Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2024, Volume 28, Issue 1
pp. 1–19

ARTICLE

Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning

Da "Alex" Yan, Xinyang Agriculture and Forestry University

Abstract

This research examined the effects of collaborative processing of ChatGPT-generated feedback on second language (L2) writing development. The sample for the seven-week experiment consisted of 117 sophomore EFL learners and six teachers from a Chinese university. The students were divided into four groups: the control group processed ChatGPT feedback individually, while the experimental groups processed it with teacher, peer, or combined teacher-and-peer collaboration. Employing a mixed design, L2 writing task improvement (measured as the gain scores from draft writing to final products for the three during-intervention writing tasks) and learning (measured as performance for a post-intervention, and similar new writing task) of the four groups were assessed and analyzed. The findings revealed that the learners who individually processed feedback registered the most significant task improvements, whereas learners processing feedback with teacher collaboration progressed the most for subsequent learning processes. The study has pedagogical implications for the construction of a more inclusive ecology to support effective uses of latest technology in L2 classrooms.

Keywords: *feedback processing, automated feedback, generative artificial intelligence, L2 writing*

Language(s) Learned in This Study: *English*

APA Citation: Yan, D. (2024). Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology*, *28*(1), 1–19. https://hdl.handle.net/10125/73597

Introduction

Automated writing evaluation (AWE) tools have been widely applied in second language (L2) writing classroom for writing appraisal and the provision of automated feedback (AF) (Jiang & Yu, 2022). Recent technological breakthroughs, for example, artificial intelligence, natural language processing, and latent semantic analysis, have steadily enhanced the performance of AF in detecting and correcting errors (cf. the reported correction accuracy of 50% in Bai & Hu, 2017; and 85% in Guo et al., 2022). Since November 2022, ChatGPT, OpenAI's chatbot powered by the Generative Pre-trained Transformer (GPT) family of large language models (LLM), marks a new chapter of AF. Researchers also highlighted its versatile affordances to imitate humans in writing instruction and assessment (Barrot, 2023; Mizumoto & Eguchi, 2023). A benchmark study has unveiled its potential to outperform existing AWE systems in correcting grammatical errors in writing products (Wu et al., 2023). To date, investigations into ChatGPT's impact on writing assessment have been conducted in simulated or out-of-classroom settings. As Lee (2020) has called for more research in an ecologically valid context, we are facing a paucity of research on the utilization of ChatGPT for AWE in authentic L2 learning settings.

In addition, existing research on AF is posited on the assumption that the *information* of feedback, once *delivered* successfully to the learners, would result in improvement of L2 writing performance (Zhang & Hyland, 2018). Such belief has been frequently criticized by advocates of the new feedback paradigm

*** Corresponding Author:** Da "Alex" Yan, alexyan1987@outlook.com

(NFP) (Jensen et al., 2023; Nieminen et al., 2022) and the ecological perspective of language learning (Han, 2019), who argued that learners must actively seek, make sense of feedback as *processes* or *events* in a sociocultural context to which the language learners belong. However, sociocultural scaffolding or support has been downplayed in research on AF, which has been sustainedly applied in individual writing settings (Shadiev & Feng, 2023). Even for research incorporating collaborative writing with AF (e.g., Tan et al., 2022), the focus remains on the provision instead of the processing of feedback. Consequently, there exists a weak nexus between the ecological and leaner-centered perspectives on feedback and the practice of AF in L2 pedagogy, where researchers used to believe that such tools did not offer appropriate communicative channels for classroom-based scaffolding during feedback processing (Dong & Shi, 2021).

Literature Review

ChatGPT-generated Feedback: More Than ACF

Prior to the advent of AI-based feedback provision, automated corrective feedback (ACF) was used in L2 writing classrooms to automatically flag and correct grammatical errors in writing products (Shadiev & Feng, 2023). The advantages of ACF include: (a) providing prompt feedback provision; (b) offering rich metalinguistic hints; and (c) reducing the burden on instructors (Ranalli, 2018). Concurrently, ACF has frequently been criticized for: (a) low learner uptake (Bai & Hu, 2017), (b) limited cognitive engagement (Koltovskaia, 2020), (c) learners distrust (Ranalli, 2021), and (d) inability to provide individualized feedback (Barrot, 2021).

The emergence of ChatGPT offers potential to address, at least to moderate, the above issues with traditional ACF. First, ChatGPT's unprecedented volume of pre-trained language data affords L2 learners powerful assistance in handling text-based tasks inclusive of writing assessment (Mizumoto & Eguchi, 2023; Wu et al., 2023). Traditional AWE is defined as a system providing writing quality evaluation and/or corrective feedback (particularly lower-order concerns, see Fu et al., 2022) whereas ChatGPT could be used for a much wider range of writing assistance or assessment (see Barrot, 2023). Second, ChatGPT, being an AI-based chatbot, could provide personalized, interactive, and adaptive feedback (Barrot, 2023). In practice, learners could *prompt* ChatGPT with their judgmental and affective feedback on received information, with which ChatGPT would change its information generation strategies via *prompt-based* or *in-context* learning (see Oppenlaender et al., 2023). Third, contrary to the "one-shot" experience of traditional ACF providers, learners are encouraged to interactively and iteratively prompt ChatGPT for feedback on improved quality and/or specification (Yan, 2024). Advised by researchers in the field of LLM, multi-turn human-ChatGPT interactivity through trail-and-error resulted in significant performance improvement of AI systems for language-related tasks (Bang et al., 2023; Dang et al., 2022). Consequently, ChatGPT not only serves as a competitive ACF provider that L2 learners can constructively engage with, but also expands the traditional scope of ACF by offering an interactive, iterative, and human-centered AF platform (Yan & Zhang, 2024). To distinguish ChatGPT-generated feedback from ACF afforded by traditional AWE systems, the term AF is used in the present study.

Individual vs. Collaborative Processing of Feedback on L2 Writing: Previous Insights

Although abundant literature has explored the processing of feedback as an individual behavior, the recent shift towards a socio-constructivist perspective of feedback has promoted a research trend to focus on the knowledge co-construction based on feedback in a sociocultural context of learning (Bitchener & Storch, 2016). Situated in the context of L2 writing, researchers have argued that the knowledge gains reaped from the collaborative processing of feedback would be transferred to the individual development of L2 writing quality (Shi et al., 2022). Existing studies have been premised on Vygotsky's sociocultural theory (SCT) and empirical evidence obtained from research on collaborative writing. In a sociocultural L2 setting, feedback is understood as the scaffolds provided by the "more knowledgeable ones" (MKO) for the attainment of and advancement in the "Zone of Proximal Development" (ZPD) (Bitchener &

Storch, 2016). Hence, collaborative processing of feedback through sociocultural interactions (for example, peer conversation, pair correction, and so forth) has been hypothesized to be effective in improving the accuracy rate of students’ revision (Kim & Emeliyanova, 2021). Such an assumption was partially supported by the result of a meta-analysis, where Lv et al. (2021) reported a significantly larger effect size of online feedback on EFL writing task performance for collaborative tasks than individual ones (compare $g = 1.075$ and 0.76 respectively). Thus, the present study follows the framework of Villamil and Guerero (2019) to conceptualize feedback as a situated phenomenon in a sociocultural context where the interactions between the agents shape its uptake and impact on learning.

