Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From a "Heavy Weight on My Shoulders" to "Reading a Text by a Stranger": Temporal Dynamics in AI-Assisted Writing Feedback

Mert Uzun

muzun20@posta.pau.edu.tr

Pamukkale University

İnci Demir

Pamukkale University

Devrim Höl

Pamukkale University

Research Article

Keywords: AI-generated feedback, feedback timing, L2 writing, Google Gemini Automated Writing Evaluation, GenAI Tools

Posted Date: June 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9876127/v1

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

Additional Declarations: No competing interests reported.

Abstract

The integration of generative AI (GenAI) into second-language (L2) writing instruction has expanded rapidly, yet the timing of AI-generated feedback remains empirically underexplored. Grounded in Cognitive Load Theory, the Noticing Hypothesis, and Sociocultural Theory, this study examined whether immediate versus delayed Gemini-generated feedback differentially affects L2 writing development. Using a within-subjects quasi-experimental mixed-methods design, 65 first-year English Language Teaching department students at a Turkish state university completed seven weekly writing tasks across two genres (argumentative and problem-solution essay types) over a counterbalanced seven-week intervention, receiving feedback either within the same session or one week later. Drafts were scored with Hughes' (2003) analytic rubric and analyzed using repeated-measures ANOVA, and reflective papers were thematically analyzed. Both conditions produced significant improvement in both genres, but it was also found that feedback timing did not significantly moderate overall writing gains. At the sub-dimensional level, delayed feedback significantly outperformed immediate feedback only for accuracy and mechanics in argumentative essays (with vocabulary trending similarly), whereas the cognitively more demanding problem-solution genre showed no timing effects. Qualitative findings linked immediate feedback to cognitive overload and uncritical uptake, and delayed feedback to deeper noticing. The temporal benefit of AI feedback appears narrow and task-dependent, supporting AI feedback as a complement to teacher mediation.

1. Introduction

Artificial Intelligence (AI) tools and Chatbots have gained increasing prominence in education, particularly for their capacity to facilitate human-like interactions in AI-assisted writing. (Barrot, 2023; Zhou & Li, 2023). The integration of Generative AI (GenAI) tools such as Gemini, ChatGPT, Claude, etc., as automated writing evaluation (AWE), is progressively transforming writing instruction (Liang et al., 2021). Moreover, the GenAI tools not only have the capacity to assist the human evaluators by giving immediate feedback but also have the potential to reduce teachers' workload in writing assessment (Wang et al. 2026). In traditional L2 writing classes, teachers provide the majority of feedback. (Han & Hyland, 2019). However, with the recent development of GenAI tools, this traditional view has changed to become an AI-supported writing feedback model (Han & Li, 2024). Consequently, AWE has become a central focus in the field of L2 writing with this evolution and development (Hirvela & Belcher, 2024). The impact of GenAI on writing showed that AI integration significantly enhances grammatical accuracy, vocabulary diversity, and overall writing quality (Barrot, 2023; Mekheimer, 2025). Furthermore, GenAI tools can also produce content-level feedback (Dai et al. 2023; Xiao & Zhi, 2023), which may be an assessor throughout the writing process. This capacity has made AI-assisted writing feedback a central area of inquiry in L2 writing.

The pedagogical significance of this technological shift is underscored by converging evidence from the empirical literature. Research has consistently demonstrated that GenAI-assisted feedback interventions can produce meaningful gains in grammatical accuracy, vocabulary diversity, and overall writing quality

(Barrot, 2023; Liu et al., 2025 ; Mekheimer, 2025; Wang et al., 2026). Despite this expanding body of scholarship, a substantive gap persists in the literature: the temporal dimension of feedback delivery has received insufficient empirical attention in the higher educational settings. Prior research has examined feedback quantity, quality, and modality, yet the question of when AI-generated feedback is most effectively delivered relative to learner writing performance remains unresolved. This gap has implications in theory as the timing of input is not a neutral variable and has significant effects on cognitive processing and social mediation of L2 growth. Barrot (2023) explicitly called for more examination of GenAI's potential in L2 writing contexts. Liu et al. (2025) similarly urged researchers to test AI tools under classroom conditions. The present study seeks to respond directly to these calls.

Grounded in the convergent explanatory power of Schmidt's (1990) Noticing Hypothesis, Sweller's (1988) Cognitive Load Theory, and Vygotsky's (1978) Sociocultural Theory, this study aims to investigate the optimal timing of AI-generated feedback in L2 writing. Each framework illuminates a distinct dimension of the feedback timing question: the Noticing Hypothesis foregrounds the role of conscious attention in L2 writing, Cognitive Load Theory introduces the constraint of working memory capacity on feedback processing, and Sociocultural Theory situates feedback within a mediational, scaffolding paradigm that is inherently sensitive to temporal positioning. Taken together, these frameworks provide a principled theoretical architecture for examining whether immediate or delayed Gemini-generated feedback produces differential outcomes in L2 writing development.

Accordingly, this quasi-experimental study addresses a gap in the literature by examining the optimal timing of AI-generated feedback to investigate L2 writing developments as a function of immediate versus delayed feedback delivery. Immediate feedback was operationalized as feedback provided within the same instructional session following task submission, delayed feedback was delivered one week after task completion, allowing a 'cooling-off' period designed to promote deeper cognitive processing and metacognitive reflection (Ferris, 2003; Michaud et al., 2025). Both feedback conditions were administered using Google Gemini (v. 1.5) through a standardized prompt protocol.

The study addresses the following research questions:

RQ 1: Does the timing of Gemini-generated feedback (immediate vs. delayed) significantly affect L2 learners' overall writing quality over time?

RQ 2: Are there differential effects of immediate versus delayed AI feedback on specific sub-dimensions of writing quality, including grammatical accuracy, vocabulary use, coherence, and content development?

RQ 3: How do L2 learners perceive the effectiveness and usability of immediate versus delayed Gemini-generated feedback in supporting their writing development?

These questions are addressed through a quasi-experimental intervention for 6 weeks, in which first-year English Language Teaching (ELT) program students received Gemini-generated feedback at different temporal intervals following weekly writing tasks.

2. Literature Review

2.1 Feedback in L2 writing

Feedback is widely recognized as shaping learners' linguistic accuracy, content development, and motivation (Hyland & Hyland, 2006; Nguyen & Chu, 2024). Research shows that written corrective feedback (WCF) can improve grammatical accuracy and overall writing quality (Guo, 2023; Lu, 2023; Nguyen & Chu, 2024). However, the literature also highlights the potential challenges and negative effects of feedback, such as emotional distress or demotivation (Mao & Lee, 2025; Shao, 2015). These effects vary with learner variables such as working memory capacity, aptitude, motivation, beliefs about feedback, and classroom environment (Ahmadian & Vasylets, 2021; Bakri, 2016). In sum, while feedback is broadly beneficial for L2 writing, its effectiveness varies as a function of feedback type, delivery mode, learner characteristics, and classroom context (Ahmadian & Vasylets, 2021; Hyland & Hyland, 2006; Nguyen & Chu, 2024). Beyond feedback type, the field has examined the source of feedback delivery. Teacher feedback remains the most extensively researched and pedagogically valued (Hyland & Hyland, 2006). However, the emergence of automated feedback systems has expanded the delivery mode landscape. Research comparing teacher and AWE feedback suggests that integrated approaches produce deeper cognitive and behavioral engagement than teacher-only feedback (Cheng & Zhang, 2024), pointing to the motivational potential of AI feedback systems in writing instruction.

2.2 The Temporal Dimension of Feedback Delivery

The timing of feedback delivery has emerged as a critical factor influencing the effectiveness of WCF in L2 writing (Fu & Li, 2022). Research demonstrates that both immediate and delayed feedback can facilitate L2 development, but their relative benefits may depend on characteristics, task types, and the nature of errors addressed (Michaud et al., 2025; Kourtali & Borges, 2023). In the psychology field, Skinner's (1953) behaviorism, which considers feedback as a tool that corrects errors and encourages correct behaviors, supports instant feedback. Immediate feedback is often associated with short-term gains in accuracy and engagement when learners can revise while their memory of the writing process is still fresh. On the other hand, the delayed feedback may promote deeper processing and metalinguistic reflection, especially for complex or semantic features (Michaud et al., 2025; Kourtali & Borges, 2023). Recent empirical work has increasingly examined this timing question. Michaud et al. (2025) found that delayed WCF fostered deeper metalinguistic reflection, while immediate WCF demonstrated a significant advantage for long-term accuracy gains and error reduction on delayed post-tests.

Further in the literature, Fu and Li (2022) examined how feedback timing affects learning outcomes. A systematic review of 20 studies investigating effective feedback timing in SLA concluded that there is no definite answer to the question of when errors in L2 should be treated; both immediate and delayed feedback groups outperform no-feedback groups, suggesting that the presence of feedback is more consequential than its precise timing. Crucially, the timing of GenAI-generated feedback remains absent from this body of literature, a gap the present study addresses.

