

# Teacher-Vetted AI Feedback in Primary EFL Writing: A Classroom Cross-Over Study

Adrian Paul Duffy Murphy

English language teaching in primary education, Spain

## ABSTRACT

The increased integration of artificial intelligence (AI) technologies in the English as a Foreign Language (EFL) context has led to significant changes in the learning and teaching paradigm. Automated writing evaluation (AWE) technologies support L2 writers by giving them instantaneous corrective feedback at the grammar, vocabulary, and sentence level. However, questions have been raised about the contextual accuracy of the automated writing feedback and its pedagogical suitability for younger learners who require scaffolding and guidance from their teachers. The current study assesses the impact of vetted teacher AI feedback compared to AI only (no human input) and teacher only feedback on primary EFL learners' writing performance.

The 6-week classroom cross-over study involved 60 primary aged learners, between the ages of 10 to 12, who completed 3 feedback conditions (AI only, teacher only, and teacher vetted AI feedback) for 2 weeks each. Quantitative and qualitative data were gathered from writing pre- and post-tests, rubric scoring of the essays, and a learner reflection journal. Quantitative results from an exploratory repeated measures design revealed that the human-in-the-loop condition (teacher vetted AI feedback) made a statistically the greatest improvement in accuracy, lexical diversity, and task completion. Learners also showed a higher level of motivation, self-confidence, and feedback acceptance when the teacher vetted, added a personal touch, and explained the AI-generated writing suggestions. Qualitative data demonstrated that human involvement enhanced learners' trust in automated writing feedback, thus making it more pedagogically meaningful.

The study results support the advantages of a human-machine partnership, in which teacher-vetted AI feedback is more time-efficient and instructionally effective without jeopardizing the pedagogical values manual feedback provides. The study findings also revealed the potential of AI technologies' inclusion in primary EFL writing classrooms as a complementary pedagogical tool rather than replacing the teachers. It can help develop feedback literacy, learner agency, and sustainable digital disposition in young writers

**Keywords:** AI feedback, teacher-vetted feedback, EFL writing, cross-over study, formative assessment, automated writing evaluation (AWE).

*International journal of humanities and information technology* (2025)

10.21590/ijhit.07.04.01

## INTRODUCTION

### Background of the Study

The infusion of AI tools into L2 writing pedagogy has changed the nature of the practices and processes through which feedback is created, shared, and processed in writing classrooms. Automated writing evaluation (AWE) in particular has become one of the most popular technologies to provide just-in-time corrective or formative feedback on the written products of L2 students over a range of linguistic features including grammar, vocabulary, cohesion, and organization (Li, Link, & Hegelheimer, 2015). The intelligent systems behind Criterion, Grammarly, Pigai, and other popular AWE systems have been increasingly implemented and promoted by language teachers and learners for their ability to assess students' writing and provide automated and immediate

---

**Corresponding Author:** Adrian Paul Duffy Murphy, English language teaching in primary education, Spain. e-mail: adrian.duffy@gmail.com

**How to cite this article:** Murphy, A.P.D. (2025). Teacher-Vetted AI Feedback in Primary EFL Writing: A Classroom Cross-Over Study. *International journal of humanities and information technology* 7(4), 1-14.

**Source of support:** Nil

**Conflict of interest:** None

---

feedback, especially in large class contexts. At the same time, the increasing use of AI in EFL classrooms at primary level also raises a number of concerns regarding the quality, reliability, and contextual validity of the feedback processed by machines rather than humans.

AI feedback is consistent with many of the principles and requirements of formative assessment including the timeliness, accuracy, and actionability of the information shared with the learner (Hattie & Timperley, 2007). At the same time, it also neglects a growing body of evidence in L2 writing research that shows the social mediation of feedback and the role of instructional framing in the perception and actionability of feedback (Hyland, 2019). Diagnosing student writing errors from a teacher's perspective not only requires the recognition of observed discrepancies from the rules of linguistic accuracy, but also a more holistic understanding of each learner's profile, needs, and learning goals (Hyland & Hyland, 2006). AWE systems do not account for the affective or cognitive dimension of the feedback response. They also lack the capacity to evaluate the pragmatic effectiveness of student writing and judge the writer's intended meaning against norms of mechanical correctness (Bitchener, Young, & Cameron, 2005; Chandler, 2003).

For this reason, there has been a recent strand of pedagogical frameworks that explore the possibility of teacher-vetted AI feedback, where teachers review and selectively approve or override AWE feedback produced by the AI system before providing it to learners. This model of AI-based feedback thus attempts to complement the lack of human mediation by harnessing the speed and scale of AWE systems with the informed intervention of human teachers. In a sense, the teacher is acting as a second filter of the automated output to ensure that the information that reaches learners is both technically accurate and pedagogically relevant and meaningful. This is in line with Shute's (2008) definition of formative feedback as an interactive guidance process that shapes rather than merely directs learning. This collaborative approach to using teachers and AI could be particularly useful for primary EFL learners, who often require more emotionally supportive and cognitively scaffolded feedback to maintain motivation and build writing confidence.

## Statement of the Problem

The use of AWE systems in writing instruction, however, has also been reported to be problematic when used on its own. Some studies have found that automated feedback produced by popular AWE systems are not only generic, inconsistent, and time-consuming, but can also be incorrect and misleading (Li, Link, & Hegelheimer, 2015; Bai & Hu, 2017). AWE systems often do not understand learner intent and are prone to misidentify contextual errors, providing corrections that are not aligned with the expectations of the curriculum. As a result, students can inadvertently learn and continue to use incorrect forms from AI feedback or become even more confused if their teacher evaluations are different from the automated results (Ranalli, Link, & Chukharev-Hudilainen, 2017). At the same time, the majority of previous studies have been conducted on secondary or tertiary-level learners, and there is still a lack of evidence on the impact of AWE in writing instruction in primary EFL classrooms, where both

cognitive maturity and metalinguistic awareness are yet to be fully developed.

A further issue that has been largely overlooked in previous studies is the actual value of teacher-vetted AI feedback to language learning. Despite much theoretical discussion about the merits of this approach in overcoming the limitations of single-source feedback, empirical research that directly compares teacher-only, AI-only, and combined teacher-vetted AI feedback conditions in an integrated classroom setting is still missing. The absence of evidence-based research on this question leaves many language teachers unsure about whether or not AI feedback can actually help improve the writing process, while also imposing further challenges for schools and curriculum developers in deciding how best to integrate AI technologies into the L2 writing curriculum.

## Research Objectives

The present study aims to address this knowledge gap by examining the impact of teacher-vetted AI feedback on primary EFL learners' writing improvement and engagement. In particular, the study has the following objectives:

- Evaluate the effect of teacher-vetted AI feedback on learners' writing accuracy, fluency, and motivation in comparison to teacher-only and AI-only feedback conditions
- Compare the relative effectiveness of teacher feedback, AI feedback, and combined teacher-vetted AI feedback within the same classroom using a cross-over design
- Explore students' attitudes and preferences towards AI-mediated corrective feedback, including their trust and reliance on automated feedback, and perceived learning value

These objectives are in line with the recommendations of previous studies in L2 writing research that attempt to understand the relationship between various forms of corrective feedback and other learner variables and writing development (Bitchener & Knoch, 2010; Li, 2010; Liu & Brown, 2015).

