Purposes*, 63, 101226. <https://doi.org/10.1016/j.jeap.2023.101226>
- Wang, L., & Lee, I. (2021). L2 learners' agentic engagement in an assessment as learning-focused writing classroom. *Assessing Writing*, 50, 100571.
<https://doi.org/10.1016/j.asw.2021.100571>

Author biodata

Junko Otoshi is a professor of Okayama University. Her research interests include L2 writing studies and intercultural communication.

Naomi Fujishima is a specially appointed professor at Kawasaki Medical School in Okayama prefecture. Her research interests include student motivation and lifelong learning as well as second language writing.



JALTCALL
Trends
Vol. 2 No. 1
2025 JALTCALL
Conference Papers
Special Issue