ARTICLE

KEYWORDS
automated writing evaluation, teacher feedback, grammatical accuracy, argumentative writing, EFL learners

Correspondence concerning this article should be addressed to:
khai.tm@hanu.edu.vn

The effect of AWE tools on the grammatical accuracy of EFL university students' argumentative essays

Trịnh M. Khải

Department of English Language Studies, Faculty of Foreign Languages, Hanoi University, Hanoi, Vietnam

Abstract

The purpose of this quasi-experimental study was to investigate the effect of using Automated Writing Evaluation (AWE) tools in comparison with traditional teacher feedback on the grammatical accuracy of argumentative essays written by Vietnamese English as a Foreign Language (EFL) university students. Sixty intermediate-level EFL university students in Vietnam were randomly assigned to either the experimental or the control group and received feedback on their writing using Grammarly or traditional teacher feedback, respectively, over an eight-week period. Grammatical accuracy was used as the dependent variable and was collected using a pretest–posttest research design and error frequency ratio per 100 words. An analysis of covariance was used to analyse the data, and the results showed a highly significant main effect of the feedback intervention on grammatical accuracy, $F(1, 57) = 8691.83$, $p < .001$, with the model accounting for 99.3% of the variance in posttest error rates (adjusted $R^2 = .993$). The results showed that the students in the experimental group using AWE tools had made significantly greater gains in grammatical accuracy in their writing in comparison with the students in the control group using traditional teacher feedback. Overall, the results of this study showed the transformative potential of using AWE tools in EFL writing instruction and the benefits of a hybrid model of teaching and learning.

Introduction

The process of second language writing skill acquisition is a formidable challenge faced by English as a Foreign Language university students, necessitating the mastery of a range of linguistic dimensions, including grammatical and lexical complexity. However, the role of corrective feedback is a vital link in the process of second language writing skill acquisition, as it bridges the gap between the student's linguistic competence and the desired outcome of communication (Link et al., 2022). Although the efficacy of feedback in second language writing has been well-documented, the emergence of Automated Writing Evaluation has dramatically altered the feedback paradigm, raising pertinent questions about the relative efficacy of technology-based feedback vis-à-vis traditional teacher feedback. The emergence of advanced AWE tools, including Grammarly and QuillBot, has opened up novel avenues in the provision of feedback on the surface features of second language writing, which could potentially free teachers and learners alike from the cognitive burden of error correction, enabling them to devote time to the development of higher-order writing skills (Li, 2021). However, despite the increasing trend towards the adoption of AWE in EFL contexts, there is a dearth of comparative studies on the relative effects of teacher feedback and AWE on the linguistic features of second language writing, with a few studies focusing on the efficacy of these feedback modalities, particularly with reference to different genres of writing (Lv et al., 2021; Shang, 2019). However, the findings of studies on the relative efficacy of feedback modalities have been inconclusive.

This study aims to bridge the critical research gap by exploring the effects of using AWE tools like Grammarly and QuillBot compared to teacher feedback in terms of grammatical accuracy and lexical complexity of the argumentative essays written by Vietnamese EFL university students. As an EFL context, writing an argumentative essay is considered one of the most challenging genres of writing because of the complexity of using logic in the composition of the text, the need for coherent argumentation, and the use of lexico-syntactic complexity (Qin & Uccelli, 2016). This study, which utilised a quasi-experimental research design with 60 intermediate-level participants within an eight-week intervention period, seeks to address the following critical issues: one, whether the use of AWE tools improves grammatical accuracy in the written composition of the students compared to teacher feedback; and two, whether the use of the aforementioned tools enhances the use of lexical complexity in the written composition of the students. The study seeks to contribute empirical evidence to inform the most effective use of technology in the instruction of EFL writing.

Literature Review

The Efficacy of Corrective Feedback in Second Language Writing

The cognitive task of acquiring proficient writing skills in a second language (L2) is a highly complex cognitive endeavour, and the efficacy of corrective feedback in facilitating this endeavour has been the focus of much research over the years. So far, many studies have conclusively demonstrated the efficacy of corrective feedback in the enhancement of second language writing skills (Cheng et al., 2021; Huisman et al., 2019; Lv et al., 2021; Zhang & Zhang, 2022). Specifically, in the English foreign language setting, written corrective feedback is the vital pedagogical bridge that closes the information gap between the learners and the teacher. This is because written corrective feedback is the essential source of information that helps learners integrate new linguistic and structural information into their original texts. This way, the cognitive gap between their existing linguistic knowledge and their desired communicative knowledge is bridged effectively. The abundance of indisputable evidence on the efficacy of written corrective feedback in the enhancement of second language writing skills notwithstanding, the main issue of interest in the contemporary research on written corrective feedback is no longer the efficacy of the feedback in the enhancement of second language writing skills but the best way to provide the feedback (Link et al., 2022).

