Original Research Article

RELC Journal
2026, Vol. 57(1) 32–47

DOI: 10.1177/00336882241268359
journals.sagepub.com/home/rel

Effects of Automated Feedback on English as a Foreign Language Learners' Writing Performance: Evidence from a Quasi-experiment

Giang Thi Linh Hoang
Faculty of English, University of Foreign Languages and International Studies, Hue University, Vietnam

Abstract

Automated writing evaluation (AWE) is increasingly used to provide formative feedback on second language (L2) students' writing. A key factor influencing the effectiveness of AWE feedback on L2 writing performance is the learners' revision behaviors as they process the feedback. Adopting a quasi-experiment, this study aims to evaluate the impacts of *Criterion* automated corrective feedback (ACF) on English as a foreign language (EFL) students' writing performance based on two measures of accuracy: overall writing accuracy and accuracy of English article usage. Learners' textual operations in response to *Criterion* ACF were examined for possible explanations for recorded gains (if any) in their writing accuracy. The main findings indicate a lack of intervention and retention effects on learners' accuracy over the semester during which *Criterion* ACF was incorporated to supplement the writing instructor's feedback on organization and content. In addition, across four writing entries conducted on *Criterion*, learners' revisions to their essays following *Criterion* ACF were primarily at the local level, dominated by addition, deletion, or substitution of individual words or short phrases rather than substantive revisions to their scripts. About one third of all *Criterion* feedback points did not result in textual changes to the first drafts, indicating a moderate uptake rate of the feedback. Implications related to formative feedback practices in the EFL writing classroom and the adaptation of *Criterion*'s technical capacities are accordingly presented.

Keywords

Automated feedback, *Criterion*, writing performance, English article system, textual operations, revisions

Corresponding author:
Giang Thi Linh Hoang, 3/10/91 Han Mac Tu St, Hue, Thua Thien Hue, Vietnam.
Email: htlgiang@hueuni.edu.vn

Introduction

Automated writing evaluation (AWE) has increasingly become ubiquitous in second language (L2) writing. Starting to attract public attention in the 1990s, AWE initially emerged as a practical solution to marking standardized writing tests, such as the replacement of one human rater in the Graduate Management Admissions Test with the Educational Testing Service (ETS) scoring engine *e-rater* or with Vantage Learning's *Intellimetric* (Warschauer and Grimes, 2008). Early automated scoring engines have since been added with editing tools aimed at formative assessments in the classroom, with the advent of several AWE programs such as *Criterion*, *Grammarly*, and *Write & Improve*. The editing tools incorporated in each AWE program generate automated feedback on grammar, lexical usage, mechanics, style, organization, and rhetorical choices in writing alongside an automated score assigned to each submitted essay. In several English as a foreign language (EFL) and English as a second language (ESL) contexts, large classes pose a great challenge for teachers' feedback provision. Therefore, a number of writing instructors have supplemented their feedback with automated feedback, especially feedback on the mechanical aspects of writing, leaving more time for teacher feedback on higher order skills such as content and organization to alleviate teacher workload and promote learners' autonomy in their writing and revision processes. Of the different error types targeted by AWE feedback, English articles have been consistently found to be a notoriously challenging aspect to be acquired among L2 learners due to their low level of treatability (e.g., Chodorow et al., 2010; Han et al., 2006; Murakami and Alexopoulou, 2016; Robertson, 2000). Most published research on written corrective feedback (CF) has examined English articles and produced mixed findings (e.g., Shintani and Ellis, 2013; Shintani et al., 2014), which warrants further AWE research into this aspect of learners' language acquisition.

Impacts of Automated Feedback on Writing Performance

Research into the effectiveness of automated feedback has sought to answer whether such feedback contributes to writing performance as evidenced in students' new compositions (retention) with mixed results. Kellogg et al.'s (2010) comparison of three feedback conditions (no feedback, intermittent feedback, and continuous feedback from *Criterion*) among 59 ESL students showed that although holistic scores were not significantly improved, students receiving continuous feedback experienced a decrease in errors of grammar, mechanics, usage, and style. In a study by Li et al. (2017), change was calculated by comparing error reduction rates from first drafts of the first paper to first drafts of subsequent papers. Their study showed that students improved in terms of one error type, *run-on sentences*, but not in the remaining eight error types where the participants experienced improved accuracy from first to revised drafts of the same essay. More recently, in a study on *Grammarly*'s effects on Filipino ESL students' writing accuracy over a semester using a quasi-experimental approach, Barrot (2021) found that the students who utilized *Grammarly* automated corrective feedback (ACF) outperformed those in the control group in the posttest judged by overall accuracy scores derived from student essays based on a weighted clause ratio approach. Also studying the effects of *Grammarly* feedback, yet focusing on students' accurate use of English articles, Ebadi et al. (2023) found that students who received both *Grammarly* and teacher form-focused feedback improved the

most in their article usage in academic writing. The authors suggest that the combined feedback condition of teacher plus *Grammarly* feedback is best for EFL learners' acquisition of the English articles.

