Regular Paper

The JALT CALL Journal ISSN 1832-4215
Vol. 18, No.3 Pages 444–467
https://doi.org/10.29140/jaltcall.v18n3.764

Castledown

Feedback precision and learners' responses: A study into ETS *Criterion* automated corrective feedback in EFL writing classrooms

Giang Thi Linh Hoang
Hue University of Foreign Languages, VIETNAM
htlgiang@hueuni.edu.vn

This study examines the implementation of *Criterion*, an automated writing evaluation system developed by ETS, as a source of diagnostic feedback on learners' linguistic performance in a Vietnamese EFL writing classroom. Thirty-eight second-year English majors had access to *Criterion* for a five-month period. Data include *Criterion* error tags on students' essays from multiple practice sessions, recorded think-aloud protocols as students engaged with the feedback for revisions, and first and revised drafts students submitted to *Criterion*. The main findings indicate *Criterion*'s satisfactory precision and capacity to trigger various engagement strategies among learners, but reservations remain due to students' modest response accuracy and lack of substantive revisions to their texts. Important implications for formative feedback practices in EFL writing classrooms and the adaptation of *Criterion*'s technical capacities are accordingly presented.

Keywords: *Criterion*, automated feedback, response accuracy, student engagement, revisions

1. Introduction

With instructional contexts being transformed by technological tools, the growing prominence of automated writing evaluation (AWE) programs in numerous teaching and learning contexts has gained research attention in the last two decades. From earlier studies which adopted a system-centric approach to evaluating the performance of AWE systems (i.e., their scoring mechanism and diagnostic feedback functions), this strand of research has diversified to include AWE impacts on students' learning in English as a second language (ESL) and foreign language (EFL) contexts. The focus in more recent studies

shifts towards students' perceptions of AWE feedback (e.g., Hoang & Kunnan, 2016; Li et al., 2017; Zhang, 2020), students' engagement with such feedback (e.g., Koltovskaia, 2020; Tian & Zhou, 2020; Zhang & Hyland, 2018), and their subsequent revisions (e.g., Chapelle et al., 2015; Guo et al., 2021; Lavolette et al., 2015).

An important part of the diagnostic feedback generated by AWE systems is their automated corrective feedback (ACF), whose impacts on students' writing accuracy in the short and long term have been investigated by only a handful of studies (e.g., Guo et al., 2021; Li et al., 2017). Given the central role of revision in the writing process (Woodworth & Barkaoui, 2020), revisions in response to automated corrective feedback merit greater attention and can be systematically examined using richer data including feedback accuracy, learners' cognitive and revision strategies to process the feedback, as well as their uptake and subsequent response accuracy. What remains under-explored in the existing body of AWE feedback research is the link between multiple data sources in longitudinal studies to provide evidence about the impact of ACF on students' learning. The current research aims to address this gap by triangulating many data sources to evaluate the impacts of the ACF generated by ETS *Criterion*, one of the most widely used AWE programs in L2 writing contexts, on the Vietnamese EFL learners' writing practice.

2. Previous literature

The following parts review relevant literature pertaining to the focus of the current research: automated feedback accuracy, the subsequent response accuracy, and student engagement with the feedback.

2.1. Feedback accuracy

Accuracy is the most well researched aspect of the automated corrective feedback generated by different AWE systems. ACF accuracy was initially investigated using two statistical indices, *precision* and *recall*. *Precision* is the number of cases which the AWE system and the human annotator agreed are errors divided by the total number of cases that the system flags as errors. Burstein et al. (2003) further explain that *precision* "is equal to the number of the system's hits divided by the total of its hits and false positives [i.e., the cases the system labels as errors but actually are not in the human judgment]" (p. 6). On the other hand, *recall* measures the rate of errors covered by the system compared to the total number of errors flagged by the human annotator. Later, researchers raise additional issues in AWE corrective feedback not yet considered by its developers when the feedback is put to use and validated as a means of classroom-based formative assessment. In a study conducted by Lavolette et al. (2015), the error codes from *Criterion* were evaluated both by considering whether a real error has been detected (correct error code) and judging if the code assigned to the tagged error is correct. In addition, Bai and Hu (2017) and

Ranalli et al. (2017) examined *Criterion*'s textual commentaries and suggested revisions for tagged errors to see if these components are correct.

The majority of classroom-based research focuses on the precision of automated corrective feedback. Among all the investigated AWE tools, *Criterion* was the most researched and also the one with best reported performances. Its documented precision is above 50% in all the studies reviewed, with most frequently found accuracy rates bunching around 75%. For example, Chodorow et al. (2010) found *Criterion* corrective feedback to be 90% precise. Dikli and Bleyle (2014) reported 72.7% precision for *Criterion* ACF, which is close to Lavolette et al.'s (2015) 75% precision. Lavolette et al. (2015) subdivided their coded categories into three types of error codes: correct, wrong, and no error (i.e., false positive) instead of reporting a summative precision statistic. Their findings include 75% correct, 14% wrong, and 11% no error of all the error tags they coded for their study. Another common finding is that accuracy varies across error types (e.g. Feng et al., 2016; Lavolette et al., 2015; Ranalli et al., 2017). For instance, Lavolette et al. (2015) showed that *Criterion* was very good at capitalization, missing comma, wrong word, and ill-formed verb errors with 85% plus precision. On the contrary, the system was worst at run-on sentences, wrong article, and spelling errors with more than half of the time being miscoded. Similarly, *Criterion* ACF was found to be between 71 and 77% precise when ten most common error types were considered in aggregate, with substantial variation across error types (Ranalli et al., 2017).

2.2. Students' response accuracy following ACF

Response accuracy indicates whether a student correctly addressed a tagged error (Guo et al., 2021). Li et al. (2015) reported students' improvements in linguistic accuracy, with statistically significant reductions in normed error rates for three out of four papers based on *Criterion* error reports of the first and revised drafts of student essays. Similarly, Li et al. (2017) documented that the use of *Criterion* feedback resulted in positive short-term effects for eight out of nine examined error types. Using *Criterion* error reports of the first and revised drafts, recent research conducted by Saricaoglu and Bilki (2021) reported significant reductions in grammatical and usage errors (e.g., garbled sentences, subject-verb agreement, missing articles, confused words, proofread this) as well as mechanics errors (e.g., missing/extra comma, spelling). When all error types are considered in aggregate, Ranalli et al. (2017) found their participants to use *Criterion* corrective feedback to successfully correct errors 55-65% of the time, very close to findings in Koltovskaia's (2020) case study which also showed a 57% error correction rate among two ESL learners using Grammarly.

