Castledown

OPEN ACCESS

Technology in Language Teaching & Learning
ISSN 2652-1687
https://www.castledown.com/journals/tltl/

Technology in Language Teaching & Learning, 8, 104008 (2026)
https://doi.org/10.29140/tltl.2026.104008

Embedding Constrained, Learner-Initiated GenAI Feedback in Collaborative L2 Writing: Targeted Grammar Gains with a Two-Week Follow-Up

NICOLAS EMERSON[a] (Corresponding author)
THOMAS HOLLAND[b]
TYLER D. MITCHELL[c]
DAVIS SHUM[d]
W. L. QUINT OGA-BALDWIN[e]

[a]*Waseda University, Japan;*
Kyushu Sangyo University, Japan
emerson@asagi.waseda.jp

[b]*Kyushu Sangyo University, Japan*
holland@mail.kyusan-u.ac.jp

[c]*Kyushu Sangyo University, Japan*
mitchell@mail.kyusan-u.ac.jp

[d]*Kyushu Sangyo University, Japan*
shum@mail.kyusan-u.ac.jp

[e]*Waseda University, Japan*
quint@waseda.jp

Abstract

Research on generative artificial intelligence (GenAI) in second-language writing has largely examined commercial tools such as Grammarly or general-purpose chatbots, typically in individual workflows. Evidence remains limited on purpose-built classroom systems that embed constrained, learner-initiated GenAI feedback within technology-mediated collaborative writing, align that feedback with explicitly taught forms, and assess outcomes beyond an immediate posttest. This study evaluated a pedagogical intervention using Collabowrite, a custom web application for collaborative story writing. Its core GenAI feature provides form-prioritized feedback as local edit suggestions with brief first-language explanations and requires learners to implement changes manually. Participants were 36 undergraduates

from four intact classes of a compulsory two-year English-as-a-foreign-language program at a private university in Japan that does not offer an English major. A one-group pretest–posttest design with a two-week delayed posttest assessed performance using three rotated parallel 40-item recognition-focused multiple-choice grammar tests targeting consistent tense, pronouns, conjunctions, and articles. Mixed-effects models indicated higher total grammar test scores at posttest and delayed posttest than at pretest, with no clear difference between posttest and delayed posttest. Because the study used a one-group pretest–posttest design without a comparison group, the observed gains cannot be attributed specifically to GenAI and may reflect regular instruction, test practice, or novelty effects. The study contributes classroom implementation evidence for a pedagogical configuration that combines explicit focus on form, meaning-focused collaborative writing, and learner-initiated, form-prioritized GenAI feedback, and provides a basis for controlled follow-up work.

Keywords: Generative artificial intelligence, collaborative writing, technology-mediated writing, written corrective feedback, second-language writing

Introduction

Generative artificial intelligence (GenAI) use in higher education is increasingly common and, for many students, is now integrated into study routines (Freeman, 2025). In language education, GenAI can provide immediate, linguistically responsive support during writing, potentially lowering barriers to revision and form-focused work (Donley, 2024). Within second-language (L2) writing, GenAI can also provide useful feedback on local writing features such as sentence variety, grammar, and vocabulary accuracy (Zhang et al., 2025). At the same time, these capabilities carry risks. When learners use AI to offload thinking and decision-making, they may bypass active engagement with learning and reduce critical thinking (Gerlich, 2025). Related work suggests that while GenAI can improve short-term task performance, its use may also encourage metacognitive laziness and reduce deep engagement with learning (Fan et al., 2025). As AI tools become more agentic (able to perform sequences of actions on users' behalf), the risk that students will use them to outsource academic work grows. Some ethicists argue that, under certain conditions, universities are morally justified in prohibiting student use of GenAI to protect learning and well-being (de Fine Licht, 2024).

However, a more productive stance is to treat this not as a referendum on technology but as a design problem: How can GenAI be incorporated into classroom language learning in ways that support development while maintaining the friction necessary for sustained learning? This problem is especially acute in compulsory English-as-a-foreign-language (EFL) settings, where lower-proficiency learners may need substantial teacher mediation to interpret and act on feedback from general-purpose chatbots (Chen & Chen, 2025). Accordingly, in these contexts, constrained, learner-initiated GenAI support may be a better fit than unstructured chatbot use, where constrained refers to feedback that is limited in scope and aligned to instructional targets, and learner-initiated refers to feedback that is generated only when learners request it.

This paper reports an educational intervention using Collabowrite, a custom web application for technology-mediated collaborative story writing that embeds this constrained, learner-initiated support within the writing activity rather than positioning GenAI as an all-purpose tutor. The system provides feedback that prioritizes the instructor-selected target grammatical form for that activity, presented as local edit suggestions with brief first-language (L1) explanations, and requires learners to evaluate and manually implement any changes. Collaborative writing involves learners working together to produce

a single text, with shared responsibility for content and language decisions (Storch, 2013). In this implementation, contributions were made through a turn-based workflow. Students worked in pairs or small groups, continuing a fictional story from the same starter sentence, discussing content and language choices, and taking turns to write the next part on their own devices, with GenAI feedback available on request during each writing turn. Learning is operationalized as gains on a targeted grammar assessment from pretest to posttest. Short-term follow-up is operationalized as delayed-posttest performance relative to pretest and posttest over a two-week interval.

Research on GenAI-supported classroom-based L2 writing has largely examined either commercial writing-support tools and automated writing evaluation systems, such as Grammarly (e.g., Mekheimer, 2025), or learner use of general-purpose chatbots such as ChatGPT for writing feedback (e.g., Hirschel & Horai, 2025; Strobl et al., 2024). In parallel, syntheses of technology-mediated collaborative writing suggest that tools that make interaction and revision visible can promote attention to form and facilitate peer feedback processes, although outcomes depend heavily on how collaboration among learners is structured and sustained (Li, 2018; Zhang & Zou, 2022). What remains underexamined is whether purpose-built classroom systems that embed constrained, learner-initiated GenAI feedback within technology-mediated collaborative writing can produce measurable gains on explicitly taught grammar targets using independent assessment.

Addressing this gap, the present study contributes classroom implementation evidence on such a pedagogical configuration and examines whether participation is associated with gains on explicitly taught grammar targets immediately after the intervention and at a two-week follow-up. Given the one-group design, the study is framed as an evaluation of a pedagogical model implemented in an authentic classroom setting, rather than evidence that any specific feature of the system, or variation in learners' use of those features, caused the observed gains.

Literature Review

Technology-Mediated Collaborative Writing

In L2 writing research, collaborative writing generally refers to two or more learners jointly composing a single text while negotiating content and language choices and sharing responsibility for the completed text (Storch, 2013, 2019). When collaborative writing is mediated by digital tools, the shared text and its revision history remain visible, making joint text production more traceable and enabling groups (and teachers) to revisit language decisions and monitor change over time (Kessler, 2009). This persistent record can support collaborative processing of language and feedback during writing and revision (Wigglesworth & Storch, 2012). Meta-analytic and synthesis work generally links technology-mediated collaborative writing to L2 learning benefits, while emphasizing that outcomes depend on how collaboration among learners is structured, including task design and accountability for the shared product (Elabdali, 2021; Zhang et al., 2021). More broadly, collaborative writing can prompt learners to deliberate about language choices, including grammatical accuracy (Storch, 2019), and technology-mediated environments may strengthen this deliberation by keeping drafts and revisions visible to the group.

Grammar-Related Processing Inside Collaborative Writing

Developing grammatical accuracy during classroom-based writing is cognitively demanding, given the need to allocate attention across meaning, form, and planning under time constraints (Kormos, 2023). Within interactionist accounts, grammatical development is supported when meaning-focused communication creates opportunities for focus on form and for noticing gaps between intended and

produced language (Long, 1991; Schmidt, 1990). In collaborative writing, such attention-to-form moments often take the form of brief, text-anchored discussions where learners identify a problem, propose alternatives, and decide what the group will commit to the shared text, commonly operationalized as language-related episodes (LREs; Swain & Lapkin, 1998). From a skill acquisition perspective, repeated cycles of retrieval, comparison, and adjustment during production and revision can support consolidation of more accurate form–meaning mappings over time (DeKeyser, 2020). In short, collaborative writing can create conditions for grammar-relevant processing, but the extent to which this occurs depends on task constraints and the availability of timely, usable feedback.

Planned Focus on Form and Targeted Written Corrective Feedback

Classroom writing tasks can be designed to foreground particular forms through planned focus on form, where learners are encouraged to use a focal structure and supported with feedback while composing and revising (Doughty & Williams, 1998; Ellis, 2001). Written corrective feedback (WCF) is a common mechanism in such designs. Meta-analytic evidence indicates that WCF can improve L2 grammatical accuracy, with stronger effects when feedback targets a limited set of forms and when learners are required to engage with feedback through revision (Kang & Han, 2015). This implies that narrow targeting can be a pedagogical strength, because it concentrates attention on forms learners are expected to apply repeatedly during production, rather than distributing effort across many loosely related issues.

GenAI-Mediated Grammar Support in L2 Writing

GenAI-mediated support in L2 writing has mainly been examined through learners' perceptions, draft-level revisions, and comparisons of feedback characteristics across conditions, with less attention to independent measures of learning on explicitly taught targets (Feng et al., 2025). Work on large language model (LLM)-based grammar feedback also suggests that design details matter, including whether feedback prompts learners to identify and revise errors themselves or simply provides direct corrections (Shin & Lee, 2025). Classroom-oriented studies also flag risks that may limit the learning value of AI writing support, including mistrust of AI responses, difficulty adopting suggestions among lower-proficiency learners, over-reliance, and mismatches between system output, learners' actual writing needs, and pedagogical aims (Jung et al., 2026). Taken together, this literature supports evaluating bounded configurations of GenAI support under classroom conditions using independent learning measures, rather than relying only on perceptions or draft-level revisions.

From Gaps to Research Questions

Prior work suggests that (a) technology-mediated collaborative writing can create opportunities for form-focused processing, (b) planned focus on form and targeted WCF can support grammatical development, and (c) GenAI can deliver rapid writing feedback but introduces risks when used without constraints. In this configuration, the shared text makes language choices visible to the group, while learner-initiated GenAI feedback can surface problems learners might otherwise miss, creating opportunities for discussion and revision focused on the taught forms. What remains underexamined is a classroom-oriented configuration that combines these strands: technology-mediated collaborative writing paired with explicit, target-focused instruction and constrained, learner-initiated GenAI feedback that prioritizes taught forms and avoids automated rewriting. Empirically, this leaves open whether participation in such an instructional package is associated with measurable gains on taught grammar targets on an independent test, and whether performance is similar at a short, delayed follow-up, consistent with calls for more longitudinal research in collaborative writing (Zhang & Plonsky, 2020).

Accordingly, the present study evaluated a pedagogical intervention using Collabowrite, a technology-mediated collaborative story-writing application with a learner-initiated GenAI feedback feature that prioritizes the lesson's target grammatical form, does not generate full-text rewrites, and requires learners to manually implement suggested changes.

Research Questions and Hypotheses

RQ1: Are grammar scores higher at posttest than at pretest following participation in the Collabowrite intervention?
H1: Scores will be higher at posttest than at pretest.