2.3 Automated Writing Evaluation and Generative AI Tools

Recently, significant advancements in GenAI and AWE tools have offered new opportunities for AI-assisted feedback in L2 writing (Karatay & Karatay, 2024; Zhang, 2020). It is important to distinguish between traditional AWE tools, such as Grammarly, Pigai, and Criterion, and GenAI models such as Gemini, Claude, and ChatGPT, which offer more advanced support in drafting, revising, and synthesizing a text.

Regarding the effectiveness of traditional AWE tools, Hassanzadeh and Fotoohnejad (2021) examined Criterion in the Iranian L2 writing classes context. The experimental study consisted of tertiary-level students divided into teacher corrective (TC) feedback and automated corrective (AC) feedback groups. After the treatment, the AC feedback group's writing scores increased significantly, which outscored the TC feedback group. Furthermore, the study conducted by Taşkıran (2022) investigated the perceptions of AI-based evaluation in writing. The Write & Improve tool was used as an instrument in the study. More than half of the distant learner participants chose to get feedback from both the AWE tool and the teacher. It was claimed that the reasons for this preference could be the challenges learners face when they are trying to understand implicit feedback, and the AWE tool focuses on only the surface-level mistakes. The study concludes that participants recognized the benefits of AWE tools. However, they also expressed that using AWE alongside the guidance of a teacher would be more effective. Similarly, Zhang (2020), in a small-scale qualitative study of three learners using Pigai, found two participants held positive views of AWE feedback while one remained neutral, with positive perceptions linked to a focus on instrumental benefits over limitations. While AWE tools are generally regarded as beneficial for L2 writing, the rapid adoption of AI-based writing assessment means their effectiveness warrants further empirical investigation.

Recent studies showed that AWE tools can guide students and foster independent learning (Naz & Robertson, 2024). It can be understood that these tools can achieve more than alleviating the burden of teachers and improving writing accuracy. One of the studies supporting this, mixed-methods study conducted by Song and Song (2023) to investigate the effects of AI use on Chinese university students' writing skills and writing motivation. Participants were assigned to the experimental group and the control group, which lasted 12 weeks of intervention. The experimental group received feedback from ChatGPT on grammar, vocabulary, coherence, and sentence structure. Moreover, the Writing Motivation Scale questionnaire was used to measure students' motivation. The findings showed the AI-assisted group demonstrated significantly higher levels of writing motivation and overall writing scores. The qualitative phase of the study showed that students felt the AI made the writing process more enjoyable and engaging. However, some students mentioned limitations of the AI, such as occasionally the tool providing suggestions that were not accurate contextually, and becoming overly reliant on the tool could hinder their critical thinking.

The study Dizon and Gold (2023) conducted aligned with the previous results. The researchers employed a quasi-experimental study to investigate how Grammarly, as an AWE tool, influences foreign language

anxiety and learner autonomy. Undergraduate students enrolled in academic writing courses at a Japanese university are required to use Grammarly for weekly writing assignments over a 16-week semester. The pre- and post-survey results indicated that integration of the tool Grammarly had a significant positive effect on increasing learner autonomy and also reduced the students' foreign language anxiety significantly. Students' reflective reports revealed that the tool has advantages in error correction, English improvement, and enhanced metalinguistic awareness; however, students mentioned the tool sometimes gives inaccurate and vague feedback, which indicates that AWE tools require critical evaluation by the user.

GenAI systems such as Gemini, ChatGPT, and Claude tools have transformed feedback provision, grading practices, and student engagement. These tools provide immediate and individualized feedback comparable in many aspects to human-generated feedback, which supports improvements in writing proficiency, motivation, and revision habits (Alsofyani & Barzanji, 2024; Escalante et al., 2023; Yavuz et al., 2025). However, only AI-generated feedback is praised for its ability to address surface-level issues (grammar, vocabulary), which highlights the importance of combining AI with teacher input to address higher-order concerns such as content development and argumentation (Tran, 2025).

Escalante et al. (2023) investigate the efficacy of using ChatGPT as an AWE tool and the student preferences regarding AI-generated versus human-generated feedback. The research focuses on university-level English learners through two distinct longitudinal studies. The researchers employed a two-part study to separate efficacy and preference. The first study utilized a six-week longitudinal quasi-experimental design; the group received personalized corrective feedback generated by GPT-4, while the control group received weekly with a human tutor. In the second study, participants received both AI-generated and human tutor feedback on their weekly writing assignment and completed a weekly preference questionnaire, which lasted six weeks. The quantitative results indicated that GenAI tools are a highly viable alternative to human tutors for supporting linguistic development, as the previous results showed. While both the experimental and control groups made similar progress in their overall writing proficiency, the researchers highlighted that AI-generated feedback offers time-saving benefits for educators. The student preferences data revealed that they were nearly divided, half preferring human feedback as they found face-to-face interaction more engaging and effective, and those who preferred AI feedback cited clarity and consistency of the feedback.

Further supporting these findings, Tiandem-Adamou (2024) conducted a quasi-experimental study integrating GenAI tools with cooperative learning over six weeks. The experimental group significantly outperformed the control group, with notable gains in coherence, cohesion, grammatical accuracy, and vocabulary. However, qualitative results highlighted that teachers' feedback is highly valued as it is more personalized and context-aware. Therefore, the results indicate the teacher's role in providing feedback on writing cannot be overlooked, and relying overly on AI-generated feedback may not always be as effective as expected.

Conversely, AWE and GenAI systems face several critical limitations and challenges. One of the drawbacks of AWE systems is their limited capacity to address higher-order writing skills such as organization, coherence, idea development, critical thinking, and creativity. Addressing challenges to GenAI integration in language education, Vaccino-Salvadore (2023) identified five key themes users need to take into consideration while using these tools: privacy of the data and security, lack of diversity and potential biases, accessibility and reliability of the tools, authenticity of the output, and academic dishonesty issues. The author mentioned that language learners might share sensitive personal thoughts, emotions, and experiences during the interaction, which raises concerns about data privacy and security. Regarding cultural representation, heavy reflection on the American socio-cultural worldview may contribute to the erosion of cultural diversity. Additionally, the AI tools may require specific hardware, software, and internet connectivity, which are not affordable or available universally. At the discourse level, AI-generated content has been criticized for lacking the emotional, social, and cultural depth of human communication, which is entirely disconnected from real-world use. Finally, and perhaps most consequentially for academic contexts, GenAI tools have concerns surrounding plagiarism, cheating, and fabricated sources (Eke, 2023). These limitations and challenges highlight the providing feedback based solely on GenAI tools.

2.4 Theoretical Frameworks

2.4.1 Cognitive Load Theory

Cognitive Load Theory (CLT), Sweller's (1988) framework posits that working memory has a finite capacity and the total cognitive load experienced during learning is composed of three interrelated types. The internal complexity of the learning material itself is referred to as intrinsic cognitive load (Sweller et al., 1998). In contrast, extraneous cognitive load is not inherent to the material but is instead generated by instructional design, such as unnecessarily complex feedback formats that consume working memory resources without contributing to learning (Paas et al., 2003). Germane cognitive load refers to the mental effort productively directed toward schema construction and automation in which the process of new information is integrated into long-term memory (Sweller & Chandler.,1991; Sweller, 1994; Van Merriënboer & Sweller, 2010). Critically, these three loads are interconnected; when intrinsic and extraneous demands are high, fewer cognitive resources remain available for the germane processing for deep learning. In the context of L2 writing feedback, this framework has direct implications for the timing and design of AI-generated feedback. Feedback that is too complicated or poorly timed might increase unnecessary burden and obstruct the schema formation that is intended to help language learners.

In the context of L2 writing feedback, CLT carries direct implications for the timing and design of AI-generated corrective input. Woo et al. (2024) explored this dynamic empirically by integrating AI (ChatGPT) into the EFL writing classroom to investigate how learning to write with a 'machine-in-the-loop' impacts ELT students' cognitive load. Students participated in a workshop where they were introduced to the basics of prompt engineering and then asked to complete a 500-word writing task. The

results showed that students experienced high cognitive load during the writing task. This high cognitive load is a crucial variable to evaluate when comparing the effects of immediate versus delayed AI-generated feedback on the drafting process. From this perspective, feedback is not only corrective information but a form of mediation that scaffolds the learner toward linguistic competencies that lie just beyond their current independent capacity. Written feedback has been shown to help students improve writing quality and skills, encourage critical reasoning, and promote learner autonomy, while social interactions during feedback activities also help students extend their ZPDs (Fithriani, 2019). From a CLT perspective applied to AI feedback, these results raise a critical question: whether immediate or delayed AI feedback more effectively positions corrective input within the learner's available working-memory capacity.

2.4.2 Schmidt’s Noticing Hypothesis

Schmidt’s (1990) Noticing Hypothesis posits that conscious attention to linguistic form is a necessary precondition for second language acquisition. Noticing enables learners to register the gap between their current interlanguage output and target-like forms; without this conscious processing, corrective feedback produces only surface-level error reduction rather than structural change in the developing L2 grammar (Ranalli, 2021).

