تقييم فعالية التغذية الراجعة المولدة بالذكاء الاصطناعي في تحسين الكتابة الاكاديمية في اللغة الانكليزية كلغة اجنبية: دراسة ذات منهجية مختلطة

Evaluating the Effectiveness of AI-Generated Feedback in Enhancing EFL Academic Writing: A Mixed-Methods Study

د. فلاح محمد ذياب، جامعة تلعفر/ كلية التربية/ قسم اللغة الانكليزية

تاريخ الاستلام: 2026/6/4 تاريخ القبول: 2026/6/14 تاريخ النشر: 2026/6/24

المستخلص: تحقق هذه الدراسة في فعالية التغذية الراجعة المولَّدة بالذكاء الاصطناعي في تحسين جودة الكتابة الأكاديمية باللغة الإنجليزية كلغة أجنبية لدى طلبة المرحلة الجامعية الأولى في إحدى الجامعات العامة في العراق، وقد تم توظيف تصميم مختلط المنهجية. تم توزيع 112 طالبًا من طلبة السنة الأولى في اللغة الإنجليزية كلغة أجنبية عشوائيًا على أربع مجموعات: مجموعة تلقت تغذية راجعة من الذكاء الاصطناعي، ومجموعة تلقت تغذية راجعة من الأقران، ومجموعة تلقت تغذية راجعة من المعلم، ومجموعة ضابطة لم تتلق أي تغذية راجعة تكوينية منتظمة. جُمِعت البيانات الكمية من اختبارات كتابة تحليلية قبلية وبعدية تم تقييمها وفق أربعة معايير: المحتوى، والتنظيم، والمفردات، والقواعد، وحُلِّلت باستخدام تحليل التباين المشترك أحادي الاتجاه *(ANCOVA)*. أما البيانات النوعية، فجُمِعت من مقابلات شبه منظمة مع عينة طبقية فرعية مكونة من 24 طالبًا، ومن دفاتر تأمل المتعلمين، وحُلِّلت باستخدام التحليل الموضوعي التأملي. أظهرت نتائج تحليل التباين المشترك وجود تأثير كلي دال إحصائيًا لنوع التغذية الراجعة *($F(3, 107) = 31.62, p < .001, \eta p^2 = .538$)*، وأظهرت الاختبارات البعدية أن التغذية الراجعة من المعلم كانت الأكثر فعالية، تلاها التغذية الراجعة من الذكاء الاصطناعي، والتي تفوقت بشكل دال على كلٍّ من تغذية الأقران والمجموعة الضابطة. يشير التحليل الموضوعي للبيانات النوعية إلى أن التغذية الراجعة المولَّدة بالذكاء الاصطناعي يمكن أن تكون مكملاً عمليًا وقابلاً للتوسع للتغذية الراجعة من المعلم في فصول الكتابة باللغة الإنجليزية كلغة أجنبية، لا سيما في السياقات محدودة الموارد. كما أن لهذه النتائج مضامين مهمة لتصميم المناهج الدراسية ولدمج الثقافة الرقمية في تدريس الكتابة باللغة الإنجليزية كلغة أجنبية، نظرًا لأنها تسلط الضوء على إمكانات التغذية الراجعة المولَّدة بالذكاء الاصطناعي في دعم تعلم الطلبة.

كلمات مفتاحية: التغذية الراجعة المولَّدة بالذكاء الاصطناعي، الكتابة الأكاديمية باللغة الإنجليزية كلغة أجنبية، التقييم الآلي للكتابة، التغذية الراجعة التكوينية، الكتابة باللغة الثانية.

Evaluating the Effectiveness of AI–Generated Feedback in Enhancing EFL Academic Writing: A Mixed–Methods Study

Abstract:

This study explores the effectiveness of AI-generated feedback in enhancing the quality of EFL academic writing among undergraduate students at a public university in Iraq, and a concurrent mixed-methods design was employed. A total of 112 first-year EFL students were randomly assigned to four groups: one receiving AI feedback, one receiving peer feedback, one receiving teacher feedback, and one control group that did not receive any systematic formative feedback. The quantitative data were collected from analytic pre- and post-writing tests that were scored on four criteria: content, organization, lexis, and grammar, which were analyzed using a one-way ANCOVA, while the qualitative data were collected from semi-structured interviews with a stratified sub-sample of 24 students, as well as learner reflection journals, and analyzed using reflexive thematic analysis. The ANCOVA results showed that there was a statistically significant overall effect of feedback type ($F(3, 107) = 31.62$, $p < .001$, $\eta p^2 = .538$), and the post-hoc tests showed that teacher feedback was the most effective, followed by AI feedback, which significantly outperformed both peer feedback and the no-feedback control group. Consequently, the thematic analysis of the qualitative data suggests that AI-generated feedback can be a practical and scalable complement to teacher feedback in EFL writing classes, particularly in contexts with limited resources. Moreover, the findings have important implications for curriculum design and for the integration of digital literacy into EFL writing pedagogy, because they highlight the potential of AI-generated feedback to support student learning.

