Journal of English Language and Education
ISSN 2597- 6850 (Online), 2502-4132 (Print)
Journal Homepage: https://jele.or.id/index.php/jele/index

OPEN ACCESS Article

Improving Grammar Accuracy in Contextual Writing Tasks Using AI Feedback Systems: A Study of EFL Learners at Science University of Al Mawaddah Warrahmah Kolaka

https://doi.org/10.31004/jele.v11i1.2006

*Vill Janna Ningzi, Andi Nurul Aulia, Suhrah, Rahmi, Muthahharah Idris[abcde]
[12345]Universitas Sains Islam Al Mawaddah Warrahmah Kolaka, Indonesia
Corresponding Author: villjannaningzi1707@gmail.com

A B S T R A C T

This study examined the effectiveness of artificial intelligence (AI) feedback systems in improving grammatical accuracy in contextual writing tasks among English as a Foreign Language (EFL) learners. Thirty undergraduate students from twelve different study programs at Universitas Sains Islam Al Mawaddah Warrahmah Kolaka participated in this research as part of a general English course. Using a mixed-methods quasi-experimental design with pretest-posttest control group framework, this study compared the effectiveness of AI-driven feedback (via Grammarly) versus conventional teacher feedback in improving grammatical accuracy within authentic writing contexts. Data were collected through written assignments, error analysis, and student perception interviews. Results demonstrated that students receiving AI feedback demonstrated statistically significant improvements in grammatical accuracy ($M = 82.5$, $SD = 6.3$) compared to the control group ($M = 74.3$, $SD = 7.1$), with $p < 0.001$. Notably, AI systems excelled at detecting and correcting morphosyntactic errors, subject-verb agreement violations, and determiner misuse. Qualitative findings revealed that students perceived AI feedback as immediate, non-judgmental, and conducive to self-correction and engagement. However, students also expressed concerns about over-reliance on automated tools. This research suggested that a hybrid approach integrating AI feedback with traditional instructor guidance offers optimal outcomes for developing grammatical competence in contextual writing. Pedagogical implications include leveraging AI tools as formative assessment instruments alongside explicit grammar instruction and peer review mechanisms.

Keywords: *AI Feedback Systems, Grammatical Accuracy, Contextual Writing, EFL Learners, and Automated Writing Evaluation.*

Article History:
Received 09th January 2026
Accepted 08th February 2026
Published 10th February 2026

INTRODUCTION

The development of English writing proficiency represents a significant challenge for EFL (English as a Foreign Language) learners, particularly in non-English speaking contexts. Writing demands the integration of multiple linguistic competencies, including syntactic accuracy, lexical appropriateness, organizational coherence, and pragmatic appropriateness (Hyland & Hyland, 2006). Among these competencies, grammatical accuracy constitutes a foundational component that directly influences the clarity, effectiveness, and acceptability of written communication (Ferris, 2011). Persistent grammatical errors not only reduce textual readability but also undermine academic credibility and professional communication quality (Bitchener, 2008). From a theoretical perspective, grammatical accuracy is closely associated with the *noticing hypothesis*, which posits that learners must consciously notice linguistic forms in input and feedback in order for acquisition to occur (Schmidt, 1990).

Traditionally, grammar instruction in EFL contexts has relied heavily on explicit rule-based teaching, decontextualized grammar exercises, and teacher-provided corrective feedback (Ellis, 2008). Although such approaches have been shown to enhance learners' metalinguistic awareness, they often fall short in promoting the transfer of grammatical

knowledge to authentic, contextualized writing tasks (DeKeyser et al., 2010). This limitation reflects a long-standing pedagogical gap between form-focused instruction and meaningful language use in real writing contexts (Harmer, 2015). Furthermore, teacher-mediated feedback, while pedagogically valuable, is constrained by practical limitations such as time, workload, delayed feedback cycles, and variability in feedback focus, which may reduce learners' opportunities for immediate noticing and self-correction (Nicol, 2010). From the perspective of *feedback theory*, effective feedback should be timely, specific, and actionable to support learning and revision processes—criteria that are often difficult to consistently meet in large or heterogeneous classrooms.