Existing studies have provided insights into the comparison between the effects of individual and collaborative processing of feedback on during-interventional L2 writing performance. In terms of the processes involved in the collaborative processing of feedback, Shi et al. (2022) identified three stages: the trustful, skeptical, and critical phases. They examined the use of evidence in L2 argumentative writing from five EFL learners who practiced collaborative processing of feedback in the one-semester study employing a content-based AWE system (that is, Virtual Writing Tutor). The study was limited in its case study design to understand the feedback processing patterns of individual writers, but the researchers described that they would face difficulties utilizing content-based AF. Mujtaba et al. (2021) conducted a 10-week comparative study to investigate the effects of individual versus collaborative processing of teacher-provided written corrective feedback on L2 writing accuracy and revision. Their study revealed that collaborative processing of corrective feedback led to writing accuracy improvement and error reduction (particularly in verb use and word choices) in revision compared to the individual processing control group. Comparatively, the students individually processing feedback had less effective uptake of the feedback messages. However, Kim and Emeliyanova (2021) study suggested different results. In their 8-week project participated in by 36 L2 learners, pair-correction control groups showcased higher error correction rates whereas the measurement of writing accuracy showed no significant difference. Yao et al. (2021) have incorporated AF with peer assessment in their longitudinal investigation into L2 writers’ mindset and motivation. The results obtained from the 15-week instructional processes with more than 500 participants indicated that the peer collaboration to process Pigai-generated corrective feedback positively impacted the formation of a growth mindset and learning motivation among the experimental group members.

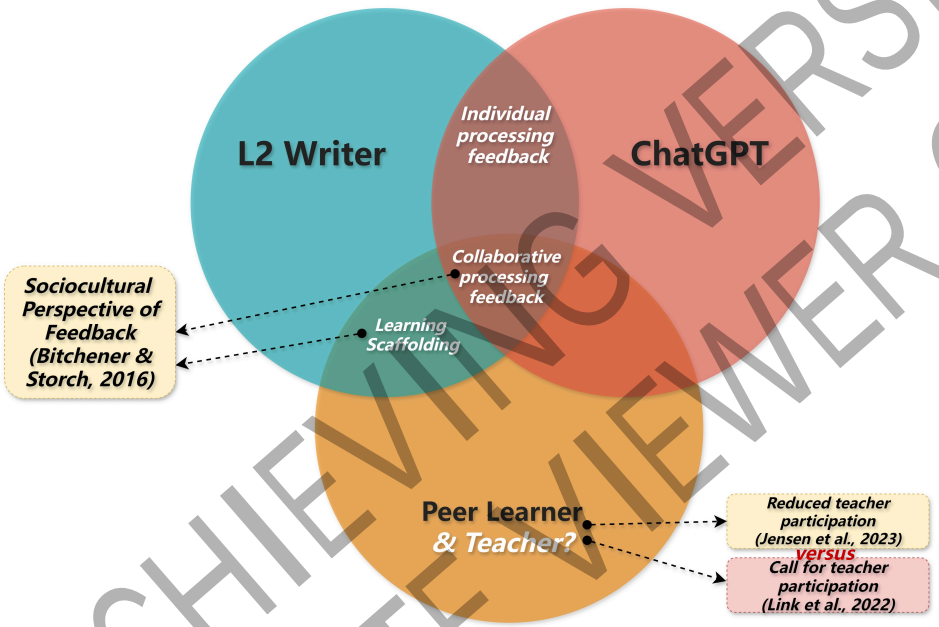
However, our understanding of the longtime effects of different feedback processing approaches on writing performance is limited. The studies by Mujtaba et al. (2021) and Kim and Emeliyanova (2021) both included a post-interventional delayed test that received no corrective feedback. Statistical analyses in Mujtaba et al. (2021) showed that students collectively processed feedback registered significant improvements in reducing errors in the delayed test (known as the post-test) compared to the first during-intervention task (known as the pre-test). Similarly, Kim and Emeliyanova (2021) also compared the first during-intervention task (known as diagnostic timed writing) and the delayed test (known as final timed writing). The results showed that no significant differences in error reduction were found between the students collectively and individually processed feedback. Taken together, existing studies, adopting a pretest-posttest design other than a longitudinal or repeated measures design, merely compared the differences in students’ performance between the first during-intervention task and the delayed test. Therefore, these studies shed limited light on the transferability of the benefits from individual and collective processing of feedback to new writing tasks and longitudinal developmental trajectories across all the measured tasks in the experiments. According to Wiliam (2018), the purpose of feedback is to improve not only immediate task *performance* but also long-term *learning*. Therefore, this study measures and investigates learning outcomes as during-intervention L2 writing *task improvement* (that is, the improvement students made after using ChatGPT-generated AF between draft and final writing artifact scores during the intervention) and progression in subsequent *learning* processes (that is, the progression students made in draft scores during and after the intervention).

Collaborative Processing of ChatGPT-Generated Feedback: Theoretical Backgrounds

To date, the collaborative processing of ChatGPT-generated feedback remains less frequently implemented and understudied. To support the exploration of the effects of collaborative processing of ChatGPT-generated feedback, a theoretical framework was built based on previous theories (Figure 1). The framework would provide theoretical support for the rationale and research design of the present study.

Figure 1

Theoretical Framework of the Study



First, the effects of collaborative processing of ChatGPT-generated feedback should be explored. We have been informed by previous literature that the cognitive challenges faced by students individually processing feedback could be solved through scaffolding from other agents, for example, teachers and peers who had higher levels of proficiency (Bitchener & Storch, 2016). Such social support and scaffolding would be equally important for learners using ChatGPT for AF, as researchers have highlighted that the effective uptake of such AF calls for higher-level cognitive abilities and AI literacy that many students don't possess (Yan, 2023, 2024). Following this line of reasoning, the collaborative processing of AF could be helpful for students to obtain more gains from the knowledge co-construction processes with MKOs in the sociocultural learning environment for better use of AI-generated feedback. Thus, the underlying major hypothesis to be tested by this research is whether collaborative processing of feedback results in better performance in task improvement and learning.

Second, the role played by teachers in the collaborative processing of ChatGPT-generated AF should be explored. Debating voices could be heard in the previous literature. For example, researchers advocating NFP argued that technology-enhanced feedback reduces the central role of teachers in providing, processing, and making use of feedback messages (for example, Jensen et al., 2023; Nieminen et al., 2022). Contrarily, such a viewpoint was not accepted by researchers in the field of L2 writing. For instance, Link et al. (2022) criticized that the "isolated use of AWE without teacher feedback is highly discouraged" (p. 25). From the sociocultural perspective of feedback, the presence of teachers in

processing feedback is worth exploring, as it has been sustainedly argued that teachers function as more capable “MKOs” due to their linguistic expertise and assessment/instructional experiences (Tian & Zhou, 2020) while peer support has frequently been doubted and distrusted (Cheng & Zhang, 2024). However, existing studies exploring the effect of collaborative processing of feedback on L2 writing (for example, Kim & Emeliyanova, 2021; Mujtaba et al., 2021) have focused only on the collaboration between individual writers and peers in the classroom. Thus, we could further expand the above-mentioned major hypotheses by including teachers as a source of sociocultural scaffolding for collaborative processing of AF.