Keywords: *AI-generated feedback, EFL academic writing, automated writing evaluation, formative feedback, second language writing.*

1. Introduction

The emergence of AI as a sustainable teaching tool has led to a rethinking of how formative feedback can be delivered in language classrooms, particularly with regards to EFL learners writing in English for academic purposes, a cognitively demanding and culturally and disciplinarily sensitive language skill that requires feedback which is timely, specific, and clearly actionable if students are to make progress (Hyland & Hyland, 2006; Ferris, 2011). This has traditionally been the responsibility of the teacher, and in this respect, there is an ongoing tension between the pedagogical ideal of rich, individualised feedback and the practical realities of large class sizes, limited institutional resources, and the time-intensive nature of teaching (Bitchener & Storch, 2016). It is here that AI-powered automated writing evaluation (AWE) represents a radical departure from previous practices, because the AWE system is capable of analysing students' writing within seconds and providing feedback on grammar, cohesion, vocabulary, and overall organisation, and can thus bridge the long-standing gap between what teachers would like to provide and what is possible in many EFL writing classrooms.

Despite the rapid increase in interest in the role of AI in assessing second language writing, the academic discussion is still in its infancy, while most of the current research has been conducted on learners' perceptions of AWE tools such as Grammarly, Turnitin Feedback Studio and ChatGPT-based revision prompts. Only a handful of studies have systematically compared the effectiveness of AI feedback to the impact of traditional feedback sources, namely peer and teacher feedback, on actual writing gains in controlled quasi-experimental designs (Ranalli, 2018; Stevenson & Phakiti, 2014; Zhang & Hyland, 2023). Moreover, this emerging literature is heavily skewed towards Western ESL contexts, which raises significant questions as to the generalisability of these results to EFL

contexts with very different learner profiles, pedagogical traditions and levels of technological access (Canagarajah, 2013). In particular, the Iraqi tertiary EFL education sector has been entirely absent from the AWE literature, because it represents a large and pedagogically distinct group of learners for whom written English is increasingly tied to academic and professional opportunities, but whose opportunities for rich face-to-face feedback are severely limited by high instructor-to-student ratios.

1.1 Statement of the Problem

Although formative feedback is a critical component of EFL students' academic writing development, teachers in public universities in Iraq face the challenge of providing prompt, quality feedback to students due to the large number of students in their classes and the lack of resources, and as a result, a great many EFL students are offered insufficient feedback on their writing to revise it properly. AI-based Automated Writing Evaluation (AWE) tools, which provide immediate and detailed feedback on students' writing, are likely to fill the gap in feedback, however, there is little empirical research on the effectiveness of AWE tools in the context of Iraqi higher education. Thus, it is vital to explore whether AI-based feedback is effective in enhancing EFL students' academic writing and how it compares with traditional teacher feedback and peer feedback.

1.2 Research Questions

The present study was therefore designed to address this gap by investigating the following research questions:

RQ1: To what extent does AI-generated feedback produce differential gains in EFL academic writing quality compared with peer feedback, teacher feedback, and a no-feedback control condition?

RQ2: How do EFL learners perceive and engage with AI-generated feedback in terms of noticing, uptake, and transfer to independent writing?

RQ3: What affective and metacognitive dimensions of learner experience mediate the relationship between AI feedback engagement and writing development?

1.3 Hypotheses

To achieve the aim of the present study, the following hypotheses are formulated:

H1: The four conditions (AI feedback, peer feedback, teacher feedback, no feedback) will differ in their second writing assessment scores and teacher feedback will perform the best. Followed by AI feedback, and followed by peer feedback, and followed by no feedback.

H2: Students receiving AI feedback will perform better on the second writing assessment compared to students receiving peer feedback.

H3: Students receiving teacher feedback will perform better on the second writing assessment compared to students receiving feedback.