RQ2: Are delayed-posttest scores higher than pretest scores at a two-week follow-up?
H2: Delayed-posttest scores will be higher than pretest scores.

Supplementary exploratory analyses using system logs of feature use during Collabowrite sessions are reported to assess whether usage–outcome associations emerge.

Methods

Study Design and Overview

This study reports Cycle 3 of an ongoing design-based research project developing Collabowrite and its accompanying pedagogy for use in compulsory university EFL classes. The design was informed by earlier iterative cycles. Cycle 1 established classroom viability, with evidence that learners found the activities enjoyable and that the AI grammar feedback operated with reasonable accuracy and was often incorporated into student revisions (Emerson, 2024). A key pedagogical insight from this work was the value of a clear weekly learning focus (e.g., a specific grammar point). In Cycle 2, learning outcomes from a focused syllabus targeting direct speech, adverbs of manner, descriptive adjectives, and story structure were assessed with pretest and posttest 15-minute timed writing tasks. However, that outcome measure proved too open-ended to sensitively capture change in the specific taught forms, making gains from the focused grammar syllabus difficult to observe. The primary aim of Cycle 3 was therefore to improve alignment between the explicit grammar syllabus and the outcome measure by assessing change in targeted grammar knowledge at an immediate posttest and a two-week delayed posttest.

Cycle 3 used a one-group pretest–posttest design with a two-week delayed posttest, implemented over an eight-week period as the third iterative cycle of the project. A parallel qualitative study examined learner experiences (not reported here). The instructional ecology integrated explicit instruction, collaborative writing, and constrained, learner-initiated GenAI support within the same turn-based writing activity, so students worked collaboratively while optionally requesting GenAI feedback during their writing turns. The present paper addresses (a) change in targeted grammar test scores across three waves (pretest, posttest, delayed posttest) and (b) supplementary exploratory associations between system-recorded feature-use metrics and grammar outcomes. Testing occurred during regular class sessions in Week 1 (pretest), Week 6 (posttest), and Week 8 (delayed posttest). The primary outcome was the total grammar score (0–40). The four target grammar forms were taught during Weeks 2–5, with instructional order counterbalanced across the four intact classes (Table 1). These forms were selected by the research team because they had been identified as recurring areas of difficulty in writing produced by previous cohorts. They were treated not as wholly new content, but as forms that students had likely already encountered in prior compulsory English study before university, making them suitable targets for review and controlled practice.

Table 1. *Data Collection and Instructional Schedule*

Phase	Pretest	Instructional sequence				Posttest	Delayed posttest
Class	**Week 1**	**Week 2**	**Week 3**	**Week 4**	**Week 5**	**Week 6**	**Week 8**
A		Consistent tense	Pronouns	Conjunctions	Articles		
B		Pronouns	Articles	Consistent tense	Conjunctions		
C		Conjunctions	Consistent tense	Articles	Pronouns		
D		Articles	Conjunctions	Pronouns	Consistent tense		

Note. The sequence in which the four target grammar forms were taught was counterbalanced across the four classes during Weeks 2–5. Consistent tense focused on stable tense use across a narrative, pronouns on personal, object, and possessive forms, conjunctions on contrast and result, and articles on definite, indefinite, and zero article use. All tests were administered during the same weeks in all classes.

Context and Course Setting

This study was conducted at a private university in Japan that does not offer an English major. At this institution, all undergraduates are required to complete a two-year general English program as a graduation requirement. This program is streamed into four proficiency levels via an initial placement test. Participants in this study were placed in the highest of the four levels and were enrolled in a four-skills course integrating reading, writing, listening, and speaking. Classes met twice weekly for 90 minutes with typical sizes of 5–15 students. During the intervention period (Weeks 2–5), Collabowrite was used in one of the two weekly classes. The other used an oral communication textbook and did not target the four focal grammar forms. In all participating classes, the regular classroom teachers, rather than external researchers, implemented the application sessions and administered the tests during scheduled class time.

Participants

The sample consisted of 36 undergraduate students (18 male, 18 female) from four intact classes. Participants were predominantly first-year students (n = 35), with one second-year student. No additional demographic data (e.g., faculty, major, or age) were collected.

At semester start, 49 students were enrolled across the four classes. Two declined participation and five did not attend any Collabowrite sessions. Of the remaining 42, six missed more than one of the three test waves (pretest, posttest, delayed posttest) and were excluded under an a priori rule. The final analytic sample was therefore 36 students. All 36 completed the pretest and posttest; five were absent at the delayed posttest.

Participants had completed the TOEIC Listening and Reading test, a widely used English proficiency test, roughly three months before the study as part of an institutional assessment. The cohort mean was 490, which has been mapped to upper A2 on the CEFR in prior work (Tannenbaum & Wylie, 2019). Thus, although participants were placed in the highest of four compulsory English levels at the institution, their proficiency was still relatively low. TOEIC scores were treated as a baseline proficiency covariate in the primary analyses.

Ethics

Informed consent was obtained from all participants after they were briefed on the study's objectives and methods. Participants were assured that participation was voluntary, that participation (or non-participation) had no effect on course grades, and that they could withdraw at any time without penalty. Data were de-identified and stored on access-restricted drives. Consent forms were written in Japanese, reviewed by Japanese native speakers, and signed by participants before the project began. Ethical review was completed by the Language Education and Research Center at the institution where the study was conducted, in accordance with institutional protocols. Under this protocol, the requirement for formal ethics committee approval was waived.

Counterbalancing and Allocation

A 4 × 4 Latin square was used to counterbalance the order in which four target grammar forms (consistent tense, pronouns, conjunctions, and articles) were taught across four intact classes (Kirk, 2013). Each class was taught by its regular instructor throughout the semester. Across Weeks 2–5, all classes received all four forms, but in different sequences (Table 1), so each form appeared once in each instructional position. Because each instructor taught all four forms within their own class, each form was delivered once by each instructor, reducing confounding between instructor and target-form content.

The Latin-square schedule was generated prior to the semester and implemented administratively across classes. Assignment occurred at the class level (intact classes), with no student-level randomization. Accordingly, while counterbalancing and covariate adjustment reduce some sources of bias (notably sequencing and teacher-by-content confounds), the design remains a one-group pretest–posttest evaluation and does not support causal attribution.

Intervention Procedure

Session Structure

Each Collabowrite session followed the same four-step pedagogical sequence aligned to the week's target form. First, instructors delivered a brief mini-lesson (approximately 10 minutes) using bilingual (Japanese and English) slides with examples intended to highlight the form–meaning mapping. Second, students completed a short worksheet activity built around a brief model story adapted from well-known Japanese folk tales familiar to students. The story was printed on the worksheet, and students then completed 10 cloze-style items targeting the same form, after which answers were checked together in class, with performance not recorded. The worksheet provided additional form-focused practice, but unlike the outcome measure, it required students to produce the target forms rather than select answers in a multiple-choice format. Third, students completed a collaborative story-writing task in Collabowrite for approximately 40 minutes. Students worked in small groups, typically in pairs or groups of three, and occasionally four, with each student using their own device. Writing was turn-based within a shared group story, with one student writing at a time and a 150-character limit per turn. Fourth, during writing and revision within their turns**,** students could request GenAI feedback that suggested local edits with brief L1 explanations, prioritized the target form when relevant, and required students to evaluate and manually implement any changes. Prior to the first intervention session, students received an orientation to the application in Week 1. This included a brief introduction to simple story structure and how to access the platform and use its core features. After the writing phase, groups also generated audio and images to accompany their stories and shared selected stories in class, but these features were not treated as part of the targeted grammar outcome mechanism. Figure 1

summarizes the session-level pedagogical sequence, the hypothesized learning processes this overall classroom sequence was intended to support, and the timing of the outcome measures.

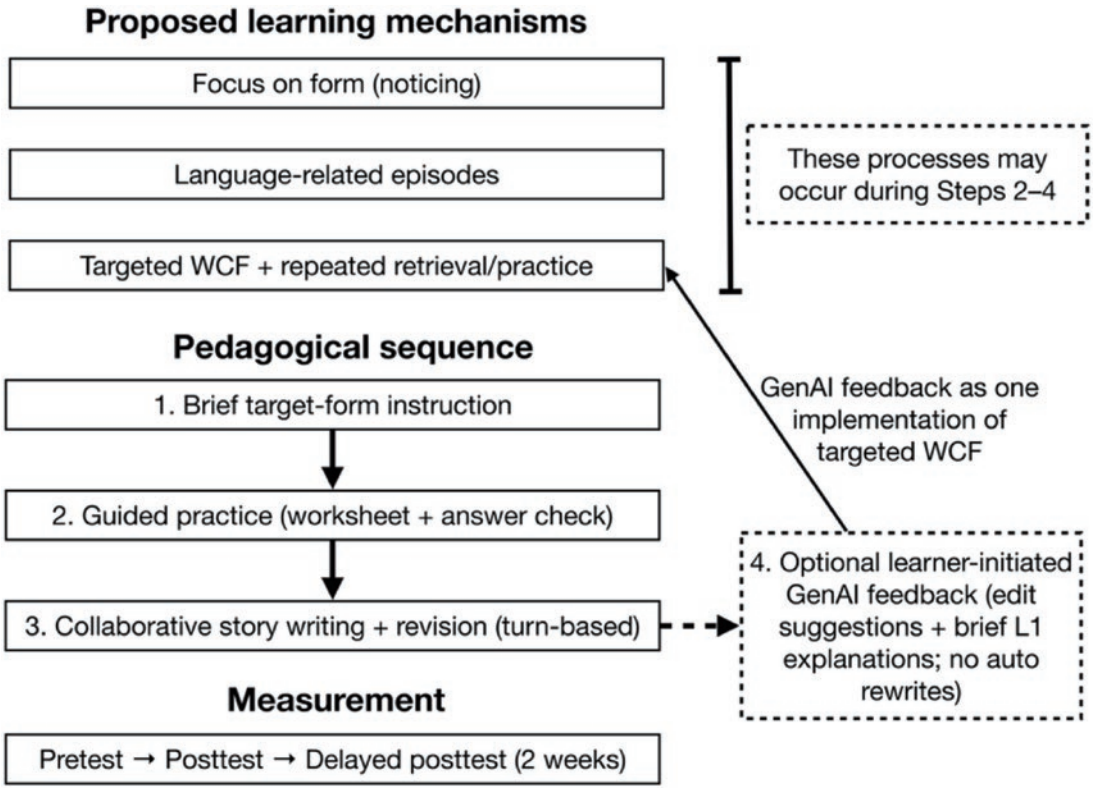

Figure 1 *Session-Level Pedagogical Sequence, Hypothesized Learning Processes, and Measurement Timeline for the Targeted Grammar Outcome*

Note. WCF = written corrective feedback.

After writing, all stories were published to the application's library. Instructors selected one or two stories that used the week's target form well and presented them using Collabowrite's presentation mode, which plays AI narration while displaying student-generated images and the story text sentence-by-sentence.

Writing Task Design and Classroom Implementation

Students worked in instructor-assigned small groups of two to four. For each session, all groups received the same starter sentence, and groups were encouraged to incorporate the session's target form during meaning-focused story production. For example, in the conjunctions-focused session, all stories began with: "Daichi and Hana loved singing. One day, they were invited to the World Karaoke Championships." Each student used their own device, and group members took turns contributing to the same shared story within the application. Turn timing was not constrained, and students could edit their own contributions during the writing phase.

