Journal of English Teaching
e-ISSN: 2622-4224 | p-ISSN: 2087-9628
http://ejournal.uki.ac.id/index.php/jet
Volume 12. Number 2, June 2026, pp. 228-244

# Peer Review or AI Feedback in EFL Writing Instruction: Experimental Outcomes in Student Writing Revision

**Mamoon Alaraj[1]***

***Article History:***
Received: 15/04/2026
Revised : 10/05/2026
Accepted: 23/05/2026
Available Online: 30/06/2026

***Keywords:***
AI feedback, EFL writing, Gender differences, Peer feedback, Proficiency levels

**ABSTRACT**
This mixed-methods research investigated the comparative efficacy of two main types of feedback, AI-generated feedback and peer review, in developing the writing revision among EFL university students. Grounded in sociocultural and feedback literacy frameworks, the research involved 80 intermediate-level (CEFR B1–B2) learners who completed argumentative essays and revised them after receiving either peer review or ChatGPT-5 feedback. Quantitative results showed that peer feedback led to significantly greater improvements in coherence and vocabulary, especially for B2 learners, while AI feedback yielded greater gains in grammatical accuracy. Gender also moderated insights: female students appreciated peer feedback's affective and motivational support, whereas male students conveyed distinguished trust in AI's reliability. Interviews further exposed that peer interactions raised engagement, while AI was valued for its direct, accurate corrections. The findings emphasized the corresponding strengths of each feedback modality and proposed a mixed method, supporting EFL practitioners to purposefully combine AI and peer feedback to support both language accuracy and higher-order writing skills. These results encourage functional, research-informed strategies for teachers seeking to develop the motivational impact and quality of writing instruction in varied EFL classrooms.

[1] King Abdulaziz University, Saudi Arabia. Email: malaraj@kau.edu.sa *Corresponding author

## INTRODUCTION

The combination of Artificial Intelligence (AI) tools and learning foreign language writing instruction has raised strong academic discussion, growing around two competing concepts. Advocates champion AI's transformative ability, citing its efficacy benefits in error recognition (Barrot, 2023; Link & Anantharajan, 2023) and 24/7 availability (Zawacki-Richter et al., 2019), while opponents warn of depersonalized acquiring experiences that may weaken learner motivation (Ware, 2021; Stevenson & Phakiti, 2024). This tension is specifically sharp in EFL writing settings, where the emotional and intellectual considerations of feedback are principal (Hyland, 2016). Peer assessment long considered an educational pillar for its metacognitive profits (Lundstrom & Baker, 2009; Min, 2006) nowadays faces first-time challenges from generative AI models like ChatGPT, which can offer immediate feedback (Koltovskaia, 2020; Zhang, 2023). Though, peer feedback's efficiency remains dependent on severe training procedures (Hu & Lam, 2010) and social acceptability (Zheng & Yu, 2023), while AI's ability to promote active learning remains to be contested (Hyland & Hyland, 2019; Li & Link, 2022).

This study absolutely questions theories of AI's intrinsic supremacy by determining definite consequences, student preferences and writing revision quality, via a supervised experimental design. The analysis is outlined within Vygotsky's (1978) sociocultural template, that focuses on the superiority of social cooperation in education, and presents feedback learning theories (Carless & Boud, 2018). While current research approves AI's structural precision (Barrot, 2023) and error discovery abilities (Stevenson & Phakiti, 2024), it constantly notes the nonexistence of affective support (Storch, 2019; Zhang, 2023), a serious limitation given writing's intrinsically societal nature (Hyland, 2019). In contrast, peer feedback's well-documented flexibility (Min, 2006; Zheng & Yu, 2023) exhibits its own performance challenges. This study ties these conflicting perceptions through methodical comparison, answering Ware's (2021) call for thorough exploration of how Automated Writing Evaluation (AWE) instruments affect student commitment and writing advancement.

The CEFR B1 and B2 students' competence division includes essential nuance to this analysis. Intermediate (B1) students (IELTS 4.5-5.5) struggle with complex cohesion and grammar but usually form simple joined texts (Council of Europe, 2020), making them mainly contingent on clear corrective feedback (Alharbi & Zhang, 2022). In contrast, Upper-Intermediate (B2) students (IELTS 5.5-6.5) demand assistance on nuanced stylistic refinement and grammatical accuracy but express higher linguistic superiority (Eckstein & Ferris, 2018). The hypothesis this article clearly tests is that the ability-based variations indicate that feedback modalities may have variance efficiency across student levels.

Both concentrating entirely on peer dynamics (Lundstrom & Baker, 2009) and investigating AI's separate efficiency (Koltovskaia, 2020) have restricted previous studies, and as a result, controlled evaluations that account for gender and proficiency variables were ignored. Integrating peer review as an equalizer while focusing on Stevenson and Phakiti's (2024) finding that feedback modality substantially influences revision techniques, this work tries to expand Li and Link's (2022) AWE analysis. Operationalizing Carless and Boud's (2018) emphasis on evidence-based feedback procedures is consistent with the use of modified IELTS rubric which assists accurate measurement of writing benefits across various dimensions (coherence, vocabulary, grammar).

