Content list available at http://ijltr.urmia.ac.ir

Iranian Journal of Language Teaching Research

ORIGINAL ARTICLE

Urmia University

How Do EFL Learners Process and Uptake Criterion Automated Corrective Feedback? Insights from Two Case Studies

Giang Thi Linh Hoang [a], Neomy Storch [b, *]

[a] *Hue University, Vietnam*

[b] *The University of Melbourne, Australia*

ABSTRACT

Research has suggested that the type of feedback learners receive can impact on whether learners understand the feedback, the extent to which they engage with it, and whether they incorporate it in their revised drafts. However, to date, only a small number of studies have investigated learner engagement with corrective feedback provided by automated writing evaluation tools, and of those few have considered in greater depth the impact of the type of automated feedback on engagement. This multiple-case study examines two EFL learners' engagement with the two forms of corrective feedback provided by Criterion categorised as generic and specific and factors that can explain the nature of their engagement. Data were collected from learners' first and revised drafts of multiple essays on Criterion, screencasts of students' think-aloud procedures while revising essays, and stimulated recall interviews. Findings indicate the learners' higher uptake rate and more successful error corrections in response to generic versus specific feedback. However, their mental effort expenditure differed when cognitively engaging with the feedback, which could be explained in terms of individual learning goals, feedback quality, and the nature of tagged errors. These findings have relevant implications for utilising automated corrective feedback in L2 writing classes.

Keywords: automated feedback; learner engagement; specific feedback; generic feedback; case study

ARTICLE HISTORY

Received: 5 Sept. 2023 **Revised version received:** 28 Dec. 2023

Accepted: 10 Jan. 2024 **Available online:** 15 Feb. 2024

* Corresponding author: School of Languages & Linguistics, The University of Melbourne, Australia
Email address: neomys@unimelb.edu.au

10.30466/ijltr.2024.121420

Introduction

The efficacy of written corrective feedback (WCF) which targets learners' linguistic errors in second language (L2) writing continues to be an important topic of investigation in the field of Applied Linguistics. From a theoretical perspective, leading interactionist theories (e.g., Long, 1996; Swain, 1993) view corrective feedback as a source of external support which triggers learners' conscious awareness about a gap in their interlanguage development. For example, in her Output Hypothesis, while stressing the importance of output in helping learners notice gaps in their production, Swain (1993) acknowledged the limitations of output alone, pointing to the supplementary role of corrective feedback in helping learners go through mental processes to modify their output more effectively. Empirical research has also provided evidence which suggests that "student engagement with written corrective feedback facilitates language acquisition and writing development" (Zhang & Hyland, 2018, p. 91). Kang and Han's (2015) meta-analysis of 22 primary studies found that WCF had a moderate to large effect on L2 writers' grammatical accuracy and that higher proficiency students benefited more from such feedback than their beginning level counterparts. This body of research has provided convincing evidence to counter one of the most compelling arguments initiated by Truscott (1996) who pointed out problems that invalidate the use of corrective feedback: the lack of systematic and consistent approaches to delivering error correction among teachers as well as the ability and willingness to pay attention to the feedback among learners.

Despite the early controversy surrounding the use of WCF, research in the field has rapidly expanded from teacher WCF to automated corrective feedback (ACF) generated by automated writing evaluation (AWE) programs. ACF is becoming a popular source of feedback in L2 writing classrooms, at most levels of education and in diverse language learning contexts. Several AWE programs tend to share two functions: (a) a scoring function which assigns holistic ratings on content (e.g., *Write & Improve*) or general performance levels (*Criterion*'s 1-6 scale, *Write & Improve*'s either IELTS nine-band scale or the six-level rubric aligned to the Common European Framework of Reference (CEFR)); and (b) diagnostic feedback on linguistic aspects of L2 writing such as grammatical accuracy, lexical choices, and mechanics (i.e., ACF). ACF emulates to some extent the corrective feedback provided by writing teachers (Mehrabi-Yazdi, 2018). For example, *Grammarly* provides either indirect feedback (with the problematic text section underlined in red) or direct explicit feedback which includes an indication of the error type and error explanation, as well as suggested revision (Koltovskaia, 2020). Another program, *Write & Improve*, generates three types of diagnostic feedback which range in degree of explicitness. Specifically, direct word-level feedback provides the most explicit feedback (location, explanation, and suggested correction); indirect word-level feedback flags a word/phrase accompanied by a brief explanation; indirect sentence level feedback uses a colour scheme to indicate the quality of each sentence in terms of its accuracy (Liu & Yu, 2022).

Criterion also generates direct and indirect feedback, but this feedback can be categorised as either specific or generic (Ranalli et al., 2017). Specific feedback appears either as a recommendation to use another word to replace a highlighted word (e.g., "You have used **your** in this sentence. You may need to use **you're** instead") or a suggested revision for a highlighted portion of the student's text (e.g., "You may need to use an article before this word. Consider using the article **a**"). In other words, specific feedback is direct while generic feedback means that the same metalinguistic explanation is provided any time an instance of an error type is detected.

To date, only a handful of studies (e.g., Liu & Yu, 2022; Ranalli, 2018) have been conducted to examine learner engagement with and revisions following the different types of ACF provided by AWE tools. As research on WCF continues to produce inconclusive findings regarding the efficacy of direct versus indirect feedback (Kang & Han, 2015), more nuanced understanding of individual students' engagement with and their revisions following ACF of different explicitness levels is needed amidst the increasing use of ACF in various L2 settings. To this end, we sought to extend this body of research by investigating individual learner engagement with the specific and generic ACF provided by *Criterion*, an AWE tool developed by ETS. A better understanding of students' engagement with different types of ACF and the potential impact this may have on L2 learners' writing development is needed to inform pedagogical decisions in ESL and EFL classrooms regarding the use of AWE systems.