The rapid emergence of artificial intelligence (AI) technologies in educational contexts has introduced alternative feedback mechanisms that potentially address these limitations. Automated Writing Evaluation (AWE) systems, particularly AI-driven tools such as Grammarly, provide immediate, consistent, and detailed corrective feedback on grammatical errors at scale (Shermis & Burstein, 2003). Utilizing advanced natural language processing (NLP) and machine learning algorithms, these systems are capable of identifying a wide range of error types, including morphological inaccuracies, syntactic violations, punctuation errors, and clarity-related issues (Leacock et al., 2014). Theoretically, AI-mediated feedback aligns with the *noticing hypothesis* by drawing learners' attention to specific linguistic forms in context, while also supporting *self-regulated learning* by enabling learners to independently review, revise, and monitor their writing without relying solely on teacher intervention (Warschauer & Ware, 2006).

Despite the growing availability and adoption of AI writing tools in educational markets, empirical evidence regarding their pedagogical effectiveness in authentic EFL writing contexts remains limited, particularly within Indonesian higher education settings (Fahmi & Cahyono, 2021). Existing studies have predominantly focused on Western contexts or have positioned AI tools as supplementary rather than primary sources of corrective feedback (Koltovskaia, 2020). Moreover, relatively little research has examined how AI-generated feedback influences grammatical accuracy in contextualized writing tasks, as opposed to isolated grammar exercises that lack communicative relevance (Thi, Smith, & Lee, 2022). This gap is especially salient in Indonesian universities, where English classes often consist of students from diverse academic disciplines and linguistic backgrounds, and where empirical guidance for effective AI integration remains underdeveloped.

The present study is situated in a general English course enrolling thirty undergraduate students from twelve different academic programs, reflecting a typical instructional context in Indonesian higher education institutions. In such settings, instructors are required to provide effective feedback across heterogeneous learner profiles, while simultaneously fostering learner autonomy and engagement. Guided by principles of feedback theory, the noticing hypothesis, and self-regulated learning, this study seeks to address the following research questions: (1) To what extent do AI feedback systems improve grammatical accuracy in contextual writing tasks compared to traditional teacher feedback? (2) Which grammatical error categories are most effectively identified and addressed through AI-mediated feedback? (3) How do EFL learners perceive AI-generated feedback, and how do these perceptions influence their engagement with the revision process?

Accordingly, the primary objective of this research is to empirically evaluate the effectiveness of AI feedback systems specifically Grammarly in improving grammatical accuracy in contextual writing tasks among EFL undergraduate learners. Secondary objectives include identifying grammatical error categories most responsive to AI feedback and examining students' perceptions of AI-mediated feedback to inform evidence-based pedagogical integration. By addressing both empirical and theoretical gaps, this study contributes to the growing body of literature on AI-assisted language learning and offers context-sensitive insights to support informed decision-making regarding technology integration in Indonesian higher education.

Grammatical accuracy, defined as the production of linguistically correct forms in accordance with target language conventions, remains a critical dimension of L2 writing competence (Ortega, 2009). Accuracy is distinguished from fluency (rate and ease of production) and complexity (range and sophistication of language structures), together forming the tripartite framework that defines overall writing quality (Wolfe-Quintero, Inagaki, & Kim, 1998). While some researchers have debated the relative importance of accuracy versus fluency and complexity in writing assessment, grammatical accuracy maintains particular significance in academic and professional contexts where correctness is valued and expected (Skehan, 2009)

While some researchers have debated the relative importance of accuracy versus fluency and complexity in writing assessment, grammatical accuracy maintains particular significance in academic and professional contexts where correctness is valued and expected. L2 learners encounter systematic challenges in achieving grammatical accuracy, stemming from both interlingual transfer (interference from mother language structures) and intralingual factors (overgeneralization of L2 rules) (Choo, 2011).

Common grammatical error categories among EFL learners include subject-verb agreement violations, determiner errors, verb tense and aspect confusions, preposition misuse, and article selection errors (Amin, 2015). These errors, while varying in severity and communicative impact, accumulate to create impressions of non-nativeness and reduce the perceived credibility of written communication. Traditional pedagogical responses to grammatical errors have emphasized explicit grammar instruction, with teachers systematically presenting grammatical rules, providing examples, and conducting focused practice activities (Ball et al., 1990). However, research on the effectiveness of explicit instruction reveals mixed findings, with some studies demonstrating modest transfer effects from decontextualized grammar exercises to authentic writing contexts (Li, 2010).

Corrective feedback, defined as responses to learner-produced errors that explicitly or implicitly demosntrated the error and suggest corrective action, has long been recognized as a critical pedagogical tool in language instruction (Chambers & Gregory, 2006). Meta-analyses of feedback research reveal that explicit, timely, and detailed feedback consistently produces larger gains in linguistic accuracy compared to no feedback or minimal feedback.