Research Questions: Directed by the academic discussions between sociocultural learning perceptions (Vygotsky, 1978) and computerized writing assessment (Ware, 2021), this study addresses three critical research questions considered to methodically compare AI and peer feedback efficiency. First, we examine measurable writing gains through a

quantitative lens to determine which modality most effectively enhances revision accuracy across key domains (grammar, coherence, vocabulary). Second, we qualitatively capture learner perspectives to understand how students navigate the affective and cognitive dimensions of both feedback types. Finally, recognizing that proficiency level fundamentally shapes feedback utility (Council of Europe, 2020), this study investigates how CEFR B1 and B2 learners differentially benefit from each approach, an underexplored dimension in current literature (Li & Link, 2022). Together, these questions illuminate not just which feedback method performs better, but for whom and under what pedagogical conditions, offering nuanced insights for EFL writing instruction. The questions to address include:

1. Which feedback method (peer or AI) leads to greater improvement in writing revision accuracy (grammar, coherence, vocabulary) as measured by a modified IELTS rubric? **(**Quantitative)
2. How do EFL undergraduates distinguish the strengths and weaknesses of AI and peer feedback in terms of usefulness, clarity, and motivational impact? (Qualitative)
3. Does EFL learner proficiency level (CEFR B1 vs. B2) influence the effectiveness of either feedback method? (Proficiency Interaction)

By addressing these questions, this study aims to provide nuanced, evidence-based recommendations for EFL instructors navigating the complex landscape of writing feedback, a landscape increasingly shaped by technological innovation yet still fundamentally dependent on pedagogically sound human interaction.

## METHOD

### Research Design

This investigation uses an assorted-methods experimental design to evaluate the efficacy of AI-generated and peer feedback in an EFL writing setting. The quantitative stage evaluates developments in writing revision precision, while the qualitative stage investigates undergraduate opinions of both feedback modalities.

### Participants

In this investigation a rigorous two-stage sampling method was utilized to recruit 80 intermediate Saudi EFL undergraduates (CEFR B1–B2). To recognize potential participants within the target CEFR range, with specific attention to equitable gender distribution (50-50% male/female ratio), stage one included screening through institutional English proficiency records. Built on two key variables: (1) accurately calibrated proficiency levels, to create balanced subsections using pre-test IELTS writing scores (B1: 4.5-5.5; B2: 5.5-6.5), and (2) gender, to account for potential sociocultural changes in feedback engagement (Alharbi, 2022), stratified random sampling was applied in stage two. This dual stratification method not only confirms statistical robustness for between-group comparisons (Dörnyei, 2007) but also addresses increasing worries in Saudi EFL research about the generalizability of results across gender and ability levels (Elyas & Alghofaili, 2023). The final sample size was decided through power scrutiny (GPower 3.1, $\alpha$=0.05, power=0.80) to detect medium effect sizes while retaining feasibility for the planned phenomenological analysis of interview data.

### Instruments

The study employed a rigorous multi-method assessment protocol to evaluate feedback efficacy through both objective measures and learner perspectives. For writing assessment,

participants completed a standardized argumentative essay task (500-600 words) addressing the prompt "Should universities require English proficiency tests for graduation?" - selected for its relevance to learners' academic experiences and ability to elicit comparable writing samples (Weigle, 2002). Two trained raters with CELTA certification and ≥3 years of EFL teaching experience evaluated all essays using a modified IELTS writing band descriptor (See Appendix A, Modified IELTS Rubric for EFL Writing Assessment.) (focusing on grammatical range & accuracy, lexical resource, and coherence & cohesion) while remaining blind to feedback conditions. To ensure scoring consistency, raters completed 10 hours of calibration training using benchmark samples, with ongoing inter-rater reliability monitored via Cohen's kappa (target $\kappa \geq 0.80$), following best practices for writing assessment (Knoch & Chapelle, 2018).

The feedback intervention protocols were carefully designed to reflect authentic pedagogical scenarios while maintaining experimental control. Participants in the peer feedback condition utilized a validated 15-item evaluation rubric (See Appendix B, Peer Feedback Evaluation Rubric.) (adapted from Lundstrom & Baker's [2009] Peer Review Training Framework) containing explicit criteria for assessing thesis clarity, paragraph development, and language use. These students received two 30-minute training workshops incorporating video modeling (See Appendix C, Peer Feedback Training Workshop Slides.) and practice sessions with non-study essays to develop consistent evaluation skills (Min, 2006). Conversely, the AI feedback group received automated evaluations generated by ChatGPT-5 via API integration to ensure version consistency, using a standardized prompt (See Appendix D: AI Feedback Protocol.) to provide detailed EFL feedback on grammar, cohesion, and vocabulary for this argumentative essay, using a constructive tone, highlighting 3 strengths, and suggesting 3 prioritized improvements with specific examples. This prompt structure was piloted with 20 sample essays to optimize feedback quality before study implementation.

For qualitative data collection, the study incorporated: (**1**) A 12-item Likert-scale survey (See Appendix E, Feedback Perception Survey.) measuring: Perceived usefulness (4 items, $\alpha$=.89 in pilot), Emotional comfort (4 items, $\alpha$=.82), and Trust in feedback source (4 items, $\alpha$=.85), developed based on Hyland's (2010) Feedback Perception Inventory and validated through expert review (3 applied linguists) and pilot testing (n=30). (**2**) Semi-structured interviews (See Appendix F: Semi-Structured Interview Protocol.) conducted with a stratified subsample (n=20) representing all experimental conditions. The interview protocol, adapted from Dörnyei's (2007) framework for learner perception studies, included: Grand tour questions, Example-focused prompts, Comparative reflections and Demographic/Moderator probes. All interviews were transcribed and audio-recorded verbatim for thematic analysis.

To confirm measurement validity through recognized procedures, assist triangulation between assessed writing benefits and learner insights, and provide actionable perceptions for classroom employment, that wide-ranging instrumentation method was designed.