Literature review: Learner engagement with feedback provided by AWE tools

As Mao and Lee (2022, p. 788) argue, "engagement is central to connecting feedback provision with learning outcomes". Yet, a relatively small number of studies have investigated learner engagement with corrective feedback provided by AWE tools (e.g., Koltovskaia, 2020; Tian & Zhou, 2020; Zhang & Hyland, 2018). Most of them, as in the case of studies investigating learners' engagement with teacher WCF, have predominantly utilized or adapted the tri-partite framework proposed by Ellis (2010), which distinguishes between cognitive, behavioural, and affective dimensions of engagement (e.g., Koltovskaia, 2020; Tian & Zhou, 2020; Zhang, 2020; Zhang & Hyland, 2018). Later, in a study on learner engagement with both teacher and automated feedback, Zhang and Hyland (2018) elaborated on each dimension in this framework: Cognitive engagement describes the cognitive and metacognitive strategies learners use to process the feedback received; Behavioural engagement refers to learners' revision actions; Affective engagement encompasses learners' emotional and attitudinal responses to the feedback. Most early studies on learner engagement with ACF examine learners' behavioural and affective engagement, leaving cognitive engagement an under-researched area. Regarding behavioral engagement (uptake of feedback), research has produced conflicting findings. For example, some studies reported very low rates of uptake of corrective feedback (11.5% in Bai & Hu, 2017 using *Pigai*) whereas other studies report an uptake rate of about 50% (Chapelle et al., 2015 using *Criterion*; Dikli, 2006 using *My Access*). Lavolette et al. (2015) reported an even higher uptake rate for *Criterion* (73%).

Studies reporting on all dimensions of engagement have shown the complex nature of engagement. As in the case of teacher feedback (e.g., Farsani & Aghamohammadi, 2021; Han, 2019; Han & Hyland, 2015; Liu & Storch, 2023), a host of individual and contextual factors shape learners' engagement. For example, the study by Zhang and Hyland (2018), conducted in EFL classes in China, investigated learners' engagement with *Pigai*, an AWE tool used by millions of students in this country. Using a case study approach, the researchers found that although the teacher and *Pigai* differed in terms of the number of errors identified and type of feedback (*Pigai* was found to identify fewer errors and tended to provide indirect feedback), what distinguished the learners' engagement with these two sources of feedback were L2 proficiency, a repertoire of learning strategies, and beliefs about learning. The study found that the more proficient student engaged deeply and fully with both teacher and AWE feedback, while the less proficient showed very limited engagement. Zhang (2020), in a follow-up case study, also found proficiency to be key to successful learner engagement with *Pigai* feedback, but other context-related factors such as teacher attitude towards AWE were also found to be important.

In a similar vein, through the use of screencasts, stimulated recalls, and semi-structured interviews with two ESL learners, Koltovskaia (2020) found that learners generally displayed positive affective engagement with *Grammarly* ACF but their behavioral engagement was impacted by their perceived accuracy of the feedback. Only 57% of the feedback provided by *Grammarly* was taken up. Koltovskaia (2020) also found proficiency to be an important variable as the more proficient learner engaged more deeply with the feedback and made more successful revisions. More recently, Ranalli (2021) used screen capture recordings, stimulated recalls, and interviews with six

EFL learners who used *Grammarly* feedback in a multiple case study and found trust to be a key factor in learners' engagement with the feedback. However, another factor was the type of feedback provided. Because *Grammarly* provides mainly direct feedback, the learners adopted a proofreading rather than a learning orientation approach when engaging with the feedback.

Ranalli (2018) attempted to investigate learner engagement with feedback provided by *Criterion*, taking into consideration the amount of information the tool provides across error types and whether the feedback provided was accurate or not. The study was large scale, conducted with 82 students recruited from intermediate and upper intermediate ESL classes who responded to an ACF-based error correction task where feedback explicitness (i.e., generic versus specific) and accuracy were controlled. The finding indicated that generic feedback resulted in fewer successful error corrections than specific feedback. Furthermore, the participants perceived generic feedback as requiring more cognitive effort and as being less helpful. Interestingly, no significant differences were found between the lower and higher proficiency groups in terms of accuracy of error correction. However, the study used an error correction task rather than students responding to feedback from *Criterion*. Furthermore, Likert scale questions were used to elicit students' self-reported data on cognitive and affective engagement. As such, the findings may not reflect the nature of learners' engagement with the type of feedback provided by *Criterion* on their writing.

Liu and Yu (2022) investigated learner engagement with the direct and indirect feedback provided by a new AWE tool entitled *Write & Improve* on a sample of 24 Chinese L2 learners of English. The authors based their conceptualization of learner engagement on the model of feedback processing and usage proposed by Gass (1997), and thus the study represents an attempt to link engagement to a theory of second language learning (cognitive perspectives). The researchers propose three key interrelated elements of learner engagement with automated feedback, including attention allocation, cognitive effort expenditure, and revision response. Combining eye-tracking data, students' verbalizations of the cognitive efforts they spent processing each feedback point, as well as the type and quality of revisions students made to their texts, Liu and Yu (2022) found that their participants spent more time and expended more cognitive effort in processing indirect than direct feedback. However, a lower proportion of indirect feedback was taken up and fewer of the revisions based on indirect feedback were correct. The findings suggest that the nature of the feedback affects learner engagement with ACF, echoing the findings of studies on teacher feedback.

The review of the literature on learner engagement with ACF highlights the paucity of research on learner engagement with different types of automated corrective feedback provided by AWE programs. Except for Ranalli (2018) and Liu and Yu's (2022) studies, research on the degree of ACF explicitness as a mediating factor in individual learners' use of the feedback is scarce and clearly merits further investigation, particularly if AWE tools are to be used to supplement teacher feedback on students' writing in L2 writing classes. Ranalli's (2018) study with 82 ESL learners strictly controlled students' revisions using an error correction task, which could control for feedback explicitness as an independent variable but this may result in a very narrow view about how the learners would actually engage with the automated feedback in a real learning situation. Liu and Yu's (2022) study addressed this shortcoming by employing eye-tracking data to track learners' attention allocation to different areas of interest on the *Write & Improve* interface as they engaged with the feedback for revisions. However, the eye-tracking technology does not provide fine-grained data related to individual learners' processing of different ACF types. Addressing these gaps, we focused particularly on the impact of generic versus specific feedback on two individual learners' engagement with ACF via richer data sources: their multiple drafts of different essays composed on *Criterion*, screen-recorded think-aloud protocols, and stimulated recall interviews. This study took place in Vietnam where the use of AWE programs such as *Grammarly* and *Write & Improve* is rapidly gaining in popularity. However, to the best of our knowledge, this is an under-represented context in the automated feedback literature, with few studies conducted with Vietnamese L2 learners of English (e.g., Hoang, 2016; Hoang & Kunnan, 2022). The three research questions that guided this study are:

1. How much of *Criterion* generic and specific corrective feedback do learners take up and incorporate in their revised drafts?
2. How successfully do the learners correct errors in response to the generic and specific corrective feedback from *Criterion*?
3. How do individual learners cognitively engage with *Criterion* generic and specific corrective feedback?

