Article

Journal of Interdisciplinary Studies in Education

Volume 15, Issue 8 (2026)
Journal of Interdisciplinary Studies in Education
ISSN: 2166-2681Print 2690-0408 Online | https://ojed.org/jise
https://doi.org/10.32674/0nwtbv54

Human, AI-Generated, and Hybrid Corrective Feedback in L2 Argumentative Writing: Effects on Learner Engagement and Writing Quality

Sindia
B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
ORCID: 0009-0006-6208-9006

Vijayakumar Selvaraj
B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
ORCID: 0000-0002-5932-3467

Sathikulameen
The New College, Chennai, India
ORCID: 0000-0002-9633-7585

Akem Solange Ojong
Drexel University, Philadelphia, United States
ORCID: 0009-0003-9590-3391

ABSTRACT

This quasi-experimental study investigated the differential effects of human teacher feedback (HTF), AI-generated feedback (AIGF), and hybrid feedback (HF) on learner engagement and argumentative writing quality among 120 undergraduate ESL students across an eight-week intervention. Engagement (behavioural, cognitive, emotional) was measured via a purpose-built questionnaire and revision-tracking logs; writing quality was scored on a five-dimension analytic rubric. Repeated-measures ANOVA and multiple regression showed that HF was associated with the greatest improvement in writing quality ($\eta^2p = .24$), HTF with the highest proportion of higher-order revisions, and AIGF with the highest revision frequency. Hybrid feedback condition membership was the strongest predictor of writing-quality improvement, followed by cognitive engagement. These findings have implications for AI-integrated writing instruction in higher education.

Keywords: AI-generated feedback, argumentative writing, learner engagement, quasi-experimental design, revision quality

© Author(s), 2026. Published by Star Scholars Press.
This article is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. https://creativecommons.org/licenses/by/4.0/

INTRODUCTION

Generative AI technologies that produce linguistically sophisticated, contextually responsive feedback on student writing have reshaped corrective feedback practices in second language (L2) writing instruction in recent years. This shift is not merely technological; it changes who provides feedback and how learners engage with non-human evaluative agents. These questions are increasingly urgent in university ESL programmes, where argumentative writing occupies a central curricular place and demands sophisticated rhetorical and intellectual skills. The literature on written corrective feedback (WCF) is extensive, having been built up over 30 years of research. However, fundamental issues remain concerning the effectiveness, selectivity, and uptake of WCF, which are discussed in the Literature Review. The depth and nature of human, AI, and hybrid feedback in terms of accuracy, as well as the behavioural, cognitive, and emotional aspects of the revision process, have been understudied. Previous studies have focused primarily on the learner's perception of automated feedback or on the accuracy outcomes of revision rather than on the actual revision process. The present study aimed to fill this gap using a quasi-experimental design with 120 undergraduate ESL learners participating in an eight-week academic writing programme with three different types of feedback. The research questions that guide the investigation are: (a) How does feedback source relate to behavioural, cognitive, and emotional engagement? (b) What are the predictions of the feedback conditions for the depth of revision and writing quality improvement? and (c) which feedback condition is likely to lead to the greatest gains in higher-order skills such as argumentative development and organisational coherence? This investigation is based on the multidimensional engagement framework of Fredricks et al. (2004) and Schmidt's (1990) Noticing Hypothesis. The Literature Review is structured around four themes (WCF, automated writing evaluation, learner engagement, and revision processes). At the same time, the Theoretical Framework, Methodology, Results, and Discussion (theoretical interpretation, deviations from previous studies, limitations, and pedagogical implications) are organised around the same themes.

THEORETICAL FRAMEWORK

This study is based on two theoretical strands. The first is Schmidt's (1990) Noticing Hypothesis, which states that L2 learning depends on the conscious noticing of linguistic forms in the input, which can be achieved through either salience or gap detection (as noted by Cheng & Zhang, 2025). This has direct implications for comparing human and AI feedback and can be significantly different in terms of their ability to highlight salient linguistic and rhetorical features for the learners. The second is Fredricks et al.'s (2004) multidimensional model of learner engagement, which consists of three dimensions: behavioural engagement (on-task participation, revision frequency), cognitive engagement (self-regulatory strategy use, metacognitive monitoring), and emotional engagement (affective response, perceived relevance, evaluative trust). Koltovskaia (2020) discovered that emotional engagement is qualitatively different for automated and human feedback, making it relevant as a comparison variable in this study's analysis. In combination, these frameworks make the feedback source not just a delivery system, but a variable that can either enhance behavioural engagement (AI feedback through speed and scale) or improve cognitive engagement (AI feedback through less dialogic scaffolding) or both (AI feedback through balancing the two). The hypothesised pathway combines both strands into a single testable model.

LITERATURE REVIEW

Written Corrective Feedback in L2 Writing

Ferris (2010) argues that the type of feedback does not directly cause grammatical learning, but depends on the developmental readiness of the learner and their metalinguistic awareness. The types of WCFs (direct correction, metalinguistic commentary) have different implications for noticing by L2 input processing, and the timing of WCF delivery further moderates these effects (Ellis, 2009; Li et al., 2026).

