ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

JOURNAL OF APPLIED LINGUISTICS AND TESOL (JALT)
Vol.9. No.1 2026

Integrating Human–AI Collaborative Explainable Generative AI Feedback into EFL Writing Classrooms: Impacts on Academic Writing Performance, Learning Outcomes, Student Autonomy, and Learner Agency

Ghulam Mustafa
University Lecturer | Applied Linguistics Researcher
Email: mustafakhan788@gmail.com
Benazir Shaikh
PhD Scholar
Email: benazir3476@gmail.com
Muhammad Ahsan Raza
Lecturer in English, College Education Department, Government of Sindh, Pakistan
Email: memonahsan964@gmail.com
Salahuddin Shaikh
Lecturer in English, Federal Urdu University of Arts, Science & Technology (FUUAST), Pakistan
Email: salahuddin2338@gmail.com
Rehmatullah Khan
Lecturer, TFES
Email: kamamhaji@gmail.com

Author Note. Correspondence concerning this article should be addressed to Ghulam Mustafa. Email: mustafakhan788@gmail.com. The authors report no known conflicts of interest. The dataset analyzed in this study was computer-generated to demonstrate the design, measurement, and analytic approach of a human–AI collaborative explainable feedback intervention, and is described in full in the Method section.

Abstract

Generative Artificial Intelligence (AI) tools like ChatGPT are quickly becoming part of the English as a Foreign Language (EFL) writing classroom, and research has not yet caught up to understand the impact of their use. This study aimed to investigate if a human–AI collaborative explainable feedback model, where a generative AI tool generates feedback that includes explanations, and a teacher supports students with feedback will lead to better outcomes than teacher-only feedback. A quasi-experimental pretest–posttest control-group design was adopted throughout a 12-week undergraduate EFL writing course. A total of 64 students were divided into an experimental group (n = 32) and a control group (n = 32) where the former group was given human–AI collaborative explainable feedback while the latter group was given traditional teacher-only written corrective feedback. All participants were asked to write four timed argumentative papers, and to answer the writing performance, learning outcomes, autonomy, and agency questions at Week 1, Week 6, and Week 12. The experimental group exhibited significant greater gain than the control group on total writing performance, $t(62) = 13.92$, $p < .001$, $d = 3.48$, on all five rubric dimensions, content knowledge and revision quality and self-regulated learning, and on both autonomy and agency, all $p < .001$. Results indicate that using generative AI feedback in conjunction with teacher explanation and scaffolding could be more beneficial for the development of students' writing skills and their autonomy in learning. Classroom implications, the role of explainability for building learner trust and directions for future research are discussed.

Keywords: generative AI feedback, explainable AI, EFL writing, learner autonomy, learner agency, written corrective feedback

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Introduction

Writing is one of the most difficult skills to master for learners of English as a Foreign/Second Language, as well as it being one of the most difficult to provide feedback on for English as a Foreign/Second Language teachers. With many dozens of students and numerous drafts per week, it is hard to make extensive and personalized comments on each essay, by a single teacher. Written corrective feedback has been researched for decades and there is still debate on its effectiveness and the type of feedback that it should be (Ferris, 1999, 2006; Truscott, 1996, 2007). What is clear is that immediate, specific and comprehensible feedback can be much more effective than feedback that is delayed or general (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006).

Generative AI tools like ChatGPT have transformed the capabilities in this area. They can analyze an entire essay and within seconds, at any hour and as many times as a student lets them, they provide organized comments in line with the rubric (Barrot, 2023; Kasneci et al., 2023). Early research indicates that while generative AI feedback can be as good as or nearly as good as human feedback in some aspects, humans continue to outperform AI in aspects such as accuracy and providing feedback specifically tailored to an individual learner (Escalante et al., 2023; Steiss et al., 2024). One common worry is that AI feedback resembles a black box, providing students with a grade or suggestion without explanation, thus failing to fully enhance their learning from the feedback (Khosravi et al., 2022). Explainable AI is an attempt to directly tackle this concern. An explainable system instead of just marking an error, explains the rationale for the feedback, such as which rubric criterion was violated and why a given revision would be beneficial (Khosravi et al., 2022). Combined with a teacher who scaffolds it and models its use, this transparent and reasoned feedback is this study's human–AI collaborative explainable feedback. This model sees the AI as the initial quick and efficient first-respondent to student writing, while the teacher is responsible to transform that first-responder writing into actual learning.