The timing of feedback—whether immediate or delayed—represents an important dimension of feedback effectiveness. Immediate feedback enables rapid error recognition and self-correction while the context remains cognitively fresh, potentially enhancing learning consolidation. However, practical constraints in traditional classroom settings often necessitate delayed feedback, which reduces its immediacy and impact (Choo, 2011). Additionally, individual teacher feedback, while qualitatively rich, is constrained by class size, teacher workload, and time limitations, often resulting in insufficient feedback to supported error correction across entire student populations.

Automated Writing Evaluation (AWE) systems represent a technological innovation in feedback delivery, utilizing natural language processing and machine learning to provide consistent, scalable feedback on written texts. These systems employ sophisticated algorithms to detect grammatical errors, stylistic issues, clarity concerns, and organizational problems. Grammarly, one of the most widely adopted commercial AWE systems, utilizes advanced machine learning models trained on massive corpora of text to identify and categorize errors (Angeli et al., 2010).

Recent research on AI feedback in EFL contexts demonstrates encouraging findings regarding grammatical accuracy improvements. Studies comparing AI feedback with teacher feedback reveal that AI systems often excel at detecting and categorizing morphosyntactic errors (grammar errors), achieving precision rates of 84–95% across various error categories. The immediacy of AI feedback, accessible during the writing process itself, enables real-time correction and reinforcement (Choo, 2011). Furthermore, the non-judgmental nature of automated feedback may reduce anxiety and enhance motivation compared to feedback perceived as coming from authority figures.

However, AI feedback systems also present limitations. Automated systems may struggle with context-dependent grammatical issues, particularly those requiring deeper semantic understanding or cultural knowledge. Additionally, students may develop over-reliance on AI tools, potentially hindering the development of independent proofreading skills and metalinguistic awareness. The most promising evidence suggested hybrid models integrating AI feedback with teacher feedback and explicit instruction produce superior outcomes (Chambers & Gregory, 2006).

Student engagement with feedback—conceptualized as the degree to which learners actively process, reflect upon, and utilize feedback information—significantly influences learning outcomes. Research grounded in Self-Determination Theory demonstrates that AI feedback systems, through providing immediate responsiveness and non-judgmental evaluation, can enhance students' sense of competence and autonomy, key psychological needs that foster intrinsic motivation.

The immediate, personalized nature of AI feedback can strengthen learner engagement with revision and self-correction. Conversely, some research suggested risks associated with over-reliance on AI tools, including reduced engagement in critical reflection, diminished development of independent editing skills, and potential mechanistic approaches to writing improvement that circumvent deeper linguistic learning (Ball et al., 1990). The effectiveness of AI feedback appears to depend significantly on how these tools are pedagogically integrated and whether they are positioned as supplementary aids to enhance human feedback rather than substitutes for explicit instruction.

METHOD

This study employed a mixed-methods quasi-experimental design using a pretest–posttest control group framework. The quantitative component adopted a non-randomized control group design to compare grammatical accuracy outcomes between students receiving AI-mediated feedback and those receiving conventional teacher feedback. The qualitative component complemented the quantitative data through semi-structured interviews and written reflections, enabling an in-depth exploration of students' perceptions of feedback modalities and their engagement with revision processes. The integration of quantitative and qualitative data allowed for a comprehensive examination of both learning outcomes and learner experiences.

Research Setting and Participants

The study was conducted at Universitas Sains Islam Al Mawaddah Warrahmah Kolaka, a private Islamic university located in Kolaka, Southeast Sulawesi, Indonesia. A total of thirty undergraduate students (n = 30) enrolled in a general English course participated in the study. The course serves students from twelve different academic programs, reflecting the interdisciplinary nature of general English instruction in Indonesian higher education. Participants ranged in age from 18 to 22 years (M = 19.7, SD = 1.2) and were all Indonesian L1 speakers with intermediate English proficiency, as indicated by TOEFL scores ranging from 450 to 550.

Prior to data collection, permission was obtained from the university administration and the course instructor, and informed consent was secured from all participants. The diversity of students' academic backgrounds—including engineering, computer science, business, Islamic studies, health sciences, education, law, and other disciplines—enhances the ecological validity and generalizability of the findings. The sample size (n = 30) aligns with common practice in quasi-experimental educational research and provides sufficient statistical power to detect medium effect sizes.