Hyland and Hyland (2006) demonstrated using naturalistic classroom data that teacher feedback can have pragmatic roles other than marking errors, such as scaffolding motivation and negotiating disciplinary identity, which have yet to be explored in the context of AI-generated feedback. In contrast to the pedagogical consensus that grammar correction in L2 writing is of little or no benefit, or even detrimental, to the learner, Truscott (1996) argues that this is due to the developmental readiness of learners and the absence of communicative pressure. However, this is contrary to the pedagogical consensus that grammar correction in L2 writing is of little or no benefit, or even detrimental, to the learner. This argument received a significant empirical response from Truscott (1996), who held that this was due to the developmental readiness of the learners and the absence of communicative pressure. Truscott's (1996) position that feedback works only under certain conditions is supported by the statistically significant and permanent accuracy gains found by Brown et al. (2026) in certain controlled designs, notably in the case of the English article system (EAS). Theoretically relevant to AI systems, which tend to provide direct correction, Cotos (2023) discovered that direct correction has more immediate effects on accuracy than indirect correction.

In contrast, indirect strategies have more long-term developmental effects. Chandler (2003) also found direct marking to be more effective than reformulation, confirming the automated systems' ability to provide surface-oriented error correction, but not for more advanced features, such as argumentative coherence, which have proven less reliable in algorithmic processes. Suzuki and Saito (2026) further show that WCF effectiveness depends on learners' metalinguistic processing of feedback rather than mere uptake. Read together, these studies converge on a qualified view of WCF effectiveness: Truscott's (1996) scepticism has been substantially revised by evidence of selective efficacy (Brown et al., 2026; Chandler, 2003), yet a genuine tension remains between accuracy gains from direct correction and the higher-order, rhetorically oriented feedback that argumentative writing requires (Cotos, 2023; Suzuki & Saito, 2026), the tension this study's comparison of feedback sources is designed to probe.

AI-Generated Feedback and Automated Writing Evaluation

Automated writing evaluation (AWE) systems have been used in L2 writing instruction for over 20 years, initially for holistic essay scoring and surface-error detection. Within large-enrolment writing courses, AWE offers immediacy, scalability, and consistency that human feedback cannot guarantee (Warschauer & Grimes, 2008). Chen and Cheng (2008) found that EFL learners' grammatical accuracy and self-revision improved after AWE feedback. However, automated commentary was not always useful without human intervention, since learners often struggled to interpret system-generated feedback, suggesting that effectiveness depends on learners' metalinguistic ability. Later studies have shifted to the use of generative AI feedback from large language models, which are more detailed and extensive than previous rule-based systems: Hartshorn and Pack (2026) investigated the use of AI-based dynamic WCF and user sentiment, while Shi and Aryadoust (2024) validated this as a growing research area. Link et al. (2022) demonstrated that the way students use automated feedback is not uniform and is influenced by their perception of its reliability; that is, more competent students may strategically avoid the feedback if they believe it is incorrect or irrelevant. Bai and Hu (2017) also discovered that the fallibility of AWE leads to emotional disengagement, which leads to learners' lack of confidence in the feedback source and the motivation to revise. Stevenson and Phakiti (2014) determined that AWE has positive effects on grammatical accuracy for drafts, while it has much smaller effects on argumentative coherence and argumentative content development. In this body of work, automated systems have consistently been found to have a positive effect on the surface accuracy dimension (Chen & Cheng, 2008; Stevenson & Phakiti, 2014), while higher-order dimensions are found to have a positive effect to a lesser extent and in inconsistent ways, and uptake is contingent on learners' trust and metacognitive disposition rather than surface accuracy alone (Bai & Hu, 2017; Link et al., 2022). This ceiling is either a natural limitation of AI feedback or an artefact of unmediated delivery, which is directly addressed by the hybrid condition in this study.

Broader work on generative AI in education reinforces both the promise and scrutiny of AI-generated feedback. Zacharis and Papadakis (2025) found that AI-assisted grading of university coursework

can approach human-level validity and reliability. However, fairness concerns persist across subgroups, underscoring the need for careful evaluation of AI's evaluative role rather than uncritical adoption. Uğraş et al. (2024) similarly reported that ChatGPT-supported instruction enriched pedagogical practice in primary education while raising concerns about over-reliance. This tension parallels the AIGF condition's mixed results in the present study.

A recent review identifies evaluative feedback and adaptive scaffolding among the recurring uses of artificial intelligence in language education (Imran et al., 2026).

Learner Engagement with Corrective Feedback

Theoretically, WCF engagement is separate from its effectiveness. In relation to revision frequency and type, Ene and Upton (2018) discovered that synchronous and asynchronous feedback had significant differences; that is, the former generated more deliberate processing and higher-order revision, which aligned with the results of Qi and Lapkin (2001) in their study of noticing in three-stage writing tasks. The time delay is what sets AI feedback (submitted) apart from human feedback (24–72 h). From the case study of computer-generated feedback, Zhang (2017) proposed three types of engagement: strategic engagers, who selectively accept linguistically significant feedback; compliance-driven engagers, who accept feedback without metalinguistic reflection; and resistant non-engagers, who accept it but ignore it, which contradicts the notion that the learners process the feedback at similar cognitive levels. Rassaei (2026) argues that deep cognitive learning requires task relevance, perceived competence, and autonomous control, which are more likely to be provided by the relational, dialogic feedback that is characteristic of experienced human teachers. The results of these studies agree that engagement is not a one-size-fits-all phenomenon.

Recent evidence further indicates that L2 learners engage with AI-generated feedback in varying cognitive, behavioural, and affective ways (Ma et al., 2026). In higher education, generative AI competence and perceived autonomy have also been associated with student engagement (Sampah et al., 2026).