Instructors supported implementation during writing by keeping a timer and a brief on-screen reminder of the week's target form visible on the projector. Instructors also monitored the shared texts and all GenAI outputs in real time via a live dashboard and prompted groups when they noticed breakdowns in comprehensibility or persistent language problems, with particular attention to the week's target form.

Fidelity

To support consistent implementation across sessions, instructors used standardized slides and worksheets for each grammar focus. All instructors attended a pre-intervention orientation covering the session sequence, recommended settings, and use of the main classroom support features. Weekly meetings were used to align delivery and resolve practical issues arising during implementation. The principal investigator also reviewed materials and instructor notes recorded in the application to confirm that the intended session sequence and core task constraints were being followed. No recorded deviations from the planned session sequence or core task constraints were identified across the implemented sessions.

Platform Design and GenAI Implementation

Learner-Initiated Grammar Feedback

Figure 2 shows the learner-initiated grammar-check feature used to support writing and revision. When learners requested feedback, the system returned local edit suggestions with brief L1 explanations to support rapid comprehension. Feedback was generated by an LLM (OpenAI's GPT-4o) via an application programming interface (API). The research team did not fine-tune the model for this study.

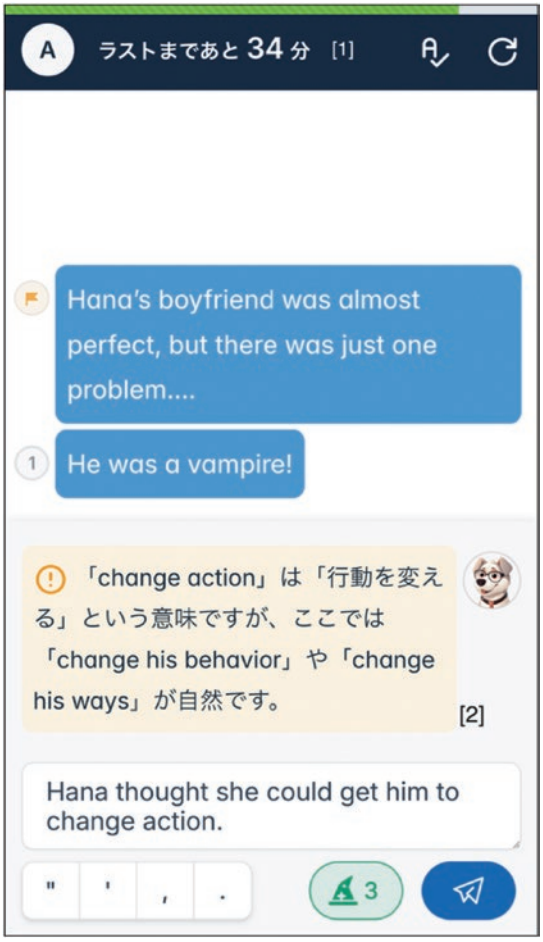

Figure 2 *Student-Facing Grammar-Check Interface Showing Story Context, Current Text, and a Local Edit Suggestion with a Brief Japanese Explanation*

Note. [1] "34 minutes remaining." [2] The Japanese feedback explains that "change his behavior" or "change his ways" would be more natural than "change action" in this context.

To constrain output, the LLM was prompted to avoid full-text rewrites, defined here as complete corrected versions of learners' submitted text, and instead provide targeted edit suggestions. Suggestions were not automatically applied in the application, and copy-and-paste was disabled in the text-entry field, so students had to implement any changes manually. Feedback was further constrained by a prioritization rule: the LLM was prompted to identify no more than three issues, prioritizing errors in the session's target grammatical form, then errors affecting meaning, and then other common errors, while preserving students' intended meaning and aiming for grammatical accuracy and natural English.

Other GenAI Features and Logged Use

The application included additional GenAI features to support story production and presentation. Although these features were not the primary focus of the design, their use was logged and examined in supplementary analyses as potential correlates of grammar outcomes. When learners entered Japanese text together with at least one English word, the grammar-check feature changed function (and appearance) to a vocabulary-lookup mode that proposed lexical options based on the current story context. This constraint was designed to keep production anchored in English while reducing reliance on external machine translation and keeping work within the application. In an optional, teacher-selected mode, an AI group member (Ducky) took a turn in sequence and proposed two candidate next turns for the group to vote on (one regeneration allowed). Each instructor enabled this mode in two of the four sessions. The feature was intended to support creativity by offering brief, context-linked continuation options, as shown in Figure 3.

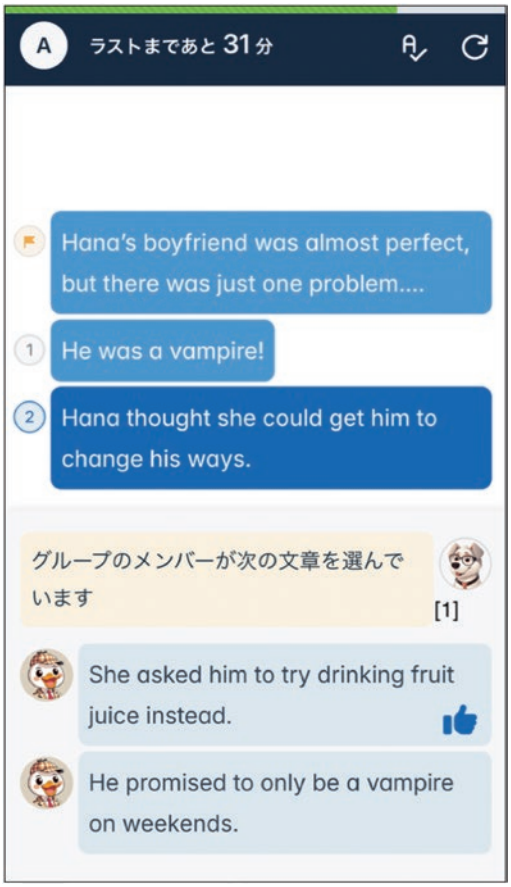

Figure 3 *Student-Facing AI Group Member Interface Showing Two Candidate Next Turns for Group Voting*

Note. [1] The Japanese text indicates that the group members are choosing the next sentence.

After writing, groups selected an AI narrator voice for audio narration, generated images from AI-generated prompts that learners could edit and voted to select a cover image (Figure 4).

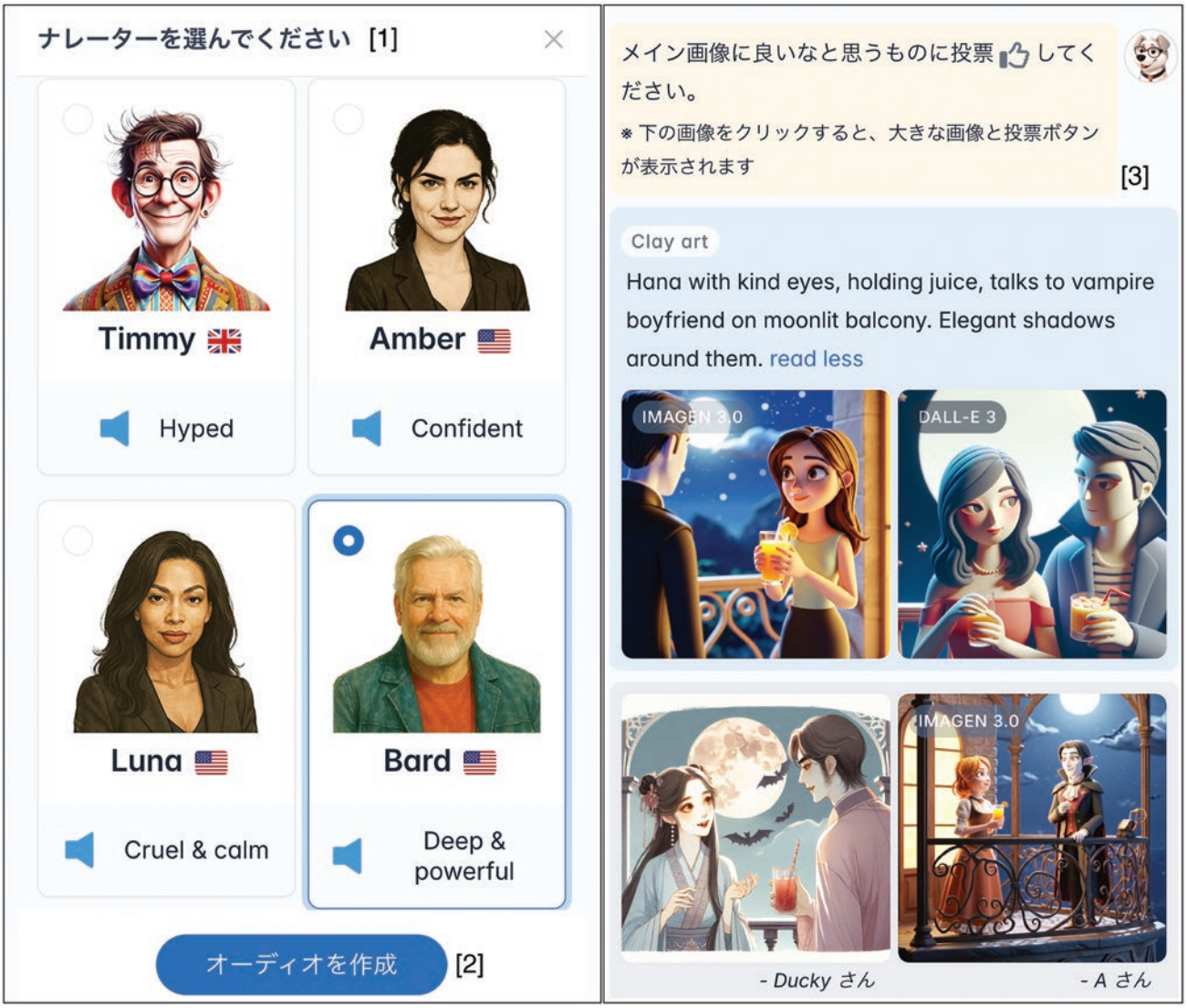

Figure 4 *Post-Writing Multimodal Features: AI Narrator Selection (Left) and Image Generation with Cover-Image Voting (Right)*

Note. [1] "Please choose a narrator." [2] "Generate audio." [3] The Japanese text instructs users to vote for the image they think would be best as the cover image and explains that clicking an image below will display a larger image and a voting button.

Safety and Monitoring

To support safe use, all text- and image-based requests were checked before submission using multiple safeguards. At the application level, requests were filtered against restricted-content rules. When a request was flagged (e.g., sexual content or excessive violence), students were prompted to rephrase and were shown a brief reason. In addition, all text-based requests were submitted to OpenAI's Moderation API prior to generation. Requests that cleared these checks were still subject to the model provider's safety systems. Instructors oversaw sessions through a live dashboard that displayed student text, AI requests, and AI outputs in real time, with flagged requests marked for review. An AI implementation summary is available on the project's Open Science Framework page (https://osf.io/cnrbm).