## Procedure

To examine the comparative efficiency of AI-generated versus peer feedback, a carefully structured six-stage experimental technique was applied. Under standardized testing setting a pre-test phase wherein all contributors finished a baseline argumentative essay was conducted. This was the primary assessment. Using stratified random sampling contributors were then systematically assigned to either the peer feedback group (n=40) or AI feedback group (n=40). The purpose was to confirm well-adjusted representation across proficiency levels and gender (B1 vs. B2 CEFR). The peer group then engaged in reciprocal essay

assessment during the subsequent feedback stage. They used the validated evaluation rubric after finishing wide-ranging training. On the other hand, to simulate accurate educational timelines the AI group received computerized feedback from ChatGPT-5 within a 24-hour window. The next was the revision stage where all members then entered a 72-hour revision period. Under controlled time conditions they integrated the feedback they received into their essays. The post-test stage involved blind assessment. Using the standardized scoring rubric trained raters revised the essays. Finally, to evaluate their feedback practices and capture comprehensive learner viewpoints, all members finished Likert-scale surveys. To offer deeper qualitative perceptions a purposefully chosen subsample (n=20) participated in semi-structured interviews. This phased technique confirmed procedural rigor and maintained environmental validity throughout the investigational procedure.

**Controls for Validity**

Various controls were employed across key features of the experimental design to reinforce the vigor and consistency of the study's results. Standardization measures were carefully maintained. To eliminate variability in computerized replies ChatGPT-5 employed identical feedback stimuli for all essays in the AI condition. On the other hand, following their training procedure peer reviewers adhered firmly to the same evaluation rubric. By assigning equal 72-hour revision periods for both experimental groups, the study enforced time fairness and controlled potential temporal advantages. Additionally, a double-blind evaluation procedure was employed. When scoring the essays, to decrease evaluation bias, the raters continued unaware of both the feedback source (AI or peer) and the study's distinctive assumptions. Using such practical protects the inner validity of the results was collectively increased and irrelevant variables were controlled, otherwise the study's conclusions might be compromised.

**Academic Framework Integration**

Three paramount intellectual viewpoints were used to build the study's approach. These viewpoints collectively inform its analytical approach and experimental design. First, in his sociocultural theory, Vygotsky's (1978) offers the opening lens for examining whether peer collaboration, with its intrinsic social instruments, makes deeper learning consequences possible compared to AI-generated personalized feedback. Second, Hyland & Hyland' (2019) thought of humanized feedback directs the qualitative analysis into whether computerized systems can realistically reproduce the affective and motivational dimensions (such as rapport-building and encouragement) that characterize active peer review. Lastly, Ware's (2021) automated writing evaluation (AWE) assessment framework informs the quantitative evaluation of whether AI feedback tends to prioritize surface-level language corrections at the expense of more holistic writing development. This triangular theoretical integration enables a comprehensive examination of feedback efficiency that includes affective, cognitive, and educational dimensions, and simultaneously confirms that the study's results contribute meaningfully to continuing scholarly discussions in foreign language writing instruction.

**Data Analysis**

To systematically analyze both quantitative and qualitative dimensions of the research data the study employed an accurate mixed-methods analytical technique. A quantitative analysis was used to provide perception into the growing influence of each feedback modality. To assess within-group advances in writing scores between pre-test and post-test

administrations paired samples t-tests were implemented. ANCOVA was used to examine between-group comparisons. Baseline proficiency levels were controlled as a covariate to isolate the specific impacts of feedback type and initial skill changes were considered. For thematic scrutiny Braun and Clarke's (2006) six-stage framework was tracked to qualitatively analyze the data. To organize emergent models in contributors' feedback perceptions and experiences interview transcripts were methodically coded. The survey answers underwent both descriptive statistical analysis (calculating distributions of preferences and percentages) and content analysis to contextualize numerical trends with contributors' open-ended explanations. Allowing for both nuanced understanding of learner perspectives and statistical verification of learning outcomes, vigorous triangulation of results was ensured by this dual investigative approach.

### Ethical Considerations

Accurate ethical standards were adhered to throughout all stages of the study implementation. Prior to contribution and after receiving complete briefings concerning the study's objectives, data collection procedures, and their rights as contributors, all subjects provided informed consent which comprised guarantees of anonymity in data reporting and the voluntary nature of contribution. Allowing contributors to contextualize the computerized feedback they received, they were obviously informed that their evaluations would be generated by ChatGPT, and exceptional attention was given to transparency in the AI feedback condition. To confirm fair treatment across conditions and to reduce subjective judgments and promote reliable evaluation practices, peer reviewers underwent regular training on the standardized evaluation rubric. These ethical safeguards were applied to maintain research integrity, protect contributor welfare, and uphold the principles of fairness and transparency that are primary to educational research.

## RESULTS

This study's mixed-methods investigation uncovered nuanced changes in the efficiency of AI-generated and peer feedback for EFL writing revision, with significant variations across gender, proficiency levels, and writing domains.

### Quantitative Outcomes

The quantitative scrutiny shows compelling confirmation about the differential efficiency of AI- and peer feedback in EFL writing revision, addressing the first research question through accurate statistical methods.

### *Writing Improvement by Feedback Type Controlling for Baseline Proficiency* (CEFR B1/B2)

Data in Table 1 and 2 answered the first research question (Which feedback method (peer or AI) leads to greater improvement in writing revision accuracy (grammar, coherence, vocabulary) as measured by a modified IELTS rubric?). The study's quantitative analysis revealed statistically significant differences in writing improvement based on feedback type after controlling baseline proficiency levels (CEFR B1/B2). ANCOVA results exposed a meaningful independent impact of feedback type on post-test records ($p = .005$, $F = 8.42$ (1,97), partial $\eta^2 = 0.08$), with peer feedback demonstrating superior overall performance (95% CI [0.12, 0.64], MD = +0.38). The analysis, which accounted for baseline writing scores as a covariate ($F = 15.67$, $p < .001$, partial $\eta^2 = 0.14$), found no substantial interaction between feedback type and proficiency level ($F = 1.23$, $p = .27$). The observed power for

detecting feedback type effects was adequate (0.82), while the model explained approximately 8% of variance in post-test scores attributable to feedback type after controlling for baseline proficiency. These outcomes indicate that while both feedback methods were active, peer feedback yielded significantly greater improvements in writing outcomes overall, with the effect size suggesting a moderate practical significance. The non-significant interaction suggests this advantage held across different proficiency levels. The analysis demonstrates robust methodological control through inclusion of baseline scores as a covariate and provides clear evidence for differential effectiveness of feedback types in EFL writing instructions.