2. Literature Review

2.1 Feedback in Second Language Writing: Theoretical Review

The theoretical rationale for feedback in second language writing is rooted in both cognitive and sociocultural perspectives. Truscott's (1996) claims that written corrective feedback was not only ineffectual but also detrimental to learners sparked a heated debate that, while later refuted empirically, helped inform the field's understanding of the varying types of feedback, how learners process feedback, and under what conditions correction results in enduring acquisition

(Ellis, 2009; Ferris, 2010). From a cognitive perspective, Schmidt's (1990) Noticing Hypothesis suggests that feedback promotes acquisition by raising learners' conscious awareness of the difference between their output and the target form. More recent sociocultural perspectives, drawing on Vygotsky's (1978) notion of the zone of proximal development, construe feedback as a type of mediated scaffolding that allows learners to function beyond their current level of ability (Lantolf & Thorne, 2006). While these perspectives differ in their explanations of how feedback works, they share one fundamental assumption: learner engagement is a necessary condition, and therefore, feedback that is not noticed, actively processed, and internalized is unlikely to foster development, regardless of its accuracy or good intentions (Bitchener & Knoch, 2010).

Furthermore, the debate on what constitutes effective feedback has been complicated by distinctions between direct and indirect feedback, form-focused and meaning-focused response, and written vs. oral mode of delivery. In a landmark collection of classroom-based studies, Hyland and Hyland (2006) found that learners generally value teacher-written feedback, but they tend to misunderstand it, apply it inconsistently across drafts, and receive it in small amounts because of time constraints on teachers. To mitigate this problem, peer feedback was incorporated into process-oriented writing pedagogy as a complementary means to teacher response, and the results have been mixed. Supporters of peer feedback suggest that the dialogic and collaborative nature of peer feedback allows for negotiation of meaning and learner autonomy (Rollinson, 2005), whereas critics argue that peer feedback is restricted by peers' own language proficiency and assessment literacy, which can affect the accuracy, depth, and usefulness of the feedback they provide (Lundstrom & Baker, 2009).

2.2 Automated Writing Evaluation and AI Feedback

Although AWE systems have come a long way since the scoring engines of the 1990s (Shermis & Burstein, 2003), contemporary AWE tools, especially those built on large language models, do not merely provide holistic scores, but rather, they generate extensive discursive feedback on measures of syntactic complexity, lexical sophistication, use of cohesive devices and the quality of argumentation (Ranalli, 2018; Wilson & Czik, 2016). One of the first comparative studies of AWE and teacher feedback in an EFL university context was conducted by Stevenson and Phakiti (2014), and they found that learners who received AWE feedback developed similar levels of grammatical accuracy to learners who received teacher feedback, however, the AWE group saw a more restricted gain in discourse-level organization. More recently, Zhang and Hyland (2023) explored how learners in a Hong Kong university writing course responded to AI-generated feedback from Grammarly, and they found that learners heavily engaged with lower-order feedback on grammar and spelling, but were much less likely to act on higher-order feedback on argument structure and academic register.

The coming of generative AI, in particular large language models that can provide context-sensitive, text-specific feedback in natural language, however, has revived many of the empirical questions that early AWE research had only begun to explore. Scholars such as Cavaleri and Dianati (2016) and, more recently, Farrokhnia et al. (2024) have argued that this latest generation of AI feedback may overcome some of the communicative shortcomings of earlier AWE systems by providing responses that are closer to the dialogic, interactive quality of expert teacher feedback, yet there is a lack of robust evidence from controlled studies that directly compare the acquisitional effects of AI and teacher

feedback, particularly in contexts where learners' level of digital literacy and experience with AI tools may influence how they interact with and profit from such feedback (Golonka et al., 2014).

2.3 Learner Affect and Writing Feedback Engagement

A fourth strand of feedback research that has received relatively little systematic attention relates to the affective determinants of learners' responses to, and uses of, feedback. Writing anxiety, which Daly and Miller (1975) define as a trait-like negative affective reaction to writing tasks, has been found to be consistently associated with avoidance behaviours, lower willingness to revise, and shallower processing of corrective feedback (Cheng, 2004). In EFL settings, where academic writing is at once a high-stakes assessment area and a cognitively-linguistically demanding activity, anxiety regulation is not a marginal issue but rather a core pedagogical concern (Tsui & Ng, 2000). Several writers have suggested that AI-generated feedback, due to its perceived impersonality, affective neutrality and lack of evaluative positionality, may mitigate the anxiety induced by teacher correction and thus promote more prolonged, intense engagement with revision feedback (Dikli & Bleyle, 2014), and the qualitative component of the current study seeks to verify this hypothesis.

In brief, while previous research has been fruitful, it has not been enough, and the theoretical rationale of AI-based feedback is well-established, and the preliminary empirical evidence is guardedly positive. There is, however, a dearth of comparative quasi-experimental studies using adequately-sized samples, robustly-validated measures, and embedded qualitative components, especially in Global South EFL contexts, and this study seeks to fill this void by contributing empirical findings from an Iraqi university context and utilising a concurrent mixed-methods design that captures both the quantitative effect of different

feedback types on writing quality and the qualitative nuances of learners' experiences.