Some studies have looked at writing performance by comparing holistic scores assigned to pre and posttest essays composed by students following the incorporation of the AWE feedback. In Reynolds et al.'s (2021) quasi-experiment, first language (L1) Chinese students were divided into two groups: one received automated feedback as original printouts of the feedback, while the other group was under the impression that the feedback they received was from the teacher (which actually was automated feedback reformatted to look like teacher feedback). The findings showed that the perceived automated feedback group improved significantly from essay 2 to essay 3, but performed worse on essay 4 over a duration of 18 weeks. It is noted that students' performance across the four essays was scored by the AWE program, *PaperRater*. Similarly, Curry and Riordan's (2021) examination of *Write & Improve* provided positive findings regarding the system's facilitation of improved writing proficiency judged by the improved automatic scores generated to the first draft of students' first writing task and the first draft of the last writing task using Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) levels as benchmarks.

Overall, previous research suggests improvement at differential levels in learners' writing performance over time following the use of automated feedback from different AWE systems. Results from these studies, however, should be interpreted with some caution because of the way accuracy change or writing proficiency was calculated. Some of these studies were based on contestable assumptions that the ACF is accurate or the AWE scoring is reliable when they chose to use the automated scoring as the standard to judge learners' writing quality and changes in accuracy scores (e.g., Curry and Riordan, 2021; Reynolds et al., 2021). This methodological choice poses a potential risk to the validity of the conclusions, as the automated feedback itself is still open to flaws and needs further improvement.

Students' Revisions Following the Automated Feedback

Frequently mentioned in the discussion about written CF is the notion of *engagement*, which is defined as the extent to which students are devoted to their learning as demonstrated in their responses to texts and attitudes to writing (Zhang and Hyland, 2018). Specifically addressing oral and written CF research, Ellis (2010) used *engagement* to indicate learners' response to the feedback they received and proposed a three-perspective approach to investigating learners' engagement with the written feedback, namely cognitive, behavioral, and affective perspectives. Most relevant to the current research is the concept of *behavioral engagement*, which is operationalized by Zhang and Hyland (2018) as students' reaction to the feedback, including their editing/revising processes or textual operations (i.e., specific actions to make changes to the text in response to the feedback).

Research into students' behavioral engagement using the automated feedback mostly showed a moderate uptake rate of AWE feedback. For example, Chapelle et al.'s (2015) examination of *Criterion* ACF shows that nearly half of all the feedback did not lead to any revised forms. Jiang and Yu (2020) framed their study around the activity theory to examine EFL students' appropriation of the automated feedback and found that students

appropriated the feedback at differential levels, used various resources (e.g., dictionary), and sought help from different community members as they engaged with the feedback. Similarly, in Barrot's (2021) study, students reported learning grammar through *Grammarly*'s metalinguistic explanations, which helped them to notice the forms and gaps in their linguistic knowledge. Liu and Yu's (2022) study examined learner engagement with *Write & Improve* by recording students' revision in their writing in response to the feedback and found that students incorporated more direct word-level feedback compared to indirect feedback. In total, about one fourth of all their revisions were changes to the content.

The literature review shows scant research on specific textual operations conducted by learners when they behaviorally engage with automated feedback, a gap that has been pointed out in previous studies (e.g., Kim and Bowles, 2019; Stevenson, 2016; Zhang and Hyland, 2018) which call for more emphasis on the process perspective in automated feedback research. Moreover, there have been few enquiries into the relationship between the recorded improvement in writing performance and students' revising processes that may lead to long-term writing proficiency. There is, therefore, a need for research that looks at revision processes and textual operations in search of evidence for the gains (or the lack thereof) in learners' writing following the use of automated feedback to inform pedagogical decisions in L2 writing. Addressing these gaps, the current study assembled data related to both observed accuracy changes in L2 learners' written scripts and their revision behaviors for possible explanations for the effects of *Criterion* ACF on L2 learners' writing. Two research questions guided this study:

1) To what extent does the use of *Criterion* ACF facilitate L2 writers' English writing proficiency in terms of overall accuracy and article usage?
2) What are students' textual operations in response to the automated feedback generated by *Criterion*?

Methodology

Context of the Study

This study was conducted among second-year English majors of a university in central Vietnam. All the students were informed of the research purpose before their consent to take part was collected. Each student received an honorarium in appreciation for their voluntary participation. Incorporated as part of the English academic writing course, *Criterion* was intended to provide diagnostic feedback on student writing. Adopting a convenience sampling approach, this quasi-experimental study used the pre, post, and delayed posttests on two intact writing classes, subdivided into an experimental and a comparison group. The comparable numbers of students in each class and their taking the same writing course taught by the same writing instructor make these classes a good fit for the research purpose of examining the effects of AWE feedback on students' writing performance. The current research focused on students' writing accuracy, with the two specific outcome variables of *overall accuracy* and *accuracy in English article usage*. *Criterion* feedback was thus limited to CF on grammar, usage, and mechanics, while the automated feedback related to style, content, and organization was not provided. This decision aims to take out the impacts of *Criterion* feedback on

style, content, and organization on experimental students' linguistic accuracy. Meanwhile, both groups received feedback on content and organization from the writing instructor, which took place after revisions in response to *Criterion* ACF had been conducted.