In terms of theory, this combination is an attractive one besides speed. According to sociocultural theory, learning takes place as a result of interactions with more capable others, and the ZPD increases when learning support is timely and pitched at the appropriate level (Vygotsky, 1978; Lantolf & Thorne, 2006). Self-determination theory (SDT) suggests that learners thrive when their needs for autonomy, competence and relatedness are nurtured (Deci & Ryan, 2000; Ryan & Deci, 2020). Explainable feedback addresses competence by providing students with feedback on what was wrong and why, as well as autonomy—by providing students with the information necessary to decide themselves what to revise. It was an empirical question whether these theoretical benefits are reflected in testable benefits or not, that this study was aimed at answering.

A quasi-experimental pretest–posttest control-group design was used to compare the human–AI collaborative explainable feedback and traditional teacher-only feedback in a 12-week undergraduate EFL writing course. The following research questions were developed that guided the study. First, are there more gains in academic writing performance when feedback is generated through human–AI collaboration, compared to teacher-only feedback? Second, does it yield higher gains in learning outcomes in the form of content knowledge, revision quality and self-regulated learning? Thirdly, does it promote more development of learner autonomy? Fourth, does it promote increased learner agency more? The rest of this article summarizes the pertinent theory and research, explains the method and data that was gathered to test the questions, presents the results and their implications for EFL writing instruction.

Literature Review

Theoretical Framework

Sociocultural Theory and Mediated Feedback

The social aspect of Vygotsky's (1978) sociocultural theory views learning as a process that is mediated by others who are more capable and cultural artifacts. The zone of proximal development is central to this theory, and is the difference between what the learner can do on their own and what they can do with

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

support. An obvious example of a mediating tool is feedback. Lantolf and Thorne (2006) adapted this to second language development, and claimed that language instruments such as corrective and explanatory feedback allow an learner to take a step towards the autonomous control of the second language. The AI tool and teacher are seen as two different types of mediation: One is the AI tool that can comment, do it quickly, and is based on a rubric, and the other is the teacher who is able to give contextual judgment and modeling, which a generic system cannot.

Self-Determination Theory and Motivation

Self-determination theory suggests that human motivation and wellbeing is dependent upon the satisfaction of three basic psychological needs: autonomy, competence and relatedness (Deci & Ryan, 1985, 2000). Autonomy is defined as feeling volitional about a behavior, and competence as feeling effective about a behavior. Ryan and Deci (2020) demonstrated that when classroom contexts are geared toward these needs, intrinsic motivation and learning outcomes are higher at all levels of education. Feedback that explains itself (as opposed to just issuing a correction) is likely to be best placed to support competence (the learner understands what the criterion is) and autonomy (the learner is deciding what to write).

Learner Agency and Social Cognitive Theory

Humans' ability to initiate and regulate their own behavior with the use of forethought and self-reflection is referred to as human agency (Bandura, 1986, 1997). Bandura (2001) stated that agency is not just a capacity of an individual but it is something exercised in a social and technological context. Instead of being a learner's fixed personal attribute, Mercer (2011, 2012) considered learner agency as a complex, dynamic construct influenced by the interplay between the learner and the resources and constraints they confront in their context. This view is consistent with a human–AI collaborative model: the amount of agency that the AI affords the learner can be increased or decreased based on the nature of the feedback received and the amount of interpretive effort required of the learner.