The temporal proximity of feedback is theoretically consequential: immediate feedback preserves working memory activation of the learner’s own production, facilitating gap recognition, while delayed feedback separates corrective information from the original linguistic event, potentially altering the noticing mechanism. ElEbyrary et al. (2024) confirmed timing as a significant moderator of learner noticing in L2 writing, while Koltovskaia (2020) cautioned that AWE feedback may be accepted without the conscious reflection that meaningful noticing requires.

2.4.3 Vygotsky’s Sociocultural Theory

Vygotsky’s (1978) Sociocultural Theory (SCT) states cognitive development within social interaction, positing that all higher psychological functions originate in dialogic engagement with more capable interlocutors before being internalized as individual mental processes. The Zone of Proximal Development (ZPD) constructs are central to the theory, which is defined as the distance between the level of independent problem-solving and the higher level achievable through collaboration with a more capable peer or expert (Panahi et al., 2013). Applied to L2 writing, feedback functions as a form of mediated scaffolding enabling learners to operate beyond their current independent capacity (Aljaafreh & Lantolf, 1994; Fithriani, 2019). Temporally, SCT implies that feedback must be delivered when the learner is cognitively and motivationally positioned to engage with the mediational input. When AI tools serve as feedback agents, the temporal positioning of feedback determines whether it effectively operates within the learner’s ZPD to scaffold linguistic development.

Taken together, these three theoretical frameworks (Cognitive Load Theory, Noticing Hypothesis, and Sociocultural Theory) converge around the temporal variable of feedback delivery while addressing distinct mechanisms through which timing may moderate L2 writing development. CLT highlights

cognitive load constraints on feedback processing; the Noticing Hypothesis foregrounds the role of temporal proximity in activating metalinguistic awareness; and SCT states the effectiveness of AI-generated feedback within the dynamic, socially constituted space of the ZPD. Their integration provides a principled and multi-layered theoretical architecture for the present study's investigation of feedback timing in AI-assisted L2 writing instruction.

2.5 Gemini as a feedback provider

Among the GenAI tools integrated into L2 writing instruction, Google Gemini remains considerably underrepresented in the empirical literature relative to ChatGPT.

Emerging studies report that Gemini enhances writing quality, supports critical thinking, and improves organization and coherence, though ethical concerns around plagiarism and academic integrity remain (Alpar, 2025; Azmi & Fithriani, 2025; Do, 2025). Despite this evidence, quasi-experimental studies examining Gemini as a primary feedback agent remain sparse. The present study contributes to this underdeveloped body of research by employing Gemini as the feedback agent. Gemini was selected for several converging reasons: its capacity to generate discourse-level feedback encompassing grammatical, lexical, and content dimensions, its institutional accessibility through a 12-month premium subscription available to higher education students, its responsiveness to standardized prompt engineering, and its multilingual processing capabilities, which render it particularly suited to EFL learner populations (Barrot, 2023; Liu et al., 2025).

3. Methodology

3.1 Research Design

This study used a within-subjects quasi-experimental mixed-methods design carried out during the spring semester of the 2025–2026 academic year. Rather than comparing two separate groups, every student experienced both feedback conditions (immediate and delayed) at different points during the study. This approach was chosen so that each student effectively served as their own basis for comparison, which removes the influence of individual differences such as prior writing ability, language proficiency, and motivation that can distort comparisons between separate groups (Dörnyei, 2007). The study examined how the *timing* of feedback from the generative AI tool Gemini affects the writing development of L2 learners, combining analysis of students' actual drafts with analysis of their written reflections to build a fuller, triangulated picture of the intervention (Creswell, 2014).).

3.2 Participants and Context

The participants were 65 English Language Teaching first-year undergraduate students enrolled in a compulsory academic writing course, taught in two intact classes. They were predominantly native Turkish speakers with English proficiency between B1 + and B2 on the Common European Framework of Reference. Before the intervention, a baseline writing sample confirmed that the two classes were comparable in writing ability, so any later differences could be attributed to the intervention rather than to

pre-existing gaps. The study was embedded in the students' regular course to keep the setting as natural as possible.

3.3 Instruments

3.3.1. Generative AI Tool (Gemini)

Feedback was generated by Google Gemini (version 1.5), a large language model that produces context-sensitive written feedback. Gemini was chosen for its strong contextual reasoning and because the institution provided every student with a 12-month premium subscription, ensuring equal access (Barrot, 2023). To keep feedback consistent across students and conditions, a single standardised prompt was prepared in advance, instructing Gemini to evaluate each essay against Hughes' (2003) rubric.

3.3.2. Academic Writing Tasks

Each student completed six academic writing tasks over the intervention — three argumentative essays and three problem-solution essays. These two genres were chosen because, at the B1+–B2 level, learners are ready to handle the discourse-level demands of organising ideas and reasoning (Coffin et al., 2003). Argumentative essays asked students to defend a position with evidence; problem-solution essays asked them to identify a problem and analyse its causes and consequences. Topics came from students' own academic and social worlds (e.g., online versus traditional education; digital addiction). Every task required a three-paragraph essay of roughly 400–450 words, producing a first draft (before feedback) and a second draft (after feedback).

3.3.3. Assessment Rubric

Student writing was evaluated using Hughes' (2003) analytic rubric. This validated instrument assesses five core competencies: Content, Organization, Vocabulary, Accuracy, and Mechanics. The five criteria and their descriptors are presented in Table 1.

Table 1
Hughes (2003) Analytic Rubric: Criteria and Descriptors

Criterion	Max Score	Descriptors
Content	40	Relevance, depth of ideas, argumentation, and substantiation of claims
Organization	20	Logical sequencing, paragraph cohesion, coherence across the text
Vocabulary	20	Range, precision, appropriateness, and register of lexical choices
Accuracy	15	Grammatical accuracy, syntactic variety, and sentence-level control
Mechanics	5	Spelling, punctuation, and capitalization conventions
Note. Hughes (2003). Total possible score = 100 points.		

3.3.4 Reflection Papers

After the intervention, each student wrote a 400–500 word reflection comparing their experience of the two conditions, prompted to consider how mentally demanding each felt, how well they noticed their errors, and how helpful the feedback was.

3.4 Data Collection Procedure

Data were collected over seven weeks. The first six weeks were the writing-and-feedback cycles, and the final week was for the reflection papers. The same instructor ran all sessions, and the timing of feedback was the only thing that differed between the two conditions.

To make sure the order of the conditions did not bias the results, the two classes received the conditions in opposite sequences (a counterbalanced schedule, shown in Fig. 1). In the immediate condition, students wrote a draft, submitted it to Gemini, and then read, processed, and acted on the feedback to produce a revised draft — all within the same lesson. In the delayed condition, students submitted their draft at the end of a lesson and received Gemini's feedback one week later, revising at the start of the following lesson. This one-week "cooling-off" gap was intended to lower the mental load of writing and give students space for reflection before revising (Ferris, 2003). Apart from this timing difference, the two conditions were identical.

Before the study began, ethical approval and written consent were obtained, and students were trained to use the standardised prompt. Because absences and technical issues meant some students missed a draft in a given week, each week's analysis included only the students who completed both drafts for that week; no student was dropped from the study as a whole for a single missing submission, and their complete drafts continued to contribute wherever available.

Phase 1 (Preparation). Ethical approval was obtained from the institutional review board, and written informed consent was secured from all participants prior to data collection. In a preparatory session, students received standardized training in prompt engineering and were issued the identical prompt script used throughout the study. To establish inter-rater reliability, two trained raters independently scored a pilot sub-sample of ten essays using Hughes' (2003) analytic rubric; agreement was strong, and discrepancies were resolved through discussion before formal scoring commenced. A baseline writing sample was then collected from both classes to verify group equivalence (see Section 3.2).

Phase 2 (Intervention). Across the six weekly cycles, each participant experienced both feedback conditions under a counterbalanced schedule. In the Immediate condition, students completed a writing task, submitted their first draft to Gemini using the standardized prompt, and received, processed, and incorporated the feedback into a revised second draft within the same instructional session. In the Delayed condition, students submitted their first draft at the end of the session and received Gemini feedback one week later; this one-week 'cooling-off' interval was designed to reduce task-induced cognitive load and promote metacognitive reflection (Ferris, 2003), with the revised second draft

produced at the start of the subsequent session. Apart from feedback timing, the two conditions were procedurally identical.