2.3. Students' engagement with automated corrective feedback

In previous research, student engagement with written feedback has been investigated using the triad of cognitive, behavioral, and affective engagement (Ellis, 2010). Of these three dimensions, probably the most thoroughly

researched in studies on learners' response to written corrective feedback is the cognitive dimension which relates to depth of processing, indicated by the level of learners' noticing of the written corrective feedback. The construct originates from Schmidt's (1993, 1995) noticing hypothesis which refers to the role of 'noticing' as a pre-requisite for the acquisition of grammatical features of a language. In examining students' uptake of feedback for revisions and how that relates to their response accuracy, Qi and Lapkin (2001) operationalized the quality/depth of feedback noticing as "perfunctory" (i.e., noticing only and without giving reasons) or "substantive" (i.e., noticing and providing reasons). Qi and Lapkin (2001) found that substantive noticing is highly conducive to correct revisions. In some other studies, cognitive and meta-cognitive strategies were investigated (e.g., Ellis, 2010; Han & Hyland, 2015; Tian & Zhou, 2020). Examples of cognitive strategies include making a mental note, memorization, and visualization, while meta-cognitive strategies can be used to regulate their engagement with the feedback such as evaluating the written corrective feedback, monitoring their use of the problematic forms, or reasoning (Han & Hyland, 2015).

What emerges from the literature is that cognitive and behavioral dimensions tend to overlap when students engage with the feedback and that students tend to exercise both cognitive and behavioral engagement strategies in each revision episode (i.e., a segment in the data when a student revises an error related to one feedback point). In the current research, through the analyses of *Criterion* error tags on learners' essays, comparison of their first and revised drafts, and think-aloud protocol recordings of students' revision processes, cognitive and behavioral dimensions in Ellis' (2010) triad engagement model are subsumed under the term engagement, which is further coded as substantive or perfunctory engagement, to see how EFL learners process *Criterion* ACF.

Most research on student engagement with AWE feedback to date adopted a multiple-case study approach. Zhang and Hyland (2018) and Zhang (2020) found that the more proficient students showed keen affective and behavioral responses to the feedback. In Zhang's (2020) study, the lower proficiency level learners were found to make primarily surface-level changes in response to form-focused feedback, while the more proficient case attended to both language and content in his revisions, resulting in changes beyond the sentence levels. Similarly, Koltovskaia (2020) found proficiency to be an important variable, with the proficient learner engaging more deeply with the feedback and making more successful revisions. The author highlights the need for accurate feedback to facilitate students' effective behavioural engagement.

Previous research has produced variable response accuracy rates for different error types, suggesting a relationship between student revision outcomes and error categories. The reviewed studies in this strand of research tend to exclusively focus on quantitative analyses of students' error corrections, but they failed (a) to pinpoint possible impacts of feedback precision on students' response accuracy and (b) to relate such revision outcomes to students' processing of the feedback and their subsequent revision actions. In other words,

studies are lacking detailed accounts of students' revising operations and the process leading up to their decisions on revisions (Link et al., 2020; Tian & Zhou, 2020). The current study addresses these gaps by investigating *Criterion* ACF along the three dimensions of feedback precision, student engagement (cognitive and behavioural) with the feedback and its subsequent impacts on response accuracy in revised texts. Findings will contribute to a systematic assessment of AWE corrective feedback for formative assessment in EFL writing classes. Four research questions guided this study.

1. What is the precision of *Criterion* ACF?
2. What is EFL tertiary learners' response accuracy following the use of *Criterion* ACF?
3. What is the correlation between feedback precision and learners' response accuracy?
4. How do EFL tertiary learners engage with *Criterion* ACF?

3. Methods

3.1. Participants

The study had a total of 38 participants who were second-year English majors at a university in central Vietnam. The vast majority of participants were female (36/38), reflecting a typical gender distribution in a lot of English departments across universities in Vietnam. Learners' age ranged from 19 to 20, with their average time of learning English being 9.6 years. They were all sophomore English majors taking the academic writing course in which students wrote problem solving, opinion, and advantage/disadvantage essays. Students' practice essays used as data for this study are of these three types. Prior to participating in this study, all the learners took a diagnostic timed writing test. Test essay length statistics revealed some outliers where certain students produced very short essays (less than two standard deviations from the group mean of 218 words). These students were excluded from the study. Analyses of writing accuracy of the 38 students eligible for data collection were conducted using a weighted clause ratio approach (see Foster & Wigglesworth, 2015). Findings show that the students' accuracy scores on the diagnostic test ranged from 0.65 to 0.97. Overall, the whole group shows decent performance with their EFL writing accuracy (M = 0.74, SD = 0.18).

Among these students, 14 were invited to take part in the think-aloud procedures which recorded students' verbalized thoughts as they worked on revisions to their essays after receiving *Criterion* ACF on grammar, usage, and mechanics. To make sure this sub-sample is representative of the whole group, selected students' accuracy scores on the diagnostic test ranged from 0.65 to 0.95, with four students scoring above 90% accurate, five scoring from 75% to 90%, and five scoring below 75%.

3.2. Data collection instruments

Data collection spanned a whole semester of 15 weeks. The class met once a week, with each session lasting 100 minutes. During the first session, the researcher spent about 20 minutes introducing *Criterion* to the students to train them on how to log into the system, compose and submit essays for assignments. They were also shown how to receive automated feedback, revise essays and resubmit them to *Criterion*. Email addresses of students enrolled in the course were then collected to create individual accounts on *Criterion*. This was followed by a homework task which required students to log into their accounts for a self-practice session on *Criterion*. The researcher came in during the second meeting to resolve problems students encountered as they started using the system. Other technical questions were answered during the first in-class practice session in the computer lab. The researcher was present during the remaining practice sessions to offer prompt technical help with *Criterion*. In general, three instruments were used to assemble data, including *Criterion* automated corrective feedback, students' submitted practice essays on *Criterion* (first and revised drafts), and 14 students' think-aloud protocols (TAPs) while revising first drafts.