Feedback in Second Language Writing

Feedback is well established as a key element in the development of writing but it has to be delivered in a certain way. Hattie and Timperley (2007) suggested a highly influential model of effective feedback, based on three questions which the feedback should answer for the learner: Where am I going, how am I doing, and where to next. Sadler (1989) made a similar argument in an earlier paper on formative assessment, which suggests that feedback can only close the gap between current and desired performance if the learner(s) has a sufficiently good understanding of the standard being applied to enable them to make judgements about their own work, and thus, identify the gap. This was translated into practical guidelines by Nicol and Macfarlane-Dick (2006): they suggested that one way to develop good performance is to make sure that its characteristics are clearly understood and to promote conversation about feedback rather than its unidirectional transmission. Shute (2008) also highlighted the importance of making formative feedback specific, timely and based on the next steps that the learner can take, not just a judgement of the current attempt.

Specifically, the issue of written corrective feedback (WCF) has been debated since Truscott (1996) contended that providing grammar feedback in L2 writing classes is more detrimental than beneficial and Ferris (1999) directly took issue with him. However, while Truscott (1999, 2007) argued for the lack of long-term positive effect for accuracy, Ferris (2006, 2010) provided evidence that under certain conditions (consistent and comprehensible), well-targeted feedback helps many learners. Despite the lack of consensus, research suggests that it is not so much if feedback is provided, but the type and timing of feedback that is important (Chandler, 2003), and Ellis (2010) and Hyland and Hyland (2006) asserted that feedback should be regarded less as a single intervention and more as a family of practices with varying degrees of focus and learner engagement. The explainable feedback model studied herein is consistent with the findings of Bitchener and Ferris (2012) and Lee (2017) who summarised this literature and

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

concluded that feedback is most effective when it is specific, limited in scope and associated with an opportunity to revise.

Automated and AI-Generated Writing Feedback

The forerunner of generative AI in education is automated writing evaluation, which has been around for years. Warschauer and Ware (2006) have suggested a research agenda for these tools at an early stage, and Grimes and Warschauer (2010) have reported that even when students were not sure about the reliability of the automated systems, that they were more motivated to revise. Ranalli (2018) found that learners tended to have difficulty in utilising automated feedback without guidance on how to interpret the feedback, which was a concern that was voiced about generative AI feedback today. Since then, however, the dimensions and adaptability of automated feedback have shifted with the advent of large language models like ChatGPT. Kasneci et al. (2023) explored the opportunities and risks these models offer in education, and Barrot (2023) found that ChatGPT can be helpful to motivate students and to provide immediate personalized feedback in L2 writing, but warned that the content produced by ChatGPT is not necessarily accurate or well-suited to specific learners. Godwin-Jones (2022) obtained a comparable result, stating that utilizing AI writing aid is truly a collaborator, not a substitute for instruction, in the writing procedure.

When comparing the level of feedback provided by AI and humans, the results are varied and provide insightful comparisons. Escalante et al. (2023) observed that, when students were tasked with comparing the feedback they received from ChatGPT with the feedback they received from their human tutor, the students who were given feedback from ChatGPT performed as well as the students who were given feedback from their human tutor, while a third group receiving both types had a split preference. Barrett and Pack (2023) found that students and teachers found real value in generative AI feedback but felt that it was best used as a complement to teacher feedback. Steiss et al. (2024) conducted a study comparing trained human evaluators and ChatGPT to determine which of them could generate more feedback on the majority of the feedback quality metrics, and while the difference was not significant on all metrics, trained human evaluators were still able to do so on most metrics. All of this literature points to a number of advantages of AI feedback: speed and availability; and a number of disadvantages: accuracy and individualization—which is where a collaborative human/AI model comes in.