Phase 3 (Post-Intervention). Within one week of the final cycle, participants submitted a reflective paper (see Section 3.3.4). All first and second drafts were then scored against the analytic rubric, and the reflective papers formed the qualitative dataset for thematic analysis

3.4.1. Data Completeness and Missing Submissions

A noteworthy limitation of the data collection process concerns incomplete submission rates. Absences, personal circumstances, and technical difficulties resulted in missing 1st Drafts, 2nd Drafts, or both for a subset of students in each class. Specifically, participants with incomplete pairs for a given week were excluded from the analysis of that week's data only, in accordance with pairwise deletion procedures recommended for repeated-measures designs with missing data (Tabachnick & Fidell, 2019). Of the 65 enrolled participants, 56 provided complete baseline writing samples sufficient for the group equivalence analysis (Class A, n = 25; Class B, n = 31). The remaining 9 participants had missing pre-intervention data and were included from the baseline comparison only; their subsequent complete submission pairs continued to contribute to the main intervention analyses where available. During the analysis of data participants were not excluded from the study entirely on the basis of one or more missing submissions; their available complete pairs continued to contribute to the analysis.

3.5 Data Analysis

For the writing scores, each student's first- and second-draft scores were compared to measure how much they improved, and this improvement was then compared between the immediate and delayed conditions, separately for each genre. This was done using a repeated-measures analysis (with class included to confirm the counterbalancing had no effect); the focus was on whether the two timing conditions led to different amounts of improvement. The reflection papers were examined through thematic analysis to identify recurring patterns in how students experienced the two conditions. Together, the two strands allowed the numerical findings and the students' own accounts to be read alongside one another (Creswell, 2014).

4. Results

The results are organized according to the study's three research questions and the research objective. The immediate and delayed feedback conditions were compared separately to examine students' progress in the writing process across two different essay types (Argumentative and Problem-Solution). The amount of progress was analyzed using 'Gain Scores', which were calculated by subtracting the first draft (pre-test) scores from the second draft (post-test) scores. Preliminary analyses addressing assumptions are reported first, followed by overall writing quality (RQ1), writing sub-dimensions (RQ2) and learner perceptions (RQ3).

4.1 Preliminary Analyses

Prior to the main analyses, the distributions of the dependent variables were inspected and judged acceptable for parametric analysis. Because each within-subjects factor (timing and draft) had only two levels, the sphericity assumption did not apply and Mauchly's test was not required. The repeated-measures ANOVA is also robust to minor departures from normality at the present sample sizes (Field, 2018).

4.2 Effects of Feedback Timing on Overall Writing Quality (RQ1)

Table 2 *Descriptive Statistics for Total Writing Scores by Genre, Timing, and Draft*

Genre	*Condition*	*Draft 1*	*Draft 2*	*Gain (Δ)*
Argumentative	*Immediate*	*77.90*	*82.69*	*+4.79*
	Delayed	*70.47*	*80.63*	*+10.16*
Problem-Solution	*Immediate*	*75.34*	*83.92*	*+8.58*
	Delayed	*70.89*	*81.56*	*+10.67*

Note. M = mean total score (max = 100). SDs for argumentative: immediate D1 = 9.61, D2 = 10.40; delayed D1 = 12.83, D2 = 10.00.

As shown in Table 4, the delayed feedback group descriptively demonstrated larger gain scores in both genres (Argumentative: Δ = +10.16 vs. +4.79; Problem-Solution: Δ = +10.67 vs. +8.58).

Table 3 *Two-Way Repeated-Measures ANOVA Summary for Overall Writing Performance*

Genre	Effect	*df*	*F*	*p*	ηp²
Argumentative (N = 49)	Draft	1, 47	26.46	< .001	.36
	Timing	1, 47	8.91	.004	.16
	Timing × Draft	1, 47	2.68	.108	.05
Problem-Solution (N = 49)	Draft	1, 47	31.86	< .001	.40
	Timing	1, 47	0.36	.550	.01
	Timing × Draft	1, 47	1.56	.218	.03

To address RQ1, two-way repeated-measures ANOVAs (Timing × Draft) were conducted separately for each genre, with feedback timing (immediate vs. delayed) and draft (first vs. second) entered as within-subjects factors and class entered as a between-subjects variable to verify that the counterbalanced

order produced no systematic effects. Because both within-subjects factors had only two levels, sphericity was satisfied by definition.

For argumentative essays (N = 49), a significant main effect of Draft confirmed that scores improved from the first to the second draft, $F(1, 47) = 26.46$, $p < .001$, $\eta^2p = .36$, a large effect. A significant main effect of Timing also emerged, $F(1, 47) = 8.91$, $p = .004$, $\eta^2p = .16$. This difference, however, reflected the lower first-draft baseline of the delayed condition (M = 70.47, SD = 12.83) relative to the immediate condition (M = 77.90, SD = 9.61) rather than a feedback-quality advantage, as the two conditions converged by the second draft (delayed M = 80.63; immediate M = 82.69). Critically, the Timing × Draft interaction was not significant, $F(1, 47) = 2.68$, $p = .108$, $\eta^2p = .05$. Although the delayed condition showed a descriptively larger gain (Δ = +10.16) than the immediate condition (Δ = +4.79), this difference in improvement rate did not reach statistical significance.

For problem-solution essays (N = 49), the pattern was comparable. A significant main effect of Draft again indicated improvement across drafts, $F(1, 47) = 31.86$, $p < .001$, $\eta^2p = .40$. Neither the main effect of Timing, $F(1, 47) = 0.36$, $p = .550$, $\eta^2p = .01$, nor the Timing × Draft interaction, $F(1, 47) = 1.56$, $p = .218$, $\eta^2p = .03$, was significant, indicating that the two feedback conditions produced equivalent improvement in this genre.

With respect to the counterbalancing check, no Timing × Class or three-way Timing × Draft × Class interactions approached significance in either genre (ps > .40), confirming that the order in which classes received the conditions did not bias the timing comparisons. The single exception was a significant Draft × Class interaction for problem-solution essays, $F(1, 47) = 4.75$, $p = .034$, $\eta^2p = .09$, indicating that the two classes improved at somewhat different rates across drafts in this genre. Because this effect was independent of feedback timing (three-way interaction, $p = .859$), it does not qualify the timing findings.

Taken together, these results indicate that both immediate and delayed Gemini-generated feedback produced significant and substantial gains in overall writing quality across both genres, but that the timing of feedback did not significantly moderate the magnitude of improvement.

4.3 Effects on Writing Sub-Dimensions (RQ2)

To address RQ2, separate two-way repeated-measures ANOVAs (Timing × Draft) were conducted for each of the five sub-dimensions of Hughes' (2003) rubric within each genre, with class entered as a between-subjects control. Full results appear in Tables 5 and 6. Descriptive statistics of the sub-dimensions are presented in Figure 2.

4.3.1 Argumentative Essays Sub Dimension Results

Table 4 *Descriptive Statistics for Sub-Dimensions: Argumentative (N = 49)*

Sub-dimension (max)	Immediate1stDraft	Immediate 2nd Draft	Delayed 1st Draft	Delayed 2nd Draft
Content (40)	30.71	32.73	27.71	31.57
Organization (20)	15.47	16.45	14.16	15.98
Vocabulary (20)	14.80	15.73	13.35	15.71
Accuracy (15)	12.33	13.04	11.12	12.76
Mechanics (5)	4.59	4.73	4.04	4.61

Note. Values are means (M).

Table 5 *Two-Way Repeated-Measures ANOVA Summary for Argumentative Essays (N = 49)*

Sub-dimension	Effect	*df*	*F*	*p*	ηp²
Content	Draft	1, 47	10.31	.002	.18
	Timing	1, 47	22.89	< .001	.33
	Timing × Draft	1, 47	1.50	.228	.03
Organization	Draft	1, 47	8.90	.005	.16
	Timing	1, 47	25.96	< .001	.36
	Timing × Draft	1, 47	1.77	.189	.04
Vocabulary	Draft	1, 47	4.33	.043	.08
	Timing	1, 47	20.82	< .001	.31
	Timing × Draft	1, 47	3.61	.064	.07
Accuracy	Draft	1, 47	8.19	.006	.15
	Timing	1, 47	26.85	< .001	.36
	Timing × Draft	1, 47	4.66	.036	.09
Mechanics	Draft	1, 47	7.93	.007	.14
	Timing	1, 47	15.65	< .001	.25
	Timing × Draft	1, 47	5.31	.026	.10

Note. Significant Timing × Draft interactions emerged for Accuracy (p = .036) and Mechanics (p = .026), both favouring the delayed condition; Vocabulary showed a non-significant trend (p = .064). Content and Organization showed no interaction.