3.2.1. *Criterion* automated corrective feedback. The automated corrective feedback generated by ETS *Criterion* was incorporated as part of the writing course for English majors who were taking their fourth semester in a four-year B.A. program. In this course, *Criterion* was intended as a supplementary diagnostic assessment tool alongside teacher feedback and each student had a *Criterion* account which allowed 24/7 access to the system. The current research used *Criterion* automated corrective feedback, but suppressed the system's feedback on organization, idea development, and style. *Criterion*'s corrective feedback is subdivided into feedback on (1) grammatical errors, (2) word usage errors, and (3) errors in writing mechanics. At the time of data collection, *Criterion* tagged errors in a total of 31 categories. For each error identified by *Criterion*, the specific word or phrase is highlighted. By dragging the cursor to this highlighted word/phrase, students can read *Criterion* explanations of the errors in a pop-up screen. Based on the feedback, students can choose to revise their essay by clicking on "Revise", which initiates a split screen with the right half being an interactive section for keying in corrections, as in Figure 1.

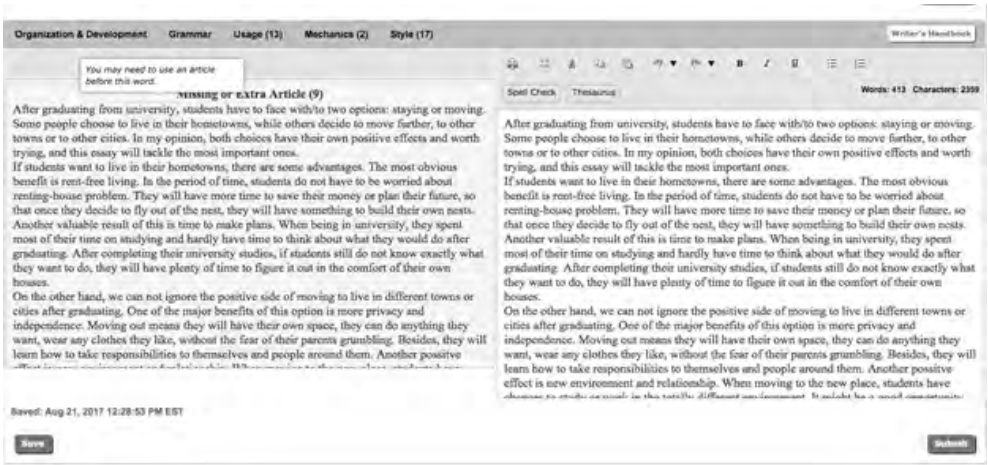

Figure 1. *Criterion*'s split screen for revision activities.

The narrow focus on corrective feedback was founded on learners' more favorable perceptions of the AWE form-focused feedback than the generic and less useful feedback related to style, content, and organization reported in previous research (e.g., Dikli, 2006; Hoang & Kunnan, 2016; Liu & Kunnan, 2016; Zhang, 2020), or empirical evidence regarding limited efficiency of AWE feedback on learners' development of higher-level writing skills (e.g., Link et al., 2020). Students could access *Criterion* corrective feedback for five months during which they had three compulsory practice sessions on the system. In each session, they composed and submitted essays to *Criterion* to receive its corrective feedback before making revisions to their writing. A total of 152 first draft essays from the homework and three in-class practice sessions, comprised of 43548 words, were collected for feedback accuracy analyses. Of all the first drafts, 2774 sentences were extracted, each including at least one error tag from *Criterion.*

3.2.2. Students' practice essay corpus. Three *Criterion* pre-loaded prompts were selected for the in-class practice sessions, including one problem-solving, one opinion, and one advantage/disadvantage essay prompts which align with the curriculum of the writing course. The practice essays collected for response accuracy analyses include a total of 228 essays (114 first and 114 revised drafts written by 38 participating students), totalling at 82130 words.

3.2.3. Students' think-aloud protocols. The students participating in the TAPs attended a demonstration session to be trained on the think-aloud procedures. On the days of in-class practice sessions, their think-aloud protocols were conducted in either English or Vietnamese depending on their personal choice. They were given a maximum of 35 minutes to revise their essays. During this time duration, students verbalized thoughts as they were revising their first drafts using *Criterion* ACF. Both students' verbalizations of thoughts and their on-screen operations as they went through the revisions were recorded using the free software OBS to provide data about their engagement with *Criterion*

ACF. All of the 14 think-aloud recordings were transcribed, checked for accuracy, and fed into NVivo for coding.

3.3. Data Analyses

3.3.1. *Criterion* feedback. Prior to coding students' response accuracy, *Criterion*'s error tags in the first drafts of the practice essays were coded for their *precision* following the verification approach used by Gamon et al. (2008, as cited in Leacock et al., 2014). This approach is "a method of simply checking the acceptability of a system's error flags or corrections compared to what the learner wrote". The choice of the verification approach was suitable for the purpose of the current study as it only determines whether the error tags generated by the system are correct or incorrect while not including the estimation of the number of errors which have not been detected by the system. With the focus of the study on learners' engagement with and subsequent revisions in response to the feedback they received from *Criterion*, the question of whether *Criterion* has missed certain errors became irrelevant. For inter-coding reliability, the author worked alongside a second coder who is a PhD candidate with 12 years' teaching EFL academic writing at the tertiary level. Both coders annotated about 10% of *Criterion* feedback points in all the first drafts of participating students' practice essays.

Adapting Lavolette et al.'s (2015) categories in coding the correctness of *Criterion*'s feedback, each feedback point was categorized as correct code (CC) to indicate that an error was correctly identified by *Criterion*, incorrect code (IC) to indicate cases when *Criterion* appropriately coded a structure as incorrect but gave it a wrong code (i.e., incorrect tagging) or when *Criterion* correctly labelled the error type but provided a wrong error message or suggestion for revision (i.e., incorrect suggestion). The third category is False positive (FP) which indicates a false alarm generated by *Criterion* when it flags a correct structure as erroneous. Precision equals the total *CCs* divided by the sum of *CCs, ICs*, and *FPs*. See Appendix A for more details about the coding scheme. Inter-coder reliability for this part of coding was 89 percent agreement. Disagreements were then discussed, resolved, and final decisions were applied to the rest of the data. Error tags were coded by error type.