The Study

To fill the above-mentioned gaps, this study sets out to explore the effects of individual versus collaborative processing of AF generated by ChatGPT in authentic L2 writing settings. Based on the above-stated hypothesis, the research answers the following two research questions (RQs):

RQ1: How do modes of collaborative (that is, teacher-scaffolded, peer-scaffolded, and teacher-peer-scaffolded) processing of ChatGPT-generated automated feedback affect students’ L2 writing task improvement by comparison with individual feedback processing?

RQ2: How do modes of collaborative (that is, teacher-scaffolded, peer-scaffolded, and teacher-peer-scaffolded) processing of ChatGPT-generated automated feedback affect students’ L2 writing learning by comparison with individual feedback processing?

Methods

The study was part of a larger research project on the application of LLM in language education, which aimed at exploring both the effects of LLM on language proficiency development and human behaviors in an AI-enhanced learning environment (for previous research outcomes, see Yan, 2023, 2024; Yan & Zhang, 2024). The present study focused on the effects of different feedback processing modes on writing performance.

Participants

117 sophomore students (M_{age} = 19.9 years) were recruited from a pool of 283 students that were enrolled in an EFL program at a Chinese university. All the students were Chinese and have been learning English as a second language on average for 10.2 years. The participant recruitment followed two criteria: (a) students’ responses to a 10-point survey which included of the following dimensions: perceived usefulness, curiosity, intention to use GenAI, and joy; and (b) grade in a previously administered language proficiency assessment developed according to China's Standards of English Language Ability (CSE). Informed by previous literature, effective uptake of technology-enhanced ACF called for higher level of language proficiency (Koltovskaia, 2020). Thus, I only recruited students whose L2 proficiency scores were higher than 76.4/100, which was equivalent to or CSE level 4-5 or CEFR B1.

To maintain students’ engagement and motivation in the project, the rated writing products were used as assignments for the grading of the course *Intermediate English Writing*, a compulsory course for all sophomore EFL students at the university. Outstanding performance in seeking, processing, and using ChatGPT-generated AF was awarded bonus marks.

After the recruitment of project participants, the following measures were taken to ensure educational equity for non-participants enrolled in the course: (a) post-project training on the use of ChatGPT for AF was arranged so that the non-participants could have access to and exposure to the technologies imparted to the participants; and (b) all the students enrolled in the writing course had the same instructional procedures, inclusive of the assessment tasks, for which the non-participants were required to use self-/peer-assessment or pre-LLM AF providers such as Grammarly to form revisions after completion of the drafts.

Additionally, six EFL teachers were recruited and trained for the study. See Table 1 for their background information and roles in the study. The training focused on: (a) strategies to use ChatGPT for AF; and (b) suggested ways to offer scaffolding to the students. Noticeably, I have put sufficient energy into training the teachers to use standard ways of using and assisting students' usage of ChatGPT (for example, how to teach students to use a standard and replicable workflow for using ChatGPT and how to prepare and provide modeling on the usage of ChatGPT) in L2 writing classrooms. Teachers were requested to provide in-class scaffolding and guidance when students faced challenges and difficulties. Written informed consents were obtained from all participants, who acknowledged the purpose, design, procedures, anonymity policies, and the principles of voluntary participation and unconditional withdrawal.

Table 1

Background Information and Roles of Teacher Participants

Pseudonyms	Gender	Age	Background	Role
Teixeira	Male	41	Ph.D. in language testing; L2 writing instructor	Coordinator, in-class teacher
Julia	Female	33	Ed.D. in educational technology; Digital humanity instructor	
Florence	Female	32	M.A. in English; L2 writing instructor	In-class teacher, rater
Mateo	Male	35	M.A. in English; L2 reading instructor	
Nyx	Female	36	M.A. in English; L2 reading instructor	
Priscilla	Female	31	M.A. in L2 pedagogy; L2 writing instructor	

ChatGPT Access and Version

The official GPT-3.5 version of ChatGPT from OpenAI was used in the study. However, due to restricted access to ChatGPT from mainland China, the researcher used a proxy application on the client side with a whitelist configuration redirecting *only* the access to official ChatGPT through an oversea proxy server. The reason for adopting the proxy configuration was to ensure an authentic experience with the official ChatGPT service, instead of using those third-party applications which risked manipulating the AI response or accessing a questionable source of information.

Newer versions of the product (for example, the GPT-4 version released on March 14, 2023) were not chosen due to the following reasons: (a) benchmark test showed that GPT-4 and GPT-3.5 had similar capabilities in providing ACF (Coyne & Sakaguchi, 2023) while the added features in GPT-4 such as multimodal data processing were not related to writing evaluation; and (b) GPT-3.5 was freely distributed with open access while GPT-4 was only for subscribers, making the former more applicable for the present study which had a rather tight budget.

ChatGPT Prompt Pattern

To overcome the potential bias in the focus and scope of ChatGPT-generated feedback due to different feedback seeking and ChatGPT prompting strategies, all the student participants were trained and required to use a uniform ChatGPT prompting pattern (Figure 2). All the prompts sent by students in the study followed a "*feedback requirement* + *rating criteria* + *essay*" pattern (the *essay* part in the prompt might be omitted if the ongoing human-AI conversation is context-informed without major formative amendments). In the training sessions, recommendations for ways to design feedback requirements and rating criteria have been displayed. In practice, students were encouraged and recommended to use multi-turn "trial-and-error" prompting for improvement of feedback quality (Dang et al., 2022). Since the prompting processes were iterative and incremental, the students had to decide and modify the components of the prompt based on the extent to which human-ChatGPT dialogs have developed and the

quality of the obtained AF messages.

Conditions

Students were randomly assigned into four groups: (a) individual processing (IP, control group, $n = 28$); (b) teacher-scaffolded (TS, experimental group 1, $n = 29$); (c) peer-scaffolded (PS, experimental group 2, $n = 29$); and (d) teacher-peer-scaffolded (TPS, experimental group 3; $n = 31$). As shown in Figure 3, the study ensured that the students were subjective to different modes of feedback processing *only* during comprehending, intaking, and integrating the received feedback (Bitchener & Storch, 2016). Thus, the performance measurement across the tasks and between the groups could reflect the effects of the different modes of feedback processing.

Figure 2

Sample Prompt Used by Learners to Seek Feedback from ChatGPT

1. FEEDBACK REQUIREMENT, TASK CONTEXT, AND OTHER INFORMATION

I would like you to provide feedback and evaluation on an essay writing by a second-year university English as a foreign language (EFL) learner. For the rating, you could consult the rating criteria that follows. For the feedback for writing improvement, try to focus on grammatical correctness, fluency, coherency, and the quality of the language. Try to give detailed feedback and explain any suggestions for improvement.