For argumentative essays (N = 49), all five sub-dimensions showed significant main effects of Draft, confirming improvement across drafts ($p \leq .043$). The Timing × Draft interaction, however, diverged between macro- and micro-level features. Neither Content ($F(1, 47) = 1.50$, $p = .228$, $\eta^2p = .03$) nor Organization ($F(1, 47) = 1.77$, $p = .189$, $\eta^2p = .04$) showed a significant interaction, indicating equivalent rates of improvement in idea development and structure under both conditions. For the micro-level features, the Accuracy interaction was significant, $F(1, 47) = 4.66$, $p = .036$, $\eta^2p = .09$, as was the Mechanics interaction, $F(1, 47) = 5.31$, $p = .026$, $\eta^2p = .10$; in both, the delayed condition improved more steeply than the immediate condition (Accuracy: delayed $\Delta = +1.63$ vs. immediate $\Delta = +0.71$; Mechanics: delayed $\Delta = +0.57$ vs. immediate $\Delta = +0.14$). The Vocabulary interaction pointed in the same direction but did not reach significance, $F(1, 47) = 3.61$, $p = .064$, $\eta^2p = .07$ (delayed $\Delta = +2.37$ vs. immediate $\Delta = +0.94$), and is best read as a non-significant trend. Thus, for argumentative writing, delayed feedback conferred a significant advantage only for the most local, form-focused features (accuracy and mechanics), with vocabulary trending similarly, while macro-level features improved equally under both timing conditions

4.3.2. Problem-Solution Essays Sub-Dimension Results

Table 6 *Descriptive Statistics for Sub-Dimensions: Problem-Solution (immediate N = 50, delayed N = 54)*

Sub-dimension (max)	Immediate 1st Draft	Immediate 2nd Draft	Delayed 1st Draft	Delayed 2nd Draft
Content (40)	29.78	33.32	28.96	33.24
Organization (20)	14.92	16.68	14.39	16.13
Vocabulary (20)	14.24	15.98	13.06	15.24
Accuracy (15)	11.90	13.14	10.72	12.67
Mechanics (5)	4.50	4.80	3.76	4.28

Note. Values are means (M).

Table 7 *Two-Way Repeated-Measures ANOVA Summary for Problem-Solution Essays (N = 50)*

Sub-dimension	Effect	*df*	*F*	*p*	ηp²
Content	Draft	1, 48	29.95	< .001	.38
	Timing	1, 48	0.47	.498	.01
	Timing × Draft	1, 48	0.01	.925	< .01
Organization	Draft	1, 48	29.46	< .001	.38
	Timing	1, 48	3.49	.068	.07
	Timing × Draft	1, 48	0.32	.574	.01
Vocabulary	Draft	1, 48	28.53	< .001	.37
	Timing	1, 48	8.70	.005	.15
	Timing × Draft	1, 48	0.04	.839	< .01
Accuracy	Draft	1, 48	43.73	< .001	.48
	Timing	1, 48	12.22	.001	.20
	Timing × Draft	1, 48	1.07	.306	.02
Mechanics	Draft	1, 48	12.29	.001	.20
	Timing	1, 48	49.89	< .001	.51
	Timing × Draft	1, 48	0.79	.379	.02

For problem-solution essays (*N* = 50), all five sub-dimensions again showed significant main effects of Draft (ps ≤ .001) but, not sub-dimension produced a significant Timing × Draft interaction (Content *p* = .925; Vocabulary *p* = .839; Organization *p* = .574; Accuracy *p* = .306; Mechanics *p* = .379). Immediate and delayed feedback therefore produced equivalent rates of improvement across every sub-dimension in this genre, paralleling the overall-score finding under RQ1. For transparency, significant Draft × Class interactions emerged for several problem-solution sub-dimensions (Vocabulary, Organization, Accuracy; *p*s ≤ .014) and a Timing × Draft × Class interaction for Accuracy (*p* = .047); because these involve class rather than the timing manipulation, they qualify neither the timing conclusions nor the overall pattern.

4.4 Learner Perceptions (RQ3)

A thematic analysis of the 48 qualitative reflection papers revealed five distinct themes regarding the temporal dynamics of AI-generated feedback in L2 writing. These themes highlight the complex interplay between feedback timing, cognitive load, Gemini as a feedback provider, reflective depth and the perceived role of AI as an educational scaffold.

4.4.1 The Cognitive Load and Emotional Engagement

The timing of AI feedback significantly influenced participants' mental effort and motivation reflecting principles of Cognitive Load Theory. Immediate feedback was frequently described as highly engaging because it accelerates the writer's momentum, though it was also heavily demanding. Participants appreciated the ability to fix mistakes while their thoughts were still active. However, processing this feedback immediately after drafting led to cognitive overload for many participants. The 18th participant described that "I felt like I was doing a double workout without any rest in between. I just wanted to finish the writing task, so seeing a lot of feedback immediately felt like a heavy weight on my shoulders". Another quote supports the load of immediate AI feedback; 22nd participant stated "It felt like finishing on a long run and then immediately being told to do another sprint"

Conversely, delayed feedback significantly reduced cognitive load and emotional pressure, though it sometimes caused a sense of disconnect. Participants reported approaching their revisions with a rested and more objective mindset. As noted by the 2nd participant "When the feedback was delivered one week later, I felt more mentally prepared and less pressured while revising. The time gap allowed me to rest and return to my writing with a clearer perspective." . This temporal distance allowed writers to evaluate their own work in a week. However, the delay also forced students to expend extra mental effort just to re-familiarize themselves with their own writing, 42nd participant noted that "I had to spend extra time and effort just to remember my original ideas and why I structured the essay that way"

4.4.2 Error Noticing and Depth of Reflection

Drawing on Schmidt's (1990) Noticing Hypothesis, the data revealed that timing fundamentally altered how students recognized and processed their linguistic errors. Under the immediate feedback condition, participants mentioned they could easily recognize and understand their errors because their contextual memory was fresh. The 14th participant mentioned "Having just written the text definitely helped me notice those errors more clearly because my ideas, sentence structures and vocabulary choices were still very fresh in my memory." However, due to cognitive fatigue, students frequently accepted these corrections blindly to finish the task quickly. The 17th participant mentioned that "I was sometimes too mentally tired to question the corrections and simply accepted them to complete the revisions quickly."

In contrast, delayed feedback induced a 'surprise' effect that fostered deeper reflection. Participants were often shocked by the careless mistakes they had made, having completely forgotten them over the week. The 17th participant reflected "I was often surprised by some of the errors Gemini pointed out. After a week, I had forgotten many of the specific choices I made, so the errors felt more obvious and sometimes embarrassing." and another quote by 19th participant supported the immediate feedback "Immediate feedback was very fresh in my mind. Delayed feedback felt a bit like reading a text written by a stranger." This distance allowed some to evaluate their essays critically, leading to a slowed, more analytical reflective process rather than a rushed correction process and the others felt distanced and estranged from their texts.

4.4.3 Differential Impacts on Writing Quality

The timing of AI feedback distinctly impacted different sub-dimensions of writing quality, specifically dividing surface-level mechanics from macro-level content. Immediate feedback was predominantly favored for improving grammatical accuracy and lexical precision. As the intended meaning was fresh, participants could instantly contextualize and adopt the corrected syntax. The 13th participant stated “I think my grammatical accuracy improved more with immediate feedback. Correcting mistakes right after writing helped me remember grammar rules better.” Another participant noted that immediate feedback was excellent for vocabulary because they could instantly apply advanced academic synonyms while their original logic was clear.

Alternatively, delayed feedback was identified as superior for macro-level restructuring, coherence and content development. The fresh perspective provided by the delayed feedback made students more willing to reorganize paragraphs and identify weak arguments. The 17th participant highlighted that by stating “delayed feedback pushed me to make more meaningful structural changes… I was more willing to undertake major restructuring because I was not attached to my original organization.” 3rd Participant supported that “for content and arguments, delayed feedback was way more helpful because analyzing ideas needs a fresh brain. After a week I realized that some of my arguments were weak.”

4.4.4 AI as a Scaffold vs The Need for Human Nuance

In the context of Vygotsky’s (1978) Zone of Proximal Development, participants’ perceptions of AI as a collaborative partner varied by timing. Immediate feedback often felt like an interactive partnership or a ‘live dialogue’ guiding participants step-by-step, whereas delayed feedback was frequently perceived as a ‘static grading report’ or a ‘mechanical correction tool.’

4.4.5. Gemini as a Feedback Provider: AI vs Human Interaction

Regardless of the timing condition, a major theme emerged regarding the limitations of AI feedback. While Gemini was highly trusted for technical language improvements. Participants consistently highlighted that AI lacks human empathy, contextual awareness and the ability to grasp deep communicative intent. The 14th participant summarized this limitation clearly by “ the biggest limitation of AI generated feedback compared to teacher feedback is the lack of genuine human understanding and emotional intelligence.” Furthermore, the AI frequently misunderstood the personal voice or irony of the writers, offering overly generic suggestions for content. As the 3rd participant emphasized, “AI doesn’t have progress… Gemini just sees the text as a code and cannot give real motivation or support like a teacher”

5. Discussion and Conclusion

5.1 RQ 1: Does the timing of Gemini-generated feedback (immediate vs. delayed) significantly affect L2 learners’ overall writing quality over time?

The first research question asked whether the timing of Gemini-generated feedback significantly affects L2 learners' overall writing quality over time. Both the immediate and delayed conditions produced statistically significant improvements from first to second drafts in both the argumentative and problem-solution genres, with large effect sizes for the draft factor. The timing of feedback, however, did not significantly moderate the rate of improvement: the Timing × Draft interaction was non-significant in both genres.