Revision Processes and Writing Quality in ESL Contexts

Not all feedback uptake and revision is of high quality. Trained peer response is found to result in qualitatively different revision patterns than untrained responses because trained reviewers' comments are more likely to lead to content-level revisions than surface revisions (Min, 2006). Hamano-Bunce (2025) identified that although feedback on AWE leads to an increase in grammatical accuracy across drafts, the improvement in content development and argumentative coherence is significantly less, which is the focus of this study's revision-type analysis: a surface/higher-order divide that is intractable. Current AI systems are not yet able to provide feedback that is consistently personally oriented and attuned to rhetorical purpose, which is most educationally valuable (Hyland & Hyland, 2006). Another challenge is the discrepancy between what teachers want to say and what learners actually hear, as noted by Lee (2014), which may be exacerbated in automated environments where feedback is produced without considering the learner's history. This body of literature indicates that the systematic surface coverage provided by AI may be paired with teachers' rhetorically focused, relationally sensitive commentary to promote conditions that are most likely to foster deep engagement and higher-order revision, a hypothesis that this study examines.

An eight-week quasi-experimental study comparing teacher feedback with AI-assisted peer feedback likewise documented differences in feedback focus and trajectories of L2 writing development (Tang et al., 2026).

RESEARCH METHOD

Research Design

This study used a quasi-experimental design to investigate the differential effects of feedback source on learner engagement, revision depth, and writing quality. Institutional constraints precluded full

randomisation; instead, three intact course sections were randomly assigned to the three feedback conditions, with baseline equivalence confirmed through pre-treatment proficiency testing (no significant group differences in initial writing quality, $F(2, 117) = 0.43$, $p = .652$). Outcome data were collected using only quantitative measures: writing-quality assessment and revision-tracking logs.

Participants and Context

The study involved 120 undergraduate ESL students who were studying first-year academic writing courses at a medium-sized public university in India. The students were randomly divided into three whole-course sections of 40 students per section, where each section received one of three feedback conditions: human teacher feedback (HTF), AI-generated feedback (AIGF), and hybrid feedback (HF). To be eligible for this course, students had to be registered in a first-year academic writing course, achieve a minimum IELTS Academic band score of 5.0 or a minimum score achieved on an institutional placement test in English, and not have received formal training in argumentative academic writing in English prior to enrolling in the course. The demographic profile of the participants is presented in Table 1. The pre-intervention equivalence across the three conditions was tested using chi-square and one-way ANOVA testing. The results showed that there was no significant difference across the three conditions in terms of gender distribution, age, and initial IELTS scores. The same equivalence testing was extended to socioeconomic background, prior AI experience, and digital literacy, variables not reported in the original submission: socioeconomic status did not differ significantly across conditions, $\chi^2(4, N = 120) = 0.18$, $p = .996$, nor did prior AI experience, $\chi^2(4, N = 120) = 0.27$, $p = .991$, nor digital literacy, $F(2, 117) = 0.70$, $p = .500$. The three conditions were therefore equivalent at baseline on all demographic and technology-readiness variables tested.

Table 1

Participant Demographic Characteristics by Feedback Condition

Variable	HTF (n = 40)	AIGF (n = 40)	HF (n = 40)
Gender, *n* (%)			
Female	24 (60.0)	22 (55.0)	23 (57.5)
Male	16 (40.0)	18 (45.0)	17 (42.5)
Age, *M* (*SD*)	19.8 (1.2)	20.1 (1.4)	19.9 (1.3)
IELTS Academic band, *M* (*SD*)	5.62 (0.41)	5.58 (0.44)	5.65 (0.39)
First language, *n* (%)			
Tamil	26 (65.0)	27 (67.5)	25 (62.5)
Urdu	9 (22.5)	8 (20.0)	10 (25.0)
Other	5 (12.5)	5 (12.5)	5 (12.5)
Socioeconomic status, *n* (%)			
Low	9 (22.5)	10 (25.0)	9 (22.5)
Middle	23 (57.5)	22 (55.0)	22 (55.0)
High	8 (20.0)	8 (20.0)	9 (22.5)
Prior AI experience, *n* (%)			
None	10 (25.0)	9 (22.5)	9 (22.5)
Occasional	19 (47.5)	19 (47.5)	18 (45.0)
Regular	11 (27.5)	12 (30.0)	13 (32.5)
Digital literacy, *M* (*SD*)	3.80 (0.39)	3.77 (0.39)	3.87 (0.39)

Note. HTF = human teacher feedback; AIGF = AI-generated feedback; HF = hybrid feedback; IELTS = International English Language Testing System; AI = artificial intelligence. IELTS scores represent the Academic module band scores at the pre-intervention assessment. Percentages are calculated within each feedback condition.

Each of the three course sections had the same teacher to minimise the influence of the teacher, and all three sections used the same curriculum, except for the feedback mechanism. In the HTF condition, teachers gave written feedback on the students' drafts within 48 hours after submission, focusing on surface and higher-order writing issues. For students under the AIGF conditions, drafts were submitted through the digital writing platform to GPT-4o, accessed via a web-based interface, which returned AI-generated comments immediately upon submission; teacher comments were withheld during the treatment period for this condition. The AI system was designed to provide corrective feedback on grammatical accuracy, coherence, and argumentative structure, but not to rewrite the student's text; the same prompt was used for all texts throughout the 8-week intervention to reduce the possibility of prompting variability. No other verification or review of accuracy was performed on the AI-generated feedback before it was given to students. The students in the HF condition received the AI-generated comments as soon as they submitted their work, and the higher-order comments by the teacher within 48 hours if not addressed by the AI system.