2. RATING CRITERIA

[*(Holistic or Partial) Rating rubric to be used in plain text*]

3. THE WRITING PRODUCT

[*The draft essay/writing product on which ChatGPT provides automated feedback*]

Figure 3

Experimental Conditions

Five stages of feedback processing (Bitchener & Storch, 2016)

<table>
<tr><th></th><th>1. Attention and Noticing of Input</th><th colspan="3">2. Comprehended Input, 3. Intake & 4. Integration</th><th>5. Output</th></tr>
<tr><th></th><th></th><th>Student</th><th>Teacher</th><th>Peer</th><th></th></tr>
<tr><td>Individual processing</td><td rowspan="4">Independently performed by the students</td><td>The students individually processed all ChatGPT-generated feedback and executed revisions based on the feedback.</td><td>N/A</td><td rowspan="2">N/A</td><td rowspan="4">Independently performed by the students</td></tr>
<tr><td>Teacher scaffolded</td><td rowspan="3">The students collaborated with other agents to process ChatGPT-generated feedback and executed revisions based on the feedback.</td><td rowspan="2">Teachers provided scaffolding on feedback seeking strategies, interpretation of the received feedback, suggestions for subsequent feedback seeking strategies, etc.</td></tr>
<tr><td>Teacher and peer scaffolded</td><td rowspan="2">Peer learners collaborated with other learners to process ChatGPT-generated feedback and offered suggestions for evaluation of the received feedback, further feedback seeking, and potential revision.</td></tr>
<tr><td>Peer scaffolded</td><td>N/A</td></tr>
</table>

The teachers scaffolded the students through (a) using modeling to offer assistance for the whole group; (b) responding to students' calls for help during the writing and revision processes; and (c) replying to students' inquiries through real-time communication platforms (e.g., WeChat, Tencent meeting). The foci of teachers' scaffolding were: (a) strategies to further prompt ChatGPT for clarification and improvement of feedback; (b) suggested ways to appreciate and potentially apply the AF messages; and (c) appropriate approaches to validate and rectify the AF messages if ambiguities and false information were identified.

The IP group was asked to self-regulate and individually process ChatGPT-generated feedback. Peer collaboration during feedback processing was prohibited. Two in-class teachers were assigned to the classroom for technical support only.

Collaborative processing of feedback (teacher, peer, or a combination) was allowed for the experimental groups. Six teachers provided face-to-face and on-demand scaffolding on feedback seeking strategies, interpretation of the received feedback, and potentially strategies to re-seek feedback from ChatGPT for TS and TPS groups. Noticeably, the revision execution processes were carried out by the writers themselves after the collaborative processing of AF.

For PS and TPS groups, students were divided into 4- or 5-member collaborative writing subgroups. Peers have been trained and encouraged to use both real-time communication platforms and collaborative office suites (that is, Shimo Docs and Nutstore Docs) for discussion of the processing of ChatGPT-generated AF. In practice, most peer collaboration and discussion were conducted as comment threads on collaborative writing websites.

Procedures

The experiment lasted for seven weeks (see Figure 4). To ensure that the experiment only compared the effects of multiple feedback processing modes, the pedagogical resources, schedules, teachers, and trainers are the same for all the participants.

The venue of all the experimental sessions was computerized language laboratories with access to ChatGPT and a collection of digital language learning resources, for example, dictionaries, thesaurus, L2

writing corpus, automatic essay grading, and so forth. It should be noticed that only ChatGPT-generated AF was used to form revision; all other resources and tools were provided only as facilitators for the teacher-student scaffolding and peer scaffolding processes.

In the learning weeks, students receive two sessions (each lasting 45 minutes) of training on ChatGPT’s usage and two sessions of self-paced practicing for each week. For example, the students were required to replicate the modeling examples teachers showed them during the training sessions.

In each of the writing and revision weeks (that is, weeks 2, 5, and 7), the students are advised to use 1 session for draft writing, 2 sessions for feedback seeking and processing, and 1 session for finalization of the writing products. All the students used ChatGPT for AF on draft writings for three tasks (T1-T3). Three weeks after the experiment, all students took a delayed test with a new topic (NT). Noticeably, all draft writing were individual, collaboration only took place during the revision of T1-T3.

Writing Tasks and Rating

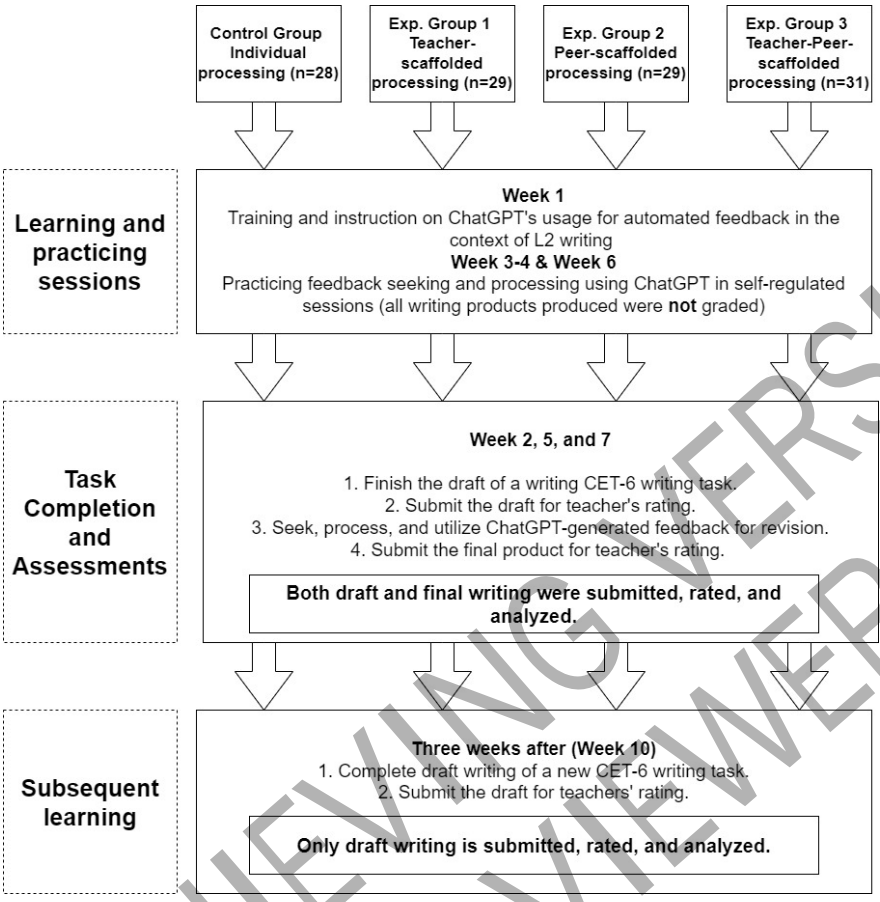
The researcher recruited a panel of five co-researchers (including three veteran teachers and two experts in educational assessments) to select appropriate writing tasks for T1-T3 and NT from College English Test Band 6 (CET-6) official test compilations through a three-round selection. In the first round, the panelists collectively clustered all the writing prompts according to the topics and corresponding perceived level of topic familiarities; then, 11 prompts with a topic related to latest social issues were chosen. In the second round, the researchers conducted a Delphi process in which the panelists individually rated the task difficulties on a scale of 10. In the final round, the three teacher experts collaboratively selected three prompts with similar task difficulties and reported to the researcher.

CET-6 writing tasks required the students to use clear, coherent, fluent English in expressing opinion on general topics, describing charts and pictures, and engaging in in-depth discussions and explanations based on outlines, charts, and other visual aids. Students were required to complete an essay of between 150 and 200 words in English following the instructions of the writing prompts. The CET-6 writing test was designed to examine the following core writing competencies: (a) idea expression; (b) discourse organization; (c) language usage; and (d) writing strategy use.