This finding aligns closely with the systematic review by Fu and Li (2022), which concluded that both immediate and delayed feedback consistently outperform no-feedback conditions, and that the presence of corrective feedback is more consequential for L2 development than its precise timing. It is likewise consistent with Michaud et al. (2025), who reported that both immediate and delayed written corrective feedback can enhance accuracy. The present results extend this conclusion to the GenAI context: when feedback is delivered by an AI tool such as Gemini, its corrective value for overall writing quality appears robust to variation in timing.

Although the delayed condition showed descriptively larger gains in argumentative writing (Δ = +10.16 vs. +4.79), this difference did not reach statistical significance (p = .108), the interaction effect size was modest (η^2p = .05), and observed power was low (.36). It is therefore possible that the study was underpowered to detect a small but genuine timing effect, and future research with larger samples may be better positioned to establish whether the descriptive delayed advantage reflects a reliable difference. On the present evidence, however, the data do not support the conclusion that delayed feedback produces greater overall gains than immediate feedback. As the sub-dimensional analyses reported below show, any differential effect of timing is better sought at the level of specific writing components than in overall scores.

5.2 RQ 2: Are there differential effects of immediate versus delayed AI feedback on specific sub-dimensions of writing quality, including grammatical accuracy, vocabulary use, coherence, and content development?

The sub-dimensional analyses revealed a theoretically meaningful divergence within the argumentative genre and a uniform pattern within the problem-solution genre. For argumentative essays, the timing of feedback significantly moderated improvement only for the most local, form-focused features: accuracy and mechanics both showed significant Timing × Draft interactions favouring the delayed condition, and vocabulary trended in the same direction without reaching significance. By contrast, the macro-level features (content and organization) improved equally under both conditions.

This pattern is consistent with Schmidt's (1990) Noticing Hypothesis. When learners returned to their argumentative texts after a one-week interval, surface-level errors that had escaped attention during drafting appear to have become more salient, supporting the deeper noticing of local form that immediate revision—conducted while contextual memory is fresh but cognitive resources are depleted—does not. This interpretation accords with ElEbyrary et al. (2024), who found feedback timing to significantly moderate learners' noticing of errors in L2 writing. Cognitive Load Theory offers a

complementary account: immediate feedback compounds the intrinsic load of drafting with feedback-processing demands, leaving fewer germane resources for the form-focused attention that micro-level correction requires (Sweller et al., 1998; Woo et al., 2024). That macro-level features improved equally under both conditions is itself informative: higher-order revision of content and structure appears to be sufficiently goal-directed to proceed productively even under the higher load of immediate feedback, consistent with the scaffolding function described within Vygotsky's (1978) Sociocultural Theory (Fithriani, 2019; Tiandem-Adamou, 2024).

The complete absence of significant interactions across all five problem-solution sub-dimensions stands in contrast to the argumentative pattern. This divergence is most plausibly explained by the higher intrinsic cognitive demands of problem-solution writing, which requires learners to define a problem and analyses its causes and consequences. The elevated baseline load of this genre appears to neutralize the delayed-feedback advantage observed for local features in argumentative essays, so that both conditions produce equivalent gains. This extends Fu and Li's (2022) observation that timing effects are task-dependent into the GenAI feedback context, and tempers the original expectation that delayed feedback would confer a general micro-level advantage: the effect is genre-bound, emerging only where the writing task itself does not already saturate working memory.

5.3 RQ 3: How do L2 learners perceive the effectiveness and usability of immediate versus delayed Gemini-generated feedback in supporting their writing development?

The thematic analysis of learner reflection papers included five interconnected themes that support quantitative findings and add theoretically grounded depth.

The first theme, cognitive load and emotional engagement, serves interpretation of quantitative data. Participants consistently described immediate feedback as engaging, however it's cognitively and emotionally distressing. Participants explained this situation as “a heavy weight on my shoulders”, "finishing a long run and then being told to do another sprint” which is closely related to extraneous load. In contrast, delayed feedback was associated with reduced pressure, greater calm and a more analytical orientation toward revision. This situation aligns with Mao and Lee’s (2025) observation that affective interference can significantly moderate feedback processing in L2 writing context.

The second theme engages with Schmidt’s (1990) Noticing Hypothesis. Participants receiving immediate feedback acknowledged that contextual memory facilitated error recognition, but cognitive fatigue leads to uncritical acceptance of corrections rather than genuine metalinguistic engagement. However, the participants receiving delayed feedback described a 'surprise effect’ which is a shock at careless mistakes forgotten over the week. This increased error salience and induced a slower, more analytical reflective process. Delayed-noticing mechanism is in line empirical evidence by ElEbyrary et al (2024) which moderates error noticing and underscores concern raised by Koltovskaia (2020) that AWE feedback may be accepted without conscious processing.

The third theme is the differential impacts across sub-dimensions. The qualitative data showed that participants received immediate feedback as more beneficial for grammatical accuracy and vocabulary, while delayed feedback was viewed as superior for content and argumentation. This self-report contrasts with the quantitative effects in argumentative essays, which showed delayed feedback producing larger gains for micro-level features.The cause may be surface-level corrections under the immediate conditions. However, the metalinguistic processing afforded by the delayed condition was not consciously attributed to the timing variable. This aligns with Escalante et al.'s (2023) finding that learner feedback preferences are shaped by self-perceived clarity and consistency rather than actual linguistic gain.

The fourth theme situates the study's findings within Vygotsky's (1978) SCT framework. Participants experienced immediate Gemini feedback as an interactive, dialog partner resembling, real-time ZPD scaffolding described by Aljaafreh and Lantolf (1994), while delayed feedback was perceived as a static, mechanical report lacking emotional and motivational qualities. This is consistent with Tran (2025) and Tiandem-Adamou (2024) who emphasize that teacher feedback is perceived as more personalized and context-aware than AI-generated alternatives. Further, participants demonstrated that lack of emotional features of GenAI feedback often causes stress. On the other hand, with this unemotional feedback lacking judgement, some students feel more comfortable. This situation reinforces Cheng and Zhang's (2024) findings that AWE is most pedagogically effective when integrated with teacher feedback rather than deployed in isolation.

The fifth and last theme explains the Gemini's limitations as a feedback provider. Participants noted that Gemini lacked empathy as other GenAI tools, contextual awareness and the ability to recognize personal voice. It failed to address higher-order concerns by offering generic suggestions. These perceptions align with the studies conducted by Vaccino- Salvadore (2023) who explained AI-generated content's limitations in replicating the emotional, social and cultural depth of human communication. Do (2025), Azmi and Fithriani (2025) similarly reported Gemini's strengths in organizational and surface-level alongside its shortcomings in content evaluation. These findings suggest that AI-only feedback environments require pedagogical scaffolding to ensure that learners engage in the active, emotional and reflective learning environment.

In conclusion, both feedback timing conditions produced significant improvements in overall writing quality, consistent with the WCF literature (Fu & Li, 2022; Michaud et al., 2025). Timing did not, however, significantly moderate the magnitude of overall improvement in either genre. Differential effects emerged only at the sub-dimensional level and only in the argumentative genre: delayed AI feedback conferred a significant advantage for the local, form-focused features of accuracy and mechanics, with vocabulary trending in the same direction, whereas macro-level features (content and organization) and all five problem-solution sub-dimensions improved equally under both conditions. These patterns suggest that the temporal benefit of AI feedback is narrow and task-dependent rather than general, surfacing only where the writing task does not already impose a high intrinsic cognitive load. Qualitatively, learners' subjective experiences of immediate and delayed feedback were systematically

shaped by cognitive load, noticing processes, and perceptions of AI as a scaffolding agent. These findings underscore the importance of designing AI feedback interventions that promote feedback literacy and metalinguistic reflection rather than passive correction uptake (Koltovskaia, 2020; Ranalli, 2021). Pedagogically, delayed Gemini-generated feedback may be particularly useful for refining local accuracy and mechanics in argumentative tasks at the B1–B2 level, but both timing conditions appear equally effective for overall writing quality. In all contexts, AI-generated feedback should be positioned as a complement to, rather than a replacement for, teacher-mediated feedback, particularly for higher-order concerns such as argumentation and coherence (Cheng & Zhang, 2024; Tran, 2025; Wang et al., 2026).

The present study has several limitations that need to be taken into consideration. The quasi-experimental within subjects design, while appropriate for controlling individual difference variables (Dörnyei, 2007), limits causal inference and generalizability. The sample was drawn from a single institution and ELT department in Türkiye, restricting transferability to other L2 contexts and proficiency levels. Furthermore, Gemini was used as the sole AI feedback agent which may cause findings to differ with other GenAI tools or prompt configurations (Barrot, 2023; Liu et al., 2025). Further research should examine feedback timing effects across a broader range of L2 proficiency levels, institutional contexts and GenAI tools to establish the boundary conditions of the timing advantage. Additionally, longitudinal designs and mixed-methods investigations would further illuminate the noticing and schema-building mechanism through which feedback timing moderators L2 writing development. As GenAI tools increasingly affect L2 writing instruction, understanding *when* as well as *how* AI feedback is delivered emerges as a critical variable in evidence-based pedagogical design.