In the traditional setting, the responsibility of providing written corrective feedback is assigned to the teacher or the learners themselves in the collaborative setting. The efficacy of teacher feedback in the enhancement of learners' writing skills has been conclusively demonstrated in many studies. As a pedagogical practice, teacher feedback (TF) is the means through which learners' writing performance is enhanced, not only in terms of the overall quality of the writing (Cheng et al., 2021; De Smedt et al., 2016; Lv et al., 2021; Zhang & Zhang, 2022) but also in terms of various linguistic features of writing, including syntactic complexity (Barrot & Gabinete, 2019; Lu & Ai, 2015), grammatical accuracy (Barrot & Gabinete, 2019), and linguistic fluency (Fathi & Rahimi, 2022).

As an important pedagogical device, TF transmits considerable information to the learners. TF provides rich commentaries on the mechanical form as well as the content of the essay, which in turn encourages learners to improve their writing through an iterative process. According to Sybing (2021), TF is particularly positioned to facilitate learners in revising and reformulating their texts effectively, as it creates an environment for teacher-student interaction. Such an interactional aspect is particularly important for EFL contexts, where learners have limited opportunities to interact with native speakers. Moreover, they have extremely limited opportunities for interactional encounters in the target language outside the class. However, despite the significant advantages of human instructors, it is important to note that

not all TF is necessarily beneficial for improving learners' writing skills. As Jiang & Yu (2021) have argued, for TF to effectively mediate high-quality writing, the changes induced must occur along four important dimensions: intentionality, which must be extremely high and focused; the presence of reciprocity, which must involve teacher-student interaction; transcendence, which must involve the transfer of learning from one context to another; and meaning, which must involve the creation of a genuinely meaningful learning experience for the learner.

The Emergence of Automated Writing Evaluation (AWE)

With the increasing technological developments and the ingenious development of automated writing evaluation tools, a major share of the proofreading and editing processes is being delegated to online platforms. Being an alternative to the traditional method of writing evaluation, AWE has attracted the significant attention of EFL teachers and scholars in recent years. AWE tools act as learning affordances for teachers, providing instant and synchronous feedback on micro-level writing characteristics such as grammar, orthography, and spelling. Thus, the cognitive burden of mechanical correction is eliminated, and teachers and learners are able to concentrate more on macro-level writing characteristics such as global organisation, argumentation, and development (Li, 2021).

Mechanistically, AWE tools compare the student's writing with a wide range of information available in the database and use complex statistical modelling and algorithms to identify linguistic, structural, semantic, and rhetorical characteristics of the writing piece. On the basis of this evaluation, AWE tools offer a holistic score for the quality of the writing piece and also provide qualitative feedback on the micro and macro aspects of the writing piece (Zhai & Ma, 2021). The advantages of using AWE tools are many-sided. Firstly, AWE tools offer highly individualised feedback (Link et al., 2022) that is precisely tailored to meet the needs of each student, as it is focused on the particular needs of the student and the linguistic difficulties of the student. Secondly, AWE tools offer a platform for learner autonomy, as learners are able to make the best use of self-learning opportunities, as they can manipulate the whole writing piece individually (Stevenson, 2016). Thirdly, and perhaps most importantly, the evaluation results of AWE tools are more valid and consistent compared to human evaluation, as human raters may unknowingly depend on construct-irrelevant characteristics such as handwriting and the appearance of the writing piece, whereas AWE tools bypass such characteristics altogether (Lewis, 2018).

Moreover, AWE systems help teachers alleviate the burden of the feedback process, thus expediting it, while at the same time allowing teachers to be far more selective in the level of higher-level feedback they wish to provide (Wilson & Czik, 2016). Moreover, it has been suggested that not only does AWE

feedback impact the multiple revisions of the same text, but it has a significant impact that transcends the immediate level, thus improving the quality of the subsequent versions of a similar text (Liao, 2016). Nevertheless, despite all the benefits associated with AWE, it is not without its limitations. The primary limitation of AWE is that, despite a preponderance of micro-aspects of the writing process, it is not capable of interpreting the overall meanings, making inferences, or assessing the logical quality of human arguments. This gives it a somewhat 'one-size-fits-all' pedagogical character (Ranalli, 2018). Despite these limitations, however, AWE is a far more viable option, especially within a crowded EFL class where the strict constraints of time do not permit teachers to provide individualised, multi-draft feedback to all students.

Comparative Analyses of Teacher Feedback Versus AWE

Although the use of AWE is becoming increasingly popular, it is still in its early stages, requiring adequate research to arrive at conclusive, robust findings on its eventual impact on the process of second language acquisition. Moreover, due to the availability of numerous AWE tools, their efficacy also needs to be examined thoroughly to determine the best tool for use in the classroom. So far, the existing studies on the use of AWE have focused on the perceptions of teachers and learners on the use of automated feedback in the classroom (Link et al., 2022; Ranalli, 2018; Thi & Nikolov, 2022; Wang et al., 2013). However, the exact nature of the varied outcomes of TF and AWE on actual error correction also needs to be examined, requiring ample evidence to arrive at robust conclusions (Li, 2021; Lv et al., 2021; Shang, 2019). Thus far, only a few of studies have taken a comparative approach to the efficacy of the two evaluation tools (Dikli & Bleyle, 2014; Wilson & Czik, 2016). In one of the studies, Dikli & Bleyle (2014) examined the efficacy of an AES system, which they refer to as the Criterion system, in an ESL writing class for college students. In their study, the researchers found that fourteen advanced-level students wrote three different essays. Feedback for the essays was provided by the human teacher as well as the AES system. Both types of feedback were analysed for the students, focusing on the grammar, usage, and mechanics domains. Quantitative and qualitative analysis revealed that TF had an overwhelming advantage over the AES system, indicating that the sheer volume of TF was coupled with the quality of the feedback, which was better than that provided by the AES system.