Instruments and Data Sources

The data sources were divided into three main sources. The first was an analytic writing-quality rubric, which was adapted from the writing-assessment rubrics already developed for other languages. It was scored on five dimensions (content development, organisational coherence, lexical sophistication, grammatical accuracy, and argumentative strength) in five bands (0–4 Insufficient to 17–20 Excellent), with a maximum holistic score of 100. The ICC was .89 for the 20 pilot scripts, with pre-data collection calibration. The raters scored the revised submission, blind to condition and time point, using three parallel argumentative prompts of similar difficulty for each time point (Weeks 2, 5, 8) discussing any dimension with a difference of more than 3 points. To obtain the main dataset, two raters independently scored all 360 participant-time submissions, which resulted in ICC(2,1) = .91. Second, a revision-tracking module integrated into the digital writing platform generated categorical data about revision type (surface-level, meaning-level, higher-order). Third, a Feedback Engagement Questionnaire (FEQ) was created for this study based on the behavioural, cognitive, and emotional engagement dimensions of Fredricks et al. (2004). However, the items were not taken from a validated instrument, and no independent test beyond internal consistency reported below was performed. Representative items were: "I revised my draft after carefully reading the feedback provided" (behavioural), "I evaluated whether each suggestion should be accepted or rejected" (cognitive), and "I trusted the source of the feedback I received" (emotional), rated on a 5-point Likert scale. Engagement was measured at Weeks 2, 5, and 8 with Cronbach's alphas calculated on the full item-level data set (N = 360 observations), of .87, .91, and .90 for the three subscales, respectively.

Analytical Procedures

Quantitative data were analysed using SPSS Statistics Version 27. The change in writing quality was evaluated using a 3 (condition) × 3 (time point) repeated-measures ANOVA, with Mauchly's test used to check for sphericity (Greenhouse-Geisser correction applied where necessary) and Bonferroni-corrected pairwise comparisons used to control the familywise error rate ($\alpha = .05$). Multiple regression was used to identify significant predictors of writing-quality gain (Week 8 minus Week 2) after mean-centring continuous predictors; feedback condition (dummy-coded, AIGF as reference), initial proficiency, and the three engagement dimensions were significant predictors. The differences in distribution for revision type were evaluated using chi-square tests. Effect sizes were measured using Cramér's V. The homogeneity of variance (Levene's, all $p > .80$) and the approximately normal within-condition distributions (HTF, $W = 0.939$, $p = .032$; HTF gain scores, $W = 0.915$, $p = .005$; see Limitations) were confirmed at most time points by assumption diagnostics computed from the participant-level dataset. The regression variance inflation factors ranged from 1.15 to 3.03 (which is below the threshold for concern of VIF > 5), and the residuals were approximately normally distributed (Shapiro-Wilk $W = 0.988$, $p = .344$) and had no obvious pattern against the fitted values.

Ethical Considerations

The university's Institutional Review Board approved this study. Informed consent was obtained, including the right to opt out of the research component without any repercussions for academic performance; the writing evaluation was part of the regular assessment of the course; and the research component did not impact grades. Data were anonymised using unique participant codes and stored on password-protected institutional systems accessible only to the research team, and data were processed according to the Declaration of Helsinki.

RESULTS

Learner Engagement Across Feedback Conditions

The descriptive statistics of the three dimensions of engagement for the three feedback conditions at the three time points are presented in Table 2. Behavioural engagement was greatest at weeks 2 and 5 in the AIGF condition, when automated feedback was delivered immediately, which may lead to rapid revision cycling. Still, by week 8, the AIGF condition's behavioural engagement was similar to the HTF condition, indicating that there was a habituation effect for the AIGF condition. The HTF and HF conditions were significantly higher in cognitive engagement scores at each assessment time point compared to the AIGF condition, and the HF condition had the highest cognitive engagement scores (week 8: M = 4.12, SD = 0.46). At week 2, emotional engagement was equivalent across the groups. Still, it increased progressively for the HTF. It decreased progressively for the AIGF, with the HTF reporting the highest level at week 8 (M = 4.01, SD = 0.49) and the AIGF reporting a significant decrease in emotional engagement during the intervention period (week 2: M = 3.52, SD = 0.61; week 8: M = 3.18, SD = 0.72).

Table 2

Descriptive Statistics for Learner Engagement Dimensions by Feedback Condition and Time Point

		M (*SD*)				
Engagement dimension	*n*	Week 2	Week 5	Week 8	Gain score	*d*
Behavioural						
HTF	40	3.42 (0.61)	3.58 (0.57)	3.71 (0.53)	+0.29	0.47
AIGF	40	3.89 (0.52)	3.94 (0.49)	3.72 (0.58)	−0.17	0.29
HF	40	3.61 (0.55)	3.84 (0.51)	3.88 (0.47)	+0.27	0.50
Cognitive						
HTF	40	3.31 (0.64)	3.67 (0.59)	3.94 (0.52)	+0.63	1.02
AIGF	40	3.28 (0.67)	3.41 (0.63)	3.35 (0.68)	+0.07	0.10
HF	40	3.45 (0.59)	3.83 (0.54)	4.12 (0.46)	+0.67	1.24
Emotional						
HTF	40	3.56 (0.58)	3.74 (0.54)	4.01 (0.49)	+0.45	0.83
AIGF	40	3.52 (0.61)	3.43 (0.64)	3.18 (0.72)	−0.34	0.48
HF	40	3.49 (0.62)	3.72 (0.55)	3.87 (0.51)	+0.38	0.66

Note. HTF = human teacher feedback; AIGF = AI-generated feedback; HF = hybrid feedback. Engagement dimensions were measured on 5-point Likert scales adapted from Fredricks et al. (2004). Gain score = Week 8 *M* − Week 2 *M*. Cohen's *d* was calculated relative to the pooled pre-intervention *SD*.