Method

Context and participants

This multiple-case study was conducted with two students who were second-year English majors taking an EFL writing course in Vietnam. The multiple-case study is inspired by findings in previous research on teacher and automated feedback which indicated that different learners may perceive and benefit differently from teacher feedback (e.g., Han & Hyland, 2015; Kim & Bowles, 2019; Storch & Wigglesworth, 2010; Zheng & Yu, 2018) and automated feedback (e.g., Zhang & Hyland, 2018). Studies involving a few dozen students, however, often fail to reveal nuanced understanding about how learners actually engage with the feedback. On the contrary, case-study research is characterized by the focus on specific learners with richer data and the researcher plays the role of the gatherer of interpretations to make sense of the data and to construct knowledge (Creswell, 2013). In addition, the selection of two cases in this study relates to what Yin (2018) suggested, "the design of multiple-case studies follows an analogous logic where the choice of the cases predicts contrasting results but for anticipatable reasons" (p. 64). In this course, students learned to write different types of argumentative essays (see Appendix for the three specific writing prompts) and *Criterion* was incorporated as part of the in-class and home assignments. The study received clearance from the Ethics Committee of the university.

The participants were Trang and Nhien (pseudonyms). At the time of data collection, both students had fulfilled B2 level requirements according to the CEFR against which the program's learning outcomes for second-year English majors were aligned. Informed by previous research which indicates that learners of different proficiency levels tend to engage differently with automated feedback (e.g., Koltovskaia, 2020; Zhang, 2020; Zhang & Hyland, 2018) and to gain more nuanced insights into the effect of feedback types on learners' engagement while minimizing the additional effect of proficiency, the cases in this research were selected based on two criteria: (a) their comparable proficiency level (upper-intermediate), as indicated by their similar study results for the writing skills in the previous semester and *Criterion* score reports on their essays submitted to the system (all receiving either a 5 or 6 on *Criterion*'s 1-6 rating scale); and (b) their similar participation levels during class meetings where they were observed to be diligent and enthusiastic class members. Neither of them had any prior experience in using automated writing evaluation tools in writing.

Data collection tools

Each student attended three in-class sessions using *Criterion* which allowed them to submit multiple drafts of the same essay. Participants' essays (first and revised drafts) on *Criterion* were collected during in-class sessions in the fifth, eighth, and eleventh weeks of the semester.

In addition, each student had one screen-recorded think-aloud protocol (TAP) as they engaged with *Criterion* ACF during revision sessions. Students received two training sessions on the use of think-aloud protocols prior to the implementation of the study. The TAPs were recorded using the free software OBS (https://obsproject.com/) which captured both students' verbalizations and on-screen operations as they revised their essays. Given the option as to which session they wanted to record their think-aloud protocol, Trang chose to conduct her TAP during the second data collection session, while Nhien in the first session. Both Trang and Nhien verbalized mainly in Vietnamese, but they sometimes switched to English. Both learners focused on verbalizing revisions in response to form-focused feedback from *Criterion* on their linguistic errors. Trang spent a total of 25 minutes revising her essay. About ten minutes of this total time were devoted to processing the automated corrective feedback generated by *Criterion*, and the remaining time was spent on self-initiated revisions on content as well as general grammatical and lexical choices in her first draft. Similarly, Nhien spent a total of 23 minutes for the think-aloud procedure. Of these 23 minutes, she spent 17.5 minutes responding to *Criterion* ACF, and the remaining time on self-initiated revisions.

Each student's TAP was followed by a stimulated recall interview. During the interviews, the first author stopped at relevant revision episodes in the video to refresh the participants' memory and asked them clarification questions about their cognitive engagement with *Criterion* corrective feedback. The durations of stimulated recall interviews for Trang and Nhien were 16.5 and 19 minutes, respectively. During these interviews, mainly open-ended questions were asked to clarify observed revisions which had minimal elaboration in the TAPs. This was done to elicit data to address the third research question: the cognitive and metacognitive strategies learners used to process the feedback they received (e.g., Why did you decide to change this to…? How did you arrive at this revised form? Why didn't you read the error message before correcting this mistake? Why didn't you revise this tagged error?). At appropriate points during the interviews, students were also asked about their perceptions of *Criterion* feedback, including their beliefs, goals, and preferences which may have contributed to observed textual revisions to further explain their overall behavioural and cognitive engagement with *Criterion* ACF (e.g., What are the main advantages and disadvantages of using *Criterion* feedback to revise your essay? What do you think about this feedback point from *Criterion*?). A summary of the data sources for the two case studies is presented in Table 1.

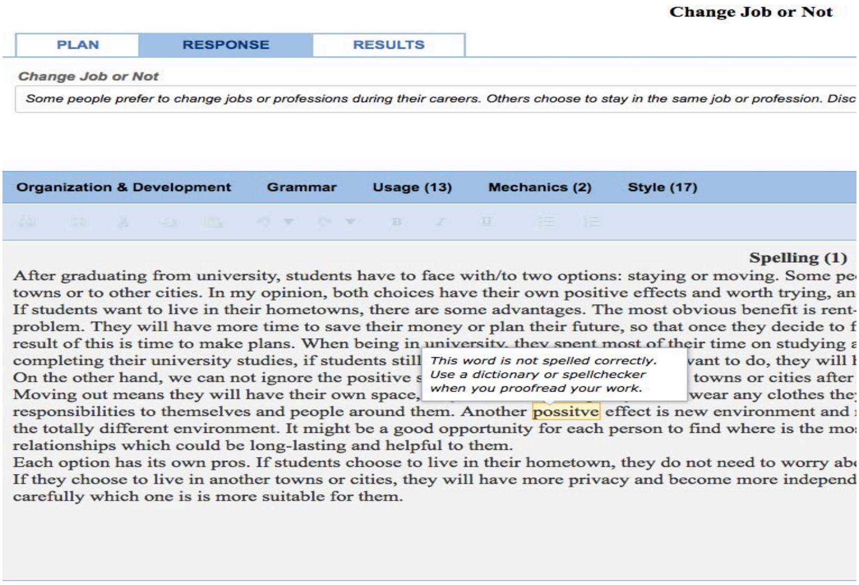
Table 1
Data Sources for the Two Participants

Case	*Session:* *Number of drafts per essay*	*Think-aloud protocol*	*Stimulated recall interview*
Trang	Session 1: 3 drafts Session 2: 3 drafts Session 3: 1 draft (no revised essay)	1 recording (Second session)	1 week after the second session
Nhien	Session 1: 2 drafts Session 2: 2 drafts Session 3: 2 drafts	1 recording (First session)	1 week after the first session

Data analyses

Criterion ACF

Although *Criterion* does provide feedback on Organization and Development, this study focused on *Criterion* feedback on grammar, usage, and mechanics (i.e., ACF). Under each feedback category tab, the scroll-down menu lists the number of error types. Once the student clicked on one error type (e.g., Subject-verb agreement, Possessive errors, Spelling, etc.), all the errors belonging to that type are highlighted in the student's essay. For each error identified, the specific word or phrase is highlighted by *Criterion*. If the student drags the pointer to this highlighted word/phrase, there will be a pop-up screen giving metalinguistic explanations of the errors to guide student corrections, as in Figure 1.