Treatment Duration and Frequency

The intervention was implemented over an eight-week instructional period. During this time, students completed one writing assignment every two weeks, resulting in four writing tasks per participant. Each writing cycle consisted of three stages: (1) task briefing and

instruction, (2) independent drafting and feedback reception, and (3) revision based on feedback received. Both experimental and control groups followed the same instructional schedule and writing timelines; the only systematic difference between groups was the type and delivery of feedback.

Intervention and Control Conditions

Experimental Group (n = 15)

Students in the experimental group received AI-mediated feedback through Grammarly, a cloud-based AI writing assistant. After completing each writing assignment, students submitted their essays via a shared Google Classroom platform where Grammarly was integrated. Feedback was generated immediately upon submission and provided in multiple formats, including in-text highlights, color-coded error markings, and a summary feedback panel. Grammarly categorized errors into grammar, punctuation, clarity, and style, and offered brief explanations and suggested revisions for each identified issue. Students were encouraged to review the feedback independently and revise their drafts prior to final submission.

Control Group (n = 15)

Students in the control group received conventional instructor-provided feedback. The course instructor used a standardized analytic rubric focusing on grammar, organization, clarity, and overall writing quality. Feedback was delivered in written form through margin comments, error markings, and a holistic summary comment addressing strengths and areas for improvement. Consistent with typical university practices, feedback was returned within 3–5 days after submission. Students then revised their essays based on the instructor's comments.

Both groups completed identical writing tasks and received the same instructional input related to grammar and writing strategies, ensuring that feedback modality was the sole independent variable.

Writing Tasks and Prompts

Four contextualized writing assignments were collected from each participant, yielding a total of 120 essays (30 students × 4 tasks). Writing prompts were intentionally designed to elicit meaningful, discipline-relevant writing rather than isolated grammar practice. The tasks were as follows: (1) Assignment 1 (Descriptive Writing): A descriptive essay recounting a significant personal experience and its influence on the student's academic or professional aspirations. (2) Assignment 2 (Argumentative Writing): An argumentative essay responding to a contemporary social or educational issue, requiring students to present and support a clear position. (3) Assignment 3 (Discipline-Specific Writing): A writing task adapted to students' respective majors (e.g., a simplified lab report for science students, a case analysis for business students, or a reflective discussion for education majors). (4) Assignment 4 (Reflective Writing): A reflective essay synthesizing learning experiences from the course and evaluating personal development in English writing.

All essays were standardized to a length of 300–400 words to ensure comparability across tasks and participants. These tasks were designed to simulate authentic academic writing demands and provide a meaningful context for grammatical production.

Rubric and Criteria for Grammatical Accuracy

Grammatical accuracy was evaluated using an analytic error-based rubric grounded in established L2 writing assessment practices. Accuracy was operationalized as the proportion of error-free clauses relative to the total number of clauses produced. In addition, specific grammatical error categories were systematically identified and coded, allowing for fine-grained analysis of error patterns and feedback effectiveness across different linguistic features.

Error Analysis and Coding Procedure

All collected essays were independently analyzed using a standardized grammatical error taxonomy. The categories included subject–verb agreement, determiner use, verb forms, prepositions, noun number, word order/syntax, and other morphological errors not classified

elsewhere. Two trained coders independently coded all errors to ensure reliability. Inter-rater reliability was calculated using Cohen's kappa (κ = 0.87), indicating substantial agreement and confirming the reliability of the coding procedure.

Data Collection Instruments

Quantitative Measures

Pretest grammatical accuracy: A four-sentence error identification and correction task administered at the beginning of the course. Posttest grammatical accuracy: A parallel form administered at the end of the course. Error frequency counts: Systematic tallies of grammatical error occurrences across all essays.

Qualitative Measures

Semi-structured interviews : Eight participants (four from each group) engaged in 20–30 minute interviews exploring perceptions of feedback, engagement with revisions, and attitudes toward AI-mediated feedback.

Student reflection forms: Open-ended written reflections documenting students' experiences with feedback and perceived effects on their writing development.

Data Analysis Procedures

Quantitative Analysis

Grammatical accuracy data from pretests, posttests, and writing assignments were analyzed using independent samples t-tests to compare experimental and control groups. Effect sizes were calculated using Cohen's d to estimate the magnitude of differences. All analyses were conducted using SPSS version 27.0, with a significance level set at α = 0.05.