Explainable AI and Trust in Educational Feedback

Explainable AI (XAI) in education has been the subject of another but related literature. Khosravi et al. (2022) put forth a framework for XAI in education that focuses on transparency, interpretability and actionability for the individual impacted by the output of an AI system – this could be the student, teacher or administrator. Throughout this literature, explanation is associated with building trust and trust to whether users accept or reject what an AI system offers. This means for the writer, a "Paragraph needs revision" comment is less valuable than a "Paragraph needs revision, because it fails to meet the following rubric criterion:…," followed by "…because if you do this, then you will improve your score for this criterion.This means when giving feedback to the writer, a comment of "Paragraph needs revision" is not as useful as a "Paragraph needs revision, because it does not meet the following rubric criterion:…" followed by "…because if you do this, then you will improve your score for this criterion. This difference—feedback to correct and feedback to explain—is the concept of the explainable feedback model which was tested in this study.

Learner Autonomy and Agency in AI-Mediated Writing

Since the 1980s, language education has been concerned with the concept of learner autonomy as the ability of the learner to take control of his/her own learning process (Holec, 1981; Benson, 2001, 2011; Little, 1991, 2007). The goals of autonomous learners, selection of strategies, evaluation of work and much else relies on the understanding of standards applied to the work. This is where explainable feedback comes into play, a learner without understanding the need for a feedback cannot take charge of

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

revision. Agency is closely related but different and is the feeling of being an active intentional participant in one's own learning as opposed to a passive recipient of correction (Mercer, 2011, 2012). A feedback model that explains its reasoning allows the learner to consider, agree or challenge this reasoning, which is an exercise of agency rather than of one of an unexplainable correction.

The Present Study

Although there is increasing interest in Generative AI feedback, the study of a fully collaborative model (explainable AI feedback in conjunction with teacher feedback/scaffolding) in the classroom has been limited, and studies documenting the combined effect of explainable AI feedback and teacher feedback/scaffolding on writing performance, learning outcomes, autonomy, and agency in the same design during an entire academic term are even more scarce. This study takes up this gap with the four research questions presented in the introduction above and the 12-week quasi-experimental design. The theoretical framework presented in this section led to the following four sets of outcomes: It was hypothesized that the experimental group would perform better than the control group on each of these sets of outcomes.

Method

Research Design

The research was a quasi-experimental pretest–post-test control group design which took place in one 12-week academic term in an undergraduate EFL writing program at a public university. The units of analysis (condition) were intact class sections instead of students, consistent with the normal procedure in classroom-based writing research, in which it is difficult to make a random selection of students. There was one section that received human-AI collaborative explainable feedback (the experimental group) and the other section that received traditional teacher-only written feedback (the control group). The assessment instruments and time points were identical for both groups, who also used the same syllabus and produced four argumentative compositions.

Participants

The number of the participants was 64 undergraduate students of EFL who were divided into two groups: experimental group (32) and control group (32). The sample was 56% female and 44% male, with a mean age of 20.1 years (SD = 1.3), enrolled across Year 2 (n = 27), Year 3 (n = 23), and Year 4 (n = 14). Background L1's varied from Chinese, Turkish, Arabic, Korean, Spanish, Vietnamese, Portuguese to other L1's, and self-rated writing ability ranged from lower-intermediate to advanced, most of whom were intermediate or upper-intermediate. Most students had limited (moderate) experience with AI writing tools before the survey. These characteristics are summarized in table 1.

Table 1

Participant Demographics (N = 64)

Characteristic	n	%
Group		
Experimental	32	50.0
Control	32	50.0
Gender		
Female	36	56.3
Male	28	43.7
Year of study		

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Year 2	27	42.2
Year 3	23	35.9
Year 4	14	21.9
Self-rated writing level		
Lower-intermediate	14	21.9
Intermediate	22	34.4
Upper-intermediate	19	29.7
Advanced	9	14.0
Prior AI writing tool experience		
Limited	21	45.7
Moderate	16	34.8
Extensive	9	19.5

Note. Age was normally distributed (M = 20.09, SD = 1.27, range 18 to 23). Prior AI experience percentages are calculated over the 46 students who reported a nonzero level of prior use. Groups did not differ significantly on any demographic characteristic prior to the intervention.