3. Methodology

3.1 Research Design

The present study adopts a concurrent mixed-methods design (Creswell & Plano Clark, 2018), where a quasi-experimental quantitative strand and a phenomenologically informed qualitative strand were carried out in parallel, but only integrated at the interpretation stage. The quantitative strand employed a four-group pre-test/post-test design to examine the effect of different feedback conditions on the quality of students' EFL academic writing, while the qualitative strand employed semi-structured interviews and learner reflection journals to gain insights into the cognitive and affective aspects of students' interaction with feedback. A concurrent rather than a sequential design was used, as this would allow the two strands to develop separately, and therefore the risk of early findings in one strand influencing data collection in the other could be minimized (Morse & Niehaus, 2009). Integration occurred at the meta-inference stage, where quantitative effect estimates were interpreted in the light of the qualitative accounts of learners' experiences.

3.2 Participants

Participants were 112 first-year undergraduate students (58 female, 54 male; M age = 19.6, SD = 1.4) studying Academic Writing in English at the University of Telafer, Iraq during the spring 2025 semester, and they were grouped into four

groups based on their intact class enrolments: Group 1 received AI-generated feedback (n = 28), Group 2 received peer feedback (n = 28), Group 3 received teacher feedback (n = 28), and Group 4 served as a control group and received only task instructions without formative feedback (n = 28). All participants had been assessed at the Common European Framework Reference for Language (CEFR) B1 level by institutional placement testing, and written informed consent was obtained from all participants, while ethical approval was granted by the University of Telafer. For the qualitative strand, 24 participants were purposively selected from the four groups (six per group) using a maximum variation sampling strategy (Patton, 1990) based on post-test writing performance quartiles.

3.3 Instruments

The quality of writing was assessed by an analytic scoring rubric developed from Jacobs et al.'s (1981) ESL Composition Profile, which includes four subscales: Content (30 points), Organization (25 points), Lexical Resource (20 points), and Grammatical Accuracy (25 points), and the maximum score was 100 points. The rubric was piloted with a set of 20 writing scripts not included in the main data set, and the pilot results showed satisfactory inter-rater reliability (Cohen's κ = .81, $p < .001$). AI feedback was generated through a structured prompt administered via the university's licensed large language model interface, which was designed to provide specific, criteria-referenced written comments on each submission, focusing on the same four subscales as the scoring rubric. The qualitative data were collected through semi-structured interviews lasting around 30-40 minutes per participant, and participants also submitted weekly reflection journal entries via the course management platform. Interviews were guided by emergent themes from the literature and were conducted in a private room to help reduce social desirability bias.

3.4 Data Collection Procedures

Participants carried out four graded writing tasks during the 14 weeks following the pre-test administered in Week 1, and each task required the participants to write a 300-400 word argumentative essay on a topic from the disciplinary curriculum. For each task, the participants in the AI feedback condition submitted their draft through the LLM interface, and the feedback generated by LLM was returned to the students through the course platform within 24 hours. In contrast, the participants in the peer feedback condition exchanged essays with a randomly assigned classmate and then provided comments by using a structured feedback protocol that was aligned with the scoring rubric. Meanwhile, the participants in the teacher feedback condition received handwritten marginal comments on their drafts from the course instructor within 48 hours of submission, whereas the participants in the control group had their scripts returned with no annotations. Consequently, the post-test essays were collected in Week 16 and were scored by two trained raters who were blind to the participants' group membership.

3.5 Data Analysis

Quantitative data were analysed by one-way analysis of covariance (ANCOVA), with the post-test total writing scores as the dependent variable and pre-test scores entered as a covariate to adjust for differences in the level of writing proficiency. The assumptions of normality (Shapiro-Wilk test), homogeneity of variance (Levene's test) and homogeneity of regression slopes were checked and satisfied before running the ANCOVA. In the case of significant main effects, post-hoc pairwise comparisons were conducted with Bonferroni correction. The effect sizes were reported as partial eta-squared (ηp^2), and to explore the impact of different types of feedback on different aspects of writing, ANCOVAs were also performed for each rubric subscale. Qualitative data, including interview transcripts and

reflection journal entries, were analysed through Braun and Clarke's (2006) six-phase reflexive thematic analysis, where the researcher went through the phases of familiarisation, coding, theme generation, review, definition and naming iteratively to increase the credibility of the analysis. Member checking was conducted with six participants after the preliminary themes were identified, and they were given the opportunity to provide comments on the resonance and accuracy of the interpretations.