Revision Depth and Type Across Feedback Conditions

The revision tracking data were analysed, and the following differences were found in the distribution of revision types by the three feedback conditions. The highest percentage of higher-order revisions (M = 42.3%, SD = 8.7) was found in the HTF condition, followed by the HF condition (M = 38.6%, SD = 9.2) and the AIGF condition (M = 21.4%, SD = 7.1). These distributional differences were

found to be statistically significant by means of chi-square analysis (χ^2 (4, N = 120) = 47.82, p <.001), with a moderate-to-large effect size (Cramér's V = .44) of the feedback condition on revision type. Although the absolute number of revisions in the AIGF condition was the highest (M = 23.7 events per draft, SD = 4.3), compared to the HTF (M = 14.2, SD = 3.8) and HF (M = 17.9, SD = 4.1) conditions, the absolute number of surface-level revisions in the AIGF condition was higher than those in the other conditions (61.3% vs. 33.7% and 42.1%, respectively).

Writing Quality Gains Across Conditions and Time Points

Writing quality scores were analysed using a 3 × 3 repeated-measures ANOVA with the three conditions (HTF, AIGF, and HF) and the three time points (Weeks 2, 5, and 8). The main effect of time was significant, F(2, 234) = 89.43, p < .001, η^2p = .43, indicating measurable improvement in overall writing quality during the intervention period across all three conditions. The main effect of condition was also significant, F(2, 117) = 18.67, p < .001, η^2p = .24, and the Condition × Time interaction was likewise significant, F(4, 234) = 12.34, p < .001, η^2p = .17, suggesting that the rate of improvement in writing quality differed significantly across the conditions. Overall, the writing quality scores were significantly higher for the HF condition at week 8 than for the HTF (p = .012) and AIGF (p < .001) conditions, but there was no significant difference between the HTF and AIGF conditions at weeks 2 and 5. By week 8, the scores differed significantly, with higher scores for the HTF condition (p = .023). Figure 1 shows the average progress of the writing quality scores for the three conditions. Table 3 presents the complete ANOVA results for each of the five dimensions of writing quality.

Content development and argumentative strength showed the most significant condition × time interaction, with the HF condition having significantly larger improvements on these dimensions than the AIGF and HTF conditions alone: content development, F(4, 234) = 16.42, p < .001, η^2p = .22; argumentative strength, F(4, 234) = 19.87, p < .001, η^2p = .25. The AIGF condition resulted in gains similar to those of the HF condition at week 5 (η^2p = .17 for the interaction effect), indicating that the systematic surface-level coverage of the automated system was comparable to that in the HF condition, but that the HF condition had additional qualitative effects on grammatical processing beyond those of the automated system alone.

Table 3

Repeated-Measures ANOVA Results for Writing Quality Dimensions by Feedback Condition

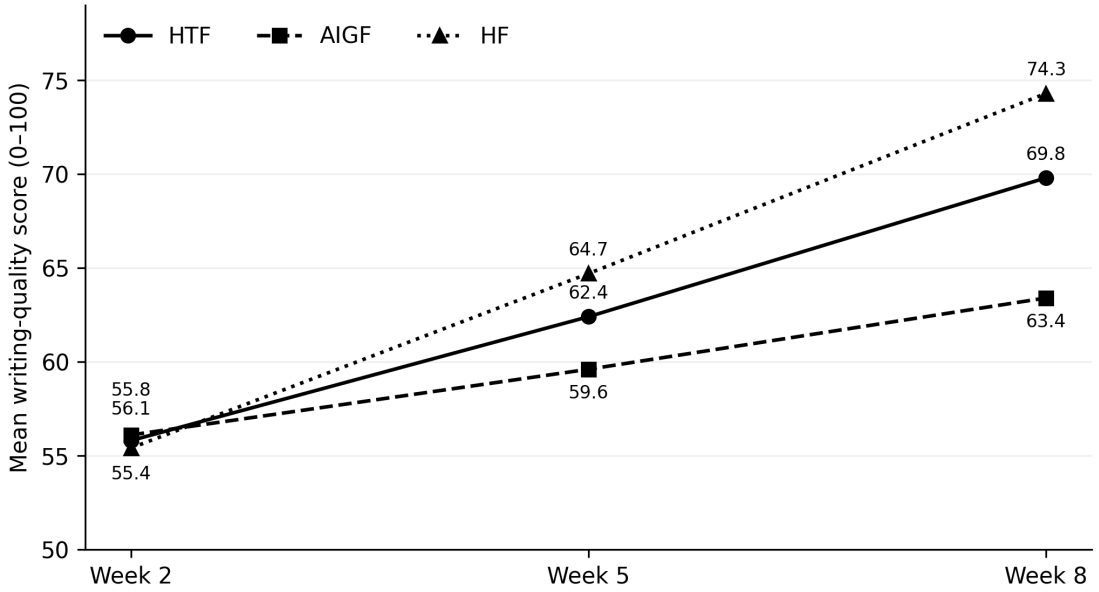
	Time			Condition			Condition × Time		
Writing quality dimension	*F*	*p*	η^2_p	*F*	*p*	η^2_p	*F*	*p*	η^2_p
Content development	76.34	< .001	.39	14.82	< .001	.20	16.42	< .001	.22
Organisational coherence	82.17	< .001	.41	11.24	< .001	.16	10.87	< .001	.16
Lexical sophistication	61.28	< .001	.34	9.43	< .001	.14	8.91	< .001	.13
Grammatical accuracy	94.56	< .001	.45	16.37	< .001	.22	12.14	< .001	.17
Argumentative strength	88.73	< .001	.43	21.08	< .001	.26	19.87	< .001	.25
Overall writing quality	89.43	< .001	.43	18.67	< .001	.24	12.34	< .001	.17

Note. N = 120. Degrees of freedom were $F(2, 234)$ for time, $F(2, 117)$ for condition, and $F(4, 234)$ for the Condition × Time interaction. η^2_p = partial eta squared.