Participants

The experimental class includes 38 students, while the comparison group has 37 students. The two groups were comparable in terms of age (ranging from 19 to 20), years of learning English, and general English proficiency level. By the time data were collected, all the students had completed and passed the four skill courses of listening, reading, speaking, and writing at the B1 CEFR level. However, pretest essay length for the comparison group averaged at 166 words, while that for the experimental group was 218. This measure alone can be indicative of some differential levels in writing skills, taking into consideration the fact that the pretest was administered during the second week of the semester for both groups under similar conditions (i.e., 45 min timing and paper-and-pen writing modality). This differential word length in the pretest essay will be taken up in the data analyses to control for pretest differences. Table 1 summarizes the information about the two groups.

The Intervention

The intervention lasted 15 weeks, with weekly meetings of 100 min each. In the second week of the semester, the experimental group attended a walk-through of *Criterion.* This was followed by a homework task that required them to log into their accounts to compose the first writing entry on *Criterion*. The researcher followed up by checking student submissions and giving further help in the following sessions to make students feel comfortable about writing on *Criterion*. On weeks five, eight, and eleven, writing practice took place in the computer lab. Other than that, students were encouraged to write to *Criterion*-loaded writing prompts of their choice as homework practice to use the automated feedback for self-revisions.

The experimental group composed essays and submitted them to *Criterion* for feedback on grammar, word usage, and mechanics to make revisions, before sending the revised drafts to the teacher also for general comments on content and organization as the comparison group. For each writing task, students were required to write and revise their drafts using *Criterion* automated feedback. *Criterion* allows students to

Table 1. Participants' demographics.

	Comparison group		*Experimental group*	
Total participants	37		38	
Gender distribution	3 males; 34 females		2 males; 36 females	
	Mean	*SD*	*Mean*	*SD*
Age	19.4	0.5	19.5	0.6
Years of learning English	10.3	2.0	9.6	2.8
Word length in pretest	166	48	218	60

draft and redraft as many times as they like. Students who did not have revised drafts for one or more of their writing entries were excluded from the study. Table 2 summarizes the procedures for this research.

The fact that the two groups did not spend similar amounts of time practicing to write is intended as part of the intervention to examine the effects of *Criterion* ACF on students' writing accuracy. Considering the same number of practice sessions and same writing prompts for both groups, the difference between the experimental and non-experimental conditions is that for each essay, the former group experienced two rounds of feedback, one being the CF from *Criterion*, which requires revisions, and the other being the teacher's general feedback on content and organization. Despite this, no revised draft is required following the teacher's feedback. Meanwhile, the comparison group did not receive any form-focused feedback. They received only one round of teacher general feedback on content and organization, which is the same as the experimental group, with no requirement for revised drafts. The extra form-focused feedback round with required revisions in response to the automated feedback creates the "experimental" condition, which allows for the measurement of the effects of *Criterion* ACF on learners' writing accuracy over the intervention period.

Data Collection and Analyses

Two instruments were used to assemble data in this research: pre, post, and delayed posttest essays and students' essays from the four writing tasks on *Criterion*.

Pre, Post, and Delayed Posttest Essays. The pre, post, and delayed posttests were 45 min in-class handwritten essays. Learner errors in a total 225 test essays were coded for

Table 2. Research procedures and data collection timeline.

Meeting	Experimental group	Comparison group
1	Introduction about the research project	
2	**Pretest**	
	Walk-through of *Criterion*	**Writing entry 1:** homework using the
	Students' account setup	paper-and-pen modality
	Writing entry 1: homework on *Criterion*	
5	**Writing entry 2:** Writing a problem-solving essay	
	Students composed and revised essays on *Criterion* in the computer lab.	Students composed essays in the paper-and-pen modality in class.
8	**Writing entry 3:** Writing an opinion essay	
	Students composed and revised essays on *Criterion* in the computer lab.	Students composed essays in the paper-and-pen modality in class.
11	**Writing entry 4:** Writing a for & against essay	
	Students composed and revised essays on *Criterion* in the computer lab.	Students composed essays in the paper-and-pen modality in class.
13	**Posttest**	
	No class meetings for four weeks in both groups	
17	**Delayed posttest**	

two kinds of accuracy. Firstly, the overall accuracy of each essay was calculated using Foster and Wigglesworth's (2016) weighted clause ratio approach to yield a more nuanced measure than coding error-free clauses while addressing the practical challenge of being consistent in coding individual L2 learner errors (see Supplementary Appendix 1). Secondly, students' accuracy of article usage was calculated following an obligatory occasion analysis (see Supplementary Appendix 2). Article use was purposefully selected for its (a) salience as the most frequently tagged error type in the corpus of essays in this research and (b) low level of treatability.