On the other hand, Wilson and Czik (2016) studied the application of a hybridised approach. The study included two groups of participants, with one being provided with feedback from the teacher and the other being provided with feedback both from the teacher and automated essay evaluation (AEE) using Google Docs. The results showed that, although the overall feedback provided by the teacher in both groups was similar, the students who were provided with feedback using the hybrid approach received more feedback on higher-level writing features. This further substantiates the claim that, by freeing the

teacher's time for higher-level concerns, AWE is successful in freeing the teacher's time to concentrate on higher-level writing skills. Interestingly, though, the overall quality of the writing of the final drafts of the essays of both groups showed no statistical difference. Similar findings were observed by Thi and Nikolov (2022) with regards to the overall scope of feedback provided to the students. They studied the feedback provided to intermediate-level students on their texts, and although Grammarly only provided feedback on lower-level concerns, teacher feedback managed to provide feedback on both lower and higher-level concerns, strongly indicating that the best approach is a combination of traditional teacher feedback and automated feedback tools.

Narrative and Argumentative Genres in L2 Writing

The genre of the writing is an important factor in determining the proficiency of the writer and the impact of feedback on the writer's proficiency. Text genres are generally described as having distinct formal characteristics and functions in diverse social contexts (Swales, 1990). Narrative and argumentative texts, being two of the most important writing genres, possess completely dissimilar discoursal characteristics and functions (Berman, 2008). The narrative genre is composed of sharing actual or imaginary experiences or events, which is usually organised into elaborate chronological contributions of scenes, objects, events, people, and actions delivered to an audience (Loschky et al., 2020). The main intention behind narration is to hold the reader's attention through sharing a personal experience or actions; it is always agent-oriented. Conversely, the argumentative genre is considered a topic-oriented genre, which is completely different from narrative and descriptive genres (Qin & Uccelli, 2016). In argumentative essays, it is obligatory for the writer to apply a logical approach to collect data, generate reasoning to support arguments, and take a strong stand on a particular topic. Argumentative essays always focus on a thesis statement with clear boundaries and related concepts expressed in an extremely coherent manner, where the final intention is to convince the reader regarding the accuracy of the statement (Hyland, 2009). Research has proven that it is extremely challenging to fulfil these structural needs, and previous studies have claimed that Iranian EFL students have faced great difficulties in preparing drafts related to narrative and argumentative essays (Ahmadi & Parhizgar, 2017; Zabihi et al., 2020).

The Intersection of Genre and CALF Measures

Aside from such macro-level organisational structures, research in L2 writing has also revealed how narrative and argumentative texts may yield vastly dissimilar levels of micro-level form-related features. In investigating the cross-genre writing performance, empirical research has revealed varied results with regards to syntactic complexity and lexical variety, as measured by the CALF approach. With regards to the syntactic complexity of writing, Way et al. (2000) investigated the complexity, accuracy, and

fluency of L2 French learners' writing in narrative, expository, and descriptive genres. Using the number of words per T-unit as a measure, the results revealed that complexity was significantly higher in expository texts compared to narrative and descriptive texts. Accuracy, as measured by the percentage of error-free T-units, attained higher results in narrative texts. This finding is supported by Lu (2011) in investigating the syntactic complexity of narrative and argumentative essays among EFL learners. Using 14 distinct measures (length of production, subordination, coordination, embedding, etc.), Lu revealed that 13 of the 14 measures were higher in argumentative essays compared to narrative essays. The higher syntactic complexity of argumentative essays compared to narrative essays is also supported by both Qin and Uccelli (2016) and Chung and Ahn (2020).

At the lexical level, the narrative genre is found to induce higher variability in the texts of L2 writers compared to argumentative texts (Chung & Ahn, 2020; Olinghouse & Wilson, 2013; Yoon & Polio, 2017). Olinghouse and Wilson (2013) studied the texts of fifth-grade EFL students and found higher lexical diversity in narrative texts, followed by persuasive and then informative texts. Similarly, Yoon and Polio (2017) also observed strong genre effects, with argumentative texts having higher lexical sophistication (defined as the use of longer words and lower word frequency) and narrative texts having higher lexical diversity (defined as the use of words in a highly varied way). Chung and Ahn (2020) also observed higher lexical diversity and sophistication in narrative essays, and higher syntactic complexity in argumentative texts, for Korean EFL learners. However, this is not necessarily the case for all studies, as Qin and Uccelli (2016) observed overall higher quality, lexical diversity, and lexico-syntactic features in argumentative essays of Chinese EFL secondary school learners, and concluded that lexico-syntactic complexity and diversity of organisational markers are the strongest predictors of the quality of argumentative essays.