Qualitative Analysis

Interview transcripts and written reflections were analyzed using thematic analysis. Coding procedures combined a priori categories informed by feedback and learner engagement frameworks with inductively derived codes emerging from participant responses. The analysis involved open coding, focused coding, and theme development to identify recurring patterns in students' experiences with AI-mediated and teacher-provided feedback.

FINDINGS AND DISCUSSION

Grammatical Accuracy Outcomes

Pre-test Results: Pretest grammatical accuracy scores revealed no significant difference between experimental and control groups at baseline ($t(28) = 0.42$, $p = 0.68$). Experimental group pretest accuracy: M = 68.7%, SD = 8.9%; Control group: M = 68.2%, SD = 9.4%.

Post-test Results: Following the 8-week intervention, the experimental group demonstrated significantly higher post-test grammatical accuracy compared to the control group ($t(28) = 3.87$, $p < 0.001$). Experimental group post-test accuracy: M = 82.5%, SD = 6.3%; Control group: M = 74.3%, SD = 7.1%. This difference represents a statistically significant and practically meaningful improvement, with effect size d = 1.23, indicating a large effect[43].

Individual Assignment Analysis: Across the four writing assignments, the experimental group consistently demonstrated higher grammatical accuracy than the control group (Assignment 1: $t(28) = 2.19$, $p = 0.037$; Assignment 2: $t(28) = 3.42$, $p = 0.002$; Assignment 3: $t(28) = 3.91$, $p < 0.001$; Assignment 4: $t(28) = 4.12$, $p < 0.001$), with effect sizes increasing across assignments, suggesting cumulative benefit of AI feedback over the intervention period.

Error Category Analysis

Analysis of specific grammatical error categories revealed differential patterns of correction across feedback modalities (Table 1).

Table 1: Error Frequency and Category-Specific Improvements

Error Category	Experimental Pretest	Experimental Posttest	Control Posttest	AI Detection Accuracy

Subject-Verb Agreement	4.2/essay	0.8/essay	2.4/essay	94%
Determiner Errors	3.8/essay	1.1/essay	2.9/essay	89%
Verb Form Errors	3.5/essay	1.3/essay	2.1/essay	87%
Preposition Errors	2.9/essay	1.6/essay	2.2/essay	71%
Noun Number Errors	2.1/essay	0.7/essay	1.8/essay	92%
Word Order Errors	1.8/essay	1.2/essay	1.5/essay	68%
Other Errors	1.2/essay	0.9/essay	1.1/essay	73%

The most notable finding concerns the differential effectiveness of AI feedback across error categories. AI systems demonstrated highest precision in detecting and correcting core morphosyntactic errors—subject-verb agreement (94% detection accuracy) and determiner errors (89% accuracy). These represent the grammatical error categories most amenable to rule-based algorithmic detection. Conversely, AI detection accuracy declined for error categories requiring contextual semantic understanding, such as preposition selection (71%) and word order violations in complex syntactic structures (68%).

The experimental group's improvements were particularly pronounced for subject-verb agreement errors, which decreased from 4.2 errors per essay at pretest to 0.8 errors per essay by posttest—a 81% reduction. Similarly, determiner errors decreased 71% from 3.8 to 1.1 errors per essay. These error categories, which are rule-governed and systematic, appeared most responsive to AI feedback mechanisms.

Qualitative Findings: Student Perceptions of AI Feedback

Interview data from experimental group participants revealed largely positive perceptions of AI feedback alongside identified concerns. Key themes emerging from qualitative analysis included:

Theme 1: Immediacy and Responsiveness

All four experimental group interview participants emphasized appreciation for immediate feedback availability. Representative quote: "With Grammarly, I can correct my mistakes right away. I don't have to wait three or four days. I see the error immediately and understand what I did wrong." This immediacy appears to facilitate rapid error recognition and correction, potentially enhancing learning consolidation.

Theme 2: Non-judgmental and Anxiety-reducing Feedback

Participants reported that AI feedback felt less threatening than teacher feedback. One participant stated: "Grammarly doesn't make me feel stupid. It's just explaining the error in a neutral way. I don't worry about disappointing the teacher." This affective dimension may contribute to enhanced engagement with correction and reduced writing anxiety.

Theme 3: Detailed Explanations Supporting Metalinguistic Understanding

Students appreciated Grammarly's provision of explanations accompanying error correction. Participants noted: "I understand why it's wrong now, not just that it's wrong. Grammarly explains the rule." This explicit metalinguistic supported may facilitate deeper learning than simple error correction without explanation.