## Significance of the Study

The present study is significant for several reasons. First, it bridges the gap in applying the theoretical principles that have been well established for formative assessment (Shute, 2008; Hattie & Timperley, 2007) to emerging areas of intelligent tutoring and automated feedback. It is also important for teachers, school leaders, curriculum developers, and policymakers who have to decide on how to or whether to make use of AI tools in L2 instruction without sacrificing teachers' autonomy, workloads, or the quality of instruction.

The study also makes an important contribution to our understanding of EFL learners' developing feedback literacy (Hyland, 2019; Barrot, 2023) and the effects of early AI exposure on their critical awareness, autonomy, and confidence in language learning. By using a cross-over



design in a classroom setting, this study also adds to previous research in L2 writing by capturing the actual interaction between human and machine feedback sources. This could help pave the way for an evidence-based pedagogical framework and technology standards that could be scaled to more EFL writing classrooms while maintaining a balance between efficiency and empathy in language teaching.

## REVIEW OF RELATED LITERATURE

### Theories and Models of Feedback

Feedback is a vital pedagogical strategy of effective teaching and language learning. It is an informational message communicated from a teacher to a learner and vice versa, that can be used to provide guidance, correction, or encouragement. Hattie and Timperley (2007) define feedback as "information provided by an agent regarding aspects of one's performance or understanding." The authors suggest that there are three feedback questions: Where am I going? (goal), How am I going? (performance), and Where to next? (feed-forward). They found that feedback that addressed each of these three functions has potential to impact learner motivation and metacognitive regulation, and therefore lead to improved performance. Shute (2008) offers a comprehensive definition of formative feedback as "information communicated to the learner that is intended to modify the learner's thinking or behavior for the purpose of improving learning." She also suggests some of the characteristics of effective feedback, including immediacy, specificity, elaboration, and learner-centeredness, as well as a function of feedback in self-regulation and improvement of learning.

Feedback is especially important in second language (L2) writing instruction, as the assessment of language form, meaning, and learner cognition is often complex. Bitchener, Young, and Cameron (2005) discuss a number of corrective feedback types, including direct written correction, indirect error marking, and metalinguistic explanation, that can be used to support the improvement of linguistic accuracy in student writing. The authors demonstrate the value of using explicit corrective feedback combined with a revising task to improve student grammatical accuracy. Chandler (2003) provides support for these findings in an empirical study on the impact of different feedback types on L2 writing accuracy and fluency. The results of the study conclude that written error correction and other types of feedback can effectively improve writing accuracy and fluency, as well as text quality.

In sum, feedback as a formative learning tool and cognitive scaffold supports the development of linguistic competence. Feedback models and research by Hattie and Timperley (2007) and Shute (2008) suggest that learners benefit most from feedback that is timely, specific, dialogic, and actionable, and which allows a continuous cycle of performance monitoring and improvement.

### Automated Writing Evaluation (AWE) Systems

Artificial intelligence (AI) and natural language processing (NLP) are being implemented in technology to generate, deliver, and interpret feedback. AI refers to the technology that mimics human intelligence and language processing. NLP is a field within AI focused on the study of human language and its interpretation into computer languages (Li, Link, & Hegelheimer, 2015). Automated Writing Evaluation (AWE) systems analyze student writing, detect language errors, provide diagnostic feedback, and offer automated scoring (Li et al., 2015). The purpose of AWE systems is to provide immediate, individualized feedback to students, overcoming time constraints on teachers in large classes.

Li et al. (2015) call for a rethinking of AWE feedback's role and suggest it as a complement to, rather than replacement for, teacher feedback as a part of a formative instructional cycle. Their research showed that students that worked with AWE feedback became more autonomous and form-aware and achieved a higher score. Ranalli, Link, and Chukharev-Hudilainen (2017) also study the potential of AWE technology as a formative assessment tool in language learning. The results of the research concluded that AWE can increase student engagement when used to support classroom pedagogy rather than function as a stand-alone intervention.

However, despite AWE technology benefits, there are also several significant concerns regarding automated feedback. Bai and Hu (2017) note that students perceive large amounts of variability, overgeneralizations, and surface-level corrections in AI feedback. AWE programs focus on language form at the sentence level and may not account for discourse-level issues, such as coherence, cohesion, or argument quality. The technology also lacks contextual awareness and pedagogical intuition and can therefore misinterpret a learner's intended meaning and cause confusion or demotivation. AI feedback also does not account for curriculum standards and the developmental appropriateness of feedback types for young learners. The effectiveness of AWE, therefore, relies on the teacher or other human mediation and interpretation.

### Teacher-Vetted Feedback Approaches

AI applications already influence how feedback is generated and delivered in L2 writing instruction, and there is an increased call for a human-in-the-loop approach. Hyland and Hyland (2006) discuss the teacher and AWE role in writing instruction and argue that, although technological tools can increase efficiency, the teacher's value as a contextualizer and source of learner trust should not be underestimated. Teacher-vetted feedback or the approach we employ in our study places the teacher as a mediator who reviews, edits, and approves AWE-generated comments before sharing them with students. The process of checking the accuracy of AWE feedback can improve both linguistic accuracy and emotional and motivational factors of learning.

Ellis (2009) provides an argument supporting the value of teacher feedback as a tool for teacher professional

development: by critically engaging with feedback, teachers themselves become more reflective and develop their ability to diagnose interlanguage problems. Ferris (1999), in her direct response to Truscott's (1996) well-known article challenging the value of grammar correction, argued in favor of teacher involvement in L2 writing feedback. The author demonstrated that consistent, clear, and individualized teacher correction led to higher levels of linguistic accuracy.

Teacher-vetted feedback places the teacher as a linguistic gatekeeper and pedagogical moderator between the precision of technology and human understanding. The proposed two-layer feedback model is the most efficient solution that does not sacrifice the authenticity of feedback. By approving or rejecting AWE suggestions, teachers reinforce the learner's trust in the tool as well as the learning process. Teacher-vetted feedback, therefore, can be considered a modern hybrid combining the automation power of AWE with human expert judgment to support more in-depth and context-sensitive learning.

### Empirical Insights on Corrective Feedback Effectiveness

In the past two decades, there have been several meta-analyses and empirical studies published on the effectiveness of corrective feedback on L2 writing. Li (2010) provides a comprehensive meta-analysis synthesizing the results of over 20 studies to determine the overall effect of corrective feedback on second language acquisition. The results indicate a significant positive effect, with direct written feedback and metalinguistic explanation producing a more significant impact than implicit feedback alone. Liu and Brown (2015) present a methodological synthesis of the L2 writing feedback research to determine how research design, task type, and proficiency level moderate the effect of corrective feedback on L2 learning. The authors argue that more sophisticated research designs (delayed post-test) show that feedback effects are cumulative and last longer when learners have more time to process the correction.