3.3.2. Students' response accuracy. Students' first and revised drafts during three in-class practice sessions were used for the analyses of their response accuracy. Modifying coding schemes related to revision operations and success of revisions in earlier studies (Chapelle et al., 2015; Zhang, 2020) to fit the data of the current research, response accuracy analyses were based on the revision outcome using four coding categories: *correct revision*, *incorrect revision*, *avoidance*, and *retention of the correct form*. The first two categories indicate uptake of *Criterion* ACF while the third one, *avoidance*, includes non-uptake cases of no revision in response to correct error tags or removal of sections containing tagged errors in revised drafts. Instances where students chose not to make changes to their texts after receiving false positives from *Criterion* were coded

under the fourth category, *retention of the correct form*. Table 1 provides specific examples for coding response accuracy.

For inter-coder reliability, the second coder, the PhD candidate, was familiarized with the list of codes for response accuracy. The researcher and the second coder both coded a sample of about 15% of the total revision points (i.e., each point is comprised of the first and revised texts in response to a *Criterion* error tag). Inter-coder reliability was 93% agreement. Disagreements were discussed and resolved before the remaining revision points were coded.

Table 1. Examples of coded categories for response accuracy

First draft	Revised draft	Response accuracy
Everybody **should conscious** to protect the environment to have a fresh atmosphere. *[Ill-formed verb]*	Everybody should protect the environment to have a fresh atmosphere.	Correct revision
It is a good chance for **their** to gain more experiences than they had. *[Confused words]*	It is a good chance for they to gain more experiences than they had	Incorrect revision
What is more, you will make the most of your youth. *[Missing question mark]*	What is more, you will make the most of your youth.	Retention of the correct form
I **can't not** be denied that there are advantages of changing job. *[Negation error]*	It can't not be denied that there are advantages of changing job.	Avoidance *(No revision to a correct error code)*
Moreover, **another** benefits 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. *[Determiner-noun agreement]*	Moreover, changing career helps them gain expertise in a new area and make them have more opportunities in the future.	Avoidance *(The part of sentence containing the error tag has been removed)*

Note: The sections of text highlighted by *Criterion* as erroneous are marked in bold.

3.3.3. Correlation between feedback precision and response accuracy. Statistical assumptions were checked before the correlation between feedback precision and response accuracy was examined. The normality of all variables was assessed using the Shapiro-Wilk test. Avoidance and correct revision rates are normally distributed, $W(23) = 0.94$, $p = .172$ and $W(23) = 0.96$, $p = .516$, respectively. However, the Shapiro-Wilk test indicates that all other variables do not have normal distribution, $W(23) = 0.82$, $p < .001$ for correct error code rates, $W(23) = 0.65$, $p < .001$ for false positive rates, $W(23) = 0.54$, $p < .001$ for retention of the correct form rates, and $W(23) = 0.87$, $p = .005$ for incorrect revision rates. Therefore, the non-parametric Spearman rank test of correlation was employed for analyses regarding the relationship between feedback precision and learners' response accuracy.

3.3.4. Students' engagement with *Criterion* feedback. The 14 TAPs were used as data for analysing students' engagement with the automated corrective

feedback. Previous literature has showed different interpretations of engagement with written feedback in the process of writing and revision. Theoretically and data driven, the current research adopts a hybrid approach to examining EFL learners' engagement with *Criterion* automated corrective feedback where cognitive and behavioral dimensions are subsumed under the term of engagement, with the cognitive perspective denoting "how learners attend to the corrective feedback", and the behavioral perspective referring to "whether and in what way learners … revise their written texts" (Ellis, 2010, p. 342).

TAP data were coded by each revision episode where a student processed *Criterion* corrective feedback on one error flagged by *Criterion*. Adapting operationalisations of students' engagement with feedback in previous research (Qi & Lapkin, 2001; Sachs & Polio, 2007; Storch & Wigglesworth, 2010), each revision episode (interchangeably called engagement episode in this study) is coded for the quality of engagement depending on the specific strategies (cognitive/meta-cognitive or revising strategies) adopted to process the feedback. An episode is coded "substantive engagement" if the student employed one or more cognitive/metacognitive or revision strategies to extensively engage with the feedback before deciding on a revised form. Such strategies include *drawing on stored metalinguistic knowledge*, getting cues from *Criterion* error messages, *looking up online resources*, consulting a peer or the teacher, and *translating into L1*. On the other hand, a revision episode is coded "perfunctory engagement" if one or more of the following revision actions took place, *adopting Criterion's suggested corrections without elaboration, guessing the correct form, using error tags as a trigger for self-editing but ignoring the error explanations*, or the student briefly mentioned the *Criterion* tag without attending to the error. Double coding was conducted by the researcher and the PhD candidate on 20% of total engagement episodes, which produced inter-coder agreement of 90%. Disagreements were discussed and resolved, and the final decisions were applied to the rest of the data.

4. Results

4.1. Feedback accuracy

A total of 3074 *Criterion* error tags were extracted from all of the first draft essays submitted to *Criterion* by participating students across the four practice sessions (one homework and three in-class sessions). In this corpus of student essays, *Criterion* generated automated flaggings on 24 error types. The verification of these error tags revealed that *Criterion* was satisfactorily precise in detecting errors for learners to attend to issues in their essays. Specifically, 82.2 percent of all error flags were correct codes, 11.7% were false positives, and the remaining 6.1% were incorrect error codes.

Figure 2 illustrates the breakdown of the frequencies of correct error codes, incorrect error codes, and false positives for the ten most frequent error types in the corpus of learner essays in the current research. See Appendix B for further details.

Figure 2. Frequencies for the precision of *Criterion* error flaggings.

Taking the widely accepted threshold of 80% precision for being a useful system (Quinlan et al., 2009), *Criterion* was below expectations in terms of fragments, preposition errors, confused words, extra comma, and particularly below 50% precise with missing comma errors. A lot of false positive tags were found in this learner corpus in terms of missing commas. The following example relates to a false alarm in missing comma where *Criterion* mistakes two dependent clauses for two independent clauses, resulting in its wrong diagnosis of a missing comma after "entertainment".