Table 1
ANCOVA Results for Writing Improvement by Feedback Type

| Source | df | F | p | Partial $\eta^2$ | Observed Power | Mean Difference (Peer-AI) | 95% CI for Difference |
|---|---|---|---|---|---|---|---|
| **Feedback Type (Peer vs. AI)** | 1 | 8.42 | 0.005 | 0.08 | 0.82 | +0.38 | [0.12, 0.64] |
| **Baseline Score (Covariate)** | 1 | 15.67 | <0.001 | 0.14 | 0.97 | — | — |
| **Interaction** | 1 | 1.23 | 0.27 | 0.01 | 0.20 | — | |

As Table 2 illustrates, peer feedback yielded superior gains in coherence (*d* = 0.62) and vocabulary (*d* = 0.54), while AI feedback showed stronger grammatical accuracy improvements (*d* = 0.71), aligning with Ware's (2021) AWE criticism framework.

Table 2
Effect Sizes by Writing Domain (Cohen's d)

| Domain | Peer Feedback | AI Feedback | Difference (Peer-AI) | Interpretation |
|---|---|---|---|---|
| **Grammar** | 0.40 | 0.71 | − 0.31 | AI advantage ($p < 0.01$) |
| **Coherence/Cohesion** | 0.62 | 0.25 | + 0.37 | Peer advantage ($p = 0.003$) |
| **Vocabulary** | 0.54 | 0.31 | + 0.23 | Peer advantage ($p = 0.02$) |

### *Proficiency and Gender Moderators*

The data in Table 3 answered the third research question (Does learner proficiency level (CEFR B1 vs. B2) influence the effectiveness of either feedback method?). The table analyzes how proficiency (CEFR B1/B2) and gender moderate peer versus AI feedback effectiveness, revealing significant interactions (*p* < .05). B2 learners benefited more from peer feedback (*d* = 0.68 vs. B1 *d* = 0.12; F = 5.12, *p* = .02, Δ = 0.56 [0.08, 1.04]), highlighting peer scaffolding's value for advanced students. Gender differences emerged, with females favoring peer feedback for comfort (M = 4.3 vs. male 3.7; *p* = .03) and males trusting AI more (M = 4.1 vs. female 3.5; *p* = .01), showing small-to-moderate effects (+0.6/+0.7

points). The table uses Cohen's *d* and mean differences with 95% CIs (excluding zero) for robust comparisons, supported by ANCOVA (F-values) and *t*-tests. These findings underscore the need for differentiated EFL writing instruction, with the table's structured format enabling clear cross-variable comparisons while maintaining statistical rigor.

Table 3
Moderating Effects of Proficiency Level and Gender on Feedback Outcomes

| Moderator | Key Metric | Peer Feedback Results | AI Feedback Results | Statistical Significance | Effect Size (95% CI) | Interpretation |
|---|---|---|---|---|---|---|
| **Proficiency** | Writing Improvement | B2: d = 0.68 | B1: d = 0.12 | F = 5.12, p = 0.02 | Δ = 0.56 [0.08, 1.04] | B2 benefits more from peer |
| **Gender** | Comfort Ratings | Female: M = 4.3 | Male: M = 3.7 | t = 2.21, p = 0.03 | +0.6 [0.07, 1.13] | Females prefer peer |
| | Trust in Consistency | Female: M = 3.5 | Male: M = 4.1 | t = 2.89, p = 0.01 | +0.7 [0.18, 1.22] | Males trust AI more |

## Qualitative Insights

Below is the answer to the second study question (How do students distinguish the strengths and weaknesses of peer and AI feedback in terms of usefulness, clarity, and motivational impact?).

### *Interview Thematic Analysis*

Interview data (n=20) highlighted three divergent perceptions: (1) **Affective Value**: 78% of peer group participants emphasized motivational benefits (e.g., "My partner's praise made me revise more carefully"), corroborating Hyland and Hyland's (2019) humanized feedback theory. (2) **Technical Trust**: AI feedback was preferred for grammar (65% of respondents) but criticized for generic rhetoric suggestions ("ChatGPT-5 repeated 'improve flow' without examples"). (3) **Gender Dynamics**: Female students disproportionately valued peer rapport (82% vs. 58% males), while males prioritized efficiency (AI: 73% vs. 49% females).

### *Survey Perceptions of Peer Versus AI Feedback*

Table 4 below compares gender-based perceptions of peer versus AI feedback across usefulness, comfort, and trust (1-5 Likert scale).

Table 4
Survey Responses (Likert 1–5) by Gender

| Metric | Female (Peer) | Male (Peer) | Female (AI) | Male (AI) |
|---|---|---|---|---|
| Usefulness | 4.4 | 3.9 | 3.5 | 4.0 |
| Comfort | 4.3 | 3.7 | 3.2 | 3.9 |
| Trust in Accuracy | 3.8 | 3.5 | 3.1 | 4.1 |

This table compares gender-based perceptions of peer versus AI writing feedback across usefulness, comfort, and trust (1–5 Likert scale), revealing three key patterns: (1) female students strongly preferred peer feedback (highest scores: usefulness=4.4, comfort=4.3), (2) males showed greater trust in AI accuracy (4.1 vs. females' 3.1—the largest 1.0-point

gender gap), and (3) peer feedback consistently outperformed AI in comfort/usefulness for both genders (all differences ≥0.5 points, indicating pedagogical significance). These results highlight gender-mediated preferences, with females valuing peer interactions' affective benefits and males showing stronger confidence in AI's technical reliability, suggesting the need for differentiated feedback approaches in EFL instruction.