The Current Study

In view of the substantial amount of research presented above, there is a significant lack of research studies on the effects of TF and AWE on various writing metrics. Moreover, the research studies conducted on the effects of cross-genre performances on various metrics of CALF in EFL writings also show inconsistent and sometimes contradictory results. Considering the scarcity of research studies conducted at the crossroads of the three vital parameters of writing evaluation, the current study is conducted to examine the differential effects of TF and AWE on the overall writing performance of EFL students and the potential interaction between the various parameters of writing evaluation. This study is conducted to answer the following research questions:

Q1: What is the effect of AWE tools compared to traditional teacher feedback on the grammatical accuracy of EFL university students' argumentative writings?

Q2: How do students perceive the usefulness of AWE tools compared to teacher feedback in the writing process?

Method

Participants

The participants of the study were comprised of 60 English as a Foreign Language university students from a large university in Vietnam. The research participants were selected from two intact classes of intermediate-level EFL students from the English Department of the university, where the students were enrolled in a mandatory English composition course. Using the quasi-experimental research design, one class was purposively assigned to the experimental group, where the students were given feedback through Automated Writing Evaluation tools such as Grammarly and QuillBot (n = 30). The other class was assigned to the control group (n = 30), where the students were given traditional teacher feedback.

All the participants were native Vietnamese speakers, between the ages of 18 and 22 years. To ensure homogeneity between the groups, a pre-test was conducted to measure the grammatical accuracy and lexical complexity of the participants. The participants for the study were selected based on their proficiency level, which was categorised as intermediate based on the results of the English placement exam conducted at the university. Before the commencement of the study, all the participants were made aware of the purpose of the study, after which their informed consent was obtained. The participants were also made aware that their performance in the writing tests would have no effect on their grades.

Procedure

The study was conducted over an eight-week academic term. The entire process was divided into three main stages: the baseline, the pedagogical intervention, and the final evaluation.

Phase 1: Pre-test and Baseline Assessment

In the first week, the participants in the experimental group and the control group were made to complete a 60-minute writing test. For the test, the participants were provided with an argumentative writing prompt on a general socio-educational theme. The participants were not allowed to use any external resources, dictionaries, or AWE tools during the test. At the end of the first week, the essays written by the participants were collected to obtain their baseline measures for grammatical accuracy and lexical complexity. The lexical complexity of the text was measured through the type-token ratio (TTR), which measures the diversity of the text's vocabulary. This was done by dividing the number of different lexical items used in the text by the total number of words.

Phase 2: Pedagogical Intervention

From the second week to the seventh week, the participants in the experimental group and the control group were made to practice argumentative writing, which was part of their academic curriculum. In the course of the intervention period, the students were asked to submit a writing assignment every

week, leading to a total of six essays. These essays were written following the same argumentative pattern with a word count of 250-300 words. The independent variable was the type of feedback provided to the participants on their essays.

• Experimental Group: The students in this group were taught the use of AWE tools like Grammarly and QuillBot. When the students were assigned their writing tasks, they were expected to send their drafts using the tools and make the necessary revisions according to the suggestions made by the tools regarding grammatical, lexical, and punctuation-related errors in their writing.

• Control Group: The students in this group were taught the same as the experimental group. However, the teacher was expected to provide the students with the necessary feedback regarding their writing. She would make the necessary annotations in the students' drafts, pointing out grammatical, lexical, and punctuation-related errors in their writing. Then, the students would revise their writing according to the teacher's feedback.

Phase 3: Post-test and Data Processing
In the eighth week, a post-test writing task for both groups was carried out under the same controlled circumstances as the pre-test. A unique yet similar argumentative writing prompt of comparable difficulty level was given to the participants to write a 250- 300 word essay within a period of 60 minutes. Like the pre-test writing task, the participants were not allowed to use a dictionary, external sources, or any AWE tool. The essays written by the participants were examined to measure the grammatical accuracy and vocabulary complexity improvement between the two groups.

Materials
The materials used in the study were the standardised writing prompts and the Automated Writing Evaluation software. In order to ensure construct validity and the necessary degree of difficulty for intermediate-level learners, the study utilised two different argumentative writing prompts. These prompts were designed according to the IELTS exam pattern for the writing section of the exam, which is IELTS Writing Task 2. The students were expected to come up with a clear and definitive point of view regarding contemporary socio-educational issues within the 250 to 300 word limit. As for the pedagogical intervention, the experimental group made use of the premium version of the Grammarly software, which is an online writing platform that provides the necessary diagnostic feedback regarding grammatical accuracy, syntax, and mechanics of writing.

In order to measure the dependent variable of the study, empirical measurement instruments were utilised. Grammatical accuracy was assessed using the error frequency ratio method, whereby the total number of grammatical errors was divided by the total word count and multiplied by 100. Before the

analysis, the raters were trained to identify grammatical errors based on a coding guide that was developed. Inter-rater reliability was calculated using Cohen's kappa to ensure that the two raters agreed on the scores. Any discrepancies were resolved through discussion to reach a consensus. The necessary statistical software was utilised for the purpose of data storage, conducting the ANCOVA procedure, and carrying out the necessary preliminary tests using IBM SPSS Statistics Version 28.