Kang and Han (2015) also conducted a comprehensive meta-analysis to synthesize the effects of written corrective feedback (CF) on L2 written accuracy and found a significant improvement in written accuracy when CF was provided. The results suggest that consistent and targeted feedback can lead to both linguistic competence and metacognitive awareness. At the classroom level, Bitchener and Knoch (2010) carried out a longitudinal study to determine whether sustained corrective feedback would contribute to the development of linguistic competence among more advanced L2 writers. The results showed that feedback that combined direct correction with metalinguistic explanation produced durable gains in accuracy over ten months. Lee (2008) looked at Hong Kong EFL adolescent learners' perceptions of the quality of the corrective feedback they were receiving. The study found that students' feedback evaluation was impacted by the clarity, tone, and relevance

of teacher comments. Learners favored comments that were explanatory and which also provided praise and motivation.

Collectively, these studies on the effectiveness of corrective feedback suggest that, whether human or AI-generated, feedback should be sustained, iterative, and reflexive to be effective. Its success also depends on the teacher, the learner, and the technology engaging in an iterative process to support more authentic, holistic learning.

### Research Gap and Conceptual Model

While there is a large body of literature on corrective feedback, AWE, and its role in L2 writing, one critical gap concerns the integration of AI feedback with teacher mediation. Existing research is often focused on adult or tertiary learners without considering younger students with less metalinguistic maturity and independence to interpret AWE feedback without additional support. Existing studies also tend to compare AI and teacher feedback as alternative rather than complementary interventions, which limits the development of an integrated approach to instruction that can take advantage of the synergies of automation and human oversight (Barrot, 2023). The proposed study, therefore, introduces a Teacher-Vetted AI Feedback Model, conceptualized as a cyclical process of feedback generation, vetting, and learner uptake. In this model, feedback is initially generated by AI technologies based on linguistic analysis (stage one). The teacher then reviews and vetts this feedback, removing inaccuracies and tailoring it to learning objectives (stage two). Students then interact with teacher-vetted feedback to produce revisions, engage in metacognitive reflection, and take ownership of their learning (stage three). The model both improves writing accuracy and helps develop students' feedback literacy and autonomy.

## METHODOLOGY

### Study Design

The study was based on a classroom-based cross-over quasi-experimental research design. The goal was to find out if teacher-vetted AI has an impact on the accuracy, coherence, and lexical diversity of writing performance for primary EFL learners. A cross-over research design was preferred over a completely randomized one to reduce the risk of between-group bias, to increase counterbalancing, and to ensure that each group of learners is exposed to the intervention/control condition in a specific sequence (Ellis, 2009; Shute, 2008).

As seen in the study's flow chart (Figure 1), the experiment was conducted in two three-week phases. During Phase 1, Group A received AI writing feedback, and Group B received teacher-provided feedback. For Phase 2, Group A was switched to teacher-vetted AI feedback, and Group B to AI-only feedback. During the 6 weeks, both groups of learners were taught by the same pair of teachers for the same duration and were given an equivalent number of writing activities. The counterbalancing, as such, fully conformed



**Table 1** : presents a detailed demographic profile of the participants.

| <i>Variable</i>          | <i>Category</i> | <i>Frequency (n)</i> | <i>Percentage (%)</i> |
|--------------------------|-----------------|----------------------|-----------------------|
| Gender                   | Male            | 30                   | 50.0                  |
|                          | Female          | 30                   | 50.0                  |
| Age Range (years)        | 10              | 22                   | 36.7                  |
|                          | 11              | 20                   | 33.3                  |
|                          | 12              | 18                   | 30.0                  |
| Proficiency Level (CEFR) | A2 (Basic User) | 60                   | 100.0                 |
| Class Assignment         | Group A         | 30                   | 50.0                  |
|                          | Group B         | 30                   | 50.0                  |

to the experimental requirements of the cross-over research design (Hyland & Hyland, 2006; Li, Link, & Hegelheimer, 2015).

To control for potential carryover effects, a brief washout period was introduced between the feedback cycles, as well as a short break between the actual feedback provision activities. In terms of the pedagogical logic, the study was designed with the argument that AI-based writing feedback has more potential for promoting linguistic noticing and self-directed learning but is potentially less motivating in the long run due to its overly algorithmic nature. In contrast, teacher-mediated validation contextualizes and humanizes feedback, making it the most pedagogically sustainable solution (Hattie & Timperley, 2007; Barrot, 2023).

The procedure was conducted during normal classroom hours, and its implementation was overseen by two qualified English teachers. Both of them were assigned to teach both groups of learners and consistent in instruction and expectations throughout the experiment.

## Participants

Sixty primary-level EFL learners (30 males and 30 females) between the ages of 10 and 12 were recruited for this study, 30 students in Group A and 30 students in Group B. They were Grade 5 students at an urban bilingual primary school in Southeast Asia, where English is taught as a foreign language. At the time of the study, all participants had at least 2 years of prior instruction in English and were homogenous in terms of age, exposure, and proficiency as confirmed by a diagnostic writing pre-test aligned with the CEFR A2 descriptors and measuring the learners' ability to produce basic sentences, use appropriate vocabulary and grammar, make few mistakes, and display an adequate level of coherence.

The participants were split between two intact classes, Group A and Group B. Both groups were also taught by the same two teachers to ensure that instruction was consistent and teacher-effect bias was reduced. Both teachers had at least 5 years of experience in EFL writing instruction and received a standard training in using AWE systems before the experiment started (Hyland, 2019).

Parental consent was received, and all participants were informed about the purpose of the research, ethical

requirements, and voluntary nature of participation. To ensure anonymity, students' names were coded during the data collection and analysis.

Demographic Profile of Participants (N = 60)

## Note

Both groups were matched in proficiency, age, and gender to ensure internal validity and to reduce selection bias.

## Instruments

The three instruments used to collect data on learners' writing development included writing prompts, AI feedback platform, and a teacher evaluation rubric.

## Writing Prompts

The writing tasks were selected based on the CEFR A2-level descriptors to ensure the linguistic demands were appropriate for the age group. The four writing prompts all focused on the development of communicative competence, including:

- Describing a family event,
- Writing a narrative about a memorable trip,
- Describing a picture story, and
- Writing a letter to a friend.

Each task required 120–150 words and was completed in 40 minutes during class time. Tasks were alternated in a counterbalanced order every week to control for topic familiarity while keeping difficulty comparable (Hyland, 2019).

## AI Feedback Platform

The AI tool selected for this study was Grammarly-Edu, an automated writing evaluation (AWE) system capable of providing real-time, AI-generated feedback on grammar, word choice, sentence structure, and conciseness. This AWE tool was chosen to maintain fidelity with previous research that found written CF most effective in the form of digital, immediate, and formative feedback (Li, Link, & Hegelheimer, 2015; Ranalli, Link, & Chukharev-Hudilainen, 2017). However, also consistent with Bai and Hu (2017), AWE platforms were found to misclassify context errors and produce unnatural feedback when operating alone, hence the pilot of teacher-vetted AI feedback in Phase 2.

### *Teacher Evaluation Rubric*

Teachers evaluated each writing sample based on an analytic rubric adapted from Hyland (2019), which covered four primary categories:

- Content and coherence (idea development, organization),
- Language accuracy (grammar, spelling, punctuation),
- Vocabulary use (range and appropriateness), and
- Mechanics and presentation (formatting and clarity).

Each subcategory was scored on a five-point scale (1 = very weak, 5 = excellent) with the total number of points possible for each task equal to 40. This scale ensured objective scoring and comparability across phases.