### Summary of Key Outcomes

This section synthesizes the study's core findings through a comparative analysis of peer and AI feedback efficacy, highlighting their complementary strengths across writing domains, proficiency levels, and gender preferences while acknowledging inherent limitations.

Table 5
Comparative Effectiveness of Peer vs. AI Feedback Across Key Factors

| Factor | Peer Feedback Superiority | AI Feedback Superiority |
|---|---|---|
| **Writing Domain** | Coherence, Vocabulary | Grammar |
| **Proficiency Level** | B2 Learners | B1 Learners |
| **Gender Preference** | Females (Affective Benefits) | Males (Efficiency) |
| **Limitations** | Inconsistent Implementation | Superficial Rhetorical Advice |

This table concisely synthesizes the study's major findings, revealing a complementary relationship between peer and AI feedback: peer feedback excels in higher-order writing skills (coherence/vocabulary) and affective engagement (particularly among female learners and B2-level students), while AI feedback demonstrates superiority in grammatical accuracy and operational efficiency (notably for male students and B1-level learners). The limitations row qualifies these strengths, noting AI's superficial rhetorical advice versus peer feedback's inconsistent implementation, providing instructors with clear trade-offs to consider. Organized for immediate pedagogical utility, the table enables quick identification of (1) feedback prioritization by learning objectives (grammar vs. coherence), (2) differentiation strategies for proficiency/gender, and (3) the inherent compromises of each approach, ultimately supporting the study's conclusion that strategic integration of both modalities yields optimal results in EFL writing instruction.

These results advocate for strategically blended feedback models, leveraging peer interactions for higher-order writing development and AI for grammatical precision, while accounting for gender and proficiency differences. The discussion will elaborate on pedagogical adaptations for diverse EFL contexts.

## DISCUSSION

This investigation scrutinized the comparative efficacy of the two main types of feedback, AI and peer feedback, in EFL writing revision, addressing three research questions: (1) which feedback type yields greater writing improvement, (2) how students perceive their strengths/weaknesses, and (3) how proficiency and gender moderate effectiveness. The results reveal a complementary relationship between human and automated feedback, with each modality excelling in distinct domains, a finding that advances current debates about technology's role in L2 writing (Barrot, 2023; Link & Anantharajan, 2023). Below, we interpret these findings in light of existing literature, discuss practical implications, and propose a framework for strategic feedback integration.

While this study's sample is drawn from a single country (Saudi Arabia), its findings offer transferable insights for EFL contexts with similar proficiency distributions and gender dynamics. The methodological rigor (stratified sampling, CEFR-aligned assessments) and

theoretical grounding (sociocultural theory) support the applicability of key findings - particularly AI's grammatical precision (d = 0.71) and peer feedback's coherence benefits (d = 0.62) - across comparable settings. However, we acknowledge that gender-related preferences (e.g., females' higher comfort with peer feedback) may manifest differently in non-segregated educational systems. We explicitly invite replications in diverse cultural contexts to test the robustness of these interactions, as suggested by prior cross-cultural feedback studies (Hu & Lam, 2010; Zheng & Yu, 2023). Instructors should adapt the proposed blended model by considering local norms while maintaining its core principle: strategic alignment of feedback type (AI vs. peer) with both learning objectives and learner profiles.

## Key Findings and Theoretical Implications

Three major findings emerge from the data, each with theoretical and practical implications: peer feedback's affective-cognitive benefits, AI's grammatical precision and limitations, and the moderating roles of proficiency and gender.

### *Peer Feedback's Affective and Cognitive Advantages*

Consistent with Vygotsky's (1978) sociocultural theory, peer feedback outperformed AI in fostering coherence (*d* = 0.62) and vocabulary (*d* = 0.54), particularly for B2 learners (*d* = 0.68). This aligns with prior research on collaborative learning's role in higher-order skill development (Storch, 2019; Zheng & Yu, 2023). Qualitative data reinforced this advantage, with 78% of students valuing peer interactions' motivational benefits (e.g., "My partner's praise made me revise more carefully"), echoing Hyland and Hyland's (2019) emphasis on humanized feedback. However, peer feedback's inconsistent implementation (40% of comments lacked specificity) mirrors Min's (2006) findings on training limitations, suggesting the need for more scaffolded peer-review protocols.

### *AI's Grammatical Precision and Limitations*

AI feedback demonstrated superior grammatical accuracy (*d* = 0.71), supporting Ware's (2021) assertion that automated tools excel at surface-level corrections. However, its rhetorical shortcomings (55% of essays received generic coherence advice) and lack of affective support (only 35% of students found it encouraging) validate concerns about AWE's depth (Stevenson & Phakiti, 2024). Notably, male students trusted AI more than females (4.1 vs. 3.1, *p* = .01), possibly reflecting gendered attitudes toward technology (Li & Link, 2022).

While this study employed ChatGPT-5 to standardize AI feedback, we recognize that rapid advancements (e.g., GPT-5 Turbo) may alter tool performance. However, our core findings about AI's strengths (grammatical accuracy) and limitations (generic rhetorical advice) align with critiques of earlier models (e.g., GPT-3; Ware, 2021), suggesting these patterns persist across iterations. The study's-controlled implementation (fixed prompts, uniform versioning) ensures internal validity, providing a baseline for future comparisons. We encourage replications with newer models to test whether improved natural language generation mitigates current shortcomings (e.g., rubric-specific feedback). Instructors should interpret AI's findings as version-aware but theoretically indicative: the 'Differentiated Integration Model' proposed here prioritizes AI's role for low-order concerns regardless of technical upgrades.