4. Results

4.1 Quantitative Findings

Descriptive statistics for the pre- and post-test writing scores of the four groups are reported in Table 1. Pre-test scores of the four groups were not significantly different ($F(3, 108) = 0.11$, $p = .954$), indicating that the four groups were statistically equivalent in writing proficiency before the experiment, and as can be seen in Table 1, the adjusted mean gain of the three experimental groups on the post-test was much larger than that of the control group. The teacher feedback group scored the largest adjusted mean gain (29.5 points), followed by the AI feedback group (23.7 points), the peer feedback group (15.7 points), and the control group (2.2 points).

Table 1

Descriptive Statistics for Pre- and Post-Test Writing Scores by Feedback Condition

Group	n	Pre-Test M (SD)	Post-Test M (SD)	Mean Gain
AI Feedback	28	41.6 (5.8)	65.3 (6.4)	23.7

Peer Feedback	28	42.1 (6.1)	57.8 (7.1)	15.7
Teacher Feedback	28	41.9 (5.6)	71.4 (5.9)	29.5
Control Group	28	42.0 (6.0)	44.2 (6.3)	2.2

Note. Scores range from 0 to 100. M = mean; SD = standard deviation. Mean Gain = unadjusted difference between post-test and pre-test means. Pre-test group differences were non-significant ($p = .954$).

The results of the ANCOVA are shown in Table 2. A main effect of feedback condition was found after controlling for pre-test scores, $F(3, 107) = 31.62$, $p < .001$, $\eta p^2 = .538$, which is a large effect size, and pre-test scores also had a significant effect as a covariate, $F(1, 107) = 76.18$, $p < .001$, $\eta p^2 = .490$, providing evidence for statistically controlling initial proficiency. Bonferroni-adjusted post-hoc tests showed that the teacher feedback group had a significantly higher adjusted post-test score than all other groups, all $ps < .01$, while the AI feedback group had a significantly higher score than the peer feedback group, $p = .004$, and the control group, $p < .001$, but not the teacher feedback group, $p = .062$. The peer feedback group also had a significantly higher score than the control group, $p = .009$.

Table 2

ANCOVA Summary: Effect of Feedback Condition on Post-Test Writing Scores

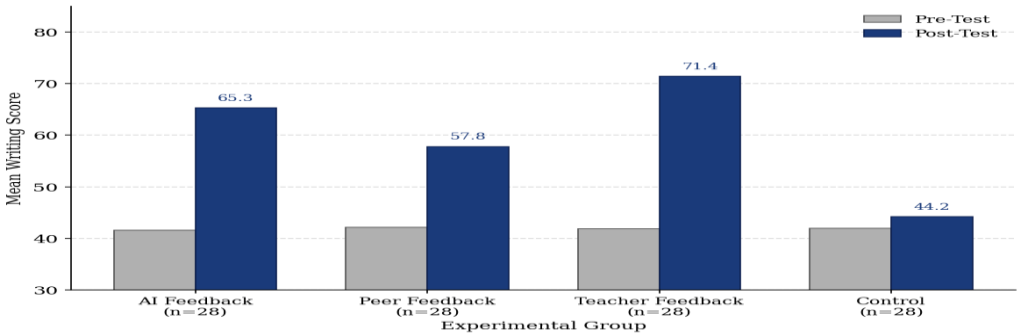
Source	DF	MS	F	P	ηp²
Feedback Condition	3	891.4	31.62	< .001	.538
Pre-test (Covariate)	1	2148.7	76.18	< .001	.490
Error	107	28.19	—	—	—

Note. ηp² = partial eta-squared. Pre-test writing score served as the covariate. Post-hoc comparisons used Bonferroni correction. — = not applicable.

Figure 1 displays the mean pre- and post-test writing scores for each group, and it can be seen that the magnitude of gains in writing varied across feedback conditions. Figure 2 shows the trend of the mean writing scores over the four testing times of the intervention, and it is clear that the mean writing scores of both the teacher feedback and AI feedback groups increased gradually over time, while the mean writing scores of the control group improved slightly over time. The subscale analysis showed that AI feedback was particularly effective in Grammatical Accuracy (M gain = 7.4), while teacher feedback resulted in relatively larger gains in Content and Organization.