Table 3 presents the repeated-measures ANOVA results for the five dimensions of writing quality, showing significant effects of time, feedback condition, and their interaction. To complement these statistical findings, Figure 1 illustrates the progression of mean writing-quality scores across Weeks 2, 5, and 8, highlighting differences in the improvement trajectories of the three feedback conditions.

Figure 1

Mean Overall Writing Quality Score Trajectories Across Three Time Points by Feedback Condition



Note. HTF = human teacher feedback; AIGF = AI-generated feedback; HF = hybrid feedback. Data points represent condition means (n = 40 per condition) on the 100-point analytic rubric at each time point.

Predictors of Writing Quality Improvement

Multiple regression analysis found that the significant predictors of writing quality gain scores (Weeks 8– 2) were as follows: The overall model was statistically significant, $F(6, 113) = 195.90$, $p < .001$, $R^2 = .91$, meaning that the set of predictors explained 91% of the variance in gain in writing quality across the entire sample. The standardised and unstandardised regression coefficients for each predictor are shown in Table 4. The strongest individual predictor was the HF condition relative to the AIGF ($\beta = .85$, $p < .001$), followed by cognitive engagement ($\beta = .30$, $p < .001$) and the HTF condition relative to the AIGF ($\beta = .24$, $p < .001$). Initial proficiency level was not a significant predictor of gain ($\beta = .02$, $p = .437$), and behavioural engagement was not a significant independent predictor ($\beta = .05$, $p = .253$). Emotional engagement was also not a significant independent predictor when it was added to cognitive engagement and condition ($\beta = .01$, $p = .818$). Thus, the feedback condition, particularly the HF condition, was the dominant predictor of writing-quality gain, with cognitive engagement contributing a substantial secondary independent effect; behavioural and emotional engagement showed no independent association with gain once condition and cognitive engagement were accounted for.

Table 4

Multiple Regression Analysis Predicting Writing Quality Gain Score (Week 8 – Week 2)

Predictor	*B*	*SE B*	β	*t*	*p*
Initial proficiency level	0.33	0.42	.02	0.78	.437
HTF condition (vs. AIGF)	2.32	0.40	.24	5.83	< .001
HF condition (vs. AIGF)	8.27	0.38	.85	21.80	< .001
Behavioural engagement	0.39	0.34	.05	1.15	.253
Cognitive engagement	2.13	0.34	.30	6.33	< .001
Emotional engagement	0.07	0.31	.01	0.23	.818

Note. $N = 120$. $R^2 = .91$, $F(6, 113) = 195.90$, $p < .001$. Feedback condition was dummy-coded with AIGF as the reference category, and continuous predictors were mean-centred. HTF = human teacher feedback; AIGF = AI-generated feedback; HF = hybrid feedback.

DISCUSSION

Interpretation of Findings

The findings from the present investigation support and build upon the theoretical framework provided by the integration of Schmidt's (1990) Noticing Hypothesis and Fredricks et al.'s (2004) engagement framework. The better writing quality achieved in the HF condition suggests that hybrid feedback can simultaneously optimise writing conditions along all three dimensions: immediate feedback from AI tools, systematic surface feedback from AI tools, and rhetorical and metacognitively scaffolded feedback from human teachers. The present data seem to indicate that this pairing directed learners' attention to two complementary levels of the text: AI feedback brought attention to salient surface features, whereas human commentary brought attention to the rhetorical purpose and argumentative structure. This distinction echoes Zamel's (1985) earlier observation that responding effectively to student writing requires attention to both surface-level correctness and the underlying composing process rather than treating written feedback as a single undifferentiated intervention. This is further supported by the finding that hybrid feedback condition membership was the strongest predictor of writing quality gain, with cognitive engagement contributing a substantial secondary effect; behavioural and emotional engagement showed no independent association with gain once condition and cognitive engagement were accounted for. The high revision rates of the AIGF condition, coupled with the relatively small gains in writing quality, are exactly what Guo (2025) predicted as a possible result of task environments that support compliance but not metacognitive processing.

Relation to Prior Research

This result, which shows that AIGF resulted in the highest revision frequency but the lowest proportion of higher-order revisions, is consistent with the theoretical prediction that AI feedback would draw the learner's attention to surface-level error patterns, thereby focusing the learner's attention on the local features of the text rather than the global features of the rhetorical context. Similar results were found by Zhang (2017) in relation to EFL writers who used AI feedback strategically and selectively, as cognitively sophisticated learners may not cope with the localised nature of AI feedback for complex argumentative writing tasks. This appears compounded by the AIGF condition's declining emotional engagement across weeks 5 and 8, consistent with the affective disengagement Bai and Hu (2017) associate with perceived feedback fallibility, and may partly explain the condition's stalled quality gains in later weeks, a mediating pathway warranting further empirical attention.