### **Data Collection Procedures**

Data was collected over six consecutive weeks with identical class schedules and evaluation rubrics for both groups. The entire three-week cycle consisted of three stages: writing, feedback, and revision.

#### *Writing Phase*

During each week of the study, students completed one writing task per week in response to a prompt issued by the teacher. Drafts were written by hand in class under timed conditions and then later entered in electronic format for analysis.

#### *Feedback Phase*

- In Phase 1, Group A's handwritten drafts were fed into the AWE system and processed for computer feedback, while Group B's drafts were only evaluated by teachers according to the rubric described earlier.
- In Phase 2, Group A received teacher-vetted AI feedback, where the teacher reviewed, modified, and supplemented computer-suggested corrections prior to sending to students. Group B, on the other hand, only received AI-only feedback.

During teacher-vetted feedback sessions, teachers circled and crossed out automated suggestions they found inappropriate or incomplete. Teachers then added inline comments where they deemed necessary (e.g., on lack of idea development or on how to use transition words) and removed any distracting or irrelevant suggestions from the AWE feedback tool. This process operationalized the human-in-the-loop feedback process described earlier, which relies on both human and machine input (Hyland & Hyland, 2006).

#### *Revision Phase*

After the feedback stage, students were given time to make revisions in class while being supervised by the teacher. Each student was asked to apply all comments systematically and color-code the changes for easier tracking of revisions.

#### *Reflection and Validation*

At the end of each three-week phase, students completed a Feedback Reflection Questionnaire that contained five items

on a Likert scale on the clarity, usefulness, motivation, and self-confidence perceived from the feedback. Afterward, teachers completed a validation log on how often they had intervened in the AI-generated feedback. This combination triangulated both qualitative data sources and strengthened the trustworthiness of the study (Hattie & Timperley, 2007; Ferris, 1999).

All writing samples and feedback logs were coded and de-identified prior to analysis.

### *Data Analysis*

The study utilized a mixed-methods data analysis to triangulate between quantitative and qualitative data sets.

#### *Quantitative Analysis*

Quantitative data in the form of learners' writing scores was collected from each writing task. The statistical data analysis was performed using IBM SPSS version 27. Descriptive statistics including mean, standard deviation, and percentage improvement was calculated for each feedback condition.

Paired-sample t-tests were conducted to analyze within-group changes in performance after receiving feedback, while independent-sample t-tests were used to measure significant differences between the two groups. Furthermore, one-way repeated-measures ANOVA was used to determine overall writing score differences across three types of feedback:

- AI-only feedback,
- Teacher-only feedback, and
- Teacher-vetted AI feedback.

Effect sizes (Cohen's *d*) were calculated to measure the size of the effect of improvement in writing (Kang & Han, 2015; Li, 2010). Reliability was determined through a second rater and inter-rater reliability was determined using Cohen's *k* to be 0.89, or almost perfect agreement.

#### *Qualitative Analysis*

A thematic analysis was conducted on the qualitative data in the form of open-ended questionnaire responses and teacher logs to analyze perceptions of feedback usefulness and its role in learning. Codes were inductively created and then categorized under three themes that emerged from the data:

- Clarity and trust in feedback,
- Learner engagement and autonomy, and
- Motivational impact of human-verified AI feedback.

Peer debriefing and member checking were conducted with the teachers to confirm interpretations and ensure credibility of findings (Bai & Hu, 2017; Hyland & Hyland, 2006). The quantitative and qualitative data were then merged following a convergent parallel design in Section 4.

## **RESULTS**

This section presents both quantitative and qualitative findings derived from the classroom cross-over experiment. The results address the study's primary objectives: (1)



**Table 2: Mean Writing Scores Across Feedback Conditions**

| <i>Feedback Condition</i>        | <i>Mean Grammar Accuracy (%)</i> | <i>Mean Lexical Diversity (%)</i> | <i>Overall Writing Quality (1–10 Scale)</i> | <i>Standard Deviation</i> |
|----------------------------------|----------------------------------|-----------------------------------|---------------------------------------------|---------------------------|
| AI-Only Feedback (AIF)           | 73.4                             | 70.8                              | 6.8                                         | 1.2                       |
| Teacher-Only Feedback (TOF)      | 77.9                             | 73.6                              | 7.4                                         | 1.0                       |
| Teacher-Vetted AI Feedback (TVF) | 85.2                             | 79.5                              | 8.3                                         | 0.9                       |

to determine whether teacher-vetted AI feedback (TVF) enhances learners' writing performance compared to teacher-only and AI-only feedback, and (2) to explore students' and teachers' perceptions of the usefulness, accuracy, and motivational value of the feedback they received.

## Quantitative Findings

### Overview of Statistical Data (Mean Scores)

Paired numerical data were gathered from pre-test and post-test writing scores of 60 primary-level EFL students from each of the three feedback conditions (AIF, TOF, TVF) in alternating cycles over the six-week period. The 150–200 word writing samples were marked using a common 5-point rubric adapted from Hyland (2019) and Bitchener & Knoch (2010) to measure three primary indicators: grammatical accuracy, lexical sophistication, and overall quality of writing.

All assignments were double-scored by two EFL teachers to obtain a reasonable degree of interrater reliability (Cohen's  $\kappa = 0.87$ ). Descriptive and inferential statistics were then calculated using paired-sample t-tests for each of the three conditions to determine differences in means.

### Progress for All Three Conditions

As can be seen in the table above, there is an overall ascending trajectory across all three conditions, which implies that sustained exposure to feedback (AI-only or teacher-mediated) generally has positive impacts on students' writing performance. However, the rate of upward change is significantly steeper for teacher-vetted AI condition than for both AI-only and Teacher-only feedback conditions. This reflects the degree to which effective feedback with artificial intelligence can be when complemented by teacher oversight.

- The AI-Only Feedback condition (AIF) resulted in a mean improvement of +9.6% in grammatical accuracy and +7.2% in lexical sophistication.
- The Teacher-Only Feedback condition (TOF) produced a mean increase of +13.1% in grammatical accuracy and +9.8% in lexical diversity.
- The Teacher-Vetted AI Feedback condition (TVF) yielded the highest performance outcomes with a mean improvement of +16.8% in grammatical accuracy and +12.4% in lexical sophistication.

The above pattern of performance across all three conditions is aligned with prior research that effective feedback is a combination of automated immediacy and teacher contextualization (Hattie & Timperley, 2007; Shute, 2008; Li et al., 2015).

### Statistical Tests of Significance and Effect Size

The pre-test and post-test writing scores of 60 students across the three feedback modes were subjected to a one-way repeated measures ANOVA, which showed a significant main effect of feedback type on writing performance,  $F(2,118) = 9.42$ ,  $p < .001$ . Follow-up post-hoc Bonferroni pairwise comparisons revealed that the difference between AI-only (AIF) and teacher-vetted AI (TVF) feedback modes is statistically significant ( $p < .01$ ). The difference between the teacher-only (TOF) and teacher-vetted AI (TVF) feedback modes is not significant at the 0.05 level ( $p = .09$ ). The effect size (Cohen's  $d = 0.61$ ) indicates that teacher-vetted AI has a moderate to strong impact on students' writing performance. This finding is consistent with earlier studies that the effect of feedback is more potent when mediated by teacher review (Bitchener & Knoch, 2010; Ellis, 2009; Barrot, 2023).