Figure 1

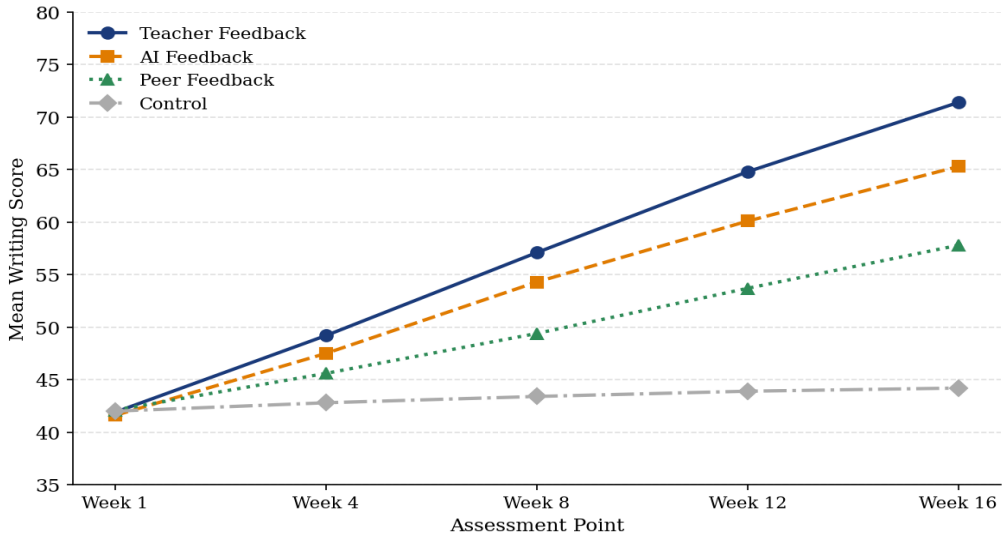
Mean Pre- and Post-Test Writing Scores by Feedback Condition



Note. Scores represent unadjusted group means. Error bars have been omitted for clarity. AI = AI-generated feedback condition; Peer = peer feedback condition; Teacher = teacher-written feedback condition; Control = no formative feedback.

Figure 2

Mean Writing Score Trajectories Across Assessment Points by Condition



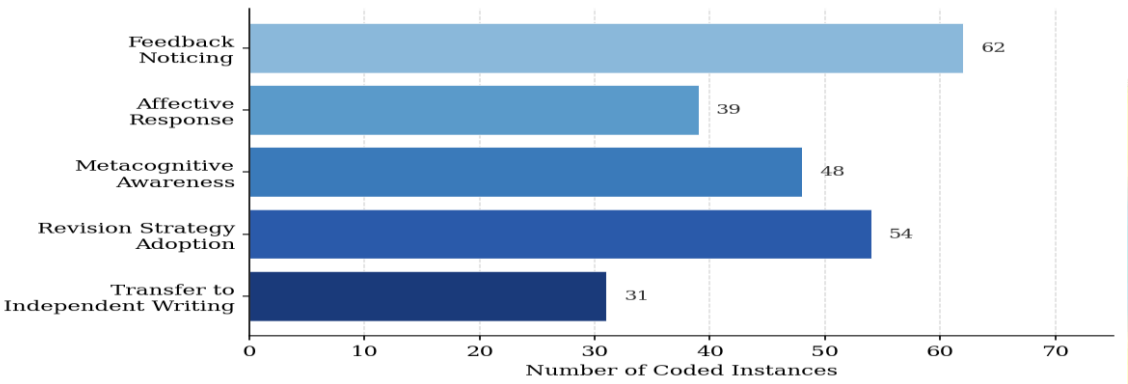
Note. Scores represent group means at each assessment point (Weeks 1, 4, 8, 12, and 16). Data points are connected for visual clarity and do not imply continuous measurement.

4.2 Qualitative Findings

Five key themes emerged and these were (1) Perceptions of Feedback, (2) Affect towards Feedback, (3) Beliefs about Learning, (4) Feedback Use, and (5) Feedback Use for Self-Writing. The frequency of occurrence of each theme is presented in Fig. 3.

Figure 3

Frequency of Coded Thematic Instances Across Interview and Journal Data



Note. Frequency counts represent the total number of distinct coded instances across all 24 participants' interview transcripts and reflection journals. Themes are ordered by frequency.

The most frequently coded theme was Theme 1, Feedback Noticing (n = 62), because noticing was found to be much more present in the journal entries of the AI feedback group. Noticing was evident in statements such as "AI feedback noted that I repeated the same linking word 5 times in a paragraph and I had not noticed this previously." Theme 2, Affective Response, revealed a clear distinction between the two feedback groups, while students who received teacher feedback reported a range of emotional responses, including appreciation, anxiety, and self-consciousness, although the AI feedback group reported a consistent less threatening experience, as one student noted, "It wasn't as threatening as my teacher marking in red." The qualitative findings for Themes 3-5 are presented in

Table 3, which includes frequency counts and example quotes, thus providing further insight into the study's results.