Student text: *Therefore, they not only cover the cost of living and entertainment[1] but also get more working experience.*

Error message: *[1]You may need to use a comma after this word.*

Also very commonly found was *Criterion*'s wrong suggestion of inserting a comma before restrictive relative clauses, again due to its failure to distinguish between coordinating and subordinating relationship between clauses.

Student text: *They have to do works[1] they don't have passion for.*

Error message: *You may need to use a comma after this word.*

Other error types with high false alarm rates are preposition errors, confused words, and fragments, with recorded rates all exceeding 15% of the total error tags in each category. Spelling, the most frequently recurring error type for learners, has 23 false positives, most of which relate to proper nouns not recognized by *Criterion*, as in:

Student text: *After a long time changing jobs, she has identified her own passion which is trading goods made from "Moringa[1] oleifera[1]".*

Error message: *This word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.*

Extra commas and ill-formed verbs are the two error types with the highest rates of incorrect error codes, at 24% and 15%, respectively. Further examination of incorrect codes related to ill-formed verbs reveals *Criterion*'s failure to correctly identify the part of speech. An example of this is provided below,

Student text: We should[1] propaganda[1] to community about effect of water pollution and its effect of health.

Error message: [1]You may have used the wrong form of this verb. Match the subject to the verb to decide whether you have used the verb correctly.

4.2. Response accuracy

Overall response accuracy rate for participating students was 54%, with the success rate being defined as the number of students' correct revisions out of the total feedback points they received. *Avoidance* accounted for 28%, while 10.5% of all learners' revisions were incorrect, leaving 7.5% for retention of originally correct forms in response to *Criterion*'s false positives.

Revision success rates varied across error types. Table 2 presents the raw counts and percentages of students' response accuracy in the ten most frequently found error types in terms of the four response categories: correct revision, incorrect revision, avoidance, and retention of the correct form.

Table 2. Response accuracy for 10 most frequently recurring error types in student essays

Error type	*n*	Correct Revision	Incorrect Revision	Avoidance	Retention of the correct form
Spelling	900	617 (68.5%)	50 (5.6%)	210 (23.3%)	23 (2.6%)
Missing or extra article	749	350 (46.7%)	90 (12%)	246 (32.9%)	63 (8.4%)
Subject verb agreement	262	174 (66.4%)	33 (12.6%)	53 (20.2%)	2 (0.8%)
Missing comma	187	66 (35.3%)	12 (6.4%)	27 (14.4%)	82 (43.9%)
Capitalization	170	95 (55.9%)	2 (1.2%)	72 (42.4%)	1 (0.6%)
Determiner–noun agreement	164	91 (55.5%)	17 (10.4%)	56 (34.1%)	0 (0%)
Ill-formed verbs	120	63 (52.5%)	23 (19.2%)	33 (27.5%)	1 (0.8%)
Confused words	116	53 (45.7%)	16 (13.8%)	39 (33.6%)	8 (6.9%)
Preposition errors	99	36 (36.4%)	13 (13.1%)	33 (33.3%)	17 (17.2%)
Extra comma	71	37 (43.7%)	7 (9.9%)	19 (26.8%)	14 (19.7%)

Judged by response accuracy rates, students were most successful at correcting errors related to spelling (68.4%), subject-verb agreement (66.4%), capitalization (55.9%), determiner-noun agreement (59%), and ill-formed verbs (52%). However, students were least likely to successfully correct errors related to

fragments, missing commas, and prepositions, all with less than 40% response accuracy rates. Most noticeably, avoidance strategies (i.e., either ignoring the correct error tags or deleting the section of texts containing the flagged errors) were frequently employed across most error types. Capitalization and fragments topped the list with more than 40% of the error flags being not attended to. Compared to the much lower rates of incorrect revisions across all error types, high avoidance rates suggest students' preference for non-uptake of *Criterion* feedback over adoption of suggested revisions they were not sure about.

4.3. Correlation between feedback precision and response accuracy

Spearman's rank correlation was computed to test the hypothesis that there is a positive relationship between feedback precision (the rate of correct error codes) and students' corresponding response accuracy rates for 24 error types. As seen in Table 3, the hypothesis was confirmed, with a fairly strong positive correlation between these two variables, $r(22) = .76$, $p < .001$. This suggests that for error types where *Criterion* had higher precision rates, students were also more likely to revise their errors successfully.

Table 3. Correlation (Spearman's rho) between correct error code rates and correct revision rates

Variable	*n*	*M*	*SD*	Correct error code
1. Correct error code	24	78.8	24.46	–
2. Correct revision	24	46.9	16.8	0.756 $p < .001$

Spearman's rank correlation was computed to test the hypothesis that there is also a positive relationship between the rates of false positives for different error categories and students' corresponding rates of retention of the correct form. Table 4 shows that the hypothesis was confirmed, with a strong positive correlation between these two variables, $r(22) = .83$, $p < .001$. The result indicates that students were able to respond appropriately by not adopting most of the wrong suggested changes.

Table 4. Correlation (Spearman's rho) between false positive rates and retention of the correct form rates

Variable	*n*	*M*	*SD*	False positive
1. False positive	24	12.5	19.6	_
2. Retention of the correct form	24	9.4	19.5	0.833 $p < .001$

To examine whether there is any relationship between the rates of incorrect

error codes and students' response accuracy, Spearman's rank correlation was conducted between incorrect error code rates and incorrect revision rates, as well as with avoidance rates. Table 5 indicates a positive correlation between incorrect error code rates and avoidance rates, $r(22) = .49$, $p = .015$, but no statistically significant correlation between incorrect error rates and incorrect revision rates, $r(22) = -.08$, $p = .720$. The findings suggest a strong relationship between the rate of incorrect error codes and that of instances when students avoid correcting their errors. In other words, students tend to avoid making changes to their texts in response to *Criterion*'s incorrect error codes.