The scores for general accuracy and accurate use of articles met the assumptions of normal distribution, linearity, and range. Scores were slightly negatively skewed for these datasets, showing that the frequent scores were more clustered towards the higher end of the scale, but no issues with kurtosis were detected. Mauchly's tests on all the sets of pretest, posttest, and delayed posttest scores for each accuracy measure also indicated that the assumption of sphericity has been met for the main effects of the variables under study (Field, 2009; Larson-Hall, 2010). Independent-samples *t*-tests were then conducted to see if there was any statistical difference between the two groups' pretest accuracy scores. Table 3 presents the results.

There was no significant difference in the two groups' pretest scores for article usage. Therefore, repeated measures analysis of variance (ANOVA) was conducted using accuracy scores obtained for the pre, post, and delayed posttests. The between-subject effect was group (with *two levels*, comparison and experimental), and the within-subject effect was time (pretest, posttest, delayed posttest). The mean difference between the posttest and pretest is interpreted as the intervention effect on the experimental group (or acquisition), and that between the delayed posttest with the posttest as retention. Regarding overall accuracy, a significant difference in the two groups' pretest scores was detected. As it is important that all analyses are adjusted for baseline levels, for this measure, gain scores were used for inferential analyses using repeated measures ANOVA. This statistical choice aims to eliminate confounds (Field, 2009), in this case being student differential baseline levels before the intervention was implemented. Two gain scores were calculated: gain score 1 was obtained by subtracting pretests from posttests to show the intervention effect, while the second gain score was posttests subtracted from delayed posttests to evaluate retention.

First and Revised Drafts on Criterion. First and revised drafts written by the experimental group from the four writing entries on *Criterion* were used for coding students' textual operations, making a total of 304 scripts. Coding student revisions involves comparing the first and revised drafts to see what changes students made to the text in response to each *Criterion* error tag. Categories from Han and Hyland's (2015) list of revision practices were adapted for initial trial-coding when the author examined students' revisions in response

Table 3. Statistical difference in two groups' pretest scores.

Accuracy measure	Independent-samples *t*-test result	Statistical difference?
Overall accuracy	$t(73) = 4.56$, $p < .001$	Yes
Article	$t(73) = .896$, $p = .373$	No

to the automated feedback and conducted coding on a portion of the data. This round of initial coding resulted in adjustments to the list of coded categories for textual revisions to produce a workable list applicable to the actual data. In the final coding scheme, six categories of students' textual operations in response to *Criterion* ACF applied: *no change*, *removal*, *addition*, *deletion*, *substitution*, and *reformulation* (see Supplementary Appendix 3). Among the six coded categories, *reformulation* refers to textual operations that revise larger portions of texts above the phrasal level, such as changes to clause/sentence structures or the introduction of new expressions to reword an idea.

Both students' grammatical accuracy and textual operations were double-coded. The double-coding of *overall accuracy* in students' test essays on the first 10% of the total clauses was challenging, and inter-coder agreement was not satisfactory. Therefore, further discussion was needed before double-coding took place on another 10% of the total extracted clauses. Coding article errors and textual operations in response to *Criterion* ACF went smoothly and double-coding of 10% of the total dataset achieved high consensus. Overall, (a) 20% of the total clauses (for the weighted clause ratio), (b) 10% of total article errors (for obligatory occasion analysis), and (c) 10% of total revision points (for textual operations) were double-coded. Inter-coder reliability statistics were (a) 81%, (b) 90.5%, and (c) 90%, respectively. All the differences were discussed and resolved before the rest of the data were coded by the author.

Results

Effects of Criterion ACF on Students' Writing Accuracy

The presentation of findings for each accuracy measure will start with descriptive statistics, followed by the inferential statistics related to the intervention and retention effects of using *Criterion* among the experimental group and how that compares to the comparison group.

Overall Accuracy. The experimental group's overall accuracy experienced little change across the pre, post, and delayed posttest, while that of the comparison group saw some improvement from the pretest to the posttest results. More details are found in Table 4.

Repeated measures ANOVA results indicated that there was no significant interaction effect between time and group on the accuracy scores for the experimental and comparison group, $F(1, 73) = .208$, $p = .649$, $\eta^2 = .003$. There was also no significant overall main effect of time, $F(1, 73) = 3.605$, $p = .062$, $\eta^2 = .047$, but a significant main effect for group on accuracy gain scores for the two groups was found, $F(1, 73) = 12.396$, $p = .005$, $\eta^2 = .101$.