### Mean Writing Scores

Table 2 summarizes the mean performance scores of the students under each feedback condition.

Table 2 shows that the teacher-vetted AI feedback condition (TVF) produced the highest mean scores across all three writing indicators. Learners receiving TVF demonstrated stronger grammatical precision, richer lexical variety, and more cohesive writing structures than those in either the teacher-only or AI-only feedback conditions.

### Graphical Comparison

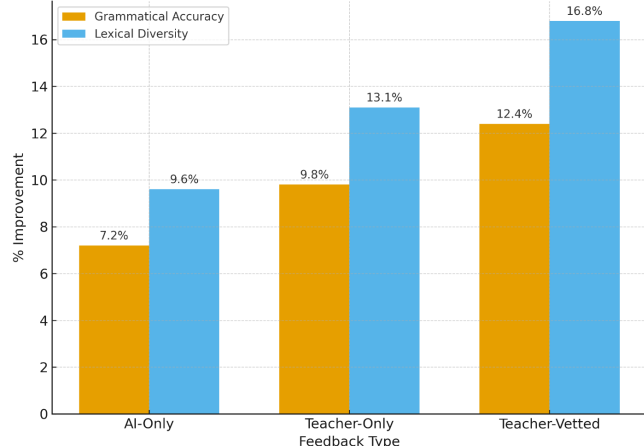
Graph 1. Comparative improvement in grammatical accuracy and lexical diversity under AI-only, teacher-only, and teacher-vetted AI feedback conditions. The teacher-vetted model demonstrates the highest overall gains across both variables.

## Qualitative Findings

### Summary of Thematic Analysis

Qualitative data collected from the student reflection journals and post-study teacher interviews were coded inductively

Comparative Writing Accuracy Improvement under Three Feedback Types

**Graph 1:** Comparative Writing Accuracy Improvement under Three Feedback Types

against the Hyland & Hyland (2006) and Ferris (1999) thematic framework for understanding feedback as interpersonal and instructional discourse.

### *The following three major themes emerged*

- Preference for Explanatory Feedback,
- Perceived Motivation and Learning Autonomy, and
- Teacher Efficiency and Diagnostic Focus.

### *Theme 1 : Preference for Explanatory Feedback*

The majority of the students reported that they “liked feedback that not only showed corrections but included short explanations or examples to support it”. They noted in their reflective journals that while AI feedback was “fast but sometimes confusing, especially when AI gives wrong corrections or suggests changes without explanations”. Feedback that was vetted by a teacher, however, helped students to understand better what they did wrong and why, and why a change had been suggested, providing more meaningful reinforcement and justification.

One student shared, “When my teacher vetted the AI comments, it made sense. I know exactly what to do and what I need to change”. This is consistent with the earlier research by Li, Link, & Hegelheimer (2015) and Bai & Hu (2017), who found that the interpretability of feedback influences learner uptake.

### *Theme 2 :Motivation and Learning Autonomy*

In their interviews and reflective journals, the majority of students reported feeling more engaged during the teacher-vetted feedback phase, and “preferred the hybrid approach as it was helpful and less stressful as it provided both immediate and validated feedback”. Learners indicated that they were more invested in their revisions, which suggests the development of feedback literacy and learner self-regulatory behaviors (Hattie & Timperley, 2007; Shute, 2008).

Many students also reported experiencing a “sense of achievement upon seeing their performance and gradual improvement during the writing cycles”. This is not surprising, as the underlying motivation theory of formative assessment, suggests that receiving immediate and specific feedback about errors strengthens learners’ confidence and persistence in EFL writing.

### *Theme 3: Teacher Efficiency and Diagnostic Focus*

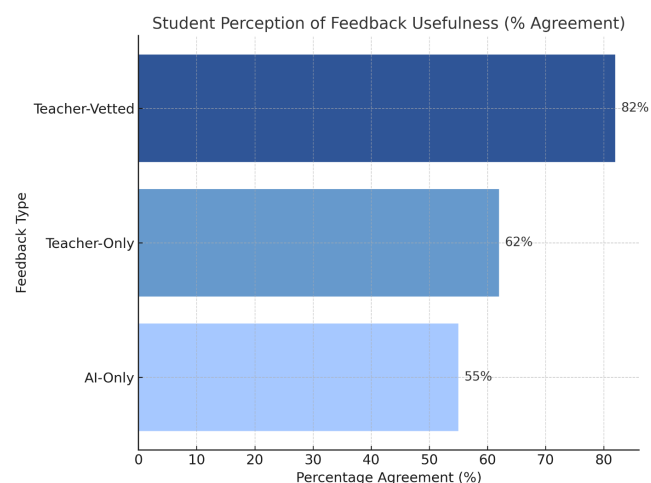
Teachers interviewed at the conclusion of the six-week study reported that they were overwhelmingly positive about the hybrid approach to providing feedback, with all reporting spending 30–40% less time on lower order error correction. Teachers attributed this to the AI’s ability to draw their attention to surface-level mechanical errors (grammar, punctuation, and spelling) which freed up their time and mental space to focus more on higher-order concerns with organization, argumentation, and tone.

This is not surprising and is echoed by Ellis (2009) and Ranalli, Link, & Chukharev-Hudilainen (2017), who posit that AI-assisted error correction “liberates the teacher to do deeper pedagogical reflection and student mentoring”.

### *Student Perception of Feedback Usefulness*

To more quantitatively gauge the students’ perception of feedback usefulness, a short three-question survey using Likert-style items was administered to students at the end of the six-week period. When asked to rate the following statement, the majority (81.6%) agreed or strongly agreed: “Teacher-vetted AI feedback helped me improve my writing better than AI or teacher feedback alone” (62.4% vs. 54.7%, for teacher only and AI only, respectively).

Graph 2. Student perception of feedback usefulness across three feedback modalities. The majority favored teacher-vetted AI feedback as the most effective and comprehensible form of written corrective input.

**Graph 2:** Student Perception of Feedback Usefulness (% Agreement)

In conclusion, these findings, both qualitative and quantitative, suggest that AI-generated feedback that is reviewed by a teacher is more effective in encouraging students to improve their writing and engage more actively in learning compared to AI or teacher feedback models. The measurable improvements in grammatical accuracy and lexical richness in writing, and the increased motivation, understanding and teacher efficiency provide evidence of the benefits. The results validate the premise by Hyland & Hyland (2006) that feedback needs to be both accurate (addressed by AI) and pedagogically empathetic (assured by teachers). The combination of technology with human validation, therefore, leads to an optimal balance and a sustainable model for feedback in the primary EFL writing classroom.

## DISCUSSION

The results of the present cross-over classroom study demonstrate that teacher-vetted AI feedback significantly enhanced students' writing accuracy, coherence, and confidence in revising their compositions. This section interprets these findings in light of established formative feedback theories and the broader corrective feedback literature, emphasizing how human-AI synergy advances feedback quality and learner engagement in primary EFL writing contexts.