Measures

Grammar Test

Targeted grammar knowledge was assessed with a 40-item multiple-choice cloze test administered at pretest, posttest, and delayed posttest. Each item required students to select the most appropriate word(s) to complete an English sentence. Responses were scored dichotomously (1 = correct, 0 = incorrect) and summed to yield a total score from 0 to 40, with no partial credit. All administrations were proctored in class under identical instructions and a uniform 15-minute time limit. To reduce practice effects, item-level feedback was not provided between waves, and students did not review test forms prior to the delayed posttest.

Items were drawn from a 120-item pool authored by the research team (30 items per targeted grammar area) and assembled into three parallel 40-item forms (Forms A, B, C) designed to balance content and approximate difficulty. Forms were rotated across students and waves using a counterbalanced schedule so that approximately equal numbers of students completed each form at each time point, reducing form and order confounds. Table 2 provides sample items for each grammar form.

Table 2. *Sample Test Items by Grammar Form*

Grammar form	Sample item	Answer choices
Consistent tense	If I don't _____ lunch, I have no energy for class.	**have**, will have, had, having
Pronouns	Ken learned some new skills and was very eager to try _____.	it, they, **them**, its
Conjunctions	We wanted to go hiking, _____ it started raining.	**but**, so, and, as a result
Articles	She is _____ European citizen.	**a**, an, the, no article

Note. Each row shows one sample item per grammar form. Correct answers are in bold.

Instrument Checks

Prior to the main study, the item pool was piloted with five first-year students enrolled in a different four-skills class at the same institution. These students did not participate in the intervention study. The pilot served as a face-validity check of item clarity and approximate difficulty; no additional demographic data were collected. Based on the pilot, pronoun items were revised because they were less challenging than intended.

Using pretest data, internal consistency was estimated with Kuder–Richardson 20 (KR-20), computed separately for each 40-item test form. Practical comparability of the three forms at pretest was examined using an equivalence approach with a ±2-point margin (5% of the 0–40 scale) and 90% confidence intervals for pairwise mean differences; variance similarity across forms was also checked using Brown–Forsythe and Levene tests. To adjust for any residual form differences, test form was included as a fixed effect in all primary models. For transparency, scores were also computed for four 10-item sets keyed to the target forms and reported descriptively (Results, Table 3). Primary inference used the 40-item total score.

Log Measures

System logs recorded writing-phase activity, including words and turns produced, grammar-check events, vocabulary-lookup events, and post-turn edits. In sessions where the optional AI group-member mode was enabled, logs also recorded the number of AI-generated turns presented to each group.

For the supplementary log analyses, three predictors were derived. Grammar-check intensity was defined as grammar checks per 100 words produced. Vocabulary lookup use was coded as any lookup versus none. AI group-member exposure was defined as the number of AI-generated turns presented to the student's group and was standardized prior to analysis.

Statistical Analysis Plan

Missing Data Handling

Analyses began with descriptive summaries at each wave and inspection of the missingness pattern. Little's test for data missing completely at random (MCAR) was conducted, with results reported in the Results section (Little, 1988). Missing values were handled using multiple imputation (MI) by chained equations with predictive mean matching (PMM). Twenty imputations were generated (k = 5 donors; fixed seed = 12345). The imputation model included time (pretest, posttest, delayed posttest), test form (A, B, C), instructional order (categorical), class, and TOEIC score. TOEIC score (when missing) and the total grammar score were imputed conditional on these variables. MI estimates were combined using Rubin's rules (Rubin, 1987; van Buuren & Groothuis-Oudshoorn, 2011).

RQ1–RQ2 Primary Models

Linear mixed-effects models were used because they are well suited to repeated-measures data, allowing dependence within students to be modeled while accommodating covariates and clustering (Singer & Willett, 2003). To estimate change in targeted grammar knowledge over time, these models were fit to the total grammar score (0–40), treating the total as approximately continuous. The primary specification included fixed effects for time and test form, TOEIC score as a covariate, and instructional order as a nuisance fixed effect. Random intercepts were included for students and for classes (four classes) to reflect repeated measures nested within classes. Because there were only four classes, the class random-intercept term was included as a conservative adjustment for clustering rather than to estimate a stable between-class variance component, and results were unchanged in a sensitivity model treating class as a fixed effect. Time was coded with the pretest as the reference level, and planned contrasts estimated (a) pretest to posttest, (b) pretest to delayed posttest, and (c) posttest versus delayed posttest.

Effect Sizes and Robustness

Missing data were handled using multiple imputation with pooled estimation following standard procedures (Enders, 2010; Rubin, 1987). Standardized effects (*d*) were computed using the pooled residual standard deviation from an MI-pooled ordinary least squares (OLS) model as a pragmatic common metric across contrasts. This standardization was used for interpretive convenience and did not affect the mixed-model inference. As model-based complements to the pretest form checks, the main effect of form and the time × form interaction were also examined. Robustness checks included (a) a complete-case version of the primary mixed model and (b) an expanded mixed model including time × form. Primary MI mixed models were estimated by maximum likelihood (ML). Additional fits using restricted maximum likelihood (REML) were used for sensitivity and for reporting variance components and an intraclass correlation coefficient (ICC).

Supplementary Exploratory Log–Outcome Analyses

To explore whether individual differences in system use covaried with grammar outcomes, exploratory OLS models with robust standard errors were fit using complete-case data. Outcomes were (a) posttest score conditional on pretest score, (b) delayed posttest score conditional on pretest score, and (c) delayed posttest score regressed on posttest score (follow-up conditional-on-posttest specification). Predictors were derived from application logs (defined in the Log measures section above), including grammar-check intensity (checks per 100 words), vocabulary-lookup use (any vs none), and AI group-member exposure (standardized AI-member turns in the group). Models adjusted for the relevant prior score (pretest or posttest, depending on the outcome) and TOEIC score, and where applicable also adjusted for test form. These analyses used complete-case data because the log-derived predictors were session-aggregated and highly skewed, and imputing them under MI would require strong, hard-to-justify assumptions given the small sample. Accordingly, these analyses are treated as exploratory and interpreted descriptively (effect estimates and confidence intervals), not as confirmatory inference. All analyses were conducted in Stata 18 (MP; StataCorp, 2023).

Results

Results Overview

Results are organized around the primary research questions. The section first describes observed performance by wave (total score and item-set scores) to contextualize RQ1–RQ2, then reports missingness patterns and pretest instrument checks that support the primary modeling approach. Next, the primary mixed-effects results addressing RQ1 (pretest-to-posttest change) and RQ2 (pretest-to-delayed posttest change at a two-week follow-up) are presented, followed by supplementary exploratory log-based analyses using application logs.

Descriptive Performance by Wave and Item Set

Table 3 summarizes observed performance on four 10-item sets corresponding to the target grammar forms (consistent tense, pronouns, conjunctions, and articles) at each wave, and Table 4 summarizes observed total scores and missingness by wave. Observed total scores increased from pretest to posttest and were similar at the delayed posttest (Table 4). Table 3 is provided as descriptive context only and is not used for inferential claims. Primary inference is based on the 40-item total score. Descriptively, pronoun items showed the highest mean scores across waves, whereas article items remained lower. However, these differences were not modeled inferentially and should be interpreted cautiously.

Table 3. *Descriptive Performance by Item Set (Items Keyed to Each Target Grammar Form) by Wave (Observed Data)*

Wave	*n*	Consistent tense *M* (*SD*)	Pronouns *M* (*SD*)	Conjunctions *M* (*SD*)	Articles *M* (*SD*)
Pretest	36	6.89 (2.09)	7.47 (2.08)	5.61 (1.99)	5.83 (1.81)
Posttest	36	7.47 (1.76)	7.86 (1.71)	7.08 (1.70)	5.97 (1.87)
Delayed posttest	31	7.61 (1.48)	8.06 (1.63)	7.45 (1.57)	5.61 (1.63)

Note. Each item set comprises 10 items keyed to the target grammar form. Scores are reported descriptively. *M* = mean, *SD* = standard deviation.

Table 4. *Descriptive Statistics and Missingness by Wave (Observed Data)*

Wave	*n*	Missing (*n*)	Missing (%)	*M*	*SD*	*SE*	95% CI	Min	Max
Pretest	36	0	0	25.81	6.40	1.07	[23.71, 27.90]	9	36
Posttest	36	0	0	28.39	4.61	0.77	[26.88, 29.89]	20	39
Delayed posttest	31	5	13.9	28.74	4.68	0.84	[27.10, 30.39]	17	36

Note. Total sample $N = 36$. Percentage missing is relative to N. Means and confidence intervals are based on observed data. M = mean, SD = standard deviation, SE = standard error, CI = confidence interval.

Missing Data and Attrition

Data were complete at pretest and posttest (0% missing at each wave, $N = 36$). At the delayed posttest, five participants were absent (13.9% missing; observed $n = 31$). Little's MCAR test was not statistically significant, $\chi^2(5) = 8.42$, $p = .134$; however, with this sample size, the test has limited power, and attrition was associated with baseline scores, suggesting MCAR is unlikely. The TOEIC covariate was missing for one participant and was imputed as part of the multiple-imputation procedure.

Delayed-posttest attrition was related to baseline performance: participants who missed the delayed posttest had lower pretest scores ($n = 5$, $M = 20.20$, $SD = 6.65$) than those who completed it ($n = 31$, $M = 26.71$, $SD = 5.99$), $t(34) = 2.22$, $p = .033$. Primary analyses therefore used MI under a missing at random (MAR) assumption, conditional on observed covariates. Given this baseline-related missingness, MI was used to reduce complete-case bias, and complete-case sensitivity checks showed the same pattern of results. Additional plots of the missingness pattern, including delayed-posttest missingness by test form and instructional order, are provided in Figure A1 in the Appendix.

Pretest Instrument Checks

Using pretest data ($N = 36$), internal consistency for the dichotomously scored items was estimated with KR-20. Reliability was high for all three forms (Form A = .85, Form B = .94, Form C = .85; Appendix Table A1); within-form zero-variance items were excluded from the relevant KR-20 computation. For transparency, KR-20 was also estimated for each 10-item set keyed to the four grammar areas within each test form. Given the short subtests and small per-form samples ($n = 10$–14), these estimates were variable and generally lower than the full-test KR-20 (range = .56–.84), with the article sets clustering between .60–.65. Accordingly, the 10-item sets were not treated as separable subscales and are reported descriptively only. A supplementary visualization of pretest reliability by form is provided in Figure A2 in the Appendix.

Pretest form comparability was examined descriptively. Mean pretest scores did not differ statistically across forms (one-way ANOVA, $F(2, 33) = 0.54$, $p = .589$), and variance-equality tests did not indicate heterogeneity (Levene and Brown–Forsythe; Appendix Table A2). As a practical check, 90% confidence intervals for pairwise pretest mean differences were compared to a ±2-point margin on the 0–40 scale; with small per-form samples, the intervals were wide and did not fall entirely within the equivalence margin (Figure 5). Accordingly, strict equivalence was not established, so test form was included as a fixed effect in all primary models to adjust for any residual form differences. Figure A5 in the Appendix shows the same equivalence check under narrower and wider margins for reference.