Abbreviations

AI: Artificial Intelligence; GenAI: Generative Artificial Intelligence; AWE: Automated Writing Evaluation; CLT: Cognitive Load Theory; L2: Second Language; WCF: Written Corrective Feedback; CEFR: Common European Framework of Reference for Languages; ELT: English Language Teaching

Declarations

Ethical approval for this study was granted by the XXX Institutional Review Ethics Committee of XXX.

Funding

This research received no specific grant from any funding agency in the public commercial, or not-for-profit sectors.

Author Contribution

Author A wrote the main manuscript text and prepared all figures, conducted the quantitative analysisAuthor B collected the data and developed the standardised feedback protocolAuthor C designed the study, developed the intervention and supervised the studyAll authors reviewed and approved the final manuscript.

Acknowledgements

The authors would like to thank the students who voluntarily participated in this study.

Data Availability

The datasets generated and/or analysed during the current study are not publicly available due to privacy and ethical restrictions but are available from the corresponding author on reasonable request.

References

1. Ahmadian, M., & Vasylets, O. (2021). The role of cognitive individual differences in L2 writing performance, written corrective feedback processing, and use. In R. M. Manchón & C. Williams (Eds.), The Routledge handbook of second language acquisition and writing (pp. 237–251). Routledge. https://doi.org/10.4324/9780429199691-17
2. Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal, 78*(4), 465–483. https://doi.org/10.2307/328585
3. Alpar, Ö. (2025). Evaluating generative AI tools for improving English writing skills: A preliminary comparison of ChatGPT-4, Google Gemini, and Microsoft Copilot. *European Journal of Educational Research, 14*(4), 1291–1308. https://doi.org/10.12973/eu-jer.14.4.1291
4. Alsofyani, A., & Barzanji, A. (2024). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. *Journal of Educational Computing Research, 63*, 431–463. https://doi.org/10.1177/07356331241307297
5. Azmi, K., & Fithriani, R. (2025). Students' perceptions of Gemini AI effectiveness in academic writing. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, 13*(1), 74–93. https://doi.org/10.24256/ideas.v13i1.6278
6. Bakri, H. (2016). The role of individual differences in second language writing corrective feedback. *Social Science Research Network*. https://doi.org/10.2139/ssrn.2843943
7. Barrot, J. S. (2023). Using ChatGPT for Second Language Writing: Pitfalls and Potentials. *Assessing Writing, 57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745
8. Cheng, X., & Zhang, L. J. (2024). Examining second language (L2) learners' engagement with AWE-teacher integrated feedback in a technology-empowered context. *The Asia-Pacific Education*

Researcher, 33, 1023–1035. https://doi.org/10.1007/s40299-024-00877-8

9. Coffin, C., Curry, M. J., Goodman, S., Hewings, A., Lillis, T. M., & Swann, J. (2003). Teaching academic writing: A toolkit for higher education. London: Routledge. http://hdl.handle.net/1802/27192.
10. Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.). SAGE.
11. Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević , D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. I*n IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100
12. Do, H. M. (2025). Potentials and pitfalls of Google Gemini in writing: Implications for educators. *Assessing Writing, 65*, Article 100955. https://doi.org/10.1016/j.asw.2025.100955
13. Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.
14. Dizon, G., & Gold, J. (2023). Exploring the effects of Grammarly on EFL students' foreign language anxiety and learner autonomy. *The JALT CALL Journal, 19*(3), 299–316. https://doi.org/10.29140/jaltcall.v19n3.1049
15. Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology, 13*, 100060. https://doi.org/10.1016/j.jrt.2023.100060
16. ElEbyary, K., Shabara, R., & Boraie, D. (2024). The differential role of AI-operated WCF in L2 students' noticing of errors and its impact on writing scores. *Language Testing in Asia, 14*(59). https://doi.org/10.1186/s40468-024-00312-1
17. 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*, 1–20. https://doi.org/10.1186/s41239-023-00425-2
18. Ferris, D.R. (2003). R*esponse To Student Writing: Implications for Second Language Students* (1st ed.). Routledge. https://doi.org/10.4324/9781410607201
19. Field, A.P. (2018) *Discovering Statistics Using IBM SPSS Statistics*. 5th Edition, Sage.
20. Fithriani, R. (2019). ZPD and the benefits of written feedback in L2 writing: Focusing on students' perceptions. The Reading Matrix: *An International Online Journal,* 19(1), 63–73.
21. Fu, M., & Li, S. (2022). The effects of immediate and delayed corrective feedback on L2 development. *Studies in Second Language Acquisition*, *44*(1), 2–34. https://doi.org/10.1017/S0272263120000388
22. Guo, T. (2023). The impact of written corrective feedback on enhancing second language writing skills: A comprehensive review. I*nternational Journal of Education and Humanities, 3*(3). https://doi.org/10.58557/(ijeh).v3i3.132
23. Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System, 126*, 103502. https://doi.org/10.1016/j.system.2024.103502

24. Han, Y., & Hyland, F. (2019). Academic emotions in written corrective feedback situations. J*ournal of English for Academic Purposes, 38*, 1–13. https://doi.org/10.1016/j.jeap.2018.12.003
25. Hassanzadeh, M., & Fotoohnejad, S. (2021). Implementing an automated feedback program for a foreign language writing course: A learner-centric study. *Journal of Computer Assisted Learning, 37*, 1494–1507. https://doi.org/10.1111/jcal.12587
26. Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.
27. Hirvela, A., & Belcher, D. D. (Eds.). (2024). *Expertise in second language writing instruction: Conceptual and empirical understandings*. Routledge.
28. Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*, 83–101. https://doi.org/10.1017/s0261444806003399
29. Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. *System. 123* https://doi.org/10.1016/j.system.2024.103332
30. Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 100450. https://doi.org/10.1016/j.asw.2020.100450
31. Kourtali, N., & Borges, L. (2023). The effects of feedback timing on L2 development in written SCMC. *Computer Assisted Language Learning, 37,* 2291–2319. https://doi.org/10.1080/09588221.2023.2171066
32. Liang, J.-C., Hwang, G.-J., Chen, M.-R. A., & Darmawansah, D. (2021). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments, 31*(7), 4270–4296. https://doi.org/10.1080/10494820.2021.1958348
33. Liu, S., Liang, C., & Yu, S. (2025). Investigating university students' intention and application of generative AI tools in L1 and L2 writing across formality: A mixed-method study. *System, 130*, Article 103831 . https://doi.org/10.1016/j.system.2025.103831
34. Lu, Y. (2023). Written corrective feedback: How can it influence adult ELLs writing performance? Journal of Education, *Humanities and Social Sciences, 8*. https://doi.org/10.54097/ehss.v8i.4367
35. Mao, Z., & Lee, I. (2025). Every advantage has its disadvantage: Side effects of teacher feedback in L2 writing. *RELC Journal.* https://doi.org/10.1177/00336882251376701
36. Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: A study on proficiency, revision frequency and writing quality. *Discover Education, 4,* Article 170. https://doi.org/10.1007/s44217-025-00602-7
37. Michaud, G., McDonough, K., & Parent, M. (2025). The impact of timing in written corrective feedback on collaborative writing and L2 accuracy development. *Canadian Journal of Applied Linguistics*. https://doi.org/10.37213/cjal.2025.34394
38. Naz, I., & Robertson, R. (2024). Exploring the feasibility and efficacy of ChatGPT3 for personalized feedback in teaching. *The Electronic Journal of e-Learning, 22*(2), 98–111. https://doi.org/10.34190/ejel.22.2.3345