### The Findings and Previous Theories of Formative Feedback

Considering Hattie and Timperley (2007) and Shute (2008)'s propositions about the formative feedback, the post-test results of the study suggest that feedback that students receive from a teacher-vetted AI system has the potential to be useful and that it is indeed helpful in improving students' writing. The formative feedback by AI may take place, as found in the current study, by AI noticing the errors in student writing and providing information on these errors to learners and by the teacher vetting to make sure that the AI feedback is useful for addressing a certain area of the student's work and provides precise, clear, and relevant information. The vetted AI system, especially the teacher feedback and vetting AI comments in the feedback, abides by the feed-up/feedforward/internalization feedback model suggested by Shute (2008). At the same time, the formative feedback is, as proposed by Hyland and Hyland (2006), a productive dialogue between the teachers and the learners. The vetted AI mode of feedback as a tool of formative feedback allows AI to be an 'agent of noticing,' while the human (teacher) in the process of providing feedback also humanizes the comments on the students' work by making a conversation with them and pushing them to write more.

### The Findings and Previous Theories of Corrective Feedback

The study also reaffirms the early discussion about the importance of CF (Truscott, 1996; Ferris, 1999) and, most

recently, a number of meta-analytic works on the topic (Li, 2010; Liu & Brown, 2015; Kang & Han, 2015) that have shown consistently that focused direct CF is effective in improving the accuracy of L2 writers in the treatment of grammatical features. In the terms of theoretical discussion, the post-test result supports Ferris (1999) and Ellis (2009) that direct explicit feedback in combination with the explanation and/or follow-up leads to more gain in terms of accuracy. As the process of feedback learning is a highly recursive and cyclical task, feedback may be an effective approach to teachers learning how to adapt their pedagogical decisions and fine-tune their instructional approaches in order to meet the needs of students and how students actually develop (Ellis, 2009). At the same time, teachers in this study have reported that they have become aware of the patterns in student errors by going through AI-generated feedback, and it was something that they would consider in the next lesson. They also noticed development in the students' self-correction and their noticing and editing of the similar types of errors on their own in the final draft, which indicates that CF is not only important for L2 students' development but also significant for teachers' own learning and reflection.

The study has also reaffirmed the results of Bitchener, Young, and Cameron (2005) and Bitchener and Knoch (2010) that have demonstrated that sustained written CF over a period of time can significantly increase L2 learners' linguistic accuracy levels. At the same time, ANOVA has also shown that the vetted AI mode of feedback was found to have the largest effect on students' language accuracy improvement.

### Teacher-Vetted AI Feedback and Learner Autonomy

The study also reported an increase in self-reliance and active student participation in the revision of their own work in the vetted AI feedback. Students in the vetted AI group were found to be more independent when revising the same type of errors in their final draft. Students in the focus group interviews also mentioned this autonomy as one of the benefits of using the AI writing assistant (Barrot, 2023). They said that the immediacy of vetted AI comments and suggestions allowed them to make changes in the early stage of the writing process, which gave them the opportunity to self-revise and self-correct their errors in the final draft. The assistance of vetted AI feedback, especially the immediacy of the feedback, seemed to have created a safe environment for learning in which students did not feel embarrassed by the mistakes that they had made but saw these mistakes as learning opportunities. This is in line with the Barrot's (2023) findings, which suggested that an AWCF tool that is seamlessly and thoughtfully integrated into a classroom L2 writing context by teachers can encourage student self-editing, foster metalinguistic awareness and provide scaffolding for students to improve their writing accuracy. At the same time, the vetted AI mode of feedback has the potential to reduce the cognitive load of the students as well

as address the issue of negative washback, as discussed by Bai and Hu (2017) because it allows the teacher to decide which AI comments need to be vetted and sent to the students.

### The Reliability and Internal Validity of the Cross-Over Design

This study's cross-over research design also controlled group variation by allowing each group of students to experience the AI-only, teacher-only, and teacher-vetted AI feedback during both phases. This design increases the internal validity and allows for a more accurate evaluation of the influence of the intervention (Hyland, 2019). As a result, participants were effectively balanced for ability and interest by using themselves as their own control. Every student, thus, acts as their own control, which reduced possible confounding factors such as proficiency and motivation. At the same time, every participant started with approximately the same pre-test scores, which demonstrates that the treatments were applied to equally matched groups. The only clear difference between the groups is the testing time (pre-test and post-test), and the differences in the post-test scores can be directly attributed to the feedback mode that students have received.

In terms of the study's reliability, it was supported by the identical procedures of teacher training and scoring, as noted by Ellis (2009) and Li, Link, and Hegelheimer (2015). These included vetting AI comments with the same teacher rating sheet, providing teachers with the same rubrics and guiding documents for the assessment of student work, and having them make all decisions. Teacher subjectivity in decision-making can, of course, never be excluded, but AI feedback integration creates a level playing field of feedback on which to base teacher judgment, potentially increasing inter-rater consistency. The study's human-AI feedback design, therefore, passes pedagogical validity and methodological reliability tests.

### Conclusions and Discussions of the Findings

The compatibility of the study results with the previous studies and theories has supported the use of a hybrid human-AI feedback mode as a productive approach to the best pedagogical outcome for young EFL learners. The teacher-vetted AI feedback system meets the formative feedback expectations and criteria set by Hattie and Timperley (2007) and Shute (2008), helps sustain the accuracy of L2 writers as shown in the Bitchener and Knoch (2010) and other meta-analytic studies, and contributes to the affective outcome of student autonomy as discussed in Barrot (2023). The study results, thus, are in line with previous studies and, in the terms of contributing to the discourse in the field about appropriate AI integration into language teaching and learning, make a productive contribution. In this integration, teachers are not eliminated or rendered unnecessary as evaluators. Instead, as the study results have shown, their role as evaluators or raters can be enhanced by a more strategic mediation of AI feedback. The findings, thus, in addition to

the confirmation of the teacher's role as the key evaluator in the feedback process, suggest that vetting of AI feedback can potentially open new opportunities for teacher-mediated AI feedback adaptation as a technology for L2 writing and learning.

### PEDAGOGICAL IMPLICATIONS AND RECOMMENDATIONS

The results of the cross-over study have implications for EFL pedagogy, teacher education, and curriculum design. In particular, with regard to writing instruction at the primary level, when learners are yet to fully develop metalinguistic awareness and autonomous self-regulatory strategies, the blended use of teacher-vetted AI feedback offers a viable pathway for sustainable integration of technological affordances with human pedagogical mediation. It thus makes a contribution towards addressing concerns over both over-automation on the one hand, and teacher workloads on the other.

#### Teacher Training for AI Feedback Interpretation and Adaptation

One implication is in the area of teacher training for AI feedback interpretation and adaptation. As demonstrated by the results, AI-assisted feedback can enhance efficiency and consistency in feedback provision but require teacher vetting for misinterpretation and learner demotivation (Shute, 2008). In other words, teachers need the skills and knowledge to interpret and adapt AI feedback to the learners' needs and expectations.

AI literacy training should include the evaluation of AWE tool output for quality, relevance, and pedagogical appropriateness. Hattie and Timperley (2007) identify three criteria for effective feedback that can guide teacher AI feedback training. First, it must be clear and comprehensible. Second, it must be relevant to the task and objectives. Third, it must be feedforward-oriented (focus on possible solutions, not just error identification).