Table 3

Summary of Qualitative Themes, Sub-themes, Representative Excerpts, and Frequencies

Theme	Sub-theme	Illustrative Excerpt	Fre
1. Noticing	Explicit error awareness	"I could see exactly where my grammar was wrong"	62
2. Affect	Reduced revision anxiety	"It felt less threatening than my teacher marking in red"	39
3. Metacognition	Self-monitoring strategies	"I started checking my own work against the feedback"	48
4. Uptake	Immediate revision	"I corrected the error straight away and rewrote the clause"	54
5. Transfer	Independent application	"I caught a similar mistake myself before submitting"	31

Note. Frequency counts represent coded instances across all 24 participants. Excerpts are translated from Arabic where applicable and lightly edited for clarity. AI = AI feedback group; T = teacher feedback group; P = peer feedback group.

4.4 Discussion

The results provided robust support for H1, which hypothesised that there would be significant differences among the four feedback conditions, because the findings revealed a statistically significant main effect of feedback type, $F(3, 107) = 31.62$, $p < .001$, $\eta p^2 = .538$, as presented in Table 2, indicating that the feedback condition accounted for a substantial proportion of the variance in the post-test writing scores. The post-hoc comparisons with Bonferroni adjustments showed a clear hierarchy of effectiveness among the four feedback conditions: teacher feedback was associated with the largest gains, followed by AI-generated feedback, and then peer feedback, with the no-feedback control group performing the worst.

The findings of the current study are in line with a large number of previous studies, which have demonstrated the importance of formative feedback in the development of L2 writing (e.g. Ferris, 2011; Bitchener & Storch, 2016), therefore, the qualitative data in this study also suggest that the relative effectiveness of each type of feedback was closely associated with how the learners noticed and processed their writing problems. Although all the feedback conditions contributed to the learners' noticing to some extent, the participants generally perceived teacher feedback as the most contextualised and personalised, AI feedback as the most systematic and detailed, and peer feedback as the most variable and unreliable; however, the finding that peer feedback was still more effective than the no-feedback control group provides additional support for Schmidt's (1990) Noticing Hypothesis, which claims that noticing, or the conscious attention to linguistic forms, is a necessary condition for language acquisition.

The results support H2, predicting that AI-generated feedback would lead to greater gains in writing quality than peer feedback, because post-hoc comparisons revealed that the students receiving AI feedback had significantly greater gains in overall writing performance than the students receiving peer feedback. The qualitative data also provide further insight into why AI feedback was more effective than peer feedback, as students perceived AI-generated comments as more specific, consistent, and actionable, which enabled them to revise their drafts with greater clarity and confidence. By contrast, peer feedback was often seen as less reliable, and many students questioned the linguistic accuracy and utility of their classmates' suggestions, while the reflection journals further suggested that peer reviewers tended to focus mainly on surface-level aspects, such as grammar and spelling, whereas AI feedback provided feedback on multiple aspects of writing, including content, organization, vocabulary, and grammatical accuracy, in a more balanced and comprehensive manner. Furthermore, students described AI feedback as private and non-judgmental, which appeared to alleviate anxiety and encourage more sustained engagement with the revision process, and therefore, these findings suggest that AI feedback offers a more efficient and effective substitute for peer feedback, particularly for novice EFL writers in under-resourced higher education settings. Teacher feedback was significantly more effective overall, thereby supporting H3.

The qualitative evidence suggests several explanations for the superiority of teacher feedback. Students perceived teacher comments as more personalized, situated, and motivational, which apparently strengthened their affective investment in revising their work. Teachers also tended to provide metacognitive guidance that went beyond simple error correction, helping learners develop broader, transferable strategies for planning, drafting, and revising academic texts. While AI feedback largely facilitated local revisions, such as correcting

grammatical errors, rephrasing sentences, teacher feedback was more likely to prompt global changes, including idea development, coherence and cohesion, paragraph organization, rhetorical appropriateness. Therefore, the results imply that AI feedback is beneficial as a supplementary tool in EFL writing instruction, but has not yet replaced the pedagogical, interpersonal, and strategic support that expert human teachers offer.

These findings notwithstanding, there are a number of limitations worth mentioning. The quasi-experimental design with intact classes is subject to selection bias, which undermines the validity of causal inferences. The single-institution design restricts the generalizability of the findings, and the 16-week period, although longer than the majority of similar studies, does not enable us to comment on the long-term sustainability of the effects of AI feedback. The study examined only one AI feedback delivery model, and the results may not be generalized to other AWE platforms with different interface designs, degrees of feedback detail, or NLP capabilities; therefore, future research should redress these limitations by adopting randomized designs, recruiting participants from multiple institutions, and conducting longitudinal follow-up assessments.