Correspondingly, the study utilized the essay rating scale proposed by Hedgcock and Lefkowitz (1992), a 100-mark rating instrument composed of five writing assessment dimensions: content, organization, grammar, vocabulary, and mechanics. The Hedgcock and Lefkowitz (1992) rating scale was used because: (a) the rating dimensions of the rating scale were in tally with the assessment objectives of the CET-6 test; and (b) the official rubric of the CET-6 writing test was a holistic rubric based on “general impression grading”, which has been criticized for its relatively low discrimination of students’ performance (Öztürk et al., 2019).

Figure 4

Procedures of the Experiment



In the present study, all the participants in the four groups were required to complete the same writing task for each of the writing assessment (i.e., T1-T3 and NT). All writing tasks were rated against the above-mentioned criteria by the 4 raters. Since all the raters were also involved in classroom instruction and scaffolding, a double-blind rating mechanism was adopted so that no student identity or grouping information was disclosed. Statistics showed that the interrater reliability was acceptable (Fleiss's κ =.84, 95%CI = [.76, .93]).

Analysis

The data were analyzed using R 4.1. I conducted standard diagnostic tests to verify the assumptions of the statistical tests. Particularly, Greenhouse-Geisser corrections were applied for violations of sphericity assessed by Mauchly's test. To answer RQ1, a 4 x 3 mixed design analysis of variance (mixed ANOVA) was performed with students' L2 writing task improvements as dependent variables and the intervention groups (between-subjects) and writing tasks (within-subjects) as independent variables. L2 writing task improvements were calculated as "gain scores" (final writing score minus draft score). The gain scores showed the extent to which the students have gained (or regressed) from their draft due to the processing of ChatGPT-generated feedback in each during-intervention writing task (that is, T1-T3). To answer RQ2, a 4x4 mixed ANOVA was conducted with students' L2 writing learning as dependent variables and the intervention groups (between-subjects) and writing tasks (within-subjects) as independent variables. L2 writing learning was measured as the development of draft writing scores for the during-intervention tasks (T1-to-T3) and the retention of the post-intervention new writing tasks (T3-to-NT). Bonferroni corrections were applied to avoid Type I errors, partial eta squared (η^2_p) was employed to assess the effect

size of the mixed ANOVA, and Cohen’s *d* was used to measure the effect size of pairwise comparisons. The interpretation of the indices of effect sizes followed the cutoffs recommended by Plonsky and Oswald (2014) and Gignac and Szodorai (2016).

Results

Descriptive statistics of rated draft writings, final products, and the corresponding task improvements are shown in Table 2**.**

Table 2

Group Means and Standard Deviations (Parenthesized) of the Draft Writing, Final Writing Scores, and Task Improvement

Conditions	T1			T2			T3			NT
	D	F	I	D	F	I	D	F	I	D
IP	61.2 (5.9)	81.2 (6.4)	20.0 (8.8)	62.5 (6.0)	84.6 (6.0)	22.1 (7.6)	61.4 (8.2)	90.2 (5.7)	28.7 (9.6)	61.9 (5.6)
PS	65.1 (6.5)	76.8 (9.4)	11.7 (11.0)	67.6 (5.9)	79.9 (5.7)	12.3 (8.2)	68.1 (4.8)	80.3 (5.0)	12.2 (7.2)	69.2 (7.8)
TS	58.8 (4.7)	71.9 (5.7)	13.1 (8.4)	65.3 (4.7)	73.5 (5.2)	8.2 (7.5)	69.4 (7.5)	81.0 (5.1)	11.6 (8.1)	71.3 (6.3)
TPS	66.2 (4.3)	73.7 (5.4)	7.5 (7.6)	69.9 (7.2)	79.6 (6.4)	9.7 (7.9)	71.3 (6.3)	83.8 (6.4)	12.5 (9.1)	73.5 (5.5)

Note: D: draft writing scores; F: final product scores; I: task improvement (F minus D).

Effects on Task Improvement

The results of the 4 (group) x 3 (tasks) mixed ANOVA performed on L2 writing task improvement, indicated by the upper half of Table 3, showed that there was a significant main effect of group ($F(3, 113) = 44.61$, $p < .001$) with a large effect size ($\eta^2_p = .542$), a significant main effect of task ($F(2,226) = 5.79$, $p = .004$) with a small effect size ($\eta^2_p = .049$), and a significant interaction effect of group x task ($F(6,226) = 2.78$, $p = .013$) with a small effect size ($\eta^2_p = .069$) on task improvement.

Appendix A shows the pairwise comparison of the effects of the between-group factor (group) on L2 writing task improvement. The results showed that the control group exhibited significantly larger improvements than the experimental groups for T1 and T2, and large ones for T3. For example, the differences between task improvement by IP and TS groups in T3 were statistically significant ($p < .001$) with a medium effect size ($d = 0.71$). Within the three experimental groups, all the pairwise comparisons of task improvements were not statistically significant. As indicated by Appendix B, all experimental group members did not make statistically significant progress in task improvement over time. However, the control group maintained a strong momentum in expanding the task improvement. For example, the IP group members’ task improvement from T2 to T3 was statistically significant ($p = .01$) with a small effect size ($d = -0.29$).

Figure 5

Developmental Trajectory of Task Improvement (A) and Learning (B).

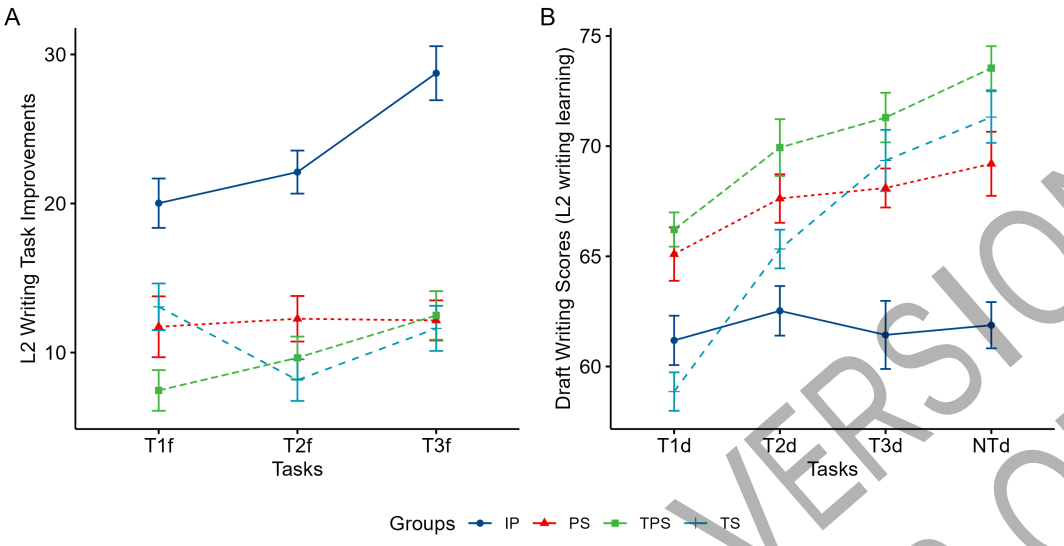


Table 3

Summary of the Mixed ANOVA Tests

Effect	df	MSE	*F*	η^2_p	*p*
Task Improvement:					
Group	3, 113	78.53	44.61	.542	<.001**
Task	2, 226	69.19	5.79	.049	.004*
Group x Task	6, 226	69.19	2.78	.069	.013*
Learning:					
Group	3, 113	38.79	37.75	.501	<.001**
Task	3, 339	38.60	21.25	.158	<.001**
Group x Task	9, 339	38.60	3.91	.094	<.001**

*Note: MSE: Mean Squared Error; * $p < .05$; ** $p < .001$.*

Effects on Writing Learning

The results of the 4 (group) x 4 (task) mixed ANOVA performed on L2 writing learning (draft writing scores), indicated by the lower half of Table 3, showed that there was a significant main effect of group ($F(3, 113) = 37.75$, $p < .001$) with a large effect size ($\eta^2_p = .501$), a significant main effect of task ($F(3,339) = 21.25$, $p < .001$) with a medium effect size ($\eta^2_p = .158$), and a significant interaction effect of group x task ($F(9, 339) = 3.91$, $p < .001$) with a small effect size ($\eta^2_p = .094$) on students' L2 writing learning.