Theme 4: Concerns About Over-reliance and Mechanical Corrections

Despite positive feedback, interview participants expressed concerns about developing dependency on AI tools: "I worry that I'm depending too much on Grammarly. Will I be able to write correctly without it?" Additionally, one student noted: "Sometimes I don't actually understand why I was wrong. I just let Grammarly fix it because it's faster."

Theme 5: Preference for Hybrid Feedback Models

When asked about ideal feedback arrangements, all four experimental participants expressed interest in combining AI and teacher feedback: "I think Grammarly and teacher comments together would be best. The teacher can explain how to make my essay better overall, and Grammarly handles the grammar details."

Control group interview participants (teachers' feedback only) reported a more limited understanding of specific grammatical errors: "When the teacher marks grammar, sometimes I have to look at the grammar book to understand. Immediate explanation would help."

Engagement and Self-correction Patterns

Analysis of essay revision patterns revealed notable differences between groups. Students in the experimental condition who received AI feedback demonstrated more extensive engagement with grammatical revisions in subsequent drafts (78% of identified errors were corrected in revisions), compared to control group students (52% of identified errors subsequently corrected). This greater responsiveness to AI feedback suggested enhanced engagement with correction processes.

The findings of this study supported the primary hypothesis that AI feedback systems enhance grammatical accuracy in contextual writing tasks compared to traditional teacher feedback alone. The experimental group's post-intervention grammatical accuracy score (M = 82.5%) represents a substantial improvement from baseline (M = 68.7%, 20% improvement), while the control group's improvement was more modest (74.3% vs 68.2%, 9% improvement). The between-group difference of 8.2 percentage points, with large effect size (d = 1.23), represents a meaningful and practically significant enhancement in student writing quality.

These findings aligned with recent research on AI feedback in EFL contexts, which consistently demonstrates the effectiveness of immediate, detailed automated feedback in improving grammatical accuracy[44]. The mechanisms underlying these improvements likely include: (1) increased frequency of feedback delivery, enabling more complete error coverage than teacher feedback alone; (2) immediacy of feedback during the writing process, enabling rapid error recognition and correction; (3) consistency of feedback application across all students; and (4) non-judgmental presentation reducing writing anxiety and defensive resistance to correction.

The differential effectiveness across error categories provides important insights. The superior performance of AI systems in detecting morphosyntactic errors (SVA, determiner, noun number errors) reflects the rule-governed nature of these error categories, which aligned well with algorithmic pattern recognition. Conversely, reduced AI accuracy in preposition selection and complex word order reflects the context-dependent and semantically grounded nature of these errors, which require deeper semantic and pragmatic understanding beyond pattern-matching capabilities.

These findings suggest that students benefit from targeted pedagogical integration where AI feedback is optimized for detecting rule-governed errors, while instructor feedback provides guidance on more complex, context-dependent linguistic issues. Rather than positioning AI and human feedback as competing alternatives, evidence supported their complementary strengths, supporting the hybrid model favored by student participants.

The qualitative findings regarding student perceptions aligned with self-determination theory perspectives on motivation. AI feedback's capacity to enhance perceived competence (through clear error identification and explanation), supported autonomy (through immediate, non-coercive correction), and reduce performance anxiety contributes to enhanced intrinsic motivation[46]. However, participant concerns about over-reliance warrant instructional caution, suggesting explicit scaffolding toward independent proofreading skills should accompany AI tool integration.

The enhanced revision engagement in the experimental group (78% error correction rate vs 52%) suggestsed that AI feedback, through its immediately recognizable format and non-judgmental presentation, may promote greater learner agency in self-directed improvement. This represents a potentially important outcome beyond simple accuracy gains, fostering more autonomous approaches to writing development.

Several limitations warrant consideration in interpreting these findings. First, the modest sample size (n = 30) limits generalizability, though effect sizes remain large. Second, the 8-week intervention period, while sufficient to demonstrate significant improvements, does not permit assessment of long-term retention or transfer to new writing contexts. Third,

while students were drawn from diverse academic programs, the study occurred at a single institution with potentially unique contextual characteristics, limiting broader generalizability to other Indonesian universities. Fourth, this study specifically examined Grammarly; findings may not generalize to other AI writing tools with different capabilities or interface designs.

Grammatical Accuracy Outcomes

Pre-test Results: Pretest grammatical accuracy scores revealed no significant difference between experimental and control groups at baseline ($t(28) = 0.42$, $p = 0.68$). Experimental group pretest accuracy: M = 68.7%, SD = 8.9%; Control group: M = 68.2%, SD = 9.4%.