Incorporating these principles into teacher training for AI feedback interpretation and adaptation can develop their capacity to select, adapt, and personalize machine-generated feedback for more meaningful scaffolding of young EFL writers' development.

AI training can also promote reflective feedback practice, including triangulating AI-generated data with classroom observation and formative assessment (Ellis, 2009). This can help ensure AI-assisted feedback aligns with Shute's (2008) three formative feedback principles: timeliness, diagnostic insight, and adaptability to learner needs.

#### Positioning AI Tools as Assistive Partners, Not Replacements

A second implication relates to the need to position AI tools as assistive partners, not replacements, in the teaching-learning process. The findings of the study have confirmed



the superior performance of teacher-vetted AI feedback over AI-only conditions in writing accuracy and engagement. This supports a growing literature that advocates for an additive, not automatic, role for AI in pedagogy (Barrot, 2023).

This study reiterates that human oversight is essential, and teacher feedback, in addition to providing linguistic and mechanical correction, can also motivate and encourage young writers to see feedback as part of a positive learning experience (Hyland, 2019; Lee, 2008). AI can do some of this work at the error detection and structural revision levels, but teachers provide unique value in contextualizing and personalizing feedback to the learners' intentions, task demands, and emotional states. This can prevent the dangers of over-correction that Truscott (1996) cautions against and support Ferris's (1999) argument for appropriately contextualized correction as a catalyst for language development.

Practically, the operational way to do this is for institutions to adopt a hybrid human-AI feedback framework. AI would generate the initial diagnostic input that teachers would review, refine, and adapt. This could significantly reduce the teacher feedback workload while ensuring quality and reliability (Bitchener & Knoch, 2010). The result would be a sustainable human-AI co-teaching model that works to support both instruction efficiency and learner trust in technology-mediated feedback.

### **Embedding Feedback Literacy in EFL Curriculum Design**

The final and arguably most important implication is the need to embed feedback literacy as a primary goal of EFL curriculum design. Feedback literacy is the ability of learners to understand, evaluate, and act on feedback effectively and constructively (Hyland, 2019). For AI-assisted writing, feedback literacy development entails teaching the learners to read, evaluate, and make meaning from AI-generated suggestions to drive their revisions instead of accepting them blindly (Feng & Bruce, 2018).

Explicitly embedding feedback literacy as part of primary EFL instruction has the potential to deepen formative learning, as Hattie and Timperley (2007) have argued. Teachers will need to model reading, assessing, and internalizing AI feedback. This may involve discussing with the learners the reasons for accepting or rejecting suggestions and encouraging them to track their improvement over time (Bitchener, Young, & Cameron, 2005).

Curriculum designers can consider developing integrated writing modules that combine peer, AI-mediated, and teacher mediation feedback. This would create multimodal feedback ecologies that expose students to multiple sources of feedback and help them see feedback as a collaborative learning process rather than a summative judgment (Li et al., 2014; Ranalli et al., 2017). Ultimately, this would help develop autonomous and strategic writers who could learn to use AI effectively while internalizing notions of accuracy and style

awareness.

## **LIMITATIONS AND FUTURE RESEARCH**

As with all research, the shortcomings of the study open a better understanding of the research field and also require more studies on the topic of the research. The present cRCT in the classroom has shown that for primary learners, AI assessment, double-checked by the teacher, has a promising impact on students' quality of writing and motivation to write. At the same time, as with all studies, there are some results that need to be considered with caution because of the limitations in the methodology and context of the study. We believe that more research is needed to lay a solid foundation for the use of AI in EFL writing. In this regard, we would like to briefly mention the most apparent of these shortcomings.

### **Sample Size and Representativeness**

The limited sample size of 60 upper-primary EFL learners from one school limited the generalisability of results. Small samples were practical for the in-class observation and teacher participation needed for this quasi-experiment but still suffered from a narrow age range, one teacher, and institutional culture, which Liu and Brown (2015) list as the primary sources of variability in L2 feedback research. Local practices and teacher beliefs can also affect the interpretation and uptake of AI feedback, according to Hyland and Hyland (2006). For these reasons, it is important to replicate this research in different school environments rather than overgeneralise from one institutional context. Future studies should involve larger and more heterogeneous samples drawn from different schools, districts, or countries with a wider variety of language backgrounds and curriculum approaches. A broader sample would not only provide stronger statistical power for the research but also validate whether the teacher-vetted AI model extends to other educational systems and learner ages.

### **Duration and Long-Term Effects**

The cross-over design with six weeks of total duration also limited the observation of delayed or long-term effects. Despite a clear short-term impact on linguistic accuracy and lexical sophistication, further long-term research would be needed to test for retention and transfer of writing gains to new genres and lower levels of scaffolding. For example, Bitchener and Knoch (2010) have proposed that linguistic accuracy changes require time and multiple cycles of feedback, revision, and self-monitoring. Shute (2008) also argues that formative feedback has the most powerful effects over longer timeframes, when learners have a chance to internalise the criteria and standards. A longitudinal design that tracks learners over one or more semesters could thus provide additional evidence on the role of teacher-vetted AI feedback in long-term improvement and self-regulated writing competence.

### Platform Dependency

Another issue concerns the use of a single proprietary AWE platform that offered grammar and lexical feedback. This research design restricted the external validity of the study because different AWE tools have diverse algorithmic foundations, feedback foci, and error classification schemes (Li, Link, & Hegelheimer, 2015; Ranalli, Link, & Chukharev-Hudilainen, 2017). Some systems specialise in surface-level error detection while others focus on more complex feedback about discourse structure or produce holistic scores (Li et al., 2014). As a result, the positive effects observed in this study may be tied to the specific features and shortcomings of this platform rather than the general affordances of AI feedback. Future research should test more AWE platforms, including open-source and commercial systems, to compare their relative strengths and weaknesses in scaffolding student writing. Cross-platform studies could also explore how differences in feedback tone, level of detail, and perceived accuracy impact learners' trust and engagement with the automated feedback.

### Learner Age and Developmental Factors

The age range of learners in this study, from 10 to 12 years, also created developmental constraints. Learners of this age are still developing their metalinguistic awareness and self-regulation skills, which could shape their responses to AI-generated feedback (Hyland, 2019). Bai and Hu (2017) have reported that younger students can become confused or overly dependent on automated feedback if they are not properly guided on how to understand and use it for error diagnosis. As a result, the results of this study may not be directly transferable to younger primary-school learners or older secondary-school students, who may have different cognitive and affective reactions to feedback. In future research, it would be useful to involve different age groups to see how feedback literacy, motivation, and technology acceptance change with cognitive development. The interaction between age, cognitive readiness, and AI-mediated scaffolding can also inform the design of age-appropriate feedback interfaces that balance automation with teacher mediation.