Table 4. Descriptive statistics of overall accuracy.

	Experimental group ($N = 38$)		Comparison group ($N = 37$)	
Time	*Mean*	*SD*	*Mean*	*SD*
Pre	0.81	0.11	0.69	0.13
Post	0.83	0.08	0.77	0.11
Delayed	0.81	0.08	0.74	0.10

The significant main effect for group was followed up by an independent-samples *t*-test, which showed that there was a significant difference in mean overall gain scores at posttest between the experimental and comparison groups, $t(62.799) = -2.798$, $p = .007$ (equal variances not assumed). The average gain score for students in the experimental group was 0.069 lower than the average gain score at posttest for comparison group students. These results suggest that the incorporation of *Criterion* ACF did not facilitate general accuracy gains for the experimental group, while the comparison group, despite the lack of access to *Criterion* ACF, significantly gained more in overall accuracy. Next, an independent-samples *t*-test was conducted to compare the two groups' gain scores at the delayed posttest. There was no significant difference in the experimental group ($M = -.0161$, $SD = .084$) and comparison group ($M = -.0262$, $SD = .108$) conditions, $t(73) = .457$, $p = .024$. These results suggest that neither group retained overall accuracy, despite the access to *Criterion* ACF as an additional resource for the experimental group.

Accuracy of English Article Use. The descriptive statistics shown in Table 5 indicate that the experimental group gained slightly in article usage from the pretest to the posttest and delayed posttest, which is also true of the comparison group.

Repeated measures ANOVA results show no significant interaction effect of time and group on the article use scores for the experimental and comparison groups, $F(2140) = 0.015$, $p = .985$, $\eta^2 = .000$. No significant effect of time was found, $F(2140) = 1.746$, $p = .178$, $\eta^2 = .024$. The main effect of group on article use accuracy scores for the two groups was not significant either, $F(1,70) = 3.027$, $p = .086$, $\eta^2 = .041$. The results suggest no significant treatment effect of *Criterion* ACF on students' article use accuracy throughout a semester, as well as the lack of retention over time.

Students' Textual Operations Following Criterion *ACF*

Of the total 3026 error tags extracted from 152 first drafts of learner writing, students' uptake rate (i.e., the quantification of their revisions that corresponded to *Criterion* error codes irrespective of the revision outcomes) was 65.1%. This means that 65.1% of all *Criterion*-tagged errors in learner first drafts were revised. The remaining 34.9% of the flagged errors were either left unchanged or the section of text containing the tagged error was removed in revised drafts. Figure 1 shows a more detailed breakdown of students' textual operations in response to *Criterion* feedback.

Table 5. Descriptive statistics of article use accuracy.

	Experimental group ($N = 38$)		Comparison group ($N = 37$)	
Time	*Mean*	*SD*	*Mean*	*SD*
Pre	0.74	0.18	0.70	0.19
Post	0.80	0.15	0.74	0.18
Delayed	0.78	0.14	0.73	0.19

It is worth noting that not all of the 31 error types that *Criterion* generates feedback on have instances found in the essay corpus of the current research. Table 6 provides raw counts of textual revisions in response to *Criterion* error tags by error type.

Examination of the error types in Table 6 reveals that *Criterion* generally targets local and surface-level errors (e.g., capitalization, spelling, missing final punctuation, …). Among the less treatable types are English articles or sentential level tags of *fragments*, *run-on sentences*, *garbled sentences*, and *proofread this*, which highlight the whole sentence suggesting students make major revisions. Therefore, it is not surprising that *Criterion* feedback tends to trigger textual revisions at the local level, mostly in the form of the substitution of a word with another word. In the following example, the student confused "there" with "their," probably due to their identical pronunciation, which was correctly revised following *Criterion* error message:

First draft: They are not keen on there[1] job anymore.

Error message: *You have used **there** in this sentence. You may need to use **their** instead.*

Revised draft: They are not keen on **their** job anymore.

No change is the second most common category found in response to *Criterion* feedback. Besides the expected *no change* response to 325 false alarms from *Criterion*, 652 out of 977 total *no-change* responses despite the correct error tags from *Criterion* make up a high non-uptake rate. Although not as commonly adopted, *addition* and *deletion* are the next two types of textual operations among these learners, at 14.6% and 7%, respectively. Like substitution, qualitative analyses of the revisions made by students show that *addition* or *deletion* took place mainly at the lexical level, as in the following example where the student deleted