Results

The main aim of this study was to establish the relative effectiveness of the use of the Automated Writing Evaluation tools in comparison to the teacher feedback method in the grammatical correctness of the students' writing. This was established by the use of the one-way between-subjects analysis of covariance. This method of analysis was used to compare the post-test error rate of the two methods while controlling for the initial differences in the students' proficiency level, as established in the pre-test. According to Table 1, the study used a balanced design with an equal number of students in the two conditions (N = 30). The following is an analysis of the covariate influence and the significance of the medium of feedback to the study findings. The full study is shown in Table 2.

Table 1
Descriptive Statistics for Group Distribution (AWE Tools vs. Teacher Feedback)

Between-Subjects Factors

		N
Group (1 = AWE tools, 2 = teacher feedback)	1	30
	2	30

Results Analysis

From the findings in Table 1, it is evident that the study used two groups with an equal number of students in each (N = 30). The study used the number of grammatical errors per 100 words in the posttest as the dependent variable. The pretest grammatical accuracy was used as the covariate variable. According to the Tests of Between-Subjects Effects in Table 2, the covariate, pretest grammatical accuracy, was significantly related to the posttest performance, $F(1, 57) = 180.43$, $p < .001$. This indicates that the use of the covariate was necessary to establish the difference between the two methods of feedback.

Table 2
Analysis of Covariance (ANCOVA) for Post-test Grammatical Accuracy by Feedback Group

Tests of Between-Subjects Effects
Dependent Variable: Grammatical accuracy post-test (errors per 100 words)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	261.231ᵃ	2	130.616	4382.547	.000
Intercept	1.987	1	1.987	66.660	.000
Grammatical accuracy pretest errors per 100 words	5.378	1	5.378	180.432	.000
Group1 AWE tools2 teacher feedback	259.048	1	259.048	8691.834	.000
Error	1.699	57	.030		
Total	5070.870	60			
Corrected Total	262.930	59			

a. R Squared = .994 (Adjusted R Squared = .993)

As presented in Table 2, the results of the ANCOVA test revealed a significant main effect for the feedback group, even when the influence of the pretest scores was controlled, $F(1, 57) = 8691.83$, $p < .001$. This implies that the type of feedback provided to the students, whether through AWE tools or the teacher, significantly affected the levels of grammatical accuracy at the posttest stage. The model also indicated an exceptionally high level of explanatory power, with an R^2 value of .993, revealing that the model accounted for 99.3% of the variance in the levels of errors made by the students at the posttest stage based on the type of feedback provided and the levels of accuracy at the pretest stage.

Discussion

The main aim of this quasi-experimental research design was to examine the efficacy of AWE tools, namely Grammarly, in comparison to traditional teacher feedback in the development of the grammatical accuracy of argumentative essays written by Vietnamese EFL university students. The results of the study provided significant evidence of the transformative potential of AI tools in the second language writing learning process, while also revealing the pedagogical boundaries between the two approaches to writing instruction. The findings of this research contribute to the ever-growing body of research responding to the need for more research comparing teacher feedback and AWE tools (Li, 2021; Lv et al., 2021; Shang, 2019).

Interpretation of Grammatical Accuracy Findings

Regarding the first research question, the ANCOVA showed that the main effect of the feedback intervention was highly significant in the posttest grammatical accuracy of the students' writing. In fact, the model showed an exceptionally high explanatory power of 99.3%, meaning that 99.3% of the total variance in the error rate of the students' writing could be explained by their pretest abilities and the type of feedback they received. This suggests that the AWE intervention was substantially more effective in correcting grammatical errors in the students' writing than the traditional teacher feedback. This is in line with the general findings in the literature that the use of corrective feedback is beneficial to the development of L2 writing abilities (Cheng et al., 2021; Huisman et al., 2019; Lv et al., 2021; Zhang & Zhang, 2022), although the present study suggests that this is particularly the case with grammatical accuracy in writing.

There are several possible reasons why the use of the AWE tool was so substantially more effective in correcting grammatical errors in the students' writing than the use of the teacher feedback. First of all, the use of the AWE tool Grammarly provided the students with synchronous, not delayed, feedback. According to the 'noticing hypothesis' in second language acquisition, learners need to 'notice' the gap between their interlanguage and the target language in order to achieve linguistic restructuring. In fact, the use of the teacher feedback in the present study was hindered by the fact that the teacher could not always provide the students with immediate feedback. In fact, the teacher often had to wait several hours or even days before the students submitted their writing. This meant that the students' cognitive engagement with the linguistic features in the writing that the teacher corrected in the feedback was often diminished when the teacher provided the feedback. In other words, the students had forgotten the linguistic features in the writing that the teacher corrected in the feedback. This is not the case with the use of the AWE tool Grammarly, which provided the students with immediate, synchronous, not delayed, feedback. Additionally, the affective filter of the learners needs to be taken into consideration. For EFL learners, being handed a paper covered in red ink by a human instructor is a daunting and discouraging experience. On the other hand, interacting with a computer program may encourage learners to go through cycles of drafting and revising without the fear of being judged by a human instructor, as discussed in the individual feedback affordances of AWE tools by Link et al. (2022) and the self-learning affordances of AWE tools by Stevenson (2016).