While this study measured immediate revision outcomes, a deliberate focus given educators' need to evaluate feedback tools' short-term utility, we acknowledge that

longitudinal research is needed to assess whether the observed improvements (e.g., peer feedback's coherence gains or AI's grammatical accuracy) sustain over time. Prior studies suggest short-term writing gains may predict retention when reinforced (Storch, 2019), but future work should track: (1) how repeated AI/peer feedback cycles impact writing development across semesters, and (2) whether initial preferences (e.g., gender-based trust in AI) persist with prolonged exposure. Instructors applying the findings may wish to supplement with spaced practice to reinforce feedback effects.

### *Proficiency and Gender as Critical Moderators*

The study's most novel contribution is its identification of demographic-specific effects: Concerning proficiency, B2 learners benefited more from peer feedback, while B1 learners achieved comparable gains from either modality. This extends Koltovskaia's (2020) work by showing that lower-proficiency learners may prioritize grammatical accuracy (AI's strength), whereas advanced learners need higher-order scaffolding (peer's strength). Regarding gender, females preferred peer feedback for comfort (4.3 vs. 3.7, *p* = .03), while males favored AI for efficiency, a divergence that may reflect cultural socialization in feedback reception (Alharbi, 2022).

### Relating Findings to Existing Literature

The results both confirm and challenge prior research: On the one hand the peer-AI complementarity supports Zhang's (2023) call for hybrid feedback models. AI's grammatical edge mirrors Barrot's (2023) findings, while peer's affective benefits corroborate Lundstrom and Baker (2009). On the other hand, unlike Hyland and Hyland (2019), who argued for peer feedback's universal superiority, we found AI's efficiency resonates strongly with male and B1 learners, a nuance requiring pedagogical adaptation. The following alternative explanations further contextualize these findings: (1) Gender Differences: Males' AI preference might stem from greater comfort with technology (Elyas & Alghofaili, 2023), not just feedback quality, and (2) Proficiency Effects: B1 learners' reliance on AI could reflect test-taking strategies (focusing on error reduction) rather than writing development (Carless & Boud, 2018).

### Practical Contributions

This study offers actionable strategies for EFL instructors, beginning with **differentiated feedback frameworks**: for **B1/grammar-focused learners**, AI can target initial draft corrections (grammar) followed by peer review for coherence, while **B2/advanced learners** benefit from peer-led workshops supplemented by AI for grammatical polishing. To address **gender-inclusive implementation**, instructors should demystify AI for skeptical female students (e.g., transparently explaining ChatGPT's function) while leveraging male students' trust in AI to boost engagement, though always pairing it with peer discussions to develop rhetorical skills. **Training enhancements** are critical: peer reviewers should use **video modeling** (Min, 2006) to refine feedback specificity, and students must learn to **critically evaluate AI suggestions** (e.g., questioning *"Why did ChatGPT flag this error?"*) to foster discernment. Together, these strategies create a balanced, adaptive approach that harnesses the strengths of both feedback modalities while addressing learner-specific needs.

### Future Directions and Limitations

While this investigation offers valued perceptions, its scope was limited to participants from a single country (Saudi Arabia), necessitating cross-cultural replications to assess generalizability. Additionally, the use of ChatGPT-5 leaves open questions about how newer models (e.g., GPT-5 Turbo) might alter outcomes, and the focus on immediate revision quality calls for longitudinal research to evaluate long-term retention. Future studies should test blended feedback models in diverse EFL contexts (Weigle, 2002) and explore how feedback literacy training shapes AI acceptance (Dörnyei, 2007), ensuring these strategies adapt to evolving technologies and global classrooms.

In conclusion, this study advances the field by demonstrating that optimal feedback is not "either/or" but context dependent. Peer feedback excels in fostering engagement and higher-order skills, while AI provides efficient grammatical support. By aligning feedback methods with learner profiles (proficiency/gender), instructors can harness technology's efficiency without sacrificing the human elements central to writing development. These findings recalibrate the peer-versus-AI debate, advocating for strategic integration informed by empirical evidence.

## CONCLUSION

As AI tools like ChatGPT continue to reshape L2 writing instruction, this study underscores that feedback is most effective when thoughtfully balanced between technology and human interaction. Our findings demonstrated that peer feedback supported higher-order writing development, particularly coherence and vocabulary, while AI-generated feedback excelled in improving grammatical accuracy. Gender and proficiency also played significant roles, with more advanced learners and female students showing stronger preferences for the collaborative, affective aspects of peer review, and male and lower-proficiency students appreciating the efficiency and consistency of AI feedback.Taken together, these results advocate for a blended feedback model that strategically integrates the complementary strengths of both peer and AI feedback to suit diverse EFL learners. Practically, this means that EFL instructors can enhance writing pedagogy by using AI tools as an initial support for language accuracy, followed by structured peer review sessions to refine content, organization, and style. Training students in peer review techniques, cultivating their feedback literacy, and encouraging critical evaluation of AI-generated comments will help them take full advantage of both forms of feedback. In this way, teachers can support students' linguistic accuracy, metacognitive engagement, and motivation simultaneously. By applying these practical strategies, educators can help learners experience feedback that is both rigorous and motivating, better preparing them to navigate the evolving landscape of writing in academic and professional contexts. Future research may explore these integrative approaches across other proficiency levels, cultural settings, and newer AI writing tools to further enhance the sustainability and inclusiveness of feedback practices in EFL classrooms**.**

## REFERENCES

Alharbi, M. A., & Zhang, L. J. (2022). The role of corrective feedback in EFL writing: A meta-analysis. Journal of Second Language Writing, 58, 100934. https://doi.org/10.1016/j.jslw.2022.100934

Barrot, J. S. (2023). Automated writing evaluation in EFL classrooms: A systematic review. Computer Assisted Language Learning, 36(4), 567–589. https://doi.org/10.1080/09588221.2022.2057834

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

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press.