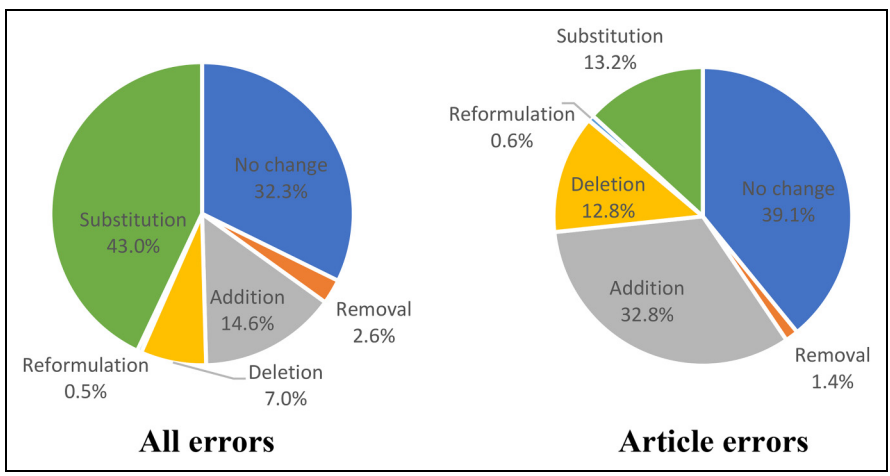

Figure 1. Students' textual operations in response to *Criterion* automated corrective feedback.

Table 6. Students' textual operations in response to *Criterion* automated corrective feedback by error type.

Error type	No change	Removal	Addition	Deletion	Reformulation	Substitution
Capitalization	71	3	10			87
Compound words	1	1		1		7
Confused words	46	11	4	3		53
Determiner–noun agreement	57	14	32	32		33
Duplicates	1	3				
Extra comma	2		5	30		3
Faulty comparisons	1			2		
Fragments		1	11	2		14
Garbled sentences			9	2	1	6
Hyphen errors	2					2
Ill-formed verbs	33	2	18	19	3	47
Missing comma	104	3	75	2		2
Missing final punctuation		6	5			
Missing or extra articles	297	11	253	94	5	89
Missing question mark	6					1
Negation errors				2		1
Possessive errors	6	1	1	3		6
Preposition errors	48	3	2	7		40
Pronoun errors	2					5
Proofread this	17		7	2	4	33
Run-on sentences	1		4			1
Spelling	223	15				664
Subject–verb agreement	54	6	7	6	1	193
Wrong article	5			5		13
TOTAL	**977**	**80**	**443**	**212**	**14**	**1300**
PERCENTAGE (%)	**32.3**	**2.6**	**14.6**	**7**	**0.5**	**43**

the possessive marker from the original text in response to *Criterion*'s possessive error tag (which is an incorrect suggestion):

> First draft: The question if money is a main indicator of people's[1] being a controversial one.
>
> Error message: *You may need to take out the apostrophe to make this word a plural noun.*
>
> Revised draft: The question if money is a main indicator of **people** being a controversial one.

Removal, another type of revision actions taken by the students, occurred in 80 instances of the total corpus. In the example below, the student chose to remove the whole phrase

"each people" instead of changing the determiner "each" or the modified noun "people" to address *Criterion* ACF:

> First draft: to sum up, self awareness of each[1] people is best solution to reduce this pollution.
>
> Error message: *You may have used the wrong determiner. Proofread the sentence to make sure that the determiner agrees with the word it modifies.*
>
> Revised draft: To sum up, self awareness is best solution to reduce this pollution.

Much less frequently found was the removal of the whole sentence and new ideas were introduced instead, as in the example below:

> First draft: Moreover, another[1] of staying in the same jobs is that it can open door for employers to learn and advance their skill or their jobs in their career.
>
> Error message: *You may have used the wrong determiner. Proofread the sentence to make sure that the determiner agrees with the word it modifies*.
>
> Revised draft: Moreover, changing career helps them gain expertise in a new area and make them have more opportunities in the future.

Reformulation, a more substantive type of textual operation, occurred least frequently in the corpus (at 0.5%). As expected, reformulation is the type of textual revision in response to less treatable error types, mostly related to *garbled sentences* and *proofread this*. The next example illustrates a student's reformulation following *Criterion*'s *proofread this* tag:

> First draft: This not only affect to human's health when they use water from rivers or lakes sources[1], but also threaten many kinds of fishes, srhimps in the seas.
>
> Error message: *This part of the sentence contains an error or misspelled word that makes it hard to understand what you mean.*
>
> Revised draft: This not only affect to **people's health** when they use **water sources from rivers or lakes,** but also threaten many kinds of fishes, shrimps in the seas.

Compared to overall revision actions, students' textual operations in response to article error feedback is marked by a high rate of *no-change* responses (at 39.1%). Other actions, including *addition*, *deletion*, and *substitution*, accounted for the large majority of students' textual edits at the lexical level in response to feedback on articles. The following example illustrates the student's *addition* of the missing article as *Criterion* suggested:

> First draft: Same[1] work throughout the career is tedious to many people.
>
> Error message: *You may need to use an article here. Consider using the article **the**.*
>
> Revised draft: **The same** work throughout the career is tedious to many people.