The extremely high R^2 value of 99.3%, however, needs to be treated with some degree of caution when interpreting the results of the study, as such a strong prediction of results is rarely seen in educational research settings. One possible explanation is the homogeneous nature of the sample, as all participants

were drawn from intact intermediate-level classes at one institution. Alternatively, it is possible that the pretest scores of the participants were so close to being perfectly predictive of the results that the addition of the group variable simply explained the remaining variance with such a high degree of precision.

Comparison with Previous Research on Teacher Feedback Versus AWE

The current study's findings, therefore, both converge with and differ from existing comparative studies. Unlike Dikli and Bleyle's study (2014), which supported the efficacy of teacher feedback over AWE for advanced ESL learners in the United States, the current study's findings on intermediate Vietnamese ESL learners support the efficacy of AWE for learners who still struggle with surface-level grammatical errors. This might be because, for advanced learners, it is possible that they might need higher-level feedback, which is usually provided by teachers, whereas, for intermediate learners, they might need extensive, repeated, and intensified feedback on basic grammatical structures.

The current study's findings support Wilson and Czik's argument (2016) that it is possible to free up cognitive resources by automating lower-level revisions. Unlike Wilson and Czik's study, however, which showed that there was no statistically significant difference between conditions regarding overall quality, the current study showed that AWE had a statistically significant effect on grammatical accuracy, which is a part of overall quality. This might be because, whereas Wilson and Czik's study measured overall quality, the current study measured grammatical accuracy, and whereas they used Google Docs, which is free, Grammarly was used in the current study, which is premium and therefore might provide learners with more extensive, varied, and higher-level feedback.

The current study's findings support and expand on those by Thi and Nikolov (2022), who showed that Grammarly, which is an AWE tool, mainly addressed surface-level grammatical errors, whereas teacher feedback addressed both lower- and higher-level grammatical errors. The current study's findings show that, for intermediate learners who are still struggling with basic grammatical accuracy, surface-level feedback is what is needed to improve grammatical accuracy substantially. Such findings also directly address the need for empirical research comparing teacher feedback and AWE tools, as called for by Lv et al. (2021), with hard quantitative data supporting the overall higher efficacy of AWE feedback for grammatical accuracy for intermediate EFL learners. The study also addresses the need for research identified by Li (2021) on how the differing effects of teacher feedback and AWE tools play out in actual error correction.

Teacher Feedback and Grammatical Accuracy in Context

The marked improvement in grammatical accuracy for the AWE group should not overshadow the previously established effectiveness of teacher feedback. Indeed, numerous studies have empirically confirmed the effectiveness of teacher feedback in enhancing the overall quality of writing (Cheng et al., 2021; De Smedt et al., 2016; Lv et al., 2021; Zhang & Zhang, 2022), as well as particular aspects of

writing, such as grammatical complexity (Barrot & Gabinete, 2019; Lu & Ai, 2015) and fluency (Fathi & Rahimi, 2022). The current study does not negate the findings of the aforementioned studies but rather implies that, for the particular construct of grammatical accuracy, the use of AWE tools might provide certain benefits for the students.

In this context, the study by Jiang and Yu (2021) suggested that, in order for teacher feedback to effectively mediate the quality of the students' writing, the former should trigger change in the four critical dimensions of intentionality, reciprocity, transcendence, and meaning. However, the time constraints of the EFL classroom might not allow teachers to meet the requirements of these dimensions for the particular construct of grammatical accuracy in the students' writing. The use of AWE tools might allow teachers to meet the requirements of these dimensions for the construct of grammatical accuracy in the students' writing, as the tools would take care of the "surface-level correction," leaving teachers free to focus on the "intentional and dialogic" aspects of the construct, which cannot be replaced by the use of technology (Sybing, 2021). This is in accordance with the findings of the study by Pourdana and Asghari (2021), which suggested that the informational load of teacher feedback is rather heavy, and the use of technology might prove beneficial for the overall effectiveness of the teaching process.

Student Perception of AWE and Teacher Feedback

As far as the second research question is concerned, while it was not possible to collect any perception data, it is worth making some observations regarding how the learners might perceive the two types of feedback. Firstly, it can be noted that, despite the lack of perception data, it was possible to observe that, in comparison to the teacher feedback group, the revision cycles carried out by the AWE group were more frequent. This could suggest that, in comparison to teacher feedback, the immediacy of automated feedback might encourage a greater sense of learner autonomy, a notion that is reinforced by Stevenson (2016), who argued that the use of AWE tools would provide learners with greater opportunities for self-learning, since they would be able to manipulate the writing tasks without the need to use a human mediator.