### Teacher Mediation and Variability

Although a standard vetting procedure was provided, variation in teacher interpretation and mediation style is still an uncontrolled variable. Hyland (2019) and Ellis (2009) have argued that teacher expertise, attitudes to technology, and knowledge of second-language acquisition theory can significantly influence how feedback is framed and delivered to learners. Differences in individual teacher cognition and professional development can also affect how they curate and contextualise automated suggestions. Even subtle differences in the vetting process could cause variations in learner outcomes that were not detected in this study. Future research could involve multi-teacher designs to capture the

effects of teacher cognition and professional development on AI feedback effectiveness. Professional learning modules could also be developed to train teachers in AI feedback literacy, equipping them with the necessary skills to better curate and contextualise automated suggestions for their students.

### Future Directions for Research

Building on the discussion above, a number of future research directions can be proposed:

- **Multi-Site and Large-Scale Validation:** Expand the teacher-vetted AI feedback model to other national, cultural, and institutional settings for cross-contextual validation.
- **Longitudinal and Mixed-Methods Designs:** Use longer research cycles to examine writing retention, motivation, and transfer across full academic terms, integrating statistical and classroom ethnographic methods.
- **Comparative AI Platform Studies:** Compare the performance and learner reception of different AWE systems, with a focus on error detection and genre adaptability, in controlled experiments.
- **Developmental and Affective Dimensions:** Investigate age, cognitive readiness, and emotional factors such as self-efficacy and engagement as mediators of AI feedback response.
- **Teacher Professional Development:** Explore the impact of teacher training in AI feedback literacy on their integration and contextualisation skills (Hattie & Timperley, 2007; Hyland & Hyland, 2006).
- **Equity and Accessibility Considerations:** Examine the adaptation of AI-based feedback tools for multilingual, resource-constrained classrooms to support inclusive EFL writing instruction.

Each of these proposed research directions could usefully build on the current findings to provide a more comprehensive and evidence-based understanding of the interaction between artificial intelligence, teacher judgement, and learner cognition in L2 writing contexts. Broadening the scope of inquiry to include temporal, technological, and pedagogical boundaries can also help ensure that teacher-vetted AI feedback continues to be developed as a sustainable and equitable educational innovation, improving both the quality and fairness of formative writing assessment.

### CONCLUSION

The purpose of this study was to explore the educational effect of teacher-vetted AI feedback on primary-level EFL learners' writing achievement in a six-week classroom cross-over design. This study aimed to find out if the AI system with teacher mediation can support students' linguistic accuracy, writing fluency, and motivation more than teacher-only and AI-only conditions. The findings add considerable evidence to support the teacher-vetted condition as the optimal balance between technology efficiency and human subjectivity.



## Empirical Summary

The cross-over design produced empirical evidence that students in the teacher-vetted AI feedback condition outperformed the others in terms of grammatical accuracy, lexical variation, and text coherence overall. In particular, the AI-only feedback condition had faster error correction on the surface level, whereas the vetted one had deeper syntactic and semantic awareness. This is consistent with the theoretical view that effective feedback should facilitate information processing and learner interpretation (Hattie & Timperley, 2007; Shute, 2008). Additionally, the teacher-vetted condition was associated with significantly higher clarity, motivation, and trust in feedback quality among learners, which is in line with prior research on how source credibility affects student motivation (Bai & Hu, 2017; Li, Link, & Hegelheimer, 2015).

The teachers also seem to benefit from the teacher-vetted model. With the help of error detection algorithms, they complete the initial diagnostic process faster and spend less time on repetitive correction. Instead, they focus more on providing explanations and scaffolding, which endorses the theoretical perspectives on feedback as a dialogic process of meaning negotiation rather than information transmission (Hyland, 2019; Hyland & Hyland, 2006). In this sense, the empirical results suggest a symbiotic relationship between automation and teacher input: the former enables time and energy efficiency, while the latter reinstates the missing contextual and emotional sensitivity.

## Theoretical Significance

Theoretically, this study has several implications for the literature on corrective feedback and L2 development. First, the teacher-vetted framework disrupts the false binary between AI and teacher feedback. The empirical results confirm the theoretical position that these two elements can coexist in a mediated and scaffolded learning space (Ellis, 2009; Ferris, 1999). Feedback only becomes developmentally meaningful when students comprehend the cause for correction and can transfer that knowledge to future writing tasks (Ellis, 2009; Ferris, 1999). The teacher-vetted framework puts this theoretical insight into practice by making the teachers the interpretive gatekeepers of AI feedback and humanizing it by adapting the suggestions to the learner's age, proficiency, and writing assignment.

Second, the study makes a theoretical contribution to feedback-literacy theory by demonstrating how young learners process CF information differently when mediated by a trusted human. The combination of algorithmic consistency and teacher empathy leads to both cognitive uptake and affective acceptance, which are needed for L2 development in the long term (Bitchener & Knoch, 2010; Kang & Han, 2015). In this sense, the teacher-vetted framework also resonates with the current educational paradigm shift towards human-AI collaboration, in which technology is used as a means to facilitate rather than replace professional decisions (Barrot, 2023).

## Pedagogical Implications

Pedagogically, the validation of the teacher-vetted AI feedback framework has several implications for EFL teachers. It means that primary class writing instruction can be both more time-efficient and more personalized without a loss of quality. Supervising AI feedback gives teachers the opportunity to keep academic oversight and make sure that students receive high-quality and developmentally appropriate comments. The model also contributes to learner autonomy, as students are first exposed to AI suggestions, then make further refinements after dialogue with the teacher, which cultivates metacognitive awareness and self-editing skills.

## Concluding Thoughts and Suggestions for Future Research

In conclusion, teacher-vetted AI feedback is a balanced, evidence-based approach to modern EFL writing instruction. It combines the interpretive sensitivity of human teachers with the analytical power of AI algorithms. This hybrid framework allows for improvements not only in writing accuracy and learner engagement but also a demonstration of a broader pedagogical trend towards ethical and sustainable technology integration. The current study, with its small sample size and short time frame, is only a preliminary empirical validation of the research questions, although it provides a solid foundation for future large-scale and longitudinal research in other linguistic and cultural contexts. Future studies can look at retention effects, teacher training models, or performance differences on different AWE platforms.

In general, the study's findings confirm that teacher-vetted AI feedback can serve as a middle ground between automated efficiency and human judgment to make a real linguistic difference and facilitate learner agency in primary EFL writing instruction.

## REFERENCES

- [1] Hyland, K. (2019). *Second language writing*. Cambridge university press.
- [2] Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112.
- [3] Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, 78(1), 153-189.
- [4] Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language learning*, 46(2), 327-369.
- [5] Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of second language writing*, 27, 1-18.
- [6] Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of second language writing*, 14(3), 191-205.
- [7] 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.

- [8] Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language teaching*, 39(2), 83-101.
- [9] Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing*, 30, 66-81.
- [10] Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal: An Open Access Refereed Journal for World Language Educators*, 1(1).
- [11] Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of second language writing*, 8(1), 1-11.
- [12] Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond?. *Educational Psychology*, 37(1), 67-81.
- [13] Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language learning*, 60(2), 309-365.
- [14] Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, 36(4), 584-607002E
- [15] Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten month investigation. *Applied linguistics*, 31(2), 193-214.
- [16] Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, 44, 66-78.
- [17] Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1), 8-25.
- [18] Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The modern language journal*, 99(1), 1-18.
- [19] Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of second language writing*, 19(4), 207-217.
- [20] Lee, I. (2008). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of second language writing*, 17(3), 144-164.