The HF condition's benefits are consistent with AI tools working best as an adjunct rather than a substitute for human teaching, echoing Stevenson and Phakiti (2014), who found that computer-generated feedback alone yields modest content-development gains relative to grammatical accuracy — also observed here in the AIGF condition. A more notable departure from prior research is the size of the HTF condition's argumentative-quality gains relative to AIGF, which is larger than in earlier comparative studies, plausibly because argumentative writing's most critical criteria — logical reasoning, evidence integration, and counterargument engagement — remain areas where AI systems struggle most. The present data support Zhai and Ma's (2022) finding that proficient learners engage in strategic disengagement from automated feedback: AIGF participants scoring 6.0 or higher on initial IELTS showed greater declines in revision frequency than lower-scoring peers.

Limitations

The present findings are limited in terms of their interpretation in several ways. Because intact course sections were used instead of individually randomised assignments, there is the possibility of selection bias, as the results may have been influenced by characteristics of the course sections other than the feedback condition, such as the dynamics of the peer groups, scheduling differences, or differential

prior experiences with digital writing tools. The findings are limited in their generalisability to ESL university contexts other than the one in India, in which the background of learners' L1 and their experience in digital feedback environments may vary significantly from that of the present sample, as may the writing course cultures. The intervention lasted for eight weeks, which was sufficient to develop initial developmental trends, but would not be able to detect the longer-term outcomes of chronic exposure to hybrid feedback on writing proficiency, which would be more likely to be detected in longitudinal designs that span a full school year. Because the study did not include a qualitative strand, it did not capture participants' perspectives on why particular feedback types were more or less effective for them, a gap that future research using interviews or think-aloud protocols could address. Because generative AI systems evolve rapidly, replications using newer model versions may not reproduce the present results. Assumption diagnostics beyond sphericity have since been computed from the participant-level dataset: variance inflation factors (1.15–3.03), residual normality (Shapiro-Wilk $W = 0.988$, $p = .344$), and homogeneity of variance (Levene's, all $p > .80$) gave no cause for concern, although writing-quality scores in the HTF condition departed from normality at Week 5 and in gain scores, which should be borne in mind when interpreting HTF-specific comparisons. A further limitation concerns the intact-class design: condition and class are completely confounded. Hence, a class-adjusted treatment effect is not separately identifiable, and a class-level explanation cannot be ruled out. Relatedly, the Week 8 engagement predictors were measured concurrently with the final writing-quality assessment used to compute the gain score; thus, engagement may be reciprocally related to gain rather than purely predictive of it.

IMPLICATIONS

The theoretical implications concern the conditional nature of feedback engagement: the cognitive and affective conditions required for feedback-driven writing development do not consistently emerge from a single feedback type but are distributed across types contingent on learner proficiency, task genre, and institutional context. This extends Fredricks et al.'s (2004) framework into the feedback domain, showing that the three dimensions of engagement are not uniformly associated with feedback type; behavioural engagement can be high while cognitive and emotional engagement remain low, with consequences for how feedback-driven learning is theorised and measured. This finding also supports Schmidt's (1990) Noticing Hypothesis in a comparative feedback situation, where the potential for acquisition through feedback commentary is not determined by the surface level of noticing but rather by the depth of noticing it generates. Sequencing AI feedback at the initial drafting stage, before teacher feedback on rhetorical and argumentative aspects, aligns with Lee's (2014) priority revision model and offers a practical and hybrid design. Program administrators need to invest in training writing instructors to understand and use the data generated by AI tools and create institutional processes that integrate AI tools into a relational feedback culture rather than replacing it with one. From an AI system design perspective, feedback systems that provide more nuanced feedback on evidential quality, argumentation structure, and rhetorical effectiveness are desirable and may be within the reach of future systems.

These findings also have implications for assessment policy and the ethical governance of AI in writing instruction: institutions will need policies that distinguish formative AI-assisted feedback from summative assessment, with safeguards against evaluating students on writing shaped by feedback of uncertain accuracy. Ethical implementation further requires transparency with students about when and how AI systems are used to generate feedback, together with a clear channel through which students may contest or seek clarification on AI-generated comments.

CONCLUSION

The present study confirmed that the feedback source has a differential and multidimensional impact on ESL learners' engagement and the quality of argumentative writing at the university level, where hybrid feedback was associated with the most conducive environment for the development of writing across the entire spectrum of the examined outcomes. The superiority of the HF condition was most evident on

higher-order dimensions of writing (argumentative strength and content development); and, in theory, can be explained by the combination of immediate salience enhancement in the writing task (Noticing Hypothesis, Schmidt, 1990) with relationally scaffolded, rhetorically oriented commentary (Roothooft & Bueno-Alastuey, 2026) to maximise noticing processes and affective investment, which, in turn, facilitates meaningful L2 writing development. The results show that hybrid feedback condition membership was the most powerful single predictor of improvement in writing, with cognitive engagement contributing a substantial secondary effect; this indicates that although feedback source sets the primary conditions for improvement, the depth of learners' cognitive processing, and not merely behavioural compliance, remains an important secondary driver of writing development. The study's contribution is theoretical, in that it demonstrates that the dimensionality of engagement matters; that is, the frequency of revision behaviour, which has been widely used as a proxy for engagement depth, is an unreliable index of the cognitive and emotional investment that is most likely to be important for developing writing. The methodological transfer of the differentiation of dimensions of engagement in a feedback source comparison is applicable to feedback studies in other L2 skills and to the study of the integration of AI in instructional contexts other than writing pedagogy. The implications of this for the conditions under which noticing and sustained engagement with writing can be realised become even more important as generative AI becomes increasingly embedded in pedagogies used in higher education. The present results suggest that the conditions for noticing and sustained engagement with writing may be best realised when generative AI is not intended to replace but to augment and extend human pedagogical judgement.