Moreover, it is possible to suggest that the non-judgmental nature of automated feedback might help learners overcome their anxiety when it comes to writing, since they might be more willing to use new grammatical structures, which might otherwise be avoided when using teacher feedback, since this would be perceived as judgmental. This would, therefore, address the affective component of the process of receiving feedback, which, despite its benefits, might inadvertently create problems when using teacher feedback. This is reinforced by the work of Wang et al. (2013) and Ranalli (2018), who argued that it is essential to understand how learners perceive automated feedback, a notion that this study could not fully explore, but which would be worth examining in future research, since it would be

necessary to establish whether the objective benefits obtained in this study are matched by subjective benefits regarding the process of receiving feedback.

Pedagogical Implications

While the statistical variance obtained in this study might be seen to indicate a paradigm shift regarding how EFL teachers might need to consider the use of AWE tools in the future, it is worth pointing out that this should not be seen to suggest that teachers might be replaced by machines, but rather that a synergistic, hybrid model of pedagogy might be necessary, a notion that is reinforced by the work of Wilson and Czik (2016) and Thi and Nikolov (2022). Moreover, since it has been made clear that AWE tools are significantly more effective in the detection and correction of micro-level grammatical and mechanical errors, it is now possible to offload this laborious work onto the tools, freeing the educators to attend to higher-level writing concerns that the AI is still not able to assess effectively, such as the power of the argumentation, the sophistication of the socio-educational position, cultural contextualisation, and the emotional content of the writing. By requiring students to submit their work through the AWE tools before submitting them formally, educators are now able to ensure that they are assessing the student's ideas and not the fundamental grammatical mistakes in the work.

This strategy also caters to the limitations of EFL classrooms pointed out by Link et al. (2022), where the time and workload of the educators are such that they are not able to provide individualised feedback on multiple drafts of the work to the students. AWE tools will now help in the scaffolding of this feedback, allowing educators to provide immediate feedback to the students and freeing them to attend to the higher-level feedback that is most important. As Lewis (2018) pointed out, this also ensures that construct-irrelevant factors, such as handwriting, do not come into play in the feedback.

Limitations of the Study

In spite of the strong statistical results, the current study is limited in several ways. Firstly, the use of an intact group design, although necessary for the naturalistic conditions of the study, meant that randomisation was not possible, which could have led to several confounding variables. Secondly, the short-term intervention of eight weeks, although showing positive results, does not indicate whether the reduction in grammatical errors is an internalisation of linguistic knowledge or simply an overdependence on technology. There is certainly a concern that, as suggested by Ranalli (2018), the overuse of tools such as Grammarly can lead to an overdependence on algorithms, which in turn can cause students to rely on the technology without actually learning the linguistic rules. Thirdly, the current study utilised the premium version of Grammarly, which may not always be available to all students, particularly in less developed education systems. Fourthly, the current study was limited to argumentative essays, which, as Way et al. (2000), Lu (2011), and Qin & Uccelli (2016) have shown, can have an impact on linguistic performance. It must also be noted that the results may not generalise to other

essay types, such as narrative or descriptive writing, which have different discoursal features, as suggested by Berman (2008) and Loschky et al. (2020).

Future Research Directions

Future studies must consider longitudinal designs to ascertain if the grammatical accuracy gained through the use of AWE is retained once the technology is removed. Moreover, qualitative approaches must also be incorporated to investigate the cognitive processes of Vietnamese EFL learners in comparison to human-provided feedback. Such an approach would address the gap mentioned by Link et al. (2022) and also the one mentioned by Ranalli (2018) regarding the perceptions of EFL learners towards the use of AI. Exploring how students perceive the usefulness and credibility of feedback for different proficiency levels may also shed light on the development of guidelines for the optimum use of AWE in varied EFL settings. Comparative studies on the efficacy of AWE for varied writing genres, such as narrative and argumentative writing, may also reveal boundary conditions for the optimum use of AWE, following the lead of cross-genre studies by Way et al. (2000), Lu (2011), and Qin and Uccelli (2016). Lastly, the study should examine the possible erosion of "authorial voice" in the use of automated tools, as well as the interplay between the two sources of feedback in the development of various facets of writing competence. According to Zhai and Ma (2021), the study of the processes of AWE in extracting linguistic and rhetorical features through statistical models could help in the more sophisticated use of these tools in the teaching of writing.

Conclusion

The quasi-experimental study examined the impact of the use of Automated Writing Evaluation tools versus teacher feedback in the grammatical accuracy of argumentative writing among Vietnamese EFL university students. The study found that there was a highly significant main effect for the feedback condition, with the use of the AWE tools resulting in significantly higher grammatical error reduction than teacher feedback. The statistical model used in the study explained 99.3% of the post-test accuracy variance, controlling for the pre-test proficiency level, which showed the potent effect of automated feedback in the correction of grammatical errors among intermediate learners.

The advantage of AWE tools may lie in their immediacy, consistency, and non-judgmental nature, which allow writers to notice grammatical rules in real time and revise their work in an iterative manner without the anxiety of a judgmental feedback system. Nevertheless, these results do not suggest a replacement of the teacher with the technology. Rather, they suggest a hybrid model of teaching and learning, in which AWE tools are used to attend to the mechanical and grammatical errors, allowing the teacher to deploy their expertise in the higher-level issues of writing, argumentation, rhetorical effect, and the writer's voice, which are still beyond the capabilities of the current AI systems. The limitations of the study are its short duration, homogeneity of the sample, exclusive focus on one kind

of writing, and the exclusive use of premium software, which may not be accessible to a wider audience. Nevertheless, the results provide strong support for the efficacy of the strategic integration of AWE tools in the teaching of writing, allowing the full potential of the combination of the technological and the pedagogical to be optimised.