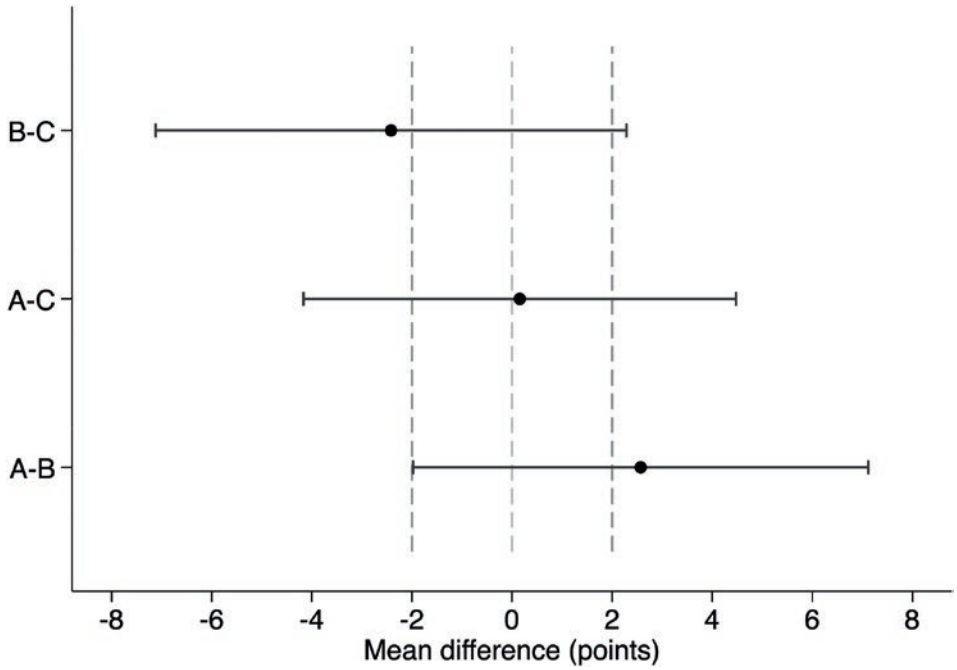

Figure 5 *Pretest Form Mean Differences with 90% Confidence Intervals and ±2-Point Equivalence Margin*

Note. Points are mean differences between forms; horizontal bars show 90% confidence intervals; vertical dashed lines indicate ±2 points.

Observed distributions by time are shown for context in Figure A3 in the Appendix.

RQ1: Are Grammar Scores Higher at Posttest Than at Pretest?

RQ1 asked whether grammar scores were higher at posttest than at pretest. In the primary MI mixed-effects model, adjusting for test form, TOEIC pretest, and instructional order, with random intercepts for student and class, posttest scores were higher than pretest scores by 2.80 points (95% CI [1.12, 4.48], $p = .001$; Table 5). On the 0–40 scale, this corresponds to about 3 additional items correct on average. The standardized effect was $d = 0.56$ (Appendix Table A3), which is considered a small effect in L2 research (Plonsky & Oswald, 2014). Figure 6 shows the same pretest-to-posttest increase, consistent with H1.

Table 5. *Planned Time Contrasts from the Multiple-Imputation Mixed-Effects Model*

Comparison	Difference (points)	*SE*	95% CI	*t*	*p*	Effect size (*d*)
Pretest to posttest	2.80	0.86	[1.12, 4.48]	3.26	.001	0.56
Pretest to delayed posttest	3.10	0.90	[1.35, 4.86]	3.46	.001	0.62
Posttest versus delayed posttest	0.31	0.90	[−1.46, 2.07]	0.34	.735	—

Note. Estimates are from a mixed-effects model (ML) fit across 20 imputed datasets and pooled using Rubin's rules; test statistics are MI-pooled *t* values. The model included fixed effects (time, form, TOEIC pretest, instructional order) and random intercepts (student, class). Effect sizes (*d*) are based on the pooled residual standard deviation from an equivalent OLS model. CI = confidence interval; ML = maximum likelihood; OLS = ordinary least squares; *SE* = standard error.

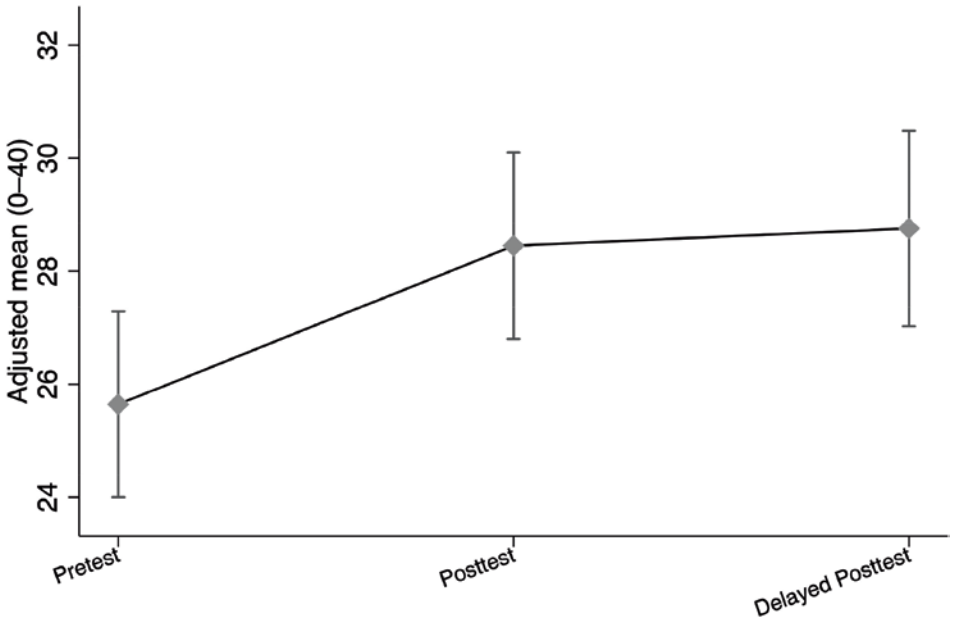

Figure 6 *Adjusted Means Across Time*

Note. Adjusted means were calculated from an ordinary least squares regression model for visualization, and conclusions match the primary multiple-imputation mixed-effects contrasts. Estimates were combined across 20 imputed datasets using Rubin's rules. The model included time, test form, TOEIC pretest score, and instructional order. Error bars represent 95% confidence intervals.

RQ2: Are Delayed Posttest Scores Higher Than Pretest Scores at the Two-Week Follow-Up?

RQ2 asked whether delayed posttest scores were higher than pretest scores at the two-week follow-up. In the primary MI mixed-effects model, delayed posttest scores were higher than pretest scores by 3.10 points (95% CI [1.35, 4.86], $p = .001$; Table 5). Delayed posttest scores were similar to posttest scores: the difference between posttest and delayed posttest was 0.31 points and was not statistically reliable (95% CI [−1.46, 2.07], $p = .735$; Table 5). The confidence interval for the difference between posttest and delayed posttest includes both a small decrease and a small increase over the two-week interval.

Because delayed-posttest nonresponse was associated with lower baseline scores (see the "Missing Data and Attrition" section), the delayed-wave estimates should be interpreted cautiously. A complete-case sensitivity check ($n = 31$ with all three waves) showed the same pattern: delayed posttest scores remained higher than pretest scores (+2.38 points, $p = .006$), and delayed posttest scores did not differ from posttest scores (−0.12 points, $p = .891$; Appendix Table A6).

Sensitivity Analyses

The primary conclusions were unchanged in a complete-case reanalysis limited to students with data at all three waves ($n = 31$; Appendix Table A6). In this sensitivity check, posttest scores were higher than pretest scores (difference = 2.50 points, $p = .004$), delayed posttest scores were higher than pretest scores (difference = 2.38 points, $p = .006$), and posttest and delayed posttest scores did not differ (difference = −0.12 points, $p = .891$). The time × form interaction was not statistically significant (Appendix Table A4), indicating no evidence that estimated change over time differed by test form. Adjusted means by time within test forms are shown in Figure A4 in the Appendix. A REML sensitivity fit indicated substantial between-student variability (student-level ICC = .49; Appendix Table A5), supporting the use of random intercepts for students, and the class-level variance component was estimated close to zero.

Supplementary Analyses: Application Logs and Grammar Outcomes

Log-Based Usage Descriptives

Supplementary analyses examined whether variation in learners' use of Collabowrite was associated with grammar outcomes. Attendance was generally high, with students attending a median of 4 sessions (interquartile range [IQR] = 3–4). Across sessions, students produced a median of 69.6 words per student per session (IQR = 58.8–75.6) over 5.0 turns (IQR = 4.4–6.1). Use of learner-initiated support was selective: students initiated a median of 7.0 grammar checks per student per session (IQR = 2.4–9.4), whereas vocabulary lookups were uncommon overall (median = 0.0 per student per session, IQR = 0.0–0.9). Post-turn edits were also observed (median = 0.8 edits per student per session, IQR = 0.3–1.1). These variables were treated as indicators of exposure and behavioral engagement within the intervention, not as direct measures of grammar learning. Together, these descriptives provide context for the exploratory log–outcome analyses (Table S1 in the AI implementation summary).

Exploratory Log–Outcome Models

Because these models use complete-case data and exploratory predictors derived from skewed logs, estimates are reported as descriptive associations and should not be over-interpreted on the basis of *p*-values. Log–outcome models tested associations between usage metrics and grammar outcomes after adjusting for baseline score, TOEIC pretest, and test form (robust SE; complete-case; Table S2 in the AI implementation summary), including a follow-up specification (delayed posttest conditional on posttest). Grammar-check intensity (checks per 100 words produced) showed no clear association with posttest, delayed posttest, or follow-up outcomes across model blocks. AI group-member exposure (standardized Ducky turns in the group) likewise showed no consistent association with outcomes. Lookup use (any lookup vs none) was associated with higher immediate posttest scores in the full model (b = 3.02, 95% CI [0.28, 5.77], $p = .032$), but the association was not clearly present at delayed posttest (b = 3.12, 95% CI [−0.51, 6.75], $p = .089$) and did not predict delayed scores conditional on posttest (b = 0.95, 95% CI [−2.14, 4.03], $p = .532$). These analyses are exploratory and descriptive. *p*-values were not adjusted for multiple comparisons, and results should not be interpreted causally given the small sample and observational design.

Discussion

This study examined whether participation in a pedagogical intervention using a custom web application for collaborative story writing, with optional form-prioritized, learner-initiated GenAI feedback, was associated with gains in targeted grammar knowledge at posttest and whether delayed posttest scores remained higher than pretest scores at a two-week follow-up.

RQ1 asked whether grammar scores were higher at posttest than at pretest. In the primary MI mixed-effects model, posttest scores were higher than pretest scores (difference = 2.80 points, 95% CI [1.12, 4.48], $d = 0.56$, 95% CI [0.09, 1.03]), consistent with H1. On the 0–40 scale, this corresponds to roughly 2–3 additional items correct on average. However, the confidence interval for *d* is wide, indicating imprecision in the estimated magnitude.

RQ2 asked whether delayed posttest scores were higher than pretest scores at the two-week follow-up. Delayed posttest scores were higher than pretest scores (difference = 3.10 points, 95% CI [1.35, 4.86], $d = 0.62$, 95% CI [0.14, 1.10]), consistent with H2. Posttest and delayed scores did not differ reliably (difference = 0.31 points, 95% CI [−1.46, 2.07]), but this should not be read as evidence of equivalence or retention, because the interval is compatible with a small decline or a small improvement.