Post-test Results: Following the 8-week intervention, the experimental group demonstrated significantly higher post-test grammatical accuracy compared to the control group ($t(28) = 3.87$, $p < 0.001$). Experimental group post-test accuracy: M = 82.5%, SD = 6.3%; Control group: M = 74.3%, SD = 7.1%. This difference represents a statistically significant and practically meaningful improvement, with effect size $d = 1.23$, indicating a large effect[43].

Individual Assignment Analysis: Across the four writing assignments, the experimental group consistently demonstrated higher grammatical accuracy than the control group (Assignment 1: $t(28) = 2.19$, $p = 0.037$; Assignment 2: $t(28) = 3.42$, $p = 0.002$; Assignment 3: $t(28) = 3.91$, $p < 0.001$; Assignment 4: $t(28) = 4.12$, $p < 0.001$), with effect sizes increasing across assignments, suggesting cumulative benefit of AI feedback over the intervention period.

Error Category Analysis

Analysis of specific grammatical error categories revealed differential patterns of correction across feedback modalities (Table 1).

Table 1: Error Frequency and Category-Specific Improvements

Error Category	Experimental Pretest	Experimental Posttest	Control Posttest	AI Detection Accuracy
Subject-Verb Agreement	4.2/essay	0.8/essay	2.4/essay	94%
Determiner Errors	3.8/essay	1.1/essay	2.9/essay	89%
Verb Form Errors	3.5/essay	1.3/essay	2.1/essay	87%
Preposition Errors	2.9/essay	1.6/essay	2.2/essay	71%
Noun Number Errors	2.1/essay	0.7/essay	1.8/essay	92%
Word Order Errors	1.8/essay	1.2/essay	1.5/essay	68%
Other Errors	1.2/essay	0.9/essay	1.1/essay	73%

The most notable finding concerns the differential effectiveness of AI feedback across error categories. AI systems demonstrated highest precision in detecting and correcting core morphosyntactic errors—subject-verb agreement (94% detection accuracy) and determiner errors (89% accuracy). These represent the grammatical error categories most amenable to rule-based algorithmic detection. Conversely, AI detection accuracy declined for error categories requiring contextual semantic understanding, such as preposition selection (71%) and word order violations in complex syntactic structures (68%).

The experimental group's improvements were particularly pronounced for subject-verb agreement errors, which decreased from 4.2 errors per essay at pretest to 0.8 errors per essay by posttest—a 81% reduction. Similarly, determiner errors decreased 71% from 3.8 to 1.1 errors per essay. These error categories, which are rule-governed and systematic, appeared most responsive to AI feedback mechanisms.

Qualitative Findings: Student Perceptions of AI Feedback

Interview data from experimental group participants revealed largely positive perceptions of AI feedback alongside identified concerns. Key themes emerging from qualitative analysis included:

Theme 1: Immediacy and Responsiveness

All four experimental group interview participants emphasized appreciation for immediate feedback availability. Representative quote: "With Grammarly, I can correct my mistakes right away. I don't have to wait three or four days. I see the error immediately and understand what I did wrong." This immediacy appears to facilitate rapid error recognition and correction, potentially enhancing learning consolidation.

Theme 2: Non-judgmental and Anxiety-reducing Feedback

Participants reported that AI feedback felt less threatening than teacher feedback. One participant stated: "Grammarly doesn't make me feel stupid. It's just explaining the error in a neutral way. I don't worry about disappointing the teacher." This affective dimension may contribute to enhanced engagement with correction and reduced writing anxiety.

Theme 3: Detailed Explanations Supporting Metalinguistic Understanding

Students appreciated Grammarly's provision of explanations accompanying error correction. Participants noted: "I understand why it's wrong now, not just that it's wrong. Grammarly explains the rule." This explicit metalinguistic supported may facilitate deeper learning than simple error correction without explanation.

Theme 4: Concerns About Over-reliance and Mechanical Corrections

Despite positive feedback, interview participants expressed concerns about developing dependency on AI tools: "I worry that I'm depending too much on Grammarly. Will I be able to write correctly without it?" Additionally, one student noted: "Sometimes I don't actually understand why I was wrong. I just let Grammarly fix it because it's faster."

Theme 5: Preference for Hybrid Feedback Models

When asked about ideal feedback arrangements, all four experimental participants expressed interest in combining AI and teacher feedback: "I think Grammarly and teacher comments together would be best. The teacher can explain how to make my essay better overall, and Grammarly handles the grammar details."