Change Job or Not

Some people prefer to change jobs or professions during their careers. Others choose to stay in the same job or profession. Disc

After graduating from university, students have to face with/to two options: staying or moving. Some pe
towns or to other cities. In my opinion, both choices have their own positive effects and worth trying, an
If students want to live in their hometowns, there are some advantages. The most obvious benefit is rent-
problem. They will have more time to save their money or plan their future, so that once they decide to f
result of this is time to make plans. When being in university, they spent most of their time on studying a
completing their university studies, if students still … vant to do, they will h
On the other hand, we can not ignore the positive s … towns or cities after
Moving out means they will have their own space, … wear any clothes the
responsibilities to themselves and people around them. Another possitve effect is new environment and
the totally different environment. It might be a good opportunity for each person to find where is the mo
relationships which could be long-lasting and helpful to them.
Each option has its own pros. If students choose to live in their hometown, they do not need to worry ab
If they choose to live in another towns or cities, they will have more privacy and become more independ
carefully which one is is more suitable for them.

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

Figure 1. Example Screenshot of Criterion Feedback in Pop-up Screen

Sometimes *Criterion* generates incorrect messages in the form of false positives. Specifically, an error tag is considered a false positive if *Criterion* incorrectly identifies a correct form as an error in students' essays, as in:

Student text:	In addition, you remarkably increase your earning power in another company which appreciates your[1] ability and strength.
Criterion:	[1]You have used **your** in this sentence. You may need to use **you're** instead.

Based on *Criterion*'s metalinguistic explanations, the current research used Ranalli et al.'s (2017) generic-specific feedback distinction to analyze the type of feedback the students received. The next sections elaborate on the analyses related to learners' uptake and engagement episodes with each feedback point from *Criterion* in relation to the type of feedback, taking into account whether the feedback was correct or incorrect.

Students' first and final drafts on Criterion

The comparison between the first and revised drafts for each essay was used to code: (1) whether the learner addressed the error pointed out (uptake) and (2) whether such uptake led to a correct or incorrect revision. In the first instance, two broad categories were applied: *Uptake* vs. *No uptake*. *Uptake* was then coded as either a correct revision or incorrect revision. Analysis for uptake and accuracy of revision took into account when false positives and fallible feedback (i.e., error codes with incorrect error labels/suggested revisions) were provided to see if each individual learner responded appropriately to these instances of incorrect feedback. Furthermore, *No uptake* included *avoidance* behaviour (i.e., instances when students refrained from addressing the correctly tagged error and thus, not taking up *Criterion* ACF by either deleting sections of the text containing the error or not making any change), and *no uptake due to false positives* (i.e., cases where students did not revise their texts after processing false positives from *Criterion*). Table 2 provides examples of how students' revised drafts were coded for uptake behaviour.

Table 2
Examples of Coded Categories for Learners' Uptake Behaviors

First draft	*Revised draft*	*Coded uptake*
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, shrimps in the seas. [*Proofread this!*]	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.	*Uptake* [Correct revision of the section highlighted by *Criterion*]
Another reason make[1] me find money is not such essential as many people have always thought is health. [*Subject-verb agreement*]	Another reason makes me find money is not such essential as many people have always thought is health.	*Uptake* [Incorrect revision]
It can't not[1] be denied that there are advantages of changing job. [*Negation errors*]	It can't not be denied that there are advantages of changing job.	*No uptake* [Avoidance]
However, is it actually right when they think that "only[1] people who earn a lot of money are successful"? [*Missing or extra article*]	However, is it actually right when they think that "only people who earn a lot of money are successful"?	*No uptake* [Due to false positives]

The first author and a PhD candidate in Applied Linguistics who had taught EFL academic writing at tertiary level for 12 years double-coded 20% of the uptake instances in students' first and revised drafts. Inter-coder agreement for students' uptake of *Criterion* ACF was high, at 97%. A few minor disagreements were resolved before the remaining part of coding took place.

Students' TAPs and stimulated recall interviews

All of the students' TAPs and stimulated recalls were transcribed, then translated into English by the first named author, except for parts of the recordings where the two students verbalised in English. A colleague working in the division of Translation and Interpretation helped examine the accuracy of translation and clarity of expression in English while cross-checking with the source texts. For cited examples in this paper, back translation was conducted to ensure loyalty to the students' original verbalizations.

TAP transcripts were then coded for revision episodes, each of which corresponded to the complete processing of one *Criterion*-tagged error and was examined for students' depth of feedback processing (i.e., cognitive engagement) to decide whether they allocated high or minimal level of mental effort to process the feedback. An engagement episode was coded as *high level of mental effort* if the learner demonstrated efforts in evaluating the received feedback and understanding the error message, using self-regulatory learning strategies such as making reference to their prior linguistic knowledge, online resources, or searching for clues in *Criterion* metalinguistic explanations to revise the tagged error. In contrast, *minimal level of mental effort* was recorded if the learner adopted the suggested changes (in specific feedback) without reference to the above sources, made a revision without reading *Criterion* metalinguistic explanations, or simply skipped the error after looking at the *Criterion* tag. In this sense, our study also draws on cognitive perspectives of language learning, focusing on allocation and quality of attention.

Coding for level of cognitive engagement is a highly inferential process (Sachs & Polio, 2007), and coders' judgement based on students' TAP verbalization may not necessarily reflect true depth of feedback processing. Therefore, the stimulated recall interviews were thematically coded for each revision episode to cross-check the levels of cognitive engagement found in the think-alouds (i.e., whether the students' comments in the interview *confirmed* or *contradicted* the level of mental effort they verbalized in the TAP). Students' answers to more probing questions in the interviews revealed explanatory factors that had not emerged in the TAPs to account for students' textual revisions and were coded as *reasons* for students' uptake or rejection of the feedback (i.e., beliefs about the role of revisions, trust level in *Criterion* ACF, learning goals, beliefs about feedback). Double coding was conducted on 20% of the revision episodes extracted from the TAP and stimulated recall transcripts. Inter-coder agreement was 90%. All the disagreements were discussed and resolved before the remaining data were coded.