Interpretation of follow-up performance should be cautious because delayed-posttest absences were associated with lower pretest scores, making the delayed posttest estimate potentially optimistic.

Supplementary log analyses were largely null; vocabulary lookup showed a positive association with posttest scores but not clearly at delayed posttest, and all associations are descriptive given the observational design and small sample.

The observed pattern of gains is consistent with two strands of research that align with the instructional sequence used here: planned focus on form with targeted feedback, and collaborative writing as a context in which form–meaning decisions become open to discussion during production.

First, focus-on-form and WCF research suggests that accuracy gains are more likely when attention is narrowed to a limited set of forms and learners are pushed to engage with feedback through revision, rather than receiving broad, unfocused correction (Kang & Han, 2015; Long, 1991). In this sequence, learners received brief form-focused instruction and short practice, then wrote stories with the aim of applying the target form accurately during an extended, meaning-focused production task. This combination is compatible with evidence that explicit instruction supports targeted development, particularly when learners have opportunities to apply forms during communicative use (Norris & Ortega, 2000; Spada & Tomita, 2010).

Second, collaborative writing can increase opportunities for attention to form because language problems become public and negotiable within the context of a shared text, which can elicit LREs in which learners identify a problem, propose alternatives, and agree on what to contribute to the shared text (Swain & Lapkin, 1998; Wigglesworth & Storch, 2012). Because feedback opportunities were multi-source (peer interaction, instructor prompts, and optional learner-initiated GenAI feedback), the sequence likely increased occasions for text-anchored negotiation. However, the present design does not isolate which source drove change, and the log-based analyses provide correlational evidence only. In the present study, LREs are used as a theoretical lens for interpreting form-focused processing during the task, not as a directly observed or measured construct.

Constrained GenAI in Collaborative Writing

Taken together, the contribution is primarily at the level of design feasibility. In this classroom configuration, learner-initiated GenAI feedback that prioritizes the lesson's target form, provides local edit suggestions with brief L1 support, and does not generate full-text rewrites can be embedded within collaborative writing while keeping authorship and revision decisions with learners. Because the study used a one-group pretest–posttest design, the observed gains cannot be attributed to GenAI.

Prior synthesis work on computer-mediated collaborative writing in L2 classrooms notes that attention to language form is often relatively limited and calls for more designs that intentionally foreground form within technology-mediated collaboration (Zhang et al., 2021). In parallel, GenAI-focused synthesis work reports positive average effects while emphasizing that many studies are short-term, often conducted in informal or self-directed contexts, and frequently rely on self-reported outcomes, with effects varying by context and instructional integration (Saarela et al., 2026). This contextual sensitivity is relevant to the present study because it evaluates a bounded, teacher-guided classroom configuration in a compulsory EFL setting, using an independent targeted grammar test rather than self-report outcomes. In contrast, a meta-analysis focused on informal, self-directed GenAI-supported digital English learning found significant effects on proficiency and self-regulation but not on motivation, underscoring that outcomes vary by learning context, degree of guidance, and what is measured (Guan et al., 2024).

Recent classroom studies in Chinese higher education offer useful comparison points for grammar-related outcomes under GenAI support. Zhou and Du Preez (2025) reported improvement on a grammar-focused dimension of a timed-essay rubric when CEFR B1 learners followed a guided chatbot workflow in an English for Academic Purposes (EAP) course. This is compatible with the present findings insofar as both suggest that grammar-related gains are more plausible when GenAI is embedded in a guided instructional sequence. In an International English Language Testing System (IELTS) writing program, Chen and Chen (2025) found that lower-intermediate learners reported positive affective gains but showed no clear improvement in overall writing performance beyond syntactic complexity. They attributed this discrepancy to misalignment between open-ended chatbot affordances and learners' cognitive readiness, noting that complex, unconstrained output often overwhelmed students and reduced active engagement with form. This also helps interpret the present findings, suggesting that more open-ended chatbot use may be a weaker fit for lower-proficiency learners than constrained GenAI writing support.

Although these studies operationalized outcomes through writing-based measures rather than independent targeted grammar tests, they converge on the same constraint: grammar-related benefits from general-purpose chatbots depend strongly on proficiency, guidance, and instructional alignment. In the present study, that alignment was built through explicit mini-lessons, worksheet practice, collaborative story writing organized around a weekly target form, and learner-initiated feedback that prioritized that form where relevant. This strengthens the case for testing bounded, instruction-aligned configurations designed for compulsory EFL settings.

Against that backdrop, this study provides classroom implementation evidence for a constrained configuration, target-form–prioritized GenAI feedback embedded within collaborative writing, alongside explicit instruction and teacher and peer support, evaluated using an independent targeted grammar test with a short delayed follow-up. Whether GenAI adds value beyond the same instructional sequence implemented without GenAI requires controlled comparisons and process evidence.

Implications for Theory and Practice

Conceptually, this work supports treating GenAI as a bounded resource embedded within a classroom writing activity, not as a standalone assistant. In this configuration, GenAI is one element of a feedback ecology alongside peer interaction and teacher support. Because this GenAI support is learner-initiated and constrained in scope and output, it can channel attention toward taught forms while keeping authorship and revision decisions with learners. This emphasis on embedded use is also consistent with recent review work arguing that GenAI is most productive when integrated into staged classroom sequences with teacher guidance and peer discussion, rather than treated as a standalone tool (Rong & Yao, 2026).

Practically, this design is likely to be a better fit for compulsory EFL contexts and lower-proficiency learners than either open-ended chatbot use or commercial writing tools. This is in line with Chen and Chen's (2025) finding that more open-ended GenAI use can be a weak fit for lower-intermediate learners. A classroom-usable pattern is learner-initiated support aligned with lesson targets and limited to local edit suggestions with brief L1 explanations, without full-text rewrites, so students still decide what to change and implement revisions themselves. The goal is not to maximize correction, but to provide usable support during real writing without shifting core composing and editing work onto the system.

Conclusion

This study evaluated a classroom implementation of technology-mediated collaborative story writing in which learners could request constrained, learner-initiated GenAI feedback, delivered as local edit

suggestions with brief L1 explanations and without full-text rewrites, while working toward explicitly taught grammar targets. Within this instructional package, grammar scores were higher at posttest and were also higher at a two-week follow-up, with no clear difference between immediate and delayed-posttest performance. The main contribution is design feasibility: constrained, learner-initiated GenAI support can be embedded within collaborative writing alongside explicit instruction without turning the activity into AI-driven text generation.

Interpretation is bounded by three constraints. First, the one-group pretest–posttest design estimates package-level change and does not support attribution to GenAI or any single component. Observed gains could reflect the broader instructional sequence (mini-lesson, short practice, collaborative writing with teacher prompts) and repeated testing rather than the learner-initiated support specifically. Second, the outcome measure was a receptive, target-focused multiple-choice test optimized for sensitivity to the taught forms, not for capturing writing performance. As a result, the study cannot speak to transfer to production (for example, accuracy in new writing, revision quality, or longer-term proficiency development) beyond these taught targets. Third, inference is limited by sample size and delayed-wave missingness. Because delayed-posttest absence was associated with lower baseline performance, follow-up estimates may be optimistic if the MAR assumption is violated. Complete-case sensitivity analyses showed the same pattern of time contrasts, but baseline-related attrition means the delayed-wave results should still be interpreted cautiously.

Next steps should broaden outcomes and examine which components of the intervention are most strongly associated with gains. Future work should assess proficiency alongside targeted grammar, extend outcomes beyond receptive multiple-choice testing to production-sensitive measures, and include longer follow-ups to evaluate transfer and durability. Controlled comparisons should also isolate which design elements matter most, including learner-initiated versus automatic feedback, the contribution of short practice activities, and whether similar gains are achievable with the same instructional sequence implemented without GenAI. Finally, larger samples and richer process measures are needed to link learners' feature use, revision behavior, and interactions during writing to outcomes, helping clarify the learning processes operating in constrained, learner-initiated GenAI use during collaborative production.

Use of Generative AI

During the preparation of this work, the authors used OpenAI's ChatGPT and Google's Gemini Pro to improve the grammar and clarity of the manuscript. No AI tools were used for the generation of research hypotheses, data analysis, or the writing of conclusions. The authors reviewed and edited the content as needed and take full responsibility for the published article.

Acknowledgments

The authors would like to thank the students who participated in this study. They also thank Caleb Moon for his assistance with proofreading the manuscript.

Disclosure of Interest

The authors have no known competing financial interests or personal relationships to declare.

Ethical Statement

Informed consent was obtained from all participants after they were briefed on the study's objectives and methods. Ethical review was completed by the Language Education and Research Center at

Kyushu Sangyo University prior to the commencement of the study. Under this protocol, the requirement for formal ethics committee approval was waived.

Funding Disclosure Statement

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) [25K06610].

Data Availability Statement

Anonymized data, analysis code (Stata 18), and an AI implementation summary are available on the Open Science Framework (https://osf.io/cnrbm).

CRediT Author Statement

Nicolas Emerson: Conceptualization, Investigation, Resources, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Thomas Holland**: Conceptualization, Investigation, Writing – review & editing. **Tyler D. Mitchell**: Conceptualization, Investigation, Resources, Writing – review & editing. **Davis Shum**: Conceptualization, Investigation, Resources, Writing – review & editing. **W. L. Quint Oga-Baldwin**: Writing – review & editing, Supervision.

References

Chen, Y., & Chen, J. (2025). Generative AI in EFL writing instruction for lower-intermediate learners: Do perceived affective gains translate to improved performance? *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688251397724

de Fine Licht, K. (2024). Generative artificial intelligence in higher education: Why the 'banning approach' to student use is sometimes morally justified. *Philosophy & Technology, 37*, 113. https://doi.org/10.1007/s13347-024-00799-9

DeKeyser, R. M. (2020). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 83–104). Routledge. https://doi.org/10.4324/9780429503986-5

Donley, K. (2024). Teaching with ChatGPT as a linguistically responsive tool for multilingual learners. *Technology in Language Teaching & Learning, 6*(3), 1719. https://doi.org/10.29140/tltl.v6n3.1719

Doughty, C., & Williams, J. (Eds.). (1998). *Focus on form in classroom second language acquisition.* Cambridge University Press.