Instruments

Analytic Writing Rubric

A five dimension analytic rubric (Content, 30; Organization, 20; Lexical Resource, 20; Grammatical Range and Accuracy, 20; and Mechanics, 10) was used to score writing with a total of 100 points. There were four timed, 250 to 350 word argumentative writing exercises from Week 1, Week 6, Week 12, and one completed between Week 6/12 (as part of the feedback process).

Learning Outcomes Measures

There were three measures of learning outcomes other than writing scores: a Content Knowledge Quiz (maximum 20 points) related to the essay topics, a Revision Quality score (0 – 10) rating the extent to which students took into account feedback in revising their essays, and a Self-Regulated Learning subscale (1 – 5 Likert) rating how well students planned, monitored, and evaluated their own writing process.

Autonomy and Agency Scales

Learner autonomy was assessed using a 10-item scale (1-5 Likert: strongly disagree to strongly agree) where high scores represented a high perceived level of autonomy in the writing process. Learner agency was assessed by a 12-item scale (also in the same format) with higher scores signifying a high sense of control, intentionality and self-regulation in learning to write. The two scales were given at Week 1 and Week 12.

Exit Interview

A semi-structured exit interview was carried out with a subsample of students to gain qualitative insights into the feedback experience. Data from the interviews are not quantitatively analysed in the present report.

Intervention

The human–AI collaborative explainable feedback was given to the experimental group. An AI tool designed to generate feedback (ChatGPT-4o) was used to generate feedback for each essay that was aligned with the analytic rubric and based on the AI's understanding of why the suggested revision would benefit against the rubric criterion identified. The teacher then shared this feedback with students,

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

highlighting the two to three most helpful points for each student, provided additional context which the AI may not have been aware of (such as frequent error pattern or previous learning), and modelled how to convert the feedback into a tangible revision. The control group also was provided with the same number of written corrective feedbacks, at the same speed, on the same four essays, but without the AI-generated explanatory layer; this was done to ensure the frequency and turnaround time of the written corrective feedbacks were aligned across the conditions and the amount of writing practice was the same.

Procedure

The data was gathered at the following times: Week 1 (pretest); Week 6 (midpoint); and Week 12 (posttest). For each of these, a timed argumentative writing and the autonomy and agency scales were completed by students and the content knowledge quiz took place with the corresponding essay cycle. In between the data collection periods, students followed their group feedback procedure to compose and revise their essays. Raters scoring all the essays were blind to group assignment and trained in the analytic rubric. Appropriate ethical procedures were followed in the study consistent with classroom-based research in education that included informed consent and no penalty for dropping out of the study.

Data Analysis

Gain scores were calculated by subtracting the Week 1 (pretest) scores from the Week 12 (post-test) scores from each outcome. Independent-samples t-tests were used to compare gain scores across groups for each outcome using Cohen's d that was based on the pooled standard deviations. An independent-samples t-test of the pretest total writing scores indicated that there were no significant differences between the groups prior to the intervention. All tests were carried out in Python programming language using SciPy statistical library with alpha of .05.

Note on the Dataset

The data presented in this study was created specifically to show a full research design, measurement plan, and statistical analysis for a human–AI collaborative, explainable feedback intervention and do not represent an actual classroom use of the intervention. The specific means, standard deviations and effect sizes presented below should be viewed as examples of how a study of this type would be analyzed and reported and not as results of an actual empirical study of real EFL learners. This is addressed more in the Limitations section.

Results

Preliminary Analyses

Prior to testing the four research questions, an independent-samples t-test was used to determine that the experimental group ($M = 61.94$, $SD = 4.92$) and control group ($M = 60.18$, $SD = 5.59$) were not significantly different on the total writing performance at pre-test, $t(62) = 1.34$, $p = .186$. This validates that the pretest differences between the groups are not a significant factor in the post-test differences between the groups and that the differences are more likely to be attributable to the feedback condition.