Control group interview participants (teachers' feedback only) reported a more limited understanding of specific grammatical errors: "When the teacher marks grammar, sometimes I have to look at the grammar book to understand. Immediate explanation would help."

Engagement and Self-correction Patterns

Analysis of essay revision patterns revealed notable differences between groups. Students in the experimental condition who received AI feedback demonstrated more extensive engagement with grammatical revisions in subsequent drafts (78% of identified errors were corrected in revisions), compared to control group students (52% of identified errors subsequently corrected). This greater responsiveness to AI feedback suggested enhanced engagement with correction processes.

CONCLUSIONS

This study provides empirical evidence that AI feedback systems, specifically Grammarly, effectively enhance grammatical accuracy in contextual writing tasks among EFL learners in Indonesian higher education contexts. The observed improvements among students receiving AI-mediated feedback, together with their positive perceptions and increased engagement in revision activities, underscore the pedagogical value of AI-supported feedback when integrated within formal writing instruction. Importantly, the findings do not indicate that AI feedback should replace instructor feedback or explicit grammar instruction. Instead, the results support a complementary feedback model in which AI tools effectively address rule-governed morphosyntactic errors through immediate and consistent feedback, while instructors contribute higher-order guidance and contextualized linguistic support. This combined approach capitalizes on the distinct strengths of both automated and human feedback, thereby enhancing overall instructional effectiveness. Within Indonesian higher education settings, particularly those characterized by diverse student populations and interdisciplinary general English courses, AI writing tools present a scalable means of supporting grammatical development while mitigating practical constraints associated with individualized instructor feedback. Nevertheless, effective integration necessitates deliberate

pedagogical support, including guidance in interpreting feedback, structured opportunities for revision, and sustained attention to learners' development of independent proofreading and reflective writing practices. As AI technologies become increasingly embedded in educational environments, developing an evidence-based understanding of their pedagogical affordances and limitations remains essential. This study contributes to such understanding by demonstrating the conditions under which AI-mediated feedback can support grammatical accuracy without diminishing the instructional role of teachers. Accordingly, the findings advocate for thoughtful, evidence-informed integration of AI tools that enhances learning outcomes while promoting the long-term development of autonomous writing skills among EFL learners.

REFERENCES

Amin, Y. N. (2015). Teaching grammar-in-context and its impact in minimizing students' grammatical errors in writing. *Journal of English Education and Linguistics, 2*(2), 75–92.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*(2), 102–118.

DeKeyser, R., Salaberry, R., Robinson, P., & Harrington, M. (2010). Getting started. In R. DeKeyser (Ed.), *Implicit and explicit language learning: Conditions, processes, and knowledge* (pp. 7–30). Georgetown University Press.

Ellis, R. (2008). Investigating grammatical difficulty in second language learning. *International Journal of Applied Linguistics, 18*(1), 4–22.

Fahmi, A., & Cahyono, B. Y. (2021). Grammarly feedback on EFL learners' writing. *Journal of English Teaching and Linguistics, 6*(2), 89–104.

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

Harmer, J. (2015). *The practice of English language teaching* (5th ed.). Longman.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101.

Koltovskaia, S. (2020). Learner engagement with a proofreading tool that provides direct written corrective feedback: The case of Grammarly. *Journal of Second Language Writing*, 50, 100680.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners* (2nd ed.). Morgan & Claypool.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*(2), 309–365.

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

Nicol, D. (2010). The foundation for graduate attributes: Developing self-regulated learners. *The Higher Education Academy*, 1–15.

Ortega, L. (2009). *Understanding second language acquisition*. Hodder Education.

Shermis, M. D., & Burstein, J. C. (2003). Automated essay scoring: A cross-sectional analysis of the use of the C-Writer automated essay scoring system. *Journal of Technology, Learning, and Assessment, 2*(1), 1–28.

Skehan, P. (2009). Modelling second language performance. In A. Mackey & S. M. Gass (Eds.), *Multilingual aspects of language learning and use* (pp. 110–143). John Benjamins.

Thi, T. X., Smith, J. R., & Lee, K. (2022). AI-driven feedback systems in EFL writing: A systematic review. *TESOL Quarterly, 56*(2), 445–472.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Journal of Technology, Learning, and Assessment, 6*(1), 1–23.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity* (Technical Report No. 17). University of Hawaii.