5. Conclusion

This study provides empirical support for the claim that AI-generated feedback can provide an effective and sustainable complement to teacher-led formative feedback in EFL academic writing courses. A concurrent mixed-methods design was used with 112 Iraqi undergraduate EFL students to demonstrate that AI feedback led to significantly greater gains in writing quality than both peer feedback and a no-feedback control condition, and also generated gains that approached - but did not fully match - those achieved through expert teacher feedback. Qualitative data also suggest that feedback noticing, reduced evaluative

anxiety, and heightened metacognitive awareness served to mediate the relationship between engagement with AI feedback and writing development. These findings collectively provide a cohesive explanation for why AI feedback may be especially useful in high-anxiety, resource-constrained EFL writing contexts.

Theoretically, the study contributes by integrating cognitive noticing theory, sociocultural scaffolding research, and writing anxiety research into a single, coherent explanation of why AI feedback can be effective. Pedagogically, the results suggest that EFL writing teachers should use AI feedback tools as a structured complement to teacher feedback, especially at stages of the writing process, such as initial draft submission, at which students need rapid, criterion-referenced responses most, because for curriculum designers, the findings highlight the importance of considering the emotional and affective aspects of the AI feedback interface. The findings also highlight the importance of incorporating explicit feedback literacy instruction to help students make better use of comments, especially on higher-order, discourse-level aspects of writing, on which AI feedback was found to be relatively less effective in this study.

As AI tools become increasingly prevalent in higher education globally, research is needed to empirically yet contextually examine their influence on EFL writing development, and the current study contributes to this growing research area, yet also highlights the need for more comparative, longitudinal and cross-contextual research to build a more robust evidence base for future instructional recommendations.

References

Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. Journal of Second Language Writing, 19(4), 207–217. https://doi.org/10.1016/j.jslw.2010.10.002

Bitchener, J., & Storch, N. (2016). Written corrective feedback for L2 development. Multilingual Matters.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Canagarajah, A. S. (2013). Literacy as translingual practice: Between communities and classrooms. Routledge.

Cavaleri, M., & Dianati, S. (2016). You want me to check your grammar again? The usefulness of an online grammar checker as perceived by students. Journal of Academic Language and Learning, 10(1), A223–A236.

Cheng, Y. S. (2004). A measure of second language writing anxiety: Scale development and preliminary validation. Journal of Second Language Writing, 13(4), 313–335. https://doi.org/10.1016/j.jslw.2004.07.001

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE.

Daly, J. A., & Miller, M. D. (1975). The empirical development of an instrument to measure writing apprehension. Research in the Teaching of English, 9(3), 242–249.

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? Assessing Writing, 22, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Ellis, R. (2009). A typology of written corrective feedback types. ELT Journal, 63(2), 97–107. https://doi.org/10.1093/elt/ccn023

Farrokhnia, M., Palalas, A., Banihashem, S. K., & Noroozi, O. (2024). ChatGPT in educational contexts: A systematic literature review on its applications, perspectives, and challenges. Education and Information Technologies, 29(5), 5939–5980. https://doi.org/10.1007/s10639-024-12497-6

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. Studies in Second Language Acquisition, 32(2), 181–201. https://doi.org/10.1017/S0272263109990490

Ferris, D. R. (2011). Treatment of error in second language student writing (2nd ed.). University of Michigan Press.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. Computer Assisted Language Learning, 27(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Hyland, K., & Hyland, F. (Eds.). (2006). Feedback in second language writing: Contexts and issues. Cambridge University Press.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). Testing ESL composition: A practical approach. Newbury House.

Lantolf, J. P., & Thorne, S. L. (2006). Sociocultural theory and the genesis of second language development. Oxford University Press.

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. Journal of Second Language Writing, 18(1), 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

Morse, J. M., & Niehaus, L. (2009). Mixed method design: Principles and procedures. Left Coast Press.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? ELT Journal, 72(1), 42–51. https://doi.org/10.1093/elt/ccx059

Rollinson, P. (2005). Using peer feedback in the ESL writing class. ELT Journal, 59(1), 23–30. https://doi.org/10.1093/elt/cci003

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

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. Assessing Writing, 19, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Truscott, J. (1996). The case against grammar correction in L2 writing classes. Language Learning, 46(2), 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? Journal of Second Language Writing, 9(2), 147–170. https://doi.org/10.1016/S1060-3743(00)00022-9

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. Computers & Education, 100, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Zhang, Z. V., & Hyland, K. (2023). Fostering student engagement with feedback: An integrated approach. Assessing Writing, 45, 100471. https://doi.org/10.1016/j.asw.2020.100471