Table 5. Correlations (Spearman's rho) between incorrect error code rates and student responses

Variable	*n*	*M*	*SD*	Incorrect error code
1. Incorrect error code	24	8.8	13.3	_
2. Incorrect revision	24	16.1	15.7	−0.077 *p* = .720
3. Avoidance	24	29.3	13.7	0.491 *p* = .015

4.4. Students' engagement with Criterion ACF

Analyses of 14 students' think-aloud protocols show that all of the 270 error tags from *Criterion* were processed by the students. Among these, 183 engagement episodes (68%) were perfunctory while the remaining 87 episodes (32%) were substantive. Table 6 presents the TAP results in more detail.

Table 6. Coded engagement episodes in students' TAPs

Engagement level	Engagement strategy	Coded episodes *N* (% of all episodes)	No. of students involved*
Perfunctory engagement 183 (68%)	Using error tags as a trigger for self-editing but ignoring the error explanations	98 (36%)	14
	Noticing without attending to the error	35 (13%)	11
	Guessing the correct form	27 (10%)	10
	Adopting *Criterion*'s suggested corrections without elaboration	23 (8%)	9
Substantive engagement 87 (32%)	Looking up online resources	37 (14%)	9
	Drawing on stored metalinguistic knowledge	28 (10%)	10
	Getting cues from *Criterion* error explanations	10 (4%)	4
	Translating into L1	10 (4%)	4
	Consulting a peer or the teacher	2 (1%)	2
Total		**270 (100%)**	

*The total TAP students were 14, and each of them had different engagement strategies when processing ACF

As can be seen from Table 6, more than one-third of revisions were quickly executed using *Criterion highlighted error flags as a trigger for self-editing*. In these engagement episodes, students did not read or refer to *Criterion* metalinguistic explanations. Instead, simply looking at the highlighted error was enough for them to promptly revise their essay. Other forms of limited engagement with ACF, *guessing the correct form*, *adopting Criterion's suggested corrections without elaboration*, or *noticing without attending to the error*, were also commonly found instances in the data, as in the following revision episode on a missing comma error where the learner quickly adopted *Criterion*'s suggested correction:

Student text:	*Thus[1] some people firmly believe that it is a measure of success.*
Error message:	*You may need to use a comma after this word.*

On seeing *Criterion*'s highlighting of the word "Thus" and reading *Criterion* message, the student verbalized, "I miss some commas. *Thus*, OK I will put an extra comma here. *Thus*, comma, *some people firmly believe*…OK." This resulted in the revised sentence "*Thus, some people firmly believe that it is a measure of success*".

Turning to substantive engagement episodes, the most frequently used extensive revising strategy was looking up online resources which was recorded in 37 episodes. This is not a surprising result as *Criterion* is a web-based learning program, which allows for easy access to online resources. In the following episode, the student consulted Oxford online dictionary to double-check a false alarm from *Criterion*:

Student text:	*The main source of these [1]pollutive things is daily activities of residents living near by water source...*
Error message:	*The word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.*
TAP excerpt:	Ahh *this main source of these pollutive things. Pollutive* [checked Oxford online dictionary] *Pollutive, pollutive.* I think it's correct.

The second most common extensive engagement strategy with 28 coded episodes was *drawing on stored metalinguistic knowledge*. During these episodes, engagement with *Criterion* error feedback positively triggered learners' stored knowledge as they searched for a solution to flagged errors, as below:

Student text:	*But the problem is, how each person[1] define[1] the word "success" for their own.*
Error message:	*[1]The subject and the verb in this sentence may not agree. Reread the sentence and look closely at the subject and the verb.*

TAP excerpt: Subject-verb agreement. What's the problem? *But the problem is, how each person define the word "success" for their own.* I meant "nhưng vấn đề ở đây là làm sao mỗi người có thể định nghĩa được từ thành công cho chính họ" [Translation of his writing into Vietnamese]. For *each person. How each person. Each person. Define.* Why? Ahh, *how each person, each person* is a third person singular. *Define* should have an "s". OK. I will add "s". *Each person* is a third person singular pronoun, so there is an "s" after the verb that follows.

What emerges from the TAP analyses is that *Criterion*'s incorrect error codes and false positives were more likely to trigger substantive than perfunctory engagement among students, as in the following excerpt,

Student text: *In addition, you remarkably increase your[1] earning power in another company which appreciates your ability and strengths.*

Error message: *[1]You have used **your** in this sentence. You may need to use you're instead.*

TAP excerpt: *Increase…you have used **your** in this sentence. You may need to use **you're** instead…* I don't think I need to change this word because *your* is an adjective. *Earning power* is a noun and *your* is a possessive. I don't think I need to change this word.

In response to false positives or incorrect error codes, students most frequently resorted to the two follow-up strategies to confirm their doubts, *reflecting on stored metalinguistic knowledge* and *looking up online resources.* In each engagement episode, the error tags from *Criterion* initiated some form of self-regulation of one's writing and revision processes. Feedback evaluation was part of the revision process, and learners in the current research tended to exercise caution and feedback evaluation through extensive engagement strategies as they worked on their revisions using automated feedback.

5. Discussion

The current research is a combined system-centric and user-centric enquiry into the use of *Criterion* automated corrective feedback for formative assessment purposes in EFL writing classrooms. It examined the precision of *Criterion* ACF, and how EFL tertiary learners made use of such feedback through their response accuracy and engagement with the feedback.

With 82.2% correct error codes, the overall precision of *Criterion* ACF is slightly higher than findings in earlier research where *Criterion* was also used for classroom-based assessment. Variable findings across studies are explicable,

given the different learner groups and the number of error tags being verified in each study. However, what the findings highlight is *Criterion*'s variable performance levels across error types, indicating the system's inflexible treatments of certain error types beyond lexical levels (e.g., fragments, comma errors) or its failure to recognize proper nouns as part of its spelling error detection.

The current research also adds to previous literature (e.g., Koltovskaia, 2020; Ranalli, 2021) by providing empirical evidence highlighting the correlation between feedback precision and response accuracy. Notably, the findings highlight students' appropriate responses to false positives and incorrect error codes from *Criterion*. Across different error categories, students consistently retained texts in response to false alarms and maintained a precautious approach to dealing with incorrect taggings. Such findings corroborate previous research which shows students' disregard of inaccurate suggestions from AWE systems (e.g., Chapelle et al., 2015; Chodorow et al., 2010; Lavolette et al., 2015; Link et al., 2020). Qualitative findings further revealed that some students verbalised their resistance to making changes during engagement episodes with false alarms, which points to the issue of trust raised in Ranalli's (2021) study. On a positive note, distrust triggers students' evaluation of the provided feedback rather than their unquestioning adoption of the suggested changes. This echoes Bai and Hu's (2017) findings on Chinese EFL counterparts who were "selective in their utilization of AWE feedback and able to adjust their uptake of AWE suggestions according to the accuracy of the feedback" (p. 67). Being English majors, sufficient proficiency levels may have added to the learners' feedback literacy, as demonstrated in their capacities to evaluate the feedback, seek external support, and regulate their cognitive processes in responding to the feedback (Yu & Liu, 2021).