Results

The explicitness of Criterion ACF and learner uptake

The essays where the two learners wrote revised drafts were analysed for feedback explicitness and their revision behaviours. To this end, all the error tags generated by *Criterion* on the first drafts of these essays were initially extracted. The total word count for Trang's first drafts of the two essays which had revised drafts was 739, and that for Nhien's three essays was 1526, making a corpus of 2265 words on which *Criterion* generated a total of 84 error tags. Of all error tags, 49 were generic and 35 were specific. The two learners' uptake of *Criterion* generic and specific corrective feedback is presented in Table 3. In general, the successful error correction rate was higher for generic feedback, at 78%, while for specific feedback it was 46%. These findings are discussed in the analyses of individual cases. The learners did not accept and respond to three out of the total of 49 generic and five out of the total of 35 specific error tags. No uptake of false positive error tags was recorded for 7 and 12 instances of generic and specific feedback, respectively.

Table 3
Student Uptake and Revisions Following Criterion Generic and Specific Feedback

	Generic feedback 49 (58.3%)				*Specific feedback* 35 (41.7%)			
	Correct revision	Incorrect revision	No uptake (Avoidance)	No uptake (False positives)	Correct revision	Incorrect revision	No uptake (Avoidance)	No uptake (False positives)
Trang	1	0	0	1	7	1	1	7
Nhien	37	1	3	6	9	1	4	5
Total	38 (78%)	1 (2%)	3 (6%)	7 (14%)	16 (46%)	2 (5.7%)	5 (14.3%)	12 (34%)

Case analyses

Trang, primarily minimal mental effort expenditure

Table 4 provides the number of *Criterion*-generated tags by error type Trang received on her first drafts (including whether they were false positives) and her revision outcomes for the first two essays (the third essay did not have a revised draft). As Table 4 shows, Trang received a total of 18 error tags, averaging at nine tags per essay. Half of the tagged errors related to the use of articles. Eight of the total 18 error tags were false positives, including one generic and seven specific feedback points. Regarding her responses, Trang made one correct revision and one non-uptake of a false positive for the two generic feedback points from *Criterion*. She also responded appropriately to 14 out of the 16 specific error tags, with seven correct revisions, seven no uptake instances due to *Criterion*'s false positives, only one incorrect revision and one avoidance.

Table 4
Trang's Feedback Received and Revisions

Category	Error Type	Feedback type	False Positive	Revision Outcome
Usage	9 Missing or Extra Article	Specific	4	6 correct revisions
	2 Confused Words			5 no-uptake instances due
	1 Determiner Noun	Specific	1	to false positives
	Agreement	Generic		1 incorrect revision
Mechanics	3 Missing Comma	Specific	1	2 correct revisions
	1 Extra Comma	Specific	1	3 no-uptake instances due
	1 Spelling	Generic	1	to false positives
	1 Hyphen	Specific		1 no uptake

Across all of the TAP coded revision episodes lasting about 10 minutes, Trang primarily allocated minimal mental effort when engaging with the feedback, irrespective of whether the feedback was generic or specific. The following example demonstrates an instance where Trang quickly arrived at a revision by adopting *Criterion*'s suggested change to revise the form highlighted.

First draft: As a student who has not paid much attention to money pressure yet depends on family[1], … [Missing article error].

Criterion: [1]You may need to use an article before this word.

Trang verbalised in the TAP: "I have missing extra articles. *Family*, *yet depends on family… students who have not paid much attention to money pressure yet and depends on the family.* OK. I will change it into *the family. Depends on the family.*" Then she correctly revised the error by adding *the* before *family.* Trang's verbalization during this revision episode gave the impression that she was over dependent on *Criterion*'s suggestions and allocated minimal mental effort. However, further data from her interview reveal that Trang seemed to quickly notice-with-understanding the errors pointed out to her, as she commented in the interview, "Because I mentioned that *as a student who has not paid much attention* and I prefer *the family…* This is *the family* ***of this student I mentioned before*** [emphasis added], so I use *the* before *family.*" Trang's interview data seemed to contradict the level of mental expenditure coded in her think-aloud recording, as the change from "family" to "the family" was actually a case of a well-informed decision based on reference to the learner's stored metalinguistic knowledge, which emerged in the stimulated recall interview.

In the next example, Trang processed *Criterion*'s message on an extra comma which is a false positive. Trang simply skipped to the next error after reading the error tag and instantly refused to take up *Criterion*'s suggestion, "Extra comma, *beside that… anything…*No, I think this extra comma will need to stay here. I won't change".

First draft: Beside that[1], anything belongs to him such as his words or his way of educating his children is thought to be "successful". [Extra comma].

Criterion: [1] You may need to remove this comma.

Interestingly, both of Trang's incorrect revision and avoidance response were about specific error tags. In more detail, Trang's avoidance was in response to a *Criterion*'s hyphen error tag, a wrong label for the error below:

First draft: Everyone should raise not only the awareness but also the self[1] activities[1] so that we all have a better water resource as well as a greater life. [Hyphen error]

Criterion: [1] You may need to add a hyphen between these two words.

In this example, "self activities" should be tagged as a word choice error where another adjective (e.g., "personal") should modify "activities". *Criterion*'s wrong error label may have confused Trang, resulting in her refusal to take up this specific feedback.

Trang's non-uptake instances can be associated with some distrust in the feedback, as demonstrated in the next tagged error about a missing comma,

First draft: They might be courage, creativity, pride, kindness[1] and sometimes as simple as be able to spend lot of time with family and friends. [Missing comma]

Criterion: [1] You may need to use a comma after this word.

In this revision episode, Trang quickly adopted *Criterion*'s specific suggestion about inserting a comma after "kindness". However, she later commented in the interview that this adoption contradicted her own belief that the insertion of the comma in this instance is optional:

Researcher: So *Criterion* asks you to use a comma after "kindness" and before "and", do you agree with this feedback from *Criterion*?

Trang: Not really, because according to my former teacher, he said that with a comma or without a comma at this space is not necessary.

In citing her teacher's instruction, Trang implied that she did not perceive the authority of *Criterion* ACF very highly. When further asked why she made a revision against her own belief, Trang replied that she considered revising using *Criterion* ACF as "a further way to get a higher mark".

In summary, Trang made 16 well-justified responses to the 18 tagged errors. With almost half of *Criterion* error tags being false positives, she demonstrated competent handling of the errors through reliance on critical processing of the feedback and made appropriate changes when needed. In order to do this, however, she employed low level of mental effort and exercised some caution in adopting *Criterion* feedback. Throughout the interview, Trang demonstrated certainty about her reliance on acquired knowledge rather than *Criterion* ACF. This probably explains the absence of a revised draft in the third session.