Appendix A shows the pairwise comparison of the effects of the between-group factor (group) on L2 writing learning across time. The IP group was statistically significantly outperformed by the

experimental groups in all the tasks. For example, in the draft scores for NT, TPS group significantly ($p < .001$) excelled the IP group with a small effect size ($d = 0.66$). Statistically significant differences were not found in the pairwise comparison across the three experimental groups among the four tasks, except T1, in which TS group was significantly outperformed by PS and TPS groups with small effect sizes ($d = 0.41$ and 0.49), and T2, in which TPS group significantly ($p = .02$) surpassed the TS group with a small effect size ($d = 0.28$). As shown by Appendix B, experimental groups with teacher involvement in feedback processing (that is, the TS and TPS groups) experienced statistically significant growth in L2 writing learning for T1-T3. For example, the TS group students grew significantly ($p < .001$) from T1 to T3 with a small effect size ($d = -0.61$), indicating remarkable L2 knowledge gains during the experiment. The retention of learning progression into the new writing task (T3-NT) was not statistically significant, with experimental groups registering slightly larger effect sizes ($0.06 \leq d\text{s} \leq 0.13$, $p\text{s} > .05$).

Discussion

The study explored the effects of individual and collaborative processing of ChatGPT-generated feedback on the overall quality of L2 writing products. During the seven-week experiment, the control group members individually processed automatic feedback provided by ChatGPT while the experimental group members received peer-, teacher-, and peer-and-teacher-scaffolded collaborative processing of feedback respectively. Task improvement and learning were assessed and contrasted at multiple timepoints. In the following paragraphs, the results were interpreted and discussed against existing theoretical and empirical insights.

RQ1 sought to determine the effects of feedback processing modes on task improvement for T1-T3. The results showed that students from the control group gained and maintained the most remarkable task improvements, whereas experimental group members’ improvements were steady and less impressive. The results contradicted our knowledge on the positive effects of collaborative feedback processing on writing quality improvement (Kim & Emeliyanova, 2021). However, such differences should be interpreted in light of the different aspects involved in the provision and processing of AF. First, the improvement attained by the control group members could be explained by ChatGPT’s mighty performance in text-based tasks (Mizumoto & Eguchi, 2023; Wu et al., 2023). In tally with these recently published works on the performances of ChatGPT in assessing L2 writing products from secondary data sources, the present study offered additional insights into the impact of the tool on writing improvement in authentic pedagogical settings. Second, the dramatic improvements registered by the control group, particularly in T2 and T3, students who processed feedback in a collaborative environment would make more realistic use of AI-generated feedback. Through the lens of automated feedback processing by Liu and Yu (2022), the collaborative processing of feedback during the intake and integration stages could effectively promote learners’ strengthening their “prior knowledge” or confirming “a new use of the linguistics item” (p. 79). Comparatively, students who individually processed the feedback would make less effective attempts to match the feedback with existing knowledge or make learning advancement through hypothesis testing about the acceptability of the feedback (Bitchener & Storch, 2016). Third, the similar triviality of task improvements among the three experimental groups could be attributed to the three-phase process among collaborative writers reported by Shi et al. (2022) that involved skeptical appraisal and critical utilization of feedback. A representative example was the TS group, whose U-shape development in task improvements suggested the willingness to progressively incorporate rather than blindly use the AF.

In its entirety, the results to answer RQ1 indicated that collaborative scaffolding in L2 writing classrooms could lead to a more controlled approach to using AI-generated information. Comparatively, individually processing feedback resulted in unrealistically impressive task improvement, which would raise our ethical concerns about such usage since the quality of the revised artifacts has strayed from what the students could deliver in their drafts. This concern resembles many criticisms of unethical usage of LLM in educational settings. For example, Barrot (2023) argued that academic integrity in the L2 writing

classroom is at risk from blind usage of ChatGPT such as effortless completion of a writing task, decreasing students' creativity and critical thinking, and undetectable plagiarism. Based on the statistical results of the present study, collaborative efforts to process ChatGPT-generated information could serve as a coping strategy. However, we still need further investigation, particularly through the analysis of human behavior, discourse, and interactional patterns, into this matter.

In answering RQ2, the study tried to explore the effects of individual versus collaborative processing of ChatGPT-generated feedback on L2 writing learning. The mixed ANOVA analysis and corresponding post-hoc pairwise comparison performed on the draft scores of the three during-intervention tasks indicated that experimental groups have made more learning progress within the experiment. The results were consistent with the findings from studies by Kim and Emeliyanova (2021), Mujtaba et al. (2021), and Shi et al. (2022) that collaborative processing of feedback led to the cultivation of more competent L2 writers in longer terms. Despite the differences in feedback provision and experimental settings, the similarities in results indicated that collaborative processing of feedback would contribute to better knowledge construction and L2 writing development in subsequent learning processes. Additionally, the statistical analysis performed on draft grades revealed that the experimental groups with teacher involvement (that is, TS and TPS) outperformed the PS group whereas the IP group showed the least development in terms of L2 writing learning progression. In a sense, the results showcased that more benefit could be reaped from the integration of technology-enhanced feedback providers such as ChatGPT into a learning environment featuring teacher-student and peer interactions. Such insight echoed the claims by Koltovskaia (2020) and Zhang and Hyland (2022), where the researchers suggested a sociocultural context with teacher involvement could amplify students' use of and engagement with AF. For both groups with teacher involvement, pairwise comparison showed that the TS group has made the most significant improvement from T1 to T3, indicating the absence of peer collaboration during feedback processing resulted in better L2 writing learning progression. Such a phenomenon supported the view held by L2 researchers that the effective use of AF or AWE should be incorporated into teacher-led instruction and assessment practices (Link et al., 2022). The present research revealed that teacher presence during feedback provision could be "somewhat removed" thanks to the latest technological breakthroughs, whereas their impact on learners' *inner feedback* (see Winstone & Carless, 2019) during feedback processing remained strong and irreplaceable. The results were also in tandem with Barrot's (2023) viewpoint that human teachers were not substitutable for tasks demanding "higher-order thinking skills, contextualization, common sense, and emotions" in ChatGPT-enhanced L2 writing classrooms (p. 5).