39. Nguyen, H., & Chu, T. (2024). Written corrective feedback in second language writing: A review of research. *rEFLections, 31*(2). https://doi.org/10.61508/refl.v31i2.275269
40. Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38,* 1–4. https://doi.org/10.1207/s15326985ep3801_1
41. Panahi, P., Birjandi, P., & Azabdaftari, B. (2013). Toward a sociocultural approach to feedback provision in L2 writing classrooms: The alignment of dynamic assessment and teacher error feedback. *Language Testing in Asia, 3*(13). https://doi.org/10.1186/2229-0443-3-13
42. Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816
43. Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 129–158. https://doi.org/10.1093/applin/11.2.129
44. Shao, X. (2015). On written corrective feedback in L2 writing. *English Language Teaching, 8(*3), 155–168. https://doi.org/10.5539/elt.v8n3p155
45. Skinner, B. F. (1953). *Science and human behavior*. Macmillan.
46. Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers* in *Psychology, 14,* 1260843. https://doi.org/10.3389/fpsyg.2023.1260843
47. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
48. Sweller, J. (1994). Cognitive Load Theory, Learning Difficulty and Instructional Design. *Learning and Instruction, 4*, 295- 312. http://dx.doi.org/10.1016/0959-4752(94)90003-5
49. Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction, 8,* 351–362. https://doi.org/10.1207/s1532690xci0804_5
50. Sweller, J., Van Merriënboer , J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. https://doi.org/10.1023/a:1022193728205
51. Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson
52. Taşkıran, A. (2022). Çevrim içi dil öğrenimi için yapay zeka tabanlı otomatik yazma değerlendirmesi: Uzaktan öğrenenlerin algıları. *Kocaeli Üniversitesi Eğitim Dergisi, 5(*1), 111–129. http://doi.org/10.33400/kuje.1053862
53. Tiandem-Adamou, Y. M. (2024). Using generative artificial intelligence to support EFL students' writing proficiency in university in China. *Journal of Educational Technology and Innovation, 6*(4). https://doi.org/10.61414/jeti.v6i4.213
54. Tran, T. T. T. (2025). Enhancing EFL writing revision practices: The impact of AI- and teacher-generated feedback and their sequences. *Education Sciences, 15*(2), 232. https://doi.org/10.3390/educsci15020232

55. Vaccino-Salvadore, S. (2023). Exploring the ethical dimensions of using ChatGPT in language learning and beyond. *Languages, 8*(3), Article 191. https://doi.org/10.3390/languages8030191
56. Van Merriënboer, J., & Sweller, J. (2010). Cognitive load theory in health professional education: Design principles and strategies. Medical Education, *44*, 85–93. https://doi.org/10.1111/j.1365-2923.2009.03498.x
57. Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.
58. Wang, Y., Hu, W., & Guo, Q. (2026). GPT versus teacher feedback on English writing: An analysis of feedback characteristics and student engagement. *System, 138*. https://doi.org/10.1016/j.system.2025.103967
59. Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. Education and Information Technologies, 29, 24963–24990. https://doi.org/10.1007/s10639-024-12819-4
60. Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages, 8*(3), 212. https://doi.org/10.3390/languages8030212
61. Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*, 150–166. https://doi.org/10.1111/bjet.13494
62. Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 1–14. https://doi.org/10.1016/j.asw.2019.100439.
63. Zhou, L., & Li, J. (2023). The impact of ChatGPT on learning motivation: A study based on self-determination theory. *Education Science and Management, 1*(1), 19–29. https://doi.org/10.56578/esm010103

Figures

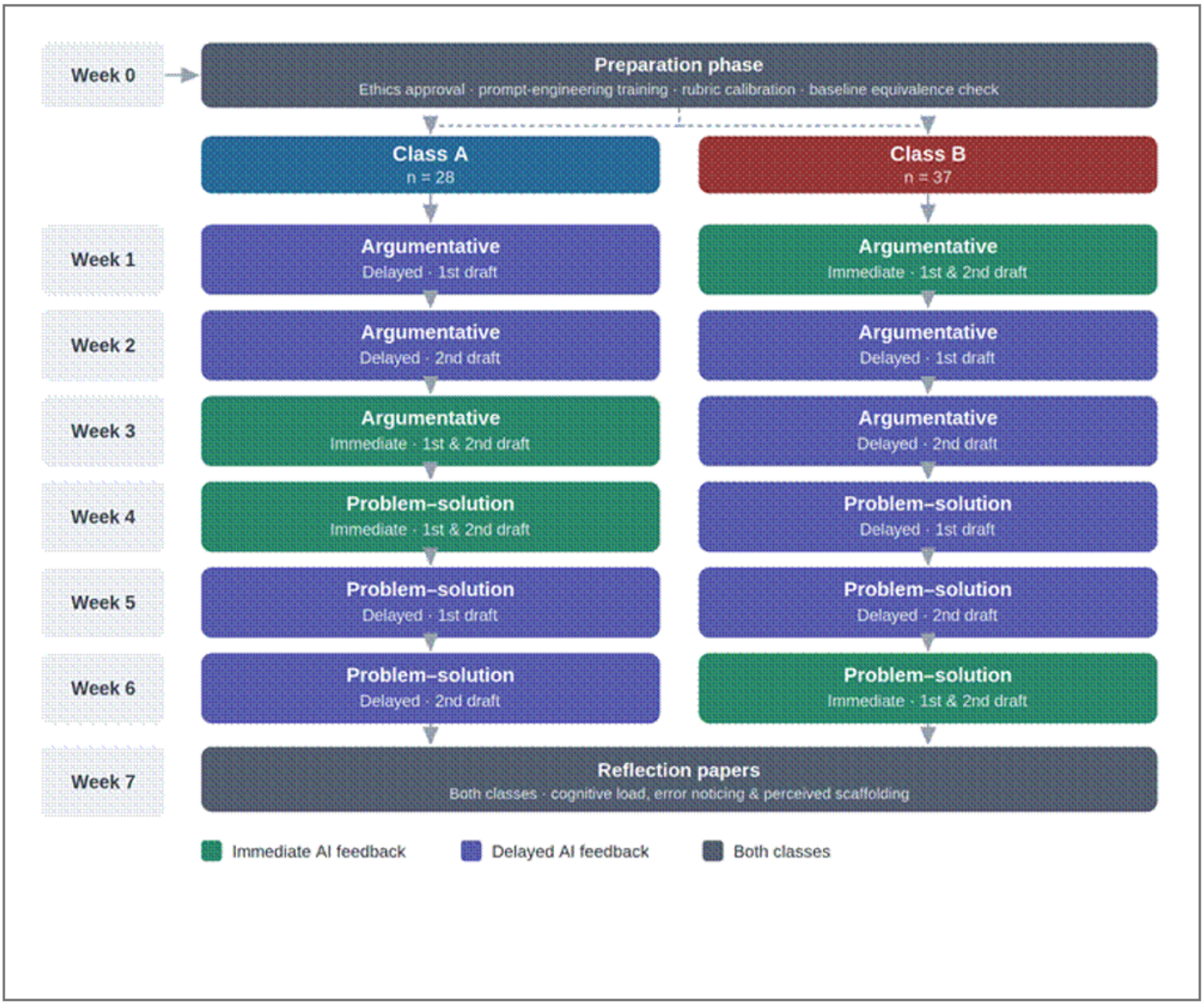


Figure 1

Comparison of Feedback Schedules for Class A and Class B

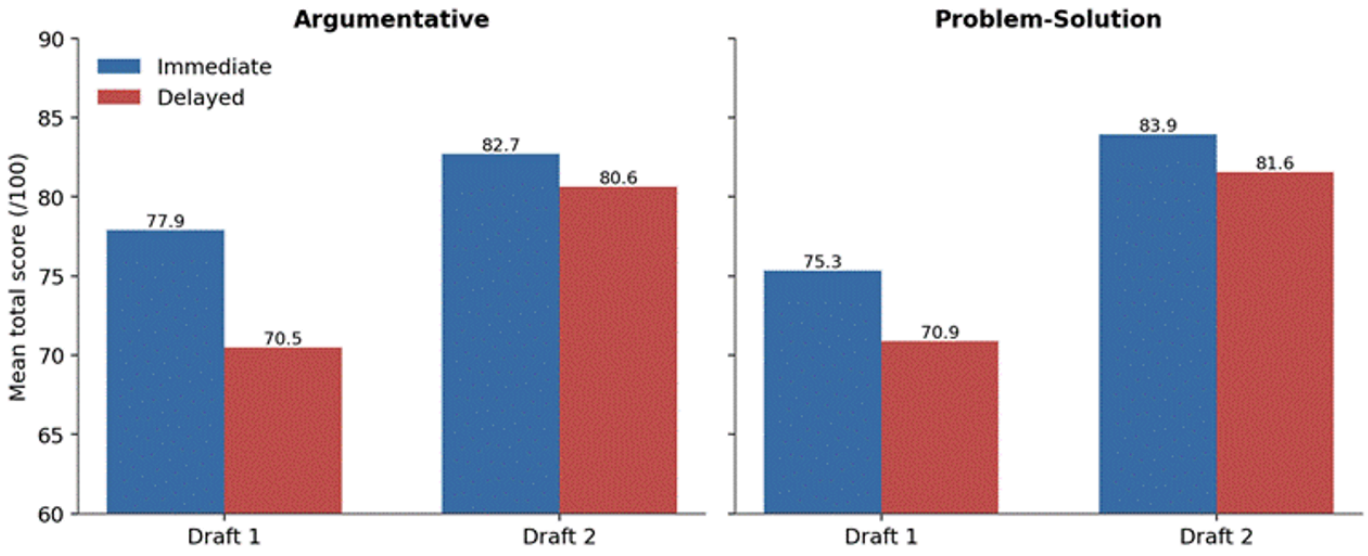


Figure 2

Mean Total Writing Scores by Genre, Feedback Timing, and Draft

Note. Total scores out of 100. Argumentative N = 49; problem-solution N = 50 (immediate) and 54 (delayed).