Nhien, mixed levels of mental effort expenditure

Table 5 provides the number of *Criterion* tags by error type Nhien received on her first drafts and her uptake behaviours across three sessions. Nhien received a total of 66 error tags, averaging 22 tags per essay. More than one third of all the error tags related to spelling, followed by errors in articles and subject-verb agreement. There were 12 false positives, with five specific and seven generic feedback points. Of the total 47 generic error flags, Nhien successfully corrected 37 errors, did not take up six of the false positives, made one incorrect revision and avoided responding to three correct error tags. Her responses to the 19 specific feedback points were of a similar pattern, with nine successful revisions, five no-uptake instances due to false positives, two incorrect revisions, and four unjustified non-uptake instances.

Table 5
Nhien's Feedback Received and Revisions

Category	Error Type	Feedback Type	False Positive	Revision Outcome
Grammar	12 Subject-Verb Agreement 2 Ill-formed Verbs 1 Fragments 1 Possessive Error	Generic Generic Generic Specific	1 1	11 correct revisions 1 no uptake 2 retentions of the correct form 2 incorrect revisions
Usage	10 Missing or Extra Article 5 Confused Words 3 Determiner Noun Agreement 2 Preposition Error 1 Wrong Article	Specific Specific Generic Generic Specific	3 1 1	11 correct revisions 5 no-uptake instances due to false positives 5 no uptake
Mechanics	26 Spelling 2 Extra Comma 1 Missing Comma	Generic Specific Specific	4 1	24 correct revisions 4 no-uptake instances due to false positives 1 no uptake

On her first essay when the think-aloud was recorded, Nhien received 17 *Criterion* error flags and she spent 17.5 minutes processing these errors. Nine revision episodes were identified as showing high level of mental effort when Nhien employed a range of different cognitive strategies to process them (six generic and three specific feedback points). In the following revision episode, Nhien expressed uncertainty about whether *law* was a countable or uncountable noun, which she double-checked using the Oxford online dictionary.

First draft: Second, having many[1] strict law is another effective way to reduce the shortage of water.

Criterion: [1]You may have used the wrong determiner. Proofread the sentence to make sure that the determiner agrees with the word it modifies.

TAP excerpt: Determiner noun agreement. Agreement of nouns. *Many*, *have many strict law, have many strict law*. Yeah. *Law* is an uncountable? I will check in the Oxford dictionary… I will check *law* to see it's a countable or uncountable noun. OK. It is a count noun. So, I use *many* here. *Second, having many strict law*. I will change this into plural noun. Hope it's correct.

Compared to Trang, Nhien adopted a more cautious approach before rejecting false positive feedback. On each occasion, she evaluated the feedback and reassessed her writing, which enabled her to respond appropriately to all the false positives in this essay. As shown in the following episode, Nhien experienced difficulty understanding *Criterion* feedback which was generic and also a false positive. Initially, she tried to figure out what the meaning of the error explanation was by googling the meaning of the word *fragment*.

First draft: Indeed, water acounts for 75 percent of humans' bodies.[1]

Criterion: [1]This sentence may be a fragment. Proofread the sentence to be sure that it has at least one independent clause with a complete subject and predicate.

TAP excerpt 1: Fragments. I don't know what fragment means. *This sentence may be a fragment, proofread this sentence to be sure that it has at least one independent clause with a complete subject and predicate.* I don't know how to correct this sentence.

Nhien went back to this false positive error code two more times after having corrected the spelling error from *acounts* to *accounts*, using Google translation to inform her non-uptake decision:

TAP excerpt 2: Go back to the first one. *Indeed, acounts for…water acounts for…*They say that it is a fragment. … [Nhien typed in the sentence on Google Translate] OK. *Thực sự thì nước chiếm 75% cơ thể người* [Nhien read the Vietnamese version of her text on Google Translate]. Is this sentence incorrect? I am not sure what is wrong with this sentence.

Nhien's only incorrect revision in this essay was in response to *Criterion* specific feedback, as below:

First draft: First of all, saving the water by raising people'[1]s aware seems to be the most actical solution.

Criterion: [1]You may need to take out the apostrophe to make this word a plural noun.

TAP excerpt: About possessive. *People.* This is the possessive error. *People, people* should be… it should not have this form. *By raising, raising….aware of people aware…. Of everyone.*

Revised draft: First of all, saving the water by raising aware of everyone seems to be the most reality solution.

Criterion neither correctly labeled this error nor offered useful information for Nhien to revise the sentence. For the remaining eight instances involving minimal mental effort, Nhien made quick revisions after seeing the error tags as she realized that she had committed errors out of carelessness, and *Criterion* feedback drew her attention to such slips. For example, she spent minimal mental effort processing specific feedback on article usage such as her quick change from "an careless way" to "a careless way", attributing this to mistyping in the first place. Similarly, she adopted *Criterion*'s correct advice almost instantly by adding *the* before noun phrases (*human being* and *water source*) to correct her text.

In her stimulated recall interview, Nhien appreciated the advantage of using *Criterion* ACF for her revisions on 'basic errors', by which she meant easily remedied errors she could self-correct combined with specific suggested revisions from *Criterion* for these surface-level errors. However, when asked about her unjustifiable no-change response to seven correct error codes from *Criterion*, she said, "I was quite stressed, so I can't think very clearly at that time". Such an emotional response may have been due to the large number of error tags Nhien had to process, compared to the much smaller number of error flags in Trang's essay. Further, when asked about her priority during the revision stage, Nhien commented, "I will focus on grammar more than the idea, because when I'm writing maybe it can be affected much by the spoken language, so I want to revise it to standard grammar for writing".

Discussion

This study set out to examine two EFL students' engagement with *Criterion* ACF based on three sources of data: multiple drafts of essays submitted to *Criterion*, their screen-recorded think-aloud protocols when engaging with the feedback for revisions, and the stimulated recall interviews.

Students' uptake and automated feedback types

Regarding overall uptake of specific versus generic feedback, both learners were English majors with upper-intermediate proficiency level and were thus less likely to accept all the feedback they received. Instead, both students exercised some caution and evaluated the quality of the feedback before arriving at uptake decisions. Trang's metalanguage to explain her revised form in the cited example about adding *the* before *family* and her resistance to adopting *Criterion*'s specific suggested deletion of the article *an* in the noun phrase *an easier and happier life* demonstrated her confidence and dependence on her own acquired language knowledge rather than the automated feedback. On the surface, Nhien seemed less confident about her own grammatical and lexical knowledge, yet she made up for this by extensive search for references online. This result finds support in Bai and Hu (2017) or Jiang and Yu's (2020) studies, both of which show high proficiency students' awareness about the limitations of the AWE system under study (i.e., *Pigai*) and their subsequent selective use of the feedback for revisions. Like Trang and Nhien in the current research, high achieving EFL learners in those studies were able to adjust their uptake level according to feedback accuracy.