Discussion

The first research question examines the extent to which *Criterion* ACF facilitates students' acquisition of English in terms of overall accuracy and article usage in L2 writing. Quantitative analyses using the pre, post, and delayed posttests generally showed no intervention or retention effects of the use of *Criterion* ACF on learners' accuracy over time. This is not in line with some previous research showing sustained accuracy gains among learners who have access to AWE form-focused feedback (e.g., Barrot, 2021; Ebadi et al., 2023). In this study, the comparison group gained more in overall accuracy despite their lack of access to *Criterion* ACF. It is worth noting that the comparison group's baseline at pretest is significantly lower than that of the experimental group, and the two groups' overall growth trajectories are thus based on gain scores to adjust for differential pretest levels. Therefore, there was much more room for improvement in the comparison group and their recorded accuracy gain from the pretest to the posttest suggests some relationship between lower baselines and chances to improve with practice over time, irrespective of the ACF. In contrast, the experimental students' high pretest accuracy scores, being more sensitive to the ceiling effect, may have left little room for improvement following the use of *Criterion* ACF. The ceiling effect may also have interacted with the lack of contingent and graduated metalinguistic explanations for less treatable errors, resulting in little change in learners' overall accuracy.

The second research question probes students' textual operations in response to *Criterion* ACF, and the results showed a moderate uptake rate of the feedback dominated by superficial and local-level revisions to the texts. The data for this question provide plausible explanations for the lack of improved L2 accuracy use. A non-uptake rate of one third of all error tags corroborate the high underuse rate of *Criterion* ACF among the English language learners reported in previous research (e.g., Chapelle et al., 2015; Lavolette et al., 2015). The particular case of English article errors adds insights into the way *Criterion* ACF was responded to among these EFL learners. The finding that almost four out of ten article error tags led to no revised form is suggestive of students' reservation to adopt the feedback. For the remaining 60% of article error flags, students took up the suggested changes in the form of deletion, replacement, or addition of the article in the sentence. The predominance of lexical-level revisions in response to *Criterion* feedback on articles finds partial explanation in the fact that *Criterion* generates feedback on two types of article errors: *missing/extra article* and *wrong article*. In missing or extra article errors, addition or deletion of an article is the system's advice to learners, through either indirect CF, as in "You may need to use an article before this word," or direct CF, such as "You may need to use an article before this word. Consider using the article **a**." For instances of wrong article errors, *Criterion* generates two error messages, "You have used **a**. You may need to use **the** instead" and "You have used **the**. You may need to use **a** instead," both of which direct learners to the substitution of one article with another. However, the feedback message is not elaborated on beyond these suggestions. For a less treatable error type, such as article usage, the lack of reinforcement through relevant metalinguistic explanations may limit students' processing of the feedback to the short-term task of error correction. If more elaborate metalinguistic explanations were provided, learners' attention could be drawn to certain gaps in their interlanguage development, which potentially facilitates L2 acquisition (Heift and Hegelheimer, 2017).

The lack of acquisition or retention of target language forms after four interventions among the experimental students does not support Flekenstein et al.'s (2023: 7) meta-analysis of AWE feedback on students' performance, which found that "long interventions of more than two sessions showed a significant effect." Two insights from the current research should be noted to explain this divergence. Firstly, the data from students' behavioral engagement with the feedback via their textual operations reveal predominantly superficial revisions in response to the automated feedback, which seems to foreshadow the lack of observed gains in either of the studied accuracy measures, probably due to limited processing of the feedback. This is further evidenced in the few recorded instances of reformulation in students' revision practices, even in response to sentence-level error tags of *proofread this*, *fragments*, or *garbled sentences*. Secondly, surface-level errors, which form the large majority of *Criterion* ACF, tend to be easily remedied. Considering the students' high proficiency level as English majors, many errors detected by *Criterion* in this corpus of tertiary learner essays were probably slips due to typing under time constraints rather than true gaps in their grammatical knowledge. In the high-impact testing conditions of the pre, post, and delayed posttests with the handwriting modality, students are more likely to write carefully to avoid such surface-level errors.

Implications and Limitations

A few implications based on the discussion are worth consideration. Firstly, what this research has revealed is *Criterion*'s failure to initiate students' substantive revisions via strategic utilization of the available feedback. Therefore, writing instructors should incorporate strategy instruction as part of familiarizing learners with AWE tools and their automated feedback functions, as AWE feedback can only benefit learners who are able to self-regulate their learning through skillful use of various resources. Supplementary oral feedback sessions during class hours can also be conducted for learners to bring up clarification questions after they have engaged with the ACF in the early stages of *Criterion* implementation.

On the part of *Criterion*, the lack of accuracy gains among learners and their humble uptake rate may indicate the need for more accurate and elaborate feedback. *Criterion* developers' choice to err on the side of *precision* over *recall* (Chodorow et al., 2010) is still highly relevant if learners' trust in the ACF is to be improved. In addition, it is expected that more meaningful algorithms are added to *Criterion* so that it can detect higher textual-level error types, add elaboration on currently generic error flags for *proofread this* and *garbled sentences*, or provide more metalinguistic explanations for less treatable errors.