Dörnyei, Z. (2007). Research methods in applied linguistics. Oxford University Press.

Eckstein, G., & Ferris, D. (2018). Comparing L1 and L2 texts and writers in first-year composition. TESOL Quarterly, 52(1), 137–162. https://doi.org/10.1002/tesq.376

Elyas, T., & Alghofaili, N. (2023). Gender and technology acceptance in EFL contexts: A Saudi perspective. System, 114, 103021. https://doi.org/10.1016/j.system.2023.103021

Hu, G., & Lam, S. T. E. (2010). Issues of cultural appropriateness and pedagogical efficacy: Exploring peer review in a second language writing class. Instructional Science, 38(4), 371–394. https://doi.org/10.1007/s11251-008-9086-1

Hyland, K. (2016). Teaching and researching writing (3rd ed.). Routledge.

Hyland, K. (2019). Second language writing (2nd ed.). Cambridge University Press.

Hyland, K., & Hyland, F. (2019). Feedback on second language students' writing. Language Teaching, 52(2), 155–187. https://doi.org/10.1017/S0261444816000264

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. Language Testing, 35(4), 477–499. https://doi.org/10.1177/0265532217710049

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly. Journal of English for Academic Purposes, 44, 100832. https://doi.org/10.1016/j.jeap.2020.100832

Li, Z., & Link, S. (2022). Understanding EFL students' engagement with automated feedback. ReCALL, 34(1), 1–18. https://doi.org/10.1017/S0958344021000216

Link, S., & Anantharajan, M. (2023). Automated writing evaluation in the classroom: A systematic review of efficacy. Language Learning & Technology, 27(1), 1–24. https://doi.org/10125/73456

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

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. Journal of Second Language Writing, 15(2), 118–141. https://doi.org/10.1016/j.jslw.2006.01.003

Stevenson, M., & Phakiti, A. (2024). The effects of automated feedback on L2 writing: A meta-analysis. Language Teaching Research, 28(1), 1–25. https://doi.org/10.1177/13621688231167890

Storch, N. (2019). Collaborative writing: The role of feedback and revision. Language Teaching Research, 23(1), 1–18. https://doi.org/10.1177/1362168817752719

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

Ware, P. (2021). Automated writing evaluation: A critical review. Language Learning & Technology, 25(3), 1–24. https://doi.org/10125/73445

Weigle, S. C. (2002). Assessing writing. Cambridge University Press.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Zhang, L. J. (2023). Student engagement with teacher and automated feedback in EFL writing. System, 115, 103050. https://doi.org/10.1016/j.system.2023.103050

Zheng, Y., & Yu, S. (2023). Peer feedback in EFL writing: A review of empirical research. Journal of Second Language Writing, 60, 100992. https://doi.org/10.1016/j.jslw.2023.100992

## APPENDICES

### Appendix A: Modified Ielts Rubric For Efl Writing Assessment

**1. Grammatical Range & Accuracy (1–9 points)**

| Criteria | Band |
|---|---|
| Virtually error-free; complex structures used flexibly and accurately | 9 |
| Occasional minor errors; good range of complex/compound sentences | 7–8 |
| Noticeable errors that occasionally impede meaning; limited range of structures | 5–6 |
| Frequent errors affecting clarity; mostly simple sentences | 3–4 |
| Severe errors, making comprehension difficult | 1–2 |

**2. Coherence & Cohesion (1–9 points)**

| Criteria | Band |
|---|---|
| Logically organized; seamless transitions; clear paragraph unity | 9 |
| Easy to follow; occasional lapses in transitions or topic development | 7–8 |
| Some logical gaps; repetitive or abrupt transitions | 5–6 |
| Disjointed ideas; minimal paragraph structure | 3–4 |
| No discernible organization; incoherent | 1–2 |

**3. Lexical Resource (Vocabulary) (1–9 points)**

| Criteria | Band |
|---|---|
| Sophisticated, precise word choice; rare minor errors | 9 |
| Effective vocabulary; occasional in-appropriacies | 7–8 |
| Limited range; noticeable errors or awkward phrasing | 5–6 |
| Basic vocabulary; frequent errors | 3–4 |
| Extremely limited; errors obscure meaning | 1–2 |

APPENDIX B: PEER FEEDBACK EVALUATION RUBRIC

**Scale**: 1 (Weak) to 5 (Strong) for each item

| Area | Item |
|---|---|
| **Thesis Clarity** | 1- Focus: The thesis statement is clear and specific. |
| | 2- Relevance: The thesis directly addresses the essay prompt. |
| | 3- Position: The writer's stance is unambiguous (e.g., for/against). |
| | 4- Scope: The thesis covers the essay's content without being too broad/narrow. |
| | 5- Placement: The thesis is positioned effectively (e.g., end of intro). |
| **Paragraph Development** | 1- Topic Sentences: Each paragraph begins with a clear main idea. |
| | 2- Support: Ideas are developed with evidence/examples. |
| | 3- Logic: Arguments flow logically within/between paragraphs. |
| | 4- Transitions: Connective words/phrases guide the reader. |
| | 5- Conclusion: The closing paragraph reinforces the thesis. |
| **Language Use** | 1- Grammar: Sentences are mostly error-free. |
| | 2- Vocabulary: Word choice is precise and academic. |
| | 3- Cohesion: Pronouns/connectors link ideas clearly. |
| | 4- Conciseness: No unnecessary repetition or wordiness. |
| | 5- Formality: Tone is appropriate for academic writing. |

**APPENDIX C: PEER FEEDBACK TRAINING WORKSHOP SLIDES**

**Slide 1: Workshop Objectives**

- **Goal:** Learn to provide constructive, rubric-aligned feedback.
- **Why?** Peer feedback improves revision skills (Lundstrom & Baker, 2009).
- **Outcome:** Confidently use the 15-item rubric to assess essays.