However, it should also be noted that the post-hoc pairwise comparisons didn't find statistically significant differences for the retention of learning progression (T3-to-NT). The results could be attributed to the nature of L2 learning. It has been reported by existing studies that L2 learners, restricted by their linguistic competence, were unable to significantly improve their writing quality in a short span of time (for example, Burnell et al., 2023). Hence, to a certain degree, the data should be interpreted as an identification of the possible directions for the long-time L2 writing development of learners using different feedback processing approaches.

The present study delivered implications for research on and practice of educational feedback in the context of AI. The study pointed out that a human-centered perspective towards the application of AI in education should be encouraged and upheld in research on the impact of AI on education. Recently, researchers have been keen to explore the effects of ChatGPT on different dimensions of educational outcomes. However, in many studies, learners remained in a passive position as recipients of *feedback-as-information*. The study showed that learners, when put in a more active position handling *feedback-as-process*, would benefit more from the application of AF in pedagogy. Hence, in line with the views held by Winstone et al. (2022), researchers should investigate how AF is processed rather than delivered, from a human-centered perspective and in authentic learning environments. Taking the research findings pertinent to the two research questions together, collaborative processing of AF not only elicits students' critical application of AF but also could promote the development of L2 writing proficiency.

Correspondingly, a more inclusive instructional and feedback ecology, consisting of teacher, students, and technology-enhanced resources, for the effective uptake of AF should be encouraged and recommended in language classrooms. On the one hand, new tools and technological breakthroughs have afforded us abundant information, whose effective utilization demanded joint intelligence from all the actors involved in a pedagogical setting. On the other hand, such technologies, particularly ChatGPT and its AI-powered likes, have offered great flexibility, dynamics, and interactivity for users to communicate with the machine. Without a sociocultural and collaborative learning environment, such captivating affordances risked turning into negative effects on learning, for example, false or *hallucinated* information (Bang et al., 2023), blind reliance on AF (Koltovskaia, 2020), or simply plagiarism (Barrot, 2023).

Conclusion

The present study investigated the effects of collaborative processing of AF provided by ChatGPT on L2 writing task performance and subsequent learning. Overall, the results showed that collaborative processing of ChatGPT-generated feedback led to rationally progressive task improvements. Collaborative processing groups exhibited stronger possibilities in subsequent L2 writing development, particularly those groups with teacher collaboration during feedback processing. In line with the research findings, it is suggested that collaborative processing of AF in a sociocultural setting jointly participated by teachers and learners should be encouraged for better incorporation of AI-powered technologies in L2 classrooms.

The study was not without limitations. First, all the participants were Chinese-speaking L2 English learners recruited from a single university. The relatively homogeneous backgrounds (particularly L2 competence and prior L2 experiences) restricted the generalizability of the findings. However, given the infancy of ChatGPT’s application in L2 classrooms, such strategies facilitated training and experimental implementation. For future studies, researchers are encouraged to verify the generalizability of the present study by replicating the experiment with a more diversified sample background. Second, the study applied version 3.5 of ChatGPT with a uniform prompting pattern. Considering the information generation power of ChatGPT, the above choices might limit the quality and diversity of received feedback. For a more inclusive investigation into the effects of ChatGPT-generated feedback on L2 learning, follow-up researchers are invited to develop and apply multiple prompting patterns with updated version of the tool in their studies. Third, the study is limited by its methodological choice to use only the quantitative approach. Considering the lack of understanding of ChatGPT’s impact and usage in L2 writing settings, subsequent research is advised to adopt qualitative or mixed method approaches to continue the exploration of the details related to the usage of LLM in L2 classrooms, for example, the interactive patterns in human-AI and peer interaction during feedback seeking and processing.

Acknowledgements

This research was supported by Young Researcher Program of Xinyang Agriculture and Forestry University [QN2022049] and Xinyang Philosophy and Social Sciences Planning Project [2024JY029].

References

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity*. arXiv. https://doi.org/10.48550/ARXIV.2302.04023

Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, 36(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 1–6. https://doi.org/10.1016/j.asw.2023.100745

Bitchener, J., & Storch, N. (2016). Written corrective feedback for L2 development. Multilingual Matters. https://doi.org/10.21832/9781783095056

Burnell, K., Pratt, K., Berg, D. A. G., & Smith, J. K. (2023). The influence of three approaches to feedback on L2 writing task improvement and subsequent learning. *Studies in Educational Evaluation*, *78*, 101291. https://doi.org/10.1016/j.stueduc.2023.101291

Cheng, X., & Zhang, L. J. (2024). Engaging secondary school students with peer feedback in L2 writing classrooms: A mixed-methods study. *Studies in Educational Evaluation*, *81*, 101337. https://doi.org/10.1016/j.stueduc.2024.101337

Coyne, S., & Sakaguchi, K. (2023). *An analysis of GPT-3's performance in grammatical error correction*. arXiv. https://doi.org/10.48550/arXiv.2303.14342

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). *How to prompt? Opportunities and challenges of zero- and few-Shot learning for human-AI interaction in creative applications of generative models*. arXiv. https://doi.org/10.48550/arXiv.2209.01390

Dong, Y., & Shi, L. (2021). Using Grammarly to support students' source-based writing practices. *Assessing Writing*, *50*, 100564. https://doi.org/10.1016/j.asw.2021.100564

Fu, Q.-K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, *37*(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Guo, Q., Feng, R., & Hua, Y. (2022). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*, *35*(9), 2312–2331. https://doi.org/10.1080/09588221.2021.1879161

Han, Y. (2019). Written corrective feedback from an ecological perspective: The interaction between the context and individual learners. *System*, *80*, 288–303. https://doi.org/10.1016/j.system.2018.12.009

Hedgcock, J., & Lefkowitz, N. (1992). Collaborative oral/aural revision in foreign language writing instruction. *Journal of Second Language Writing*, *1*(3), 255–276. https://doi.org/10.1016/1060-3743(92)90006-B

Jensen, L. X., Bearman, M., & Boud, D. (2023). Feedback encounters: Towards a framework for analysing and understanding feedback processes. *Assessment & Evaluation in Higher Education*, *48*(1), 121–134. https://doi.org/10.1080/02602938.2022.2059446

Jiang, L., & Yu, S. (2022). Appropriating automated feedback in L2 writing: Experiences of Chinese EFL student writers. *Computer Assisted Language Learning*, *35*(7), 1329–1353. https://doi.org/10.1080/09588221.2020.1799824

Kim, Y., & Emeliyanova, L. (2021). The effects of written corrective feedback on the accuracy of L2 writing: Comparing collaborative and individual revision behavior. *Language Teaching Research*, *25*(2), 234–255. https://doi.org/10.1177/1362168819831406

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

Lee, I. (2020). Utility of focused/comprehensive written corrective feedback research for authentic L2 writing classrooms. *Journal of Second Language Writing*, *49*, 1–7. https://doi.org/10.1016/j.jslw.2020.100734

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Liu, S., & Yu, G. (2022). L2 learners’ engagement with automated feedback: An eye-tracking study. *Language Learning & Technology*, *26*(2), 78–105. https://doi.org/10125/73480

Lv, X., Ren, W., & Xie, Y. (2021). The effects of online feedback on ESL/EFL writing: A meta-analysis. *The Asia-Pacific Education Researcher*, *30*(6), 643–653. https://doi.org/10.1007/s40299-021-00594-6