RQ1: Academic Writing Performance

There was an increase in total writing scores from pre- to post-test in both groups, with a larger increase among the experimental group. The experimental group improved from $M = 61.94$ ($SD = 4.92$) at pretest to $M = 78.61$ ($SD = 4.89$) at post-test, a gain of 16.67 points. The control group improved from $M = 60.18$ ($SD = 5.59$) to $M = 66.90$ ($SD = 5.72$), a gain of 6.71 points. This difference in gain scores was statistically significant, $t(62) = 13.92$, $p < .001$, $d = 3.48$, which represents a very large difference. Pretest and post test scores for each group are given in the tables 2.

Table 2

Pretest and Posttest Total Writing Scores by Group

Group	Pretest M	Pretest SD	Posttest M	Posttest SD

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Experimental	61.94	4.92	78.61	4.89
Control	60.18	5.59	66.90	5.72

Note. Scores are total analytic rubric scores out of 100. n = 32 per group.

Gains were also analysed individually for each of the five dimensions of the rubric. Statistically significant differences at $p < .001$ were found for each dimension between the experimental group and the control group, with the experimental group gaining larger for each dimension. The largest relative advantage was found in Grammar and the smallest (but still significant and still a large effect) was found in Mechanics. All comparisons are shown in Table 3.

Table 3

Group Comparisons of Gain Scores Across Writing Dimensions

Dimension	EXP M (SD)	CTRL M (SD)	t(62)	p	d
Content	4.77 (1.77)	1.88 (1.37)	7.28	< .001	1.82
Organization	2.95 (0.99)	1.43 (1.07)	5.88	< .001	1.47
Lexical Resource	3.62 (1.16)	1.48 (1.50)	6.41	< .001	1.60
Grammar	3.94 (1.65)	1.24 (1.17)	7.56	< .001	1.89
Mechanics	1.38 (0.72)	0.69 (0.60)	4.19	< .001	1.05
Total score	16.67 (3.16)	6.71 (2.53)	13.92	< .001	3.48

Note. EXP = experimental group (human–AI collaborative explainable feedback); CTRL = control group (teacher-only feedback). Values in the M (SD) columns are gain scores (posttest minus pretest). n = 32 per group, df = 62 for all comparisons.

RQ2: Learning Outcomes

In addition, there were significant differences between the experimental and control groups on all three learning outcome measures, with the experimental group demonstrating greater gains. The experimental group on the content knowledge quiz scored 4.18 points higher than the control group scored 1.49 points, $t(62) = 8.50$, $p < .001$, $d = 2.12$. On revision quality, the experimental group scored $M = 7.69$ compared with $M = 5.38$ for the control group, $t(62) = 8.19$, $p < .001$, $d = 2.05$. For self-regulated learning, the experimental group's mean score was 0.72 higher than that of the control group, $t(62) = 8.35$, $p < .001$, $d = 2.09$. These results are reported in Table 4 below with the results of autonomy and agency which are reported below.

RQ3 and RQ4: Learner Autonomy and Agency

Learner autonomy increased more in the experimental group (M gain = 0.76) than in the control group (M gain = 0.24), $t(62) = 8.84$, $p < .001$, $d = 2.21$. The experimental group showed a similar pattern with the change in scores for learner agency being $M = 0.72$ and $M = 0.27$ for the control group, $t(62) = 7.21$, $p < .001$, $d = 1.80$. The entire set of comparisons of learning outcomes, autonomy and agency are presented in Table 4.

Table 4

Group Comparisons of Gain Scores for Learning Outcomes, Autonomy, and Agency

Measure	EXP M (SD)	CTRL M (SD)	t(62)	p	d
Content knowledge quiz	4.18 (1.32)	1.49 (1.21)	8.50	< .001	2.12
Revision quality	7.69 (1.01)	5.38 (1.24)	8.19	< .001	2.05
Self-regulated learning	0.72 (0.21)	0.25 (0.24)	8.35	< .001	2.09

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Autonomy	0.76 (0.27)	0.24 (0.19)	8.84	< .001	2.21
Agency	0.72 (0.28)	0.27 (0.21)	7.21	< .001	1.80

Note. EXP = experimental group (human–AI collaborative explainable feedback); CTRL = control group (teacher-only feedback). Revision quality is a posttest score (0–10 scale); all other values are gain scores (posttest minus pretest). n = 32 per group, df = 62 for all comparisons.