From a system-centric perspective, specific feedback is more likely to be fallible. This is true for the data in the current research with most false positives from *Criterion* being specific suggestions for revisions. On the learners' side, there was a higher rate of no uptake in response to specific feedback compared to that for generic feedback (14.3% vs. 6%). Yet, as Godfroid (2020) put it, "absence of evidence is not evidence of absence" (p. 67), and occasions when no revised form was recorded for the two studied cases found explanatory factors other than their ignorance of the feedback. Specifically, the data from the think-aloud protocols of both learners include instances of resistance to specific feedback points which provide incorrect remedial actions (e.g., incorrect suggestion of adding a hyphen between the two words *self* and *activities*). This corroborates earlier research on teacher written feedback conducted by Swain (2006) and Swain and Lapkin (2003) whose findings indicate lack of feedback uptake due to learners' resistance to feedback when it contradicted their beliefs.

Compared to specific feedback, generic feedback is less likely to be erroneous as it basically highlights a text section and directs learners to a general course of remedial action. This also means that in response to generic feedback, learners are required to examine the tagged error in greater depth to evaluate not only the value of the feedback but also to find out the nature of the error using either external resources or their own metalinguistic knowledge (Ferris, 2002). For low proficiency learners, this may cause some challenge due to their tendency to rely on the suggestions in the feedback. However, no longer novice writers, the two learners in this study were able to conduct self-directed learning when cognitively engaged with the automated feedback. The unfocused, generic, non-dialogic, and fallible nature of AWE feedback (Mehrabi-Yazdi, 2018; Ranalli, 2018), albeit a shortcoming of current AWE feedback sources, offers high achieving students the opportunity to rely on their mobilization of various reference sources and revising strategies to deal with tagged errors.

Students' cognitive engagement and automated feedback types

The two learners demonstrated different levels of mental effort when processing *Criterion* ACF. For Trang, there seemed to be little difference in her cognitive engagement patterns with either specific or generic feedback, as all of the engagement episodes were marked with minimal effort to arrive at the revised forms. On the other hand, Nhien invested a high level of mental effort in using extensive strategies such as referencing online forums and dictionaries to double check her revisions. The differential invested mental efforts between two learners do not support previous research indicating that high proficiency learners tend to engage with the feedback extensively (e.g., Koltovskaia, 2020; Zhang & Hyland, 2018). An explanation for this variable finding lies in the nature of the errors tagged by *Criterion.* Previous research points out that AWE systems address surface-level features (Hoang, 2022; Link et al., 2020) and mostly non-meaning changing errors (Tian & Zhou, 2020). Therefore, a high level of cognitive engagement seems unnecessary for learners who can quickly notice with understanding the tagged errors like Trang in the current research.

For Nhien, generic feedback seems to have coincided with greater effort expenditure, with six out of the nine episodes marked with high level of mental effort being in response to *Criterion* generic feedback. With specific feedback, on the other hand, Nhien quickly implemented revisions by adopting the suggested changes, except in the cases of false positives and fallible feedback when she often started with questioning the accuracy of her own writing, followed by using different cognitive strategies to confirm her doubts before rejecting the feedback. For this learner, generic feedback, just like false positives and fallible feedback, triggered deep feedback processing (Hassanzadeh & Fotoohnejad, 2021; Lavolette et al., 2015; Liu & Yu, 2022, Ranalli, 2018), which kickstarted a chain of cognitive strategies to address the tagged errors.

Explanatory factors for learners' feedback uptake and cognitive engagement patterns

Both learners in the current research, despite different levels of mental effort when processing *Criterion* ACF, showed equally high successful responses to the feedback. Their successful handling of the feedback could be attributed to the fact that *Criterion* ACF targets surface-level errors in student essays, and learners at intermediate and higher levels have little difficulty making superficial revisions (Jiang & Yu, 2020) if errors are correctly identified. For proficient learners, a higher level of mental effort does not give them an advantage over minimal mental effort in increasing their chances of arriving at correct revisions when dealing with simple and easily remedied errors such as spelling, punctuation, or subject-verb agreement.

The two learners' variable cognitive engagement levels and uptake of the corrective feedback can also find partial explanation in their learning goals and beliefs about the revising stage. Nhien pursued learning goals, and thus found *Criterion* ACF a highly useful source helping her to notice certain errors or gaps in her output. For this learner, error flaggings from *Criterion* started a chain of metacognitive strategies from planning, monitoring, seeking additional information from a range of sources, to evaluating revisions. This ties in with Zhang and Hyland's (2018) finding that the more highly engaged learner made use of more revision strategies to address the tagged errors. On the other hand, Trang, driven by performance goals, did not choose to deeply engage with the feedback to extend the learning opportunities it provided. She was more focused on the immediate task of correcting the errors through quick rejection or uptake of the feedback. She employed minimal mental effort and relied primarily on stored metalinguistic knowledge as a major source of reference to expedite the revision process.

Another factor worth reiterating is that the two cases received significantly different numbers of error tags from *Criterion.* Nhien received almost three times as many error codes as Trang did for each essay. During revision episodes, Nhien alluded to the detrimental impact of the cognitive overload she experienced when processing the large number of error flags. Nhien's tendency to substantively engage with the feedback and her longer revision episodes may have added to her cognitive load, resulting in her avoidance to address seven correctly tagged errors. No such issue was experienced by Trang who processed a relatively small number of feedback points quickly. Instances of cognitive overload can be partially attributed to *Criterion*'s comprehensive feedback mechanisms which can overwhelm learners with many error tags on a single draft.

Implications and conclusion

The case study approach adopted in this research provides insights into individual learners' cognitive engagement and uptake of *Criterion* ACF which is subdivided into specific and generic feedback types. A proportional relationship between students' mental effort expenditure and *Criterion* feedback explicitness was not established. Instead, the findings suggest that how much effort a learner allocates to processing a feedback point depends on her beliefs, learning goals, and the nature of the error itself. At the current stage, specific feedback seems more fallible and less usable, as evidenced in the two learners' lower uptake and successful revision rates compared to generic feedback. The sample size of a case study precludes generalisability, but AWE developers may need to consider the balance between feedback explicitness and usability to avoid distrust issues among learners.