References

Ahmadi, A., & Parhizgar, S. (2017). Coherence errors in Iranian EFL learners' writing: A rhetorical structure theory approach. *Journal of Language Horizons, 1*(1), 9–37. https://doi.org/10.22051/lghor.2017.8588.1011

Barrot, J., & Gabinete, M. (2019). Complexity, accuracy, and fluency in the argumentative writing of ESL and EFL learners. *International Review of Applied Linguistics in Language Teaching, 59*(2), 209–232. https://doi.org/10.1515/iral-2017-0012

Berman, R. A. (2008). The psycholinguistics of developing text construction. *Journal of Child Language, 35*(4), 735–771. https://doi.org/10.1017/S0305000908008787

Cheng, X., Zhang, L. J., & Yan, Q. (2021). Exploring teacher written feedback in EFL writing classrooms: Beliefs and practices in interaction. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688211057665

Chung, E. S., & Ahn, S. (2021). The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2020.1871029

De Smedt, F., Van Keer, H., & Merchie, E. (2016). Student, teacher and class-level correlates of Flemish late elementary school children's writing performance. *Reading and Writing, 29*(5), 833–868. https://doi.org/10.1007/s11145-015-9590-z

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Fathi, J., & Rahimi, M. (2022). Examining the impact of flipped classroom on writing complexity, accuracy, and fluency: A case of EFL students. *Computer Assisted Language Learning, 35*(7), 1668–1706. https://doi.org/10.1080/09588221.2020.1825097

Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education, 44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896

Hyland, K. (2009). *Academic discourse: English in a global context*. Continuum.

Jiang, L., & Yu, S. (2021). Understanding changes in EFL teachers' feedback practice during COVID-19: Implications for teacher feedback literacy at a time of crisis. *The Asia-Pacific Education Researcher, 30*(6), 509–518. https://doi.org/10.1007/s40299-021-00583-9

Lewis, S. B. (2018). Human versus automated essay scoring: A critical review. *Arab World English Journal, 9*(2). https://doi.org/10.2139/ssrn.3201916

Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System, 99*, Article 102505. https://doi.org/10.1016/j.system.2021.102505

Liao, H. C. (2016). Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach. *System, 62*, 77–92. https://doi.org/10.1016/j.system.2016.02.007

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(3), 1–30. https://doi.org/10.1080/09588221.2020.1743323

Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science, 12*(1), 311–351. https://doi.org/10.1111/tops.12455

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

Lv, X., Ren, W., & Xie, Y. (2021). The effects of online feedback on ESL/EFL writing: A meta-analysis. *The Asia-Pacific Education Researcher, 30*(6), 643–653. https://doi.org/10.1007/s40299-021-00594-6

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing, 26*(1), 45–65. https://doi.org/10.1007/s11145-012-9392-5

Pourdana, N., & Asghari, S. (2021). Different dimensions of teacher and peer assessment of EFL learners' writing: Descriptive and narrative genres in focus. *Language Testing in Asia, 11*(1), Article 6. https://doi.org/10.1186/s40468-021-00122-9

Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of Chinese EFL learners' writing performance. *Journal of Second Language Writing, 33*, 3–17. https://doi.org/10.1016/j.jslw.2016.06.001

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

Shang, H. F. (2022). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments, 30*(1), 4–16. https://doi.org/10.1080/10494820.2019.1629601

Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition, 42*, 1–16. https://doi.org/10.1016/j.comp-com.2016.05.001

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Sybing, R. (2021). Examining dialogic opportunities in teacher-student interaction: An ethnographic observation of the language classroom. *Learning, Culture and Social Interaction, 28*, Article 100492. https://doi.org/10.1016/j.lcsi.2021.100492

Thi, N. K., & Nikolov, M. (2022). How teacher and Grammarly feedback complement one another in Myanmar EFL students' writing. *The Asia-Pacific Education Researcher, 31*(6), 767–779. https://doi.org/10.1007/s40299-021-00625-2

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26*(3), 234–257. https://doi.org/10.1080/09588221.2012.655300

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal, 84*(2), 171–184. https://doi.org/10.1111/0026-7902.00060

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Yoon, H. J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly, 51*(2), 275–301. https://doi.org/10.1002/tesq.296

Zabihi, R., Mousavi, S. H., & Salehian, A. (2020). The differential role of domain-specific anxiety in learners' narrative and argumentative L2 written task performances. *Current Psychology, 39*(4), 1438–1444. https://doi.org/10.1007/s12144-018-9850-6

Zhai, N., & Ma, X. (2021). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2021.1897019

Zhang, J., & Zhang, L. J. (2022). The effect of feedback on metacognitive strategy use in EFL writing. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2022.2069822