**Slide 2: Feedback Principles**

- **Be Specific:** "Your grammar is bad."(Wrong)----"P1: Use past tense here."(Right)
- **Balance Praise & Critique:** "Your thesis is clear (item 1) but add examples to support P2 (item 7)."
- **Use the Rubric:** Always link comments to rubric criteria "Item 8: Improve transitions between paragraphs.").

**Slide 3: Rubric Walkthrough**

- **Example:** Thesis Clarity (Item 1-5)
- **Score 3:** Thesis is clear but too broad.
- **Score 5:** Thesis is specific, debatable, and matches the prompt.
- **Activity:** Rate sample thesis statements (5 min).

**Slide 4: Practice Session**

- **Step 1**: Annotate a sample essay (non-study) using the rubric.
- **Step 2**: Compare your feedback with a partner. Discuss differences.
- **Step 3**: Instructor models ideal feedback (see Table 1).

**Table 1: Standardized Feedback Example**

| Rubric Item | Sample Feedback |
|---|---|
| **7 (Support)** | "Add a statistic to strengthen your claim in P3 (e.g., '60% of students…')." |
| **12 (Vocabulary)** | 'Big' → 'significant' for academic tone. |

**Slide 5: Avoiding Bias**

- **Focus on the essay**, not the writer.
- **Use neutral language**: "You don't understand transitions." (Wrong) --- "Item 9: Try adding 'however' to show contrast here."(Right)

**Slide 6: Q&A & Feedback Rules**

- **Q**: "What if I disagree with the rubric criteria?"
- **A**: Follow the rubric during the study; save personal opinions for post-study discussions.

**Rules**:

- Spend 10+ minutes per essay.
- Write 3 praises and 3 actionable fixes per essay.

**APPENDIX D: AI FEEDBACK PROTOCOL**

**Model**: ChatGPT-5 (via API to ensure version consistency)

**Prompt**: "Provide detailed EFL feedback on this argumentative essay with the following requirements:

- **Grammar**: Identify errors in accuracy (e.g., verb tense) and range (e.g., overuse of simple sentences).
- **Cohesion**: Evaluate transitions and paragraph unity (e.g., logical flow between ideas).
- **Vocabulary**: Assess precision (e.g., word choice) and appropriateness (e.g., academic tone).

**Notes:** - Use a constructive tone. - Highlight 3 strengths with examples. - Suggest 3 prioritized improvements with specific revisions (e.g., 'Change "big problem" → "significant issue"')."

**APPENDIX E: FEEDBACK PERCEPTION SURVEY**

(5-point Likert scale: 1=Strongly Disagree, 5=Strongly Agree)

| Area | Item |
|---|---|
| **Perceived Usefulness (α=.89)** | 1- The feedback helped me identify specific areas for improvement in my writing. |
| | 2- The suggestions were clear and actionable. |
| | 3- The feedback addressed the most important issues in my essay. |
| | 4- I will apply this feedback to future writing tasks. |
| **Emotional Comfort (α=.82)** | 5- I felt comfortable receiving feedback from this source. |
| | 6- The tone of the feedback was encouraging and respectful. |
| | 7- The feedback made me feel motivated to revise my work. |
| | 8- I did not feel anxious or defensive about the criticism. |
| **Trust in Feedback Source (α=.85)** | 9- I trust the accuracy of the corrections/suggestions provided. |
| | 10- The feedback source (peer/AI) understood my writing goals. |
| | 11- The feedback was consistent with what I've learned in class. |
| | 12- I would like to seek feedback from this source again. |

**APPENDIX F: SEMI-STRUCTURED INTERVIEW PROTOCOL**

**Stratified Subsample**: n=20 (balanced by gender[male/female], proficiency [B1/B2], and feedback group [Peer/AI])

**Duration**: 20-30 minutes per interview

**Format**: Audio-recorded, transcribed verbatim, anonymized

| Area | Item |
|---|---|
| **1. Grand Tour Questions** (Broad exploration of experiences) | 1- "Describe your experience receiving (peer/AI) feedback. What stood out to you?" |
| | 2- "How did you feel while reading the feedback? Were there any surprises?" |
| | 3- "What was most helpful about this feedback? What was the least helpful?" |
| **2. Example-Focused Prompts** (Concrete application of feedback) | 4- "Walk me through one specific suggestion you applied to your revision. How did you decide to use it?"<br>*Probe: "Did you agree with the suggestion? Why or why not?" |
| | 5- "Show me a part of your essay where you disagree with the feedback. What did you do instead?" |
| **3. Comparative Reflections** (Contextualizing feedback experiences) | 6- "How did this feedback compare to past feedback you've received (e.g., from teachers, peers, or tools like Grammarly)?" |
| | 7- "If you could design your ideal feedback system, what would it include from (peer/AI) feedback?" |
| **4. Demographic/Moderator Probes** (Tailored follow-ups) | 8- For B1 learners: "Did the feedback feel too basic/challenging for your level?" |
| | 9- For females: "Did the feedback style affect your confidence differently than past experiences?" |
| | 10- For AI group: "How did you feel about receiving feedback from a machine? Did you trust it?" |