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 1–13. https://doi.org/10.1016/j.rmal.2023.100050

Mujtaba, S. M., Reynolds, B. L., Parkash, R., & Singh, M. K. M. (2021). Individual and collaborative processing of written corrective feedback affects second language writing accuracy and revision. *Assessing Writing*, *50*, 100566. https://doi.org/10.1016/j.asw.2021.100566

Nieminen, J. H., Tai, J., Boud, D., & Henderson, M. (2022). Student agency in feedback: Beyond the individual. *Assessment & Evaluation in Higher Education*, *47*(1), 95–108. https://doi.org/10.1080/02602938.2021.1887080

Oppenlaender, J., Linder, R., & Silvennoinen, J. (2023). *Prompting AI art: An investigation into the creative skill of prompt engineering*. arXiv. https://doi.org/10.48550/arXiv.2303.13534

Öztürk, N. B., Şahin, M. G., & İlhan, M. (2019). An analysis of scoring via analytic rubric and general impression in peer assessment. *Turkish Journal of Education*, *8*(4), 258–275. https://doi.org/10.19128/turje.609073

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

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*, 1–16. https://doi.org/10.1016/j.jslw.2021.100816

Shadiev, R., & Feng, Y. (2023). Using automated corrective feedback tools in language learning: A review study. *Interactive Learning Environments*, *32*, 2358–2566. https://doi.org/10.1080/10494820.2022.2153145

Shi, Z., Liu, F., Lai, C., & Jin, T. (2022). Enhancing the use of evidence in argumentative writing through collaborative processing of content- based automated writing evaluation feedback. *Language Learning & Technology*, *26*(2), 106–128. https://doi.org/10125/73481

Tan, S., Cho, Y. W., & Xu, W. (2022). Exploring the effects of automated written corrective feedback, computer-mediated peer feedback and their combination mode on EFL learner's writing performance. *Interactive Learning Environments*, *31*(10), 7276–7286. https://doi.org/10.1080/10494820.2022.2066137

Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System*, *91*, 1–14. https://doi.org/10.1016/j.system.2020.102247

Villamil, O. S., & Guerrero, M. C. M. (2019). Sociocultural theory: A framework for understanding the socio-cognitive dimensions of peer feedback. In F. Hyland & K. Hyland (Eds.), *Feedback in second language writing: Contexts and Issues* (2nd ed.) (pp. 25–44). Cambridge University Press. https://doi.org/10.1017/9781108635547.004

Wiliam, D. (2018). Feedback: At the heart of – But definitely not all of – Formative assessment. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 3–28). Cambridge University Press. https://doi.org/10.1017/9781316832134.003

Winstone, N., Boud, D., Dawson, P., & Heron, M. (2022). From feedback-as-information to feedback-as-process: A linguistic analysis of the feedback literature. *Assessment & Evaluation in Higher Education*, *47*(2), 213–230. https://doi.org/10.1080/02602938.2021.1902467

Winstone, N., & Carless, D. (2019). Interweaving internal and external feedback: A learning-focused approach. In *Designing Effective Feedback Processes in Higher Education* (pp. 115–131). Routledge.

Wu, H., Wang, W., Wan, Y., Jiao, W., & Lyu, M. (2023). *ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical error correction benchmark*. arXiv. https://doi.org/10.48550/ARXIV.2303.13648

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, *28*(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Yan, D. (2024). Feedback seeking abilities of L2 writers using ChatGPT: A mixed method multiple case study. *Kybernetes*. Advanced online publication. https://doi.org/10.1108/K-09-2023-1933

Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities and Social Sciences Communications*, *11*(1), 1–14. https://doi.org/10.1057/s41599-024-03543-y

Yao, Y., Wang, W., & Yang, X. (2021). Perceptions of the inclusion of Automatic Writing Evaluation in peer assessment on EFL writers' language mindsets and motivation: A short-term longitudinal study. *Assessing Writing*, *50*, 100568. https://doi.org/10.1016/j.asw.2021.100568

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90–102. https://doi.org/10.1016/j.asw.2018.02.004

Zhang, Z., & Hyland, K. (2022). Fostering student engagement with feedback: An integrated approach. *Assessing Writing*, *51*, 100586. https://doi.org/10.1016/j.asw.2021.100586

Appendix A. Pairwise comparison of the effects of the between-subjects factors on task improvement and learning

		IP - PS	IP - TPS	IP - TS	PS - TPS	PS - TS	TPS - TS
T1-I	*d*	0.33	0.5	0.27	0.17	-0.05	-0.23
	p	<.001**	<.001**	.03*	0.41	0.95	0.11

T2-I	*d*	0.45	0.57	0.63	0.12	0.19	0.07
	p	<.001**	<.001**	<.001**	0.95	0.29	0.97
T3-I	*d*	0.69	0.69	0.71	-0.01	0.02	0.04
	p	<.001**	<.001**	<.001**	0.97	0.96	0.94
T1-D	*d*	-0.26	-0.33	0.15	-0.07	0.41	0.49
	p	.04*	<.001**	0.65	0.87	<.001**	<.001**
T2-D	*d*	-0.3	-0.44	-0.17	-0.14	0.14	0.28
	p	.01*	<.001**	0.49	0.96	0.91	.02*
T3- D	*d*	-0.35	-0.53	-0.41	-0.17	-0.07	0.1
	p	<.001**	<.001**	<.001**	0.42	0.94	0.97
NT-D	*d*	-0.41	-0.66	-0.53	-0.25	-0.12	0.13
	p	<.001**	<.001**	<.001**	0.06	0.97	0.95

*Note: D: draft; I: task improvement; d: Cohen’s d; p: Bonferroni adjusted p-value; * p < .05; ** p < .001.*

Appendix B. Pairwise comparison of the effects of the with-subjects factors on task improvement and learning.

	IP		**PS**		**TPS**		**TS**	
	d	***p***	***d***	***p***	***d***	***p***	***d***	***p***
T1I - T2I	-0.09	.98	-0.02	.96	-0.10	.92	0.21	.08
T1I - T3I	-0.36	<.001**	-0.02	.94	-0.22	.06	0.06	.98
T2I - T3I	-0.29	.01*	0.01	.97	-0.13	.48	-0.16	.29
T1D - T2D	-0.08	.97	-0.16	.54	-0.25	.06	-0.41	<.001**
T1D - T3D	-0.01	.98	-0.17	.39	-0.31	.01*	-0.61	<.001**
T1D - NTD	-0.04	.89	-0.24	.07	-0.45	<.001**	-0.74	<.001**
T2D - T3D	0.06	.94	-0.03	.93	-0.08	.94	-0.22	.12
T2D - NTD	0.04	.97	-0.09	.89	-0.22	.14	-0.35	<.001**
T3D - NTD	-0.02	.98	-0.06	.95	-0.13	.86	-0.11	.96

*Note: D: draft; I: task improvement; d: Cohen’s d; p: Bonferroni adjusted p-value; * p < .05; ** p < .001.*

About the Author

Da “Alex” Yan is a senior lecturer at Xinyang Agriculture and Forestry University, China. His research interests include computer-assisted language learning, peer feedback, the formative use of rubrics, and human-GenAI interactions in feedback.

E-mail: alexyan1987@outlook.com

ORCiD: https://orcid.org/0000-0002-1265-9772