The current research has some limitations to be acknowledged. Firstly, beside the baseline requirement of four writing entries on *Criterion*, the amount of students' self-practice on *Criterion* was not strictly controlled in this research, which may result in possible differential exposure levels among the 38 experimental students. Future studies can adopt case-study research that examines individual mediating factors for more nuanced understanding of how individual learners' use of the feedback impacts their writing development. Also, the study is relatively small scaled with two intact classes of 75 students, hindering generalization of the findings to other teaching contexts, especially for English learners of other language backgrounds whose L1s may be more or less

comparable to English. In addition, the intervention lasted one semester with four required writing entries, which may not allow enough exposure for learners to become skillful users of *Criterion* feedback and internalize the target language forms. Future research may aim to address this shortcoming through more intensive intervention with a higher frequency of required tasks to enhance exposure to the feedback.

Declaration of conflicting interests

The author has no conflicts of interest to declare.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Melbourne International Research Scholarship, The University of Melbourne, Australia.

ORCID iD

Giang Thi Linh Hoang https://orcid.org/0000-0001-6890-7950

Supplemental Material

Supplemental material for this article is available online.

References

Barrot JS (2021) Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning* 36(4). DOI: 10.1080/09588221.2021.1936071.

Chapelle CA, Cotos E and Lee JY (2015) Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing* 32(3): 385–405.

Chodorow M, Gamon M and Tetreault J (2010) The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing* 27(3): 419–436.

Curry N and Riordan E (2021) Intelligent CALL systems for writing development: investigating the use of write & improve for developing written language and writing skill. In: *CALL Theory Applications for Online TESOL Education. Advances in Educational Technologies and Instructional Design (AETID)*. Hershey: IGI Global, 252–273.

Ebadi S, Gholami M and Vakili S (2023) Investigating the effects of using *Grammarly* in EFL writing: The case of articles. *Computers in the Schools* 40(1): 85–105.

Ellis R (2010) Epilogue: A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition* 32: 335–349.

Field A (2009) *Discovering Statistics Using SPSS*. 3rd ed. CA: Sage Publications Ltd.

Flekenstein J, Liebenow LW and Meyer J (2023) Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence* 6: 1–11.

Foster P and Wigglesworth G (2016) Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics* 36: 98–116.

Han N, Chodorow M and Leacock C (2006) Detecting errors in English article usage by non-native speakers. *Natural Language Engineering* 12(2): 115–129.

Han Y and Hyland F (2015) Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of Second Language Writing* 30: 31–44.

Heift T and Hegelheimer V (2017) Computer-assisted corrective feedback and language learning. In: Nassaji H and Kartchava E (eds) *Corrective Feedback in Second Language Teaching and Learning: Research, Theory, Applications, Implications*. New York: Routledge, 51–65.

Jiang L and Yu S (2020) Appropriating automated feedback in L2 writing: experiences of Chinese EFL student writers. *Computer Assisted Language Learning* 35(7): 1329–1353.

Kellogg RT, Whiteford AP and Quinlan T (2010) Does automated feedback help students learn to write? *Journal of Educational Computing Research* 42(2): 173–196.

Kim HR and Bowles M (2019) How deeply do second language learners process written corrective feedback? Insights gained from think-alouds. *TESOL Quarterly* 53(4): 913–938.

Larson-Hall J (2010) *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge.

Lavolette E, Polio C and Kahng J (2015) The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology* 19(2): 50–68.

Li Z, Feng H and Saricaoglu A (2017) The short and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *CALICO Journal* 34(3): 355–375.

Liu S and Yu G (2022) L2 learners' engagement with automated feedback: An eye-tracking study. *Language Learning & Technology* 26(2): 78–105.

Murakami A and Alexopoulou T (2016) L1 influence on the acquisition order of English grammatical morphemes. *Studies in Second Language Acquisition* 38(3): 365–401.

Reynolds BL, Kao CW and Huang Y (2021) Investigating the effects of perceived feedback source on second language writing performance: A quasi-experimental study. *Asia-Pacific Education Researcher* 30(6): 585–595.

Robertson D (2000) Variability in the use of the English article system by Chinese learners of English. *Second Language Research* 16(2): 135–172.

Shintani N and Ellis R (2013) The comparative effect of metalinguistic explanation and direct written corrective feedback on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing* 22(3): 286–306.

Shintani N, Ellis R and Suzuki W (2014) Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning* 64(1): 103–131.

Stevenson M (2016) A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition* 42: 1–16.

Warschauer M and Grimes D (2008) Automated writing assessment in the classroom. *Pedagogies: An International Journal* 3(1): 22–36.

Zhang Z and Hyland K (2018) Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing* 36: 90–102.