It is also worth stressing *Criterion*'s capacity to draw learners' attention to targeted linguistic forms in their essays with all the feedback being noticed and processed either perfunctorily or substantively. Strategic learners could potentially make use of the error tags and *Criterion* revision platforms to enhance self-regulation, as in-built features on this system and similar AWE tools were found to generally promote learner autonomy (Stevenson, 2016). Similar to the perceptions among learners in Li et al.'s (2015) study, participants in this research referred to external websites for example sentences containing relevant words or looked up grammatical rules related to an error being processed, which exemplifies the positive effects of *Criterion* ACF on the development of self-regulatory revision strategies and increased grammatical awareness. Learners' attention was accordingly drawn to certain gaps in their interlanguage development, which potentially facilitates L2 acquisition (Heift & Hegelheimer, 2017).

However, students' engagement episodes mostly indicate the superficial nature of the errors being flagged by *Criterion*, as reported in earlier research (Chen & Cheng, 2006; Dikli & Bleyle, 2014; Li et al., 2015; Warschauer & Grimes, 2008). Accordingly, when revising essays on *Criterion* using the automated error tags, students in the current research tended to make superficial changes at

lexical and sentential levels such as inserting/deleting single words and punctuation marks, correcting a misspelled word, or adding plural and third person singular suffixes. The absence of substantial revisions related to content suggests that the system's over-emphasis on form-related issues may inhibit discourse level revisions. In addition, with learners' overall response accuracy of 54% following *Criterion* error tags and an additional 7.5% appropriate response to false positives (i.e., retention of correct forms), the result is close to Ranalli et al.'s (2017) 55-60% successful revision rate for 82 ESL learners in a US midwestern university. Using a more cautious approach to interpreting this statistic, Ranalli et al. (2017) adopted Manchón's (2011) distinction between *learning to write* (LW) and *writing to learn* (WL) as a baseline for assessing the value of an AWE program as a learning tool. In this distinction, a 55-60% correct revision rate could be considered insufficient support for revising practices among learners if the goal was LW which stresses writing skill development and better written products.

From a system-centric perspective, the data suggests some correlation between feedback precision and students' response accuracy. Additionally, students' choice of avoidance over adoption of suggested changes when processing incorrect automated error taggings indicates the need for sustained efforts towards precise feedback. *Criterion* developers' choice to err on the side of *precision* over *recall* (Burstein et al., 2003; Chodorow et al., 2010) is still highly relevant if learners' trust in the system's automated corrective feedback 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. With improved technical capacities, *Criterion* feedback can initiate more meaningful and substantive revisions among English language learners. From a user-centric perspective, writing instructors can help learners make the best use of AWE feedback by addressing the issues related to the system's low perceived authority and enhancing learners' feedback literacy. Supplementary oral feedback sessions during class hours can be provided so that learners can bring up clarification questions after they have engaged with *Criterion* ACF in the early stages of AWE implementation. Furthermore, strategy training sessions should be embedded when students can learn cognitive/metacognitive strategies for revisions or share reference sources they find most helpful to seek further information for the error codes received.

6. Limitations and recommendations for future research

The current study has some limitations to be acknowledged. Firstly, the sample size is quite limited, resulting in a modest essay corpus. Future studies can aim for larger samples that include every error category covered by *Criterion* to produce more generalizable findings regarding the system's performance on EFL essays. In addition, research on AWE feedback can trial different modes of automated feedback provision on specific learner populations. Students can be placed in different experimental groups, each with a different condition (e.g., total access to *Criterion* feedback including *grammar*, *usage*, *mechanics*,

style, *organisation* and *development*, form-focused feedback only, or focused feedback groups with access to one of the targeted language forms). Comparing the effects of these conditions can reveal more nuanced information about different ways to implement the use of automated feedback for target groups of learners. Secondly, the TAP data collected from 14 learners were examined in totality to investigate the whole group's engagement patterns rather than digging into how individual learner characteristics impact engagement and response accuracy. Future research can aim for richer case study data by factoring in individual learner characteristics such as proficiency levels, response accuracy rates, or feedback uptake and retention to provide deeper insight into the impact of AWE feedback on L2 writing. Finally, the suppressed automated discourse level feedback on content, organisation, and style in this research may have biased learners' engagement with the feedback to some extent. Therefore, a few comments and discussion related to learners' revision practices should be taken with precaution.

Acknowledgements

I would like to acknowledge the vital contribution of the Educational Testing Service for providing access to *Criterion* for this study.

References

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology, 37*(1), 67–81. http://dx.doi.org/10.1080/01443410.2016.1223275

Burstein, J., Chodorow, M., & Leacock, C. (2003). *CriterionSM* online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence.*

Chapelle, C. A., Cotos, E., & Lee, J. Y. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, *32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Chen, C. E., & Cheng, W. (2006, May). *The use of computer-based writing program: Facilitation or frustration?* [Paper presentation] The 23rd International Conference on English Teaching and Learning, Republic of China.