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

Brown, D., Liu, Q., & Norouzian, R. (2026). Effectiveness of written corrective feedback in developing L2 accuracy: A Bayesian meta-analysis. *Language Teaching Research, 30*(3), 1357–1389. https://doi.org/10.1177/13621688221147374

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*(3), 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9

Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94–112. https://doi.org/10.64152/10125/44145

Cheng, X., & Zhang, L. J. (2025). Investigating synchronous and asynchronous written corrective feedback in a computer-assisted environment: EFL learners' linguistic performance and perspectives. *Computer Assisted Language Learning, 38*(7), 1687–1716. https://doi.org/10.1080/09588221.2024.2315070

Cotos, E. (2023). Automated feedback on writing. In O. Kruse, C. Rapp, C. M. Anson, K. Benetos, E. Cotos, A. Devitt, & A. Shibani (Eds.), *Digital writing technologies in higher education* (pp. 347–364). Springer. https://doi.org/10.1007/978-3-031-36033-6_22

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

Ene, E., & Upton, T. A. (2018). Synchronous and asynchronous teacher electronic feedback and learner uptake in ESL composition. *Journal of Second Language Writing, 41*, 1–13. https://doi.org/10.1016/j.jslw.2018.05.005

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

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Guo, X. (2025). Facilitator or thinking inhibitor: Understanding the role of ChatGPT-generated written corrective feedback in language learning. *Interactive Learning Environments, 33*(5), 3526–3544. https://doi.org/10.1080/10494820.2024.2445177

Hamano-Bunce, D. (2025). The effects of direct written corrective feedback and comparator texts on the complexity and accuracy of revisions and new pieces of writing. *Language Teaching Research, 29*(6), 2460–2479. https://doi.org/10.1177/13621688221127643

Hartshorn, K. J., & Pack, A. (2026). The effects of artificial intelligence-based dynamic written corrective feedback on second language writing and user sentiment. *RELC Journal*, *57*(1), 69–88. https://doi.org/10.1177/00336882251405498

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Imran, M., Almusharraf, N., & Shehzad, K. (2026). Artificial intelligence in school language education: A systematic review of trends, challenges, and future directions. *Journal of Interdisciplinary Studies in Education, 15*(6), 47–78. https://doi.org/10.32674/w506dk33

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. (2014). Feedback in writing: Issues and challenges. *Assessing Writing, 19*, 1–5. https://doi.org/10.1016/j.asw.2013.11.009

Li, S., Ou, L., & Lee, I. (2026). The timing of corrective feedback in second language learning. *Language Teaching, 59*(1), 29–45. https://doi.org/10.1017/S0261444824000478

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

Ma, X., Zhang, C., & Lee, I. (2026). L2 learners' engagement with AI-generated feedback on writing. *Assessing Writing, 68*, Article 101020. https://doi.org/10.1016/j.asw.2026.101020

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

Qi, D. S., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing, 10*(4), 277–303. https://doi.org/10.1016/S1060-3743(01)00046-7

Rassaei, E. (2026). The interplay between corrective feedback timing and foreign language anxiety in L2 development. *Language Teaching Research, 30*(3), 1432–1455. https://doi.org/10.1177/13621688231195141

Roothooft, H., & Bueno-Alastuey, M. C. (2026). Students' beliefs about written corrective feedback in two foreign language contexts. *Language Awareness, 35*(2), 511–531. https://doi.org/10.1080/09658416.2025.2522107

Sampah, S. N. A., Nabang, M., Krampa, E. K., Issah, M., Koduah, F., Essel, H. B., & Adu-Agyem, J. (2026). Generative AI competence and student engagement in higher education: Mediating roles of utilization, autonomy, and formal learning. *Journal of Interdisciplinary Studies in Education, 15*(2), 109–146. https://doi.org/10.32674/3s3cx241

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

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

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

Suzuki, W., & Saito, R. (2026). Cognitive and affective reactions to written corrective feedback, type of linguistic items, and second language writing accuracy improvement: An exploratory study. *Language Awareness, 1*–21. https://doi.org/10.1080/09658416.2025.2611884

Tang, J., Li, P., & Luo, R. (2026). Teacher feedback vs. AI-assisted peer feedback in L2 writing: A quasi-experimental study in a Chinese university. *PLOS ONE, 21*(6), Article e0345976. https://doi.org/10.1371/journal.pone.0345976

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

Uğraş, H., Uğraş, M., Papadakis, S., & Kalogiannakis, M. (2024). ChatGPT-supported education in primary schools: The potential of ChatGPT for sustainable practices. *Sustainability, 16*(22), Article 9855. https://doi.org/10.3390/su16229855

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22–36. https://doi.org/10.1080/15544800701771580

Zacharis, G., & Papadakis, S. (2025). Can AI grade like a human? Validity, reliability, and fairness in university coursework assessment. *Educational Process: International Journal, 19*, Article e2025591. https://doi.org/10.22521/edupij.2025.19.591

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19*(1), 79–101. https://doi.org/10.2307/3586773

Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: a systematic investigation of college students' acceptance. *Computer Assisted Language Learning, 35*(9), 2817–2842. https://doi.org/10.1080/09588221.2021.1897019

Zhang, Z. (2017). Student engagement with computer-generated feedback: A case study. *ELT Journal, 71*(3), 317–328. https://doi.org/10.1093/elt/ccw089