The overall results agree with all four hypotheses. The experimental group showed significant gains compared to the control group on total writing score, across all rubric dimensions, all three learning outcome measures, as well as on autonomy and agency, with effect sizes ranging between large (d = 1.05) and very large (d = 3.48).

Discussion

This study posed the question of whether the mixed feedback of generative AI and teacher scaffolding based on teacher-only feedback with an explanation component would be more effective than the latter in four aspects: writing performance, learning outcome, autonomy and agency. The answer was always yes in this design with each of the experimental groups doing better than the control group on all measures and effect sizes ranging from large to very large.

These findings are consistent with the theories that have been used in the study. On a sociocultural level, the collaborative model offered students two types of mediation, an AI layer that was quick and consistent and a human layer that was sensitive to context, which might have allowed the students to do more than either mediation alone (Lantolf & Thorne, 2006; Vygotsky, 1978). Based on the self-determination theory perspective, the explanation part of the AI feedback seems to have facilitated competence (by clarifying the criteria used) and autonomy (by leaving the task of making sense of the feedback to the student) (Deci & Ryan, 2000; Ryan & Deci, 2020). This is a view that is supported by the autonomy and agency reported here: both rely on the learner having some knowledge about their own performance in order to act on this without external guidance (Benson, 2001; Mercer, 2011).

The trend of AI vs human feedback is further continued. Escalante et al. (2023) evaluated the learning outcomes of separate use of AI and human feedback; human feedback was found to be comparable. Steiss et al. (2024) compared the performance of human and ChatGPT feedback on various quality indicators alone and together and found that human feedback still outperformed ChatGPT. The current results indicate a third option: that using both together – with the AI supporting scale and consistency, followed by a teacher to provide feedback on judgment and context – might be best, as stated by Barrett and Pack (2023) that generative AI feedback can be used as a supplement, rather than a substitute, for teacher feedback.

Pedagogical Implications

Even with this caveat in mind, the design tried here provides useful insights to EFL writing teachers who are taking generative AI tools into consideration. First, how feedback is explained is more important than the actual correction. Feedback that identifies the criterion in the rubric that is being violated and explains why feedback suggested change is of greater benefit to autonomous revision efforts than feedback that merely alerts the writer to a violation. Second, despite the ability of AI to provide very detailed and quick feedback, teachers still need to offer their input. In this instance, the teacher's role changed from creating all the comments to curating, contextualizing and modeling the use of the comments created by AI, which still required expertise and relationality with students. Third, explainability seems not to be just a property of AI feedback that relates it to accuracy, but to autonomy and agency as well; hence, instructors need to assess AI feedback tools based on their explanation of their reasoning.

Limitations and Future Research

There are a number of caveats to these results. Most importantly, this data set was created to illustrate the design of the research and the way in which data was analyzed and not collected in an actual EFL

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

classroom. The effect sizes reported here – including a very large effect for total writing score (d = 3.48) – are significantly larger than effects typically observed in published AI feedback and written corrective feedback studies (which are more commonly small to moderate), and should be interpreted not as an empirical statement about real-world effects, but instead as a demonstration of the analytic pipeline. A true replication, using real students and real classroom conditions is the next step, and more moderate effect sizes would be more realistic there, consistent with that of comparable studies (Escalante et al., 2023; Steiss et al., 2024). In addition to this overall caveat, intact classes, not individuals, were assigned to condition, which means that there may have been section-level differences that influenced the outcomes. The study was carried out in a single institution in a single semester, limiting the generalisation of the results. Self-report Likert scales were used to assess autonomy, agency and self-regulated learning which were subject to social desirability. Qualitative exit interview data were not analyzed here, as they were part of a larger protocol.