Elabdali, I. (2021). Are two heads really better than one? A meta-analysis of the L2 learning benefits of collaborative writing. *Journal of Second Language Writing, 52*, 100788. https://doi.org/10.1016/j.jslw.2020.100788

Ellis, R. (2001). Introduction: Investigating form-focused instruction. *Language Learning,* 51(Suppl. 1), 1–46. https://doi.org/10.1111/j.1467-1770.2001.tb00013.x

Emerson, N. (2024). AI-enhanced collaborative story writing in the EFL classroom. *Technology in Language Teaching & Learning, 6*(3), 1764. https://doi.org/10.29140/tltl.v6n3.1764

Enders, C. K. (2010). *Applied missing data analysis.* Guilford Press.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56(*2), 489–530. https://doi.org/10.1111/bjet.13544

Feng, H., Li, K., & Zhang, L. J. (2025). What does AI bring to second language writing? A systematic review (2014–2024). *Language Learning & Technology, 29*(1), 1–27. https://doi.org/10.64152/10125/73629

Freeman, J. (2025). *Student generative AI survey 2025*. Higher Education Policy Institute. https://www.hepi.ac.uk/wp-content/uploads/2025/02/HEPI-Kortext-Student-Generative-AI-Survey-2025.pdf

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Guan, L., Li, S., & Gu, M. M. (2024). AI in informal digital English learning: A meta-analysis of its effectiveness on proficiency, motivation, and self-regulation. *Computers and Education: Artificial Intelligence, 7*, 100323. https://doi.org/10.1016/j.caeai.2024.100323

Hirschel, R., & Horai, K. (2025). Exploring AI to automate EFL corrective written feedback in the first language. *Technology in Language Teaching & Learning, 7*(1), 2208. https://doi.org/10.29140/tltl.v7n1.2208

Jung, J., Joo, S., Park, H., & Han, I. (2026). Challenges of integrating ChatGPT into EFL writing: An activity theory framework. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2026.2617386

Kang, E. Y., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal, 99*(1), 1–18. https://doi.org/10.1111/modl.12189

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

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). SAGE. https://doi.org/10.4135/9781483384733

Kormos, J. (2023). The role of cognitive factors in second language writing and writing to learn a second language. *Studies in Second Language Acquisition, 45*(3), 622–646. https://doi.org/10.1017/S0272263122000481

Li, M. (2018). Computer-mediated collaborative writing in L2 contexts: An analysis of empirical research. *Computer Assisted Language Learning, 31*(8), 882–904. https://doi.org/10.1080/09588221.2018.1465981

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. B. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). John Benjamins. https://doi.org/10.1075/sibil.2.07lon

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

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Rong, M., & Yao, Y. (2026). From assistance to advancement: A systematic review of how GenAI supports L2 writing learning from the lens of activity theory. *ECNU Review of Education,* 9(2). https://doi.org/10.1177/20965311261437802

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. https://doi.org/10.1002/9780470316696

Saarela, M., Gunasekara, S., & Kumarage, P. (2026). A meta-analysis of generative AI effects on language proficiency and affective–cognitive outcomes in language learning. *Discover Computing, 29*, Article 116. https://doi.org/10.1007/s10791-026-10015-1

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Shin, D., & Lee, J. H. (2025). Leveraging LLM-based chatbots for interactional grammar feedback in L2 writing: Opportunities and challenges. *Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2549037

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*(2), 263–308. https://doi.org/10.1111/j.1467-9922.2010.00562.x

StataCorp. (2023). *Stata Statistical Software: Release 18* [Computer software]. StataCorp LLC.

Storch, N. (2013). *Collaborative writing in L2 classrooms*. Multilingual Matters. https://doi.org/10.21832/9781847699954

Storch, N. (2019). Collaborative writing. *Language Teaching, 52*(1), 40–59. https://doi.org/10.1017/S0261444818000320

Strobl, C., Menke-Bazhutkina, I., Abel, N., & Michel, M. (2024). Adopting ChatGPT as a writing buddy in the advanced L2 writing class. *Technology in Language Teaching & Learning, 6(*1), 1168. https://doi.org/10.29140/tltl.v6n1.1168

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal, 82*(3), 320–337. https://doi.org/10.1111/j.1540-4781.1998.tb01209.x

Tannenbaum, R. J., & Wylie, E. C. (2019). *Mapping the TOEIC® tests on the CEFR* [Technical report]. Educational Testing Service. https://www.ets.org/pdfs/toeic/toeic-mapping-cefr-reference.pdf

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Wigglesworth, G., & Storch, N. (2012). What role for collaboration in writing and writing feedback? *Journal of Second Language Writing, 21*(4), 364–374. https://doi.org/10.1016/j.jslw.2012.09.005

Zhang, M., Gibbons, J., & Li, M. (2021). Computer-mediated collaborative writing in L2 classrooms: A systematic review. *Journal of Second Language Writing, 54*, 100854. https://doi.org/10.1016/j.jslw.2021.100854

Zhang, M., & Plonsky, L. (2020). Collaborative writing in face-to-face settings: A substantive and methodological review. *Journal of Second Language Writing, 49*, Article 100753. https://doi.org/10.1016/j.jslw.2020.100753

Zhang, R., & Zou, D. (2022). Types, features, and effectiveness of technologies in collaborative writing for second language learning. *Computer Assisted Language Learning, 35*(9), 2391–2422. https://doi.org/10.1080/09588221.2021.1880441

Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. F. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. *Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2503890

Zhou, S., & Du Preez, G. D. (2025). Bringing Cinderella into spotlight: GenAI-assisted grammar acquisition in academic writing. *Language Learning & Technology, 29*(1), 1–25. https://doi.org/10.64152/10125/73639

Appendix

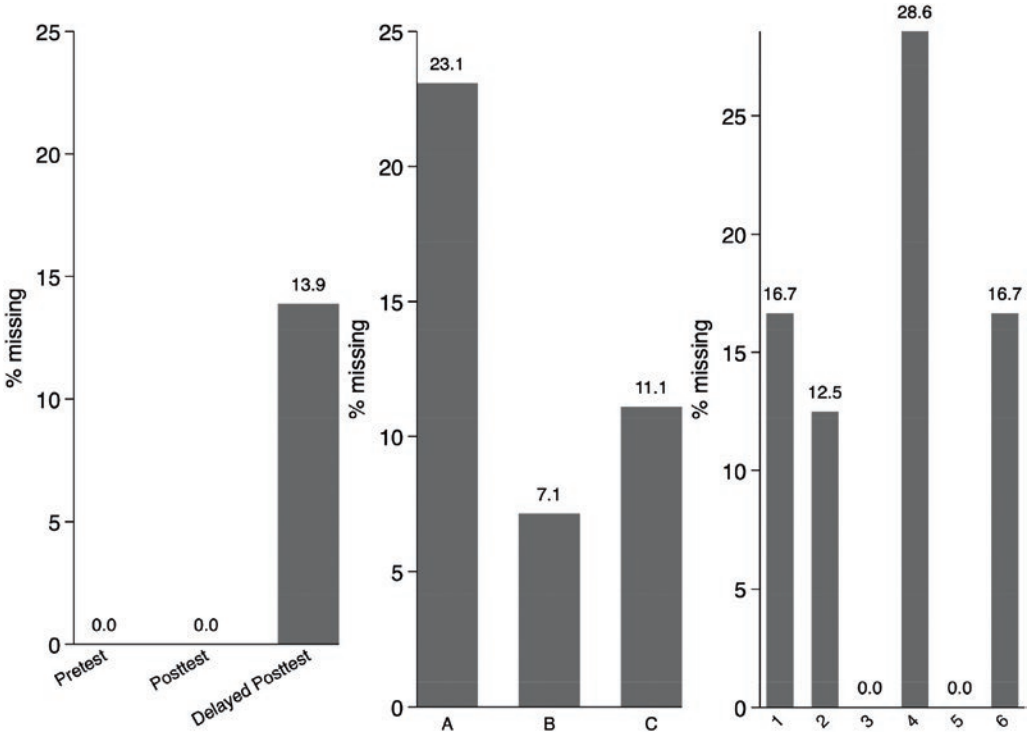

Figure A1. *Missingness by Wave; Delayed-Posttest Missingness by Form and Order*
Note: Panel 1: Percentage missing by wave. Panels 2–3: Percentage missing at the delayed posttest, by test form and assignment order. Percentages are means of the missingness indicator (1 = missing, 0 = observed) × 100.

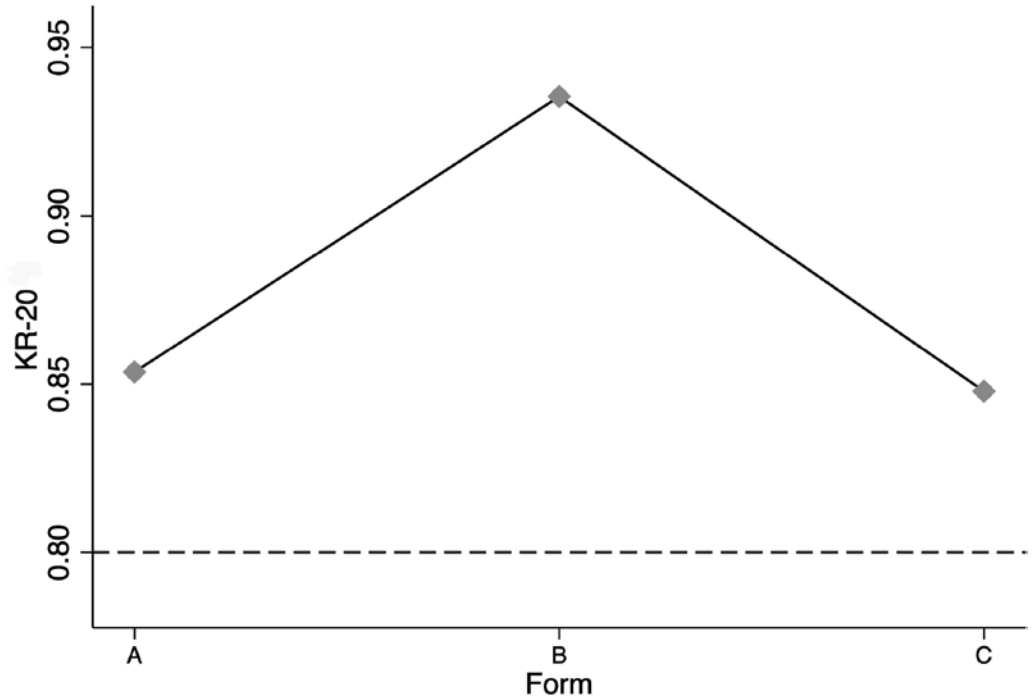

Figure A2. *Pretest Reliability by Form (KR-20)*
Note: KR-20 were computed on dichotomously scored pretest items; within-form zero-variance items were excluded. The dashed line marks the .80 adequacy benchmark.