Chodorow, M., Gamon, M., & 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. https://doi.org/10.1177/0265532210364391

Dikli, S. (2006). *Automated essay scoring in an English as a second language setting* (Doctoral dissertation, Florida State University). Retrieved from http://etd.lib.fsu.edu/theses_1/available/etd-07052007-152924/unrestricted/sd_dissertation.pdf

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

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

Feng, H. H., Saricaoglu, A., & Chukharev-Hudilainen, E. (2016). Automated error detection for developing grammar proficiency of EFL learners. *CALICO Journal, 33*(1), 49–70. https://doi.org/10.1558/cj.v33i1.26507

Guo, Q., Feng, R., & Hua, Y. (2021). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*, 1–20. https://doi.org/10.1080/09588221.2021.1879161

Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of Second Language Writing*, *30*, 31–44. https://doi.org/10.1016/j.jslw.2015.08.002

Heift, T., & Hegelheimer, V. (2017). Computer-assisted corrective feedback and language learning. In H. Nassaji and E. Kartchava (Eds.), *Corrective feedback in second language teaching and learning: Research, theory, applications, implications* (pp. 51–65). New York: Routledge

Hoang, G. T. L., & Kunnan, A. (2016). Automated essay evaluation for English language learners: A case study of *MY Access*. *Language Assessment Quarterly, 13*(4), 359–376. https://doi.org/10.1080/15434303.2016.1230121

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by *Grammarly*: A multiple case study. *Assessing Writing, 44*, 1–12. https://doi.org/10.1016/j.asw.2020.100450

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology, 19*(2), 50–68. http://dx.doi.org/10125/44417

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detections for language learners* (2nd ed.). San Rafael, CA: Morgan and Claypool.

Li, Z., Feng, H., & 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. https://doi.org/10.1558/cj.26382

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 1–30*. https://doi.org/10.1080/09588221.2020.1743323

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *CALICO Journal, 33*(1), 71–91. https://doi.org/10.1558/cj.v33i1.26380

Manchón, R. M. (2011). Situating the learning-to-write and writing-to-learn dimensions of L2 writing. In R. M. Manchón (Ed.), *Learning-to-write and writing-to-learn in an additional language* (pp. 3–15). Philadelphia, PA: John Benjamins.

Qi, D. S., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing, 10*, 277–303. https://doi.org/10.1016/S1060-3743(01)00046-7

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology, 37*(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, 1–16. https://doi.org/10.1016/j.jslw.2021.100816

Sachs, R., & Polio, C. (2007). Learners' uses of two types of written feedback on a L2 writing revision task. *Studies in Second Language Acquisition, 29,* 67–100. https://doi.org/10.1017/S0272263107070039

Saricaoglu, A., & Bilki, Z. (2021). Voluntary use of automated writing evaluation by content course students. *ReCALL,* 1–13. https://doi.org/10.1017/S0958344021000021

Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics, 13*, 206–226.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu, HI: Second Language Teaching and Curriculum Center, University of Hawai'i.

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.compcom.2016.05.001

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing: Case studies. *Studies in Second Language Acquisition, 32*, 303–334. https://doi.org/10.1017/S0272263109990532

Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System, 91*. https://doi.org/10.1016/j.system.2020.102247

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22–36. https://doi.org/10.1080/15544800701771580

Woodworth, J., & Barkaoui, K. (2020). Perspectives on Using Automated Writing Evaluation Systems to Provide Written Corrective Feedback in the ESL Classroom. *TESL Canada Journal*, 37(2), 234–247.
Yu, S., & Liu, C. (2021). Improving student feedback literacy in academic writing: An evidence-based framework. *Assessing Writing, 48*. https://doi.org/10.1016/j.asw.2021.100525
Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing, 43*. https://doi.org/10.1016/j.asw.2019.100439
Zhang, Z., V. & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36*, 90–102. http://doi.org/10.1016/j.asw.2018.02.004

Appendix A

Verification of error codes generated by Criterion

The evaluation of *Criterion* feedback precision is conducted using the verification annotation approach. Three categories applied to this part of coding:

1. Correct error code (CC): cases when an error was correctly identified by *Criterion* in terms of both the error tag and the error message. The following examples illustrate *Criterion*'s correct codes.
Example 1:

Student text: Another benefits of staying at the same job for a long time is that you will have strong work relationship.

Criterion error tag: Determiner-noun agreement

Error message: You may have used the wrong determiner. Proofread the sentence to make sure that the determiner agrees with the word it modifies.

Example 2:

Student text: It may be difficult for they to form strong relationship that endure after they stop working.

Criterion error tag: Pronoun error

Error message: You may have used the wrong pronoun.

2. Incorrect error code (IC): cases when *Criterion* appropriately coded a structure as incorrect but gave it a wrong error tag, or when *Criterion* offered a confusing error message that failed to pinpoint the nature of the problem (i.e., wrong error message/suggestion for revision).
Example 1:

Student text: First, staying in the same career, it is easy for some people to working because it dose not waste much time to start with new jobs.

Criterion error tag: Subject-verb agreement (**Wrong tag**)

Error message: The subject and the verb in this sentence may not agree. Reread the sentence and look closely at the subject and the verb. (**Wrong error message**)

Example 2:

Student text: Besides, you also have to find a person you love to get happy from 8p.m to 6a.m next day'.

Criterion error tag: Possessive errors

Error message: You may need to take out the apostrophe to make this word a plural noun. (**Wrong suggestion for revision)**

3. False positive (FP): cases when the system created a false alarm by flagging a correct structure in essays as an error. Two examples of false positives are provided below:

Example 1

Student text: The prevalence of changing jobs has been a growing concern in the past few years.

Criterion error tag: Preposition errors

Error message: You may be using the wrong preposition.

Example 2

Student text: Actually, it leads to many bad things that affect to our environment and after that are our health.

Criterion error tag: Confused words

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

Appendix B

Error tag verification for 10 most frequent error types (n>50)

Error type	n	Correct error codes	Incorrect error codes	False positives
Spelling	900	877 (97.4%)	0 (0%)	23 (2.6%)
Missing or extra article	749	604 (80.6%)	56 (7.5%)	89 (11.9%)
Subject-verb agreement	262	240 (91.6%)	14 (5.3%)	8 (3.1%)
Missing comma	187	54 (28.9%)	10 (5.3%)	123 (65.8%)
Capitalization	170	151 (88.8%)	18 (10.6%)	1 (0.6%)
Determiner noun agreement	164	159 (97%)	5 (3%)	0 (0%)
Ill-formed verbs	120	100 (83%)	18 (15%)	2 (2%)
Confused words	116	87 (75%)	8 (7%)	21 (18%)
Preposition errors	99	70 (70.7%)	8 (8.1%)	21 (21.2%)
Extra comma	71	38 (53.5%)	17 (24%)	16 (22.5%)