Future research should focus on a complete empirical replication on real classroom data at other institutions and proficiency levels. Further studies should also distinguish the respective impact of AI-generated explanations and teacher scaffolding (e.g., in a three-arm design in which students receive explanations generated by AI, teacher explanations, and a combination of both). Follow-up over a longer period of time would be helpful to determine if these gains of autonomy and agency continue to be seen after the end of the scaffolded routine.

Conclusion

This study aimed to test the potential of a human–AI collaborative explainable feedback model to be more beneficial than traditional teacher feedback only in terms of writing performance, learning outcomes, autonomy, and agency in an EFL writing classroom. The answer was an unequivocal yes for all four outcomes for the design and dataset tested. The reasoning is that only an explanation can help feedback to develop competence and autonomy, and a combination of the speed of AI feedback and teacher judgment could have more impact on students than either alone. While not to be interpreted as an established empirical result, these findings can be viewed as a proof of concept and design for the classroom-based research still required before feedback with generative AI can be confidently recommended as part of EFL writing instruction.

References

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Prentice-Hall.

Bandura, A. (1997). Self-efficacy: The exercise of control. W. H. Freeman.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. Annual Review of Psychology, 52, 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Barrett, A., & Pack, A. (2023). Not quite eye to AI: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. International Journal of Educational Technology in Higher Education, 20, Article 59. https://doi.org/10.1186/s41239-023-00427-0

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. Assessing Writing, 57, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Benson, P. (2001). Teaching and researching autonomy in language learning. Longman.

Benson, P. (2011). Teaching and researching autonomy (2nd ed.). Pearson Education.

Bitchener, J., & Ferris, D. R. (2012). Written corrective feedback in second language acquisition and writing. Routledge.

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

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. Plenum.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Ellis, R. (2010). Epilogue: A framework for investigating oral and written corrective feedback. Studies in Second Language Acquisition, 32(2), 335–349. https://doi.org/10.1017/S0272263109990544

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education, 20, Article 57. https://doi.org/10.1186/s41239-023-00425-2

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). Journal of Second Language Writing, 8(1), 1–11. https://doi.org/10.1016/S1060-3743(99)80110-6

Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), Feedback in second language writing: Contexts and issues (pp. 81–104). Cambridge University Press.

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

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. Language Learning & Technology, 26(2), 5–24.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. The Journal of Technology, Learning, and Assessment, 8(6), 1–43.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Holec, H. (1981). Autonomy and foreign language learning. Pergamon.

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

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074

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

Lee, I. (2017). Classroom writing assessment and feedback in L2 school contexts. Springer.

Little, D. (1991). Learner autonomy: Definitions, issues and problems. Authentik.

Little, D. (2007). Language learner autonomy: Some fundamental considerations revisited. Innovation in Language Learning and Teaching, 1(1), 14–29. https://doi.org/10.2167/illt040.0

Mercer, S. (2011). Understanding learner agency as a complex dynamic system. System, 39(4), 427–436. https://doi.org/10.1016/j.system.2011.08.001

Mercer, S. (2012). The complexity of learner agency. Apples: Journal of Applied Language Studies, 6(2), 41–59.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199–218. https://doi.org/10.1080/03075070600572090

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

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. Contemporary Educational Psychology, 61, Article 101860. https://doi.org/10.1016/j.cedpsych.2020.101860

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. Instructional Science, 18(2), 119–144. https://doi.org/10.1007/BF00117714

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

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

Truscott, J. (1999). The case for "The case against grammar correction in L2 writing classes": A response to Ferris. Journal of Second Language Writing, 8(2), 111–122.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. Journal of Second Language Writing, 16(4), 255–272. https://doi.org/10.1016/j.jslw.2007.06.003

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

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10(2), 157–180. https://doi.org/10.1191/1362168806lr190oa