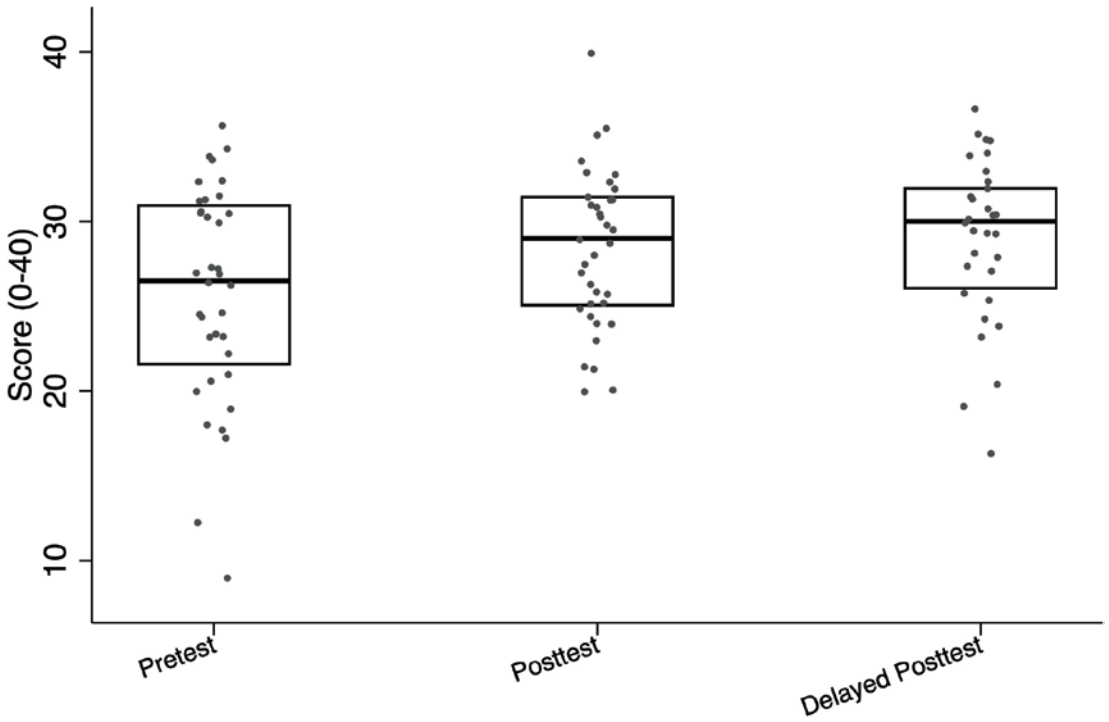

Figure A3. *Observed Distributions by Time (Box-and-Jitter)*
Note: Boxes show medians and interquartile range (IQR); whiskers extend to 1.5 × the IQR. Jittered points are individual scores (0–40). This figure is descriptive; inferential results appear in Table 5 and Table A6.

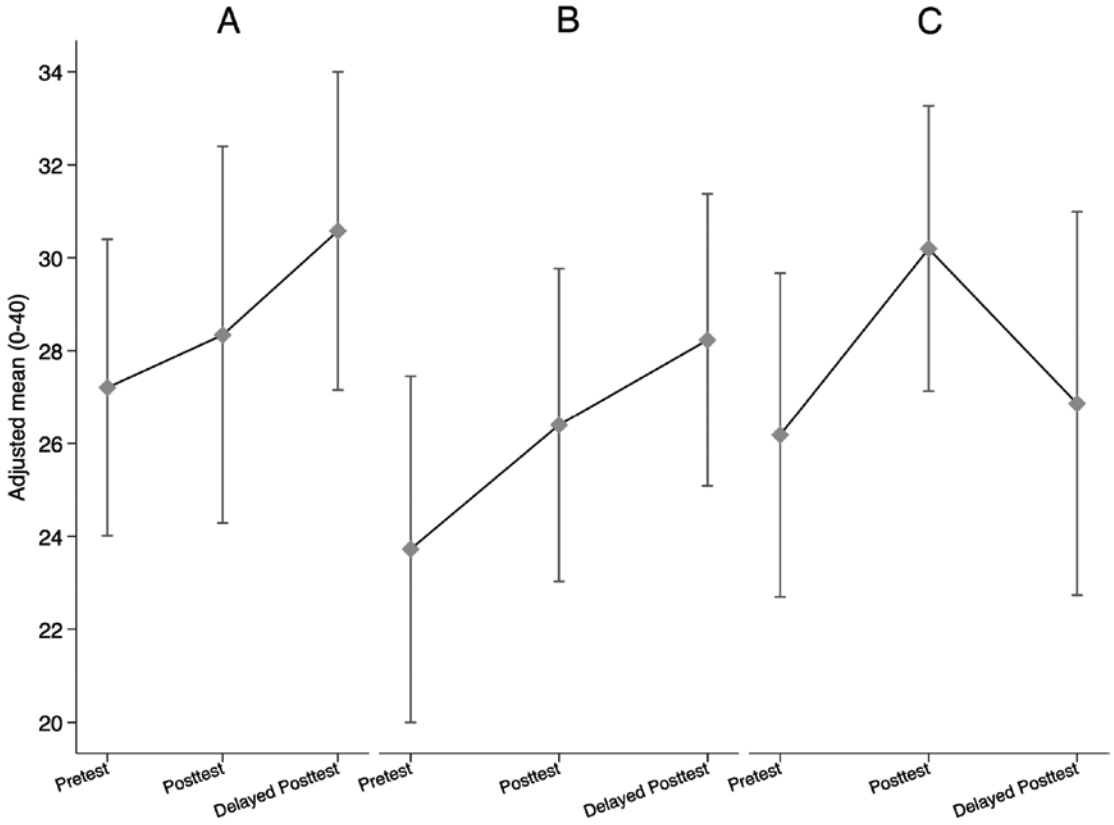

Figure A4. *Adjusted Means by Time Within Test Forms*
Note: Adjusted means are from an ordinary least squares (OLS) model fit across 20 imputed datasets. The model included the time × form interaction while controlling for TOEIC pretest score and instructional order. As noted in Table A4, the time × form interaction was not significant. Error bars represent 95% confidence intervals.

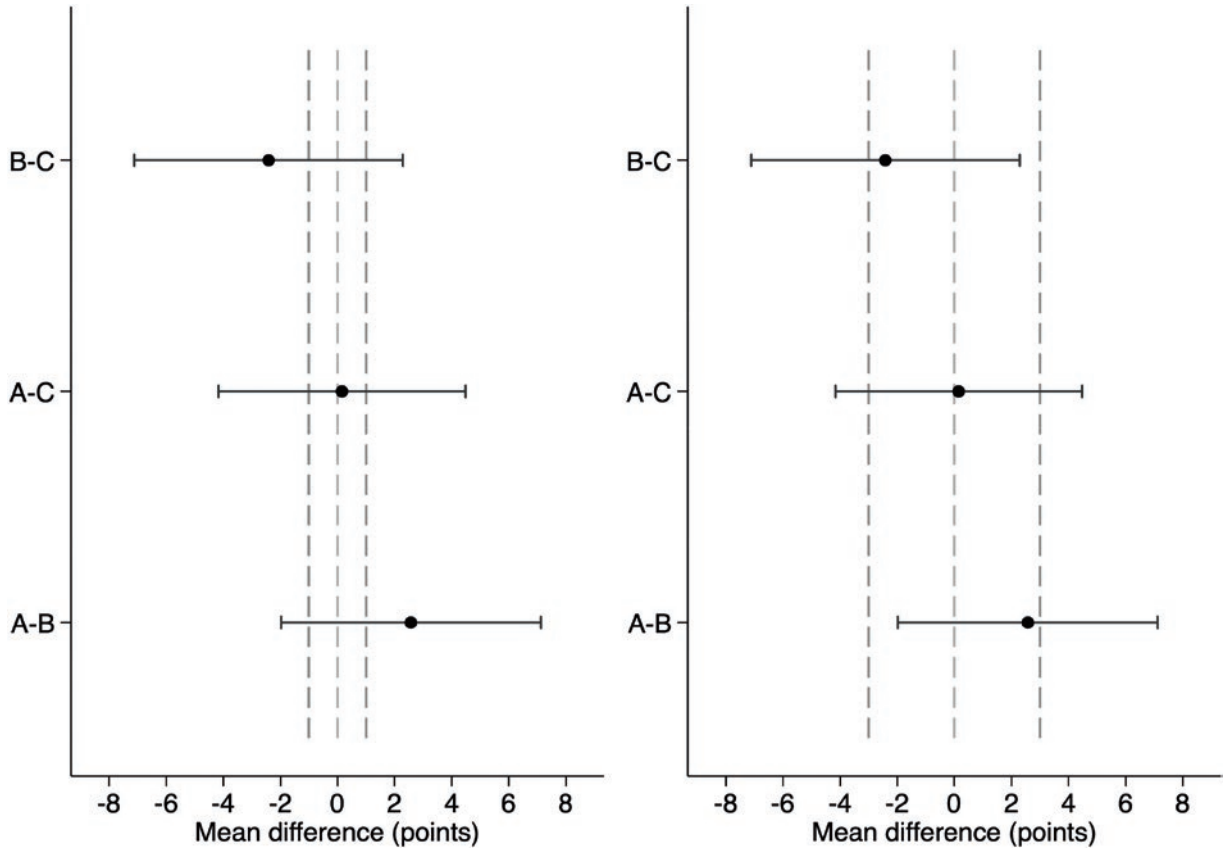

Figure A5. *Pretest Form Equivalence Under Alternative Margins (Left: ±1 Point; Right: ±3 Points)*
Note: Points show A–B, A–C, and B–C mean differences with 90% confidence intervals (per equivalence-testing convention). Dashed vertical lines indicate the reference (0) and the equivalence margins (left: ±1 point; right: ±3 points). Positive values indicate higher means for the first-named form.

Table A1. *Internal Consistency at Pretest (KR-20) by Form*

Form	*n* (students)	Items used	KR-20
A	14	37	.854
B	10	36	.935
C	12	35	.848

Note. Items were scored 0/1, and within-form zero-variance items were dropped before computing KR-20. "Items used" indicates the number of non-constant items contributing to the KR-20 estimate in each form.

Table A2. *Variance-Equality Tests Across Forms (Pretest)*

Panel A. Descriptives by form

Form	*n*	Mean	*SD*
A	14	26.571	6.186
B	10	24	7.645
C	12	26.417	5.775

Panel B. Tests of equal variances

Test	*F*	*df*	*p*
Levene (mean-based; W0)	0.6162	2, 33	.546
Brown–Forsythe (median; W50)	0.5883	2, 33	.561
Levene (10% trimmed; W10)	0.991	2, 33	.555

Note. Tests compare pretest score variances across forms. *F* shown with *df*1, *df*2. All *p* values from the *F* distribution.

Table A3. *Standardized Effects From MI-OLS (*d*) With 95% Confidence Intervals*

Contrast	*d*	95% CI
Pretest to posttest	0.56	[0.09, 1.03]
Pretest to delayed posttest	0.62	[0.14, 1.10]

Note. Standardized effect size (*d*) was computed by dividing the MI-pooled time contrast by the pooled residual standard deviation (*SD* = 5.02) across imputations. CI = confidence interval; MI-OLS = ordinary least squares model on multiply imputed data.

Table A4. *Tests of Form Main Effect and Time × Form Interaction (Mixed Model)*

Test	Wald χ^2(*df*)	*p*
Form main effect	3.73 (2)	.155
Time × form interaction	4.47 (4)	.347

Note. Wald χ^2 tests are from the primary multiple-imputation (MI) mixed-effects model. The model specification is provided in the note for Table 5.

Table A5. *Variance Components and Person-Level ICC*

Component	*SD*	Variance
Person (id)	3.63	13.21
Residual	3.67	13.50
ICC (person)	0.49	—

Note. Components are from a random-intercept REML model. Class-level variance was approximately zero and is omitted. ICC = Var(person) / [Var(person) + Var(residual)]. ICC = intraclass correlation coefficient; REML = restricted maximum likelihood; *SD* = standard deviation.

Table A6. *Time Contrasts from the Mixed-Effects Model for the Complete-Case Analysis (*n *= 31)*

Contrast	Difference (points)	95% CI	*p*
Pretest to posttest	2.50	[0.80, 4.21]	.004
Pretest to delayed posttest	2.38	[0.67, 4.10]	.006
Posttest versus delayed posttest	−0.12	[−1.83, 1.59]	.891

Note. This complete-case analysis used the same model specification as the primary analysis reported in Table 5. CI = confidence interval.