Future research can continue to look at different types of feedback currently generated by AWE systems. Just as learners vary in their beliefs, learning goals, revising strategies, or proficiency levels, automated feedback varies in its level of explicitness and feedback focus. In this research, Nhien's cognitive overload when processing comprehensive feedback from *Criterion* raises an issue to consider when devising feedback functions on AWE systems. Students should be able to switch between feedback modes where they can choose to attend to focused linguistic areas rather than having to simultaneously deal with multiple error categories. Following teacher feedback research, automated feedback explicitness is an area for future research investigating the interaction between automated feedback types and learner variables such as proficiency, motivation, and learning goals. This study has combined qualitative and quantitative data analyses, yet the sample of two cases does not allow for a systematic convergent mixed methods design. With the rapid growth of mixed methods research in the field of Applied Linguistics over the last decade (Farsani et al., 2022), future larger scale studies can expand on such an approach to triangulate and analyse data from more cases and over longer periods of time.

For writing instructors, this study reveals that students' failed revision attempts often come from fallible feedback (wrong error labels or misleading error explanations). Clearly, prior student training on the use of *Criterion* ACF and a forewarning of possible false positives and incorrect error codes are needed in order to maximize the benefits of *Criterion* ACF for EFL learners. For more proficient learners, like the two cases in the current research, feedback evaluation can be incorporated as part of writing instruction to enhance their revision of relevant grammatical rules, which could help them "learn to apply these grammatical rules in their own writing independently and to evaluate and adopt only the feedback they deem useful" (Woodworth & Barkaoui, 2020, p. 239). In the long run, this helps learners transition from reliance on traditional feedback (i.e., from teachers and peers) to effective use of automated feedback to improve revision processes and learning outcomes.

Acknowledgements

This research was funded by the University of Melbourne via the Melbourne International Research Scholarship program.

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

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

Creswell, J. W. (2013). *Qualitative inquiry & research design: Choosing among five approaches.* (3rd ed.). SAGE Publications.

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

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

Farsani, M. A., & Aghamohammadi, N. (2021). Exploring students' engagement with peer- and teacher written feedback in an EFL writing course: A multiple case study of Iranian graduate learners. *MEXTESOL Journal, 45*(4), 1-17.

Farsani, M. A., Babaii, E., Beikmohammadi, M., & Farsani, M. B. (2022). Mixed-methods research proficiency for applied linguistics: a PLS-path modelling approach. *Quality & Quantity, 56*, 3337-3362. https://doi.org/10.1007/s11135-021-01268-7

Ferris, D. R. (2002). *Treatment of error in second language student writing.* University of Michigan Press.

Gass, S. (1997). *Input, interaction, and the second language learner.* Erlbaum.

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide.* Routledge.

Han, Y. (2019). Written corrective feedback from an ecological perspective: The interaction between the context and individual learners. *System, 80*, 288–303. http://doi.org/10.1016/j.system.2018.12.009

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

Hassanzadeh, M., & Fotoohnejad, M. (2021). Implementing an automated feedback program for a foreign language writing course: A learner-centric study. *Journal of Computer Assisted Learning*, 1494-1507. https://doi.org/10.1111/jcal.12587

Hoang, T. L. G. (2022). Feedback precision and learners' response: A study into ETS *Criterion* automated corrective feedback in EFL writing classrooms. *The JALT CALL Journal, 18*(3), 444-467.

Hoang, T. L. G., & Kunnan, A. J. (2016). Automated writing instructional tool for English language learners: A case study of *MY Access*. *Language Assessment Quarterly, 13*(4), 359-376. http://dx.doi.org/10.1080/15434303.2016.1230121

Jiang, L., & Yu, S. (2020). Appropriating automated feedback in L2 writing: Experiences of Chinese EFL student writers. *Computer Assisted Language Learning*, *35*(7), 1329-1353. https://doi.org/10.1080/09588221.2020.1799824

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal, 99*(1), 1-18. https://dx-doi-org.bris.idm.oclc.org/10.1111/modl.12189

Kim, H. R., & Bowles, M. (2019). How deeply do second language learners process written corrective feedback? Insights gained from think-alouds. *TESOL Quarterly, 53*(4), 913-938. https://doi.org/10.1002/tesq

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by *Grammarly*: A multiple case study. *Assessing Writing, 44*. 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

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), *605–634*. https://doi.org/10.1080/09588221.2020.1743323

Liu, K., & Storch, N. (2023). Second language learners' engagement with written feedback. *Australian Review of Applied Linguistics. 46*(1), 4-28. https://doi.org/10.1075/aral.20029.liu

Liu, S., & Yu, G. (2022). L2 learners' engagement with automated feedback: An eye-tracking study. *Language Learning & Technology, 26*(2), 78-105. https://doi.org/10125/73480

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W.C. Ritchie & T.K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). Academic Press.

Mao, Z., & Lee, I. (2022). Researching L2 student engagement with written feedback: Insights from sociocultural theory. *TESOL Quarterly*, *56*(2), 788-798. https://doi.org/10.1002/tesq.3071

Mehrabi-Yazdi, O. (2018). Short communication on the missing dialogic aspect of an automated writing evaluation system in written feedback research. *Journal of Second Language Writing, 41*, 92–97. https://doi.org/10.1016/j.jslw.2018.05.004

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

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

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

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

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

Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *The Canadian Modern Language Review, 50*(1), 158-164. https://doi.org/10.3138/cmlr.50.1.158

Swain, M. (2006). Languaging, agency and collaboration in advanced language proficiency. In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 95–108). New York: Continuum.

Swain, M., & Lapkin, S. (2003). Talking it through: Two French immersion learners' response to reformulation. *International Journal of Educational Research*, *37*, 285–304. https://doi.org/10.1016/S0883-0355(03)00006-5

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

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, *46*, 327-369.

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. https://doi.org/10.18806/tesl.v37i2.1340

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.

Zheng, Y., & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. *Assessing Writing, 37*, 13-24. https://doi.org/10.1016/j.asw.2018.03.001

Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing, 43*, 1-14. https://doi.org/10.1016/j.asw.2019.100439

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36*, 90-102. https://doi.org/10.1016/j.asw.2018.02.004

Appendix

Task Prompts of the Three Essays Written by the Students on *Criterion*

Session	***Genre***	***Task prompt***
1	Problem solving essay	**Reducing Pollution** *There are many kinds of pollution. What can you do to help reduce one kind of pollution in your community? Use examples and specific details to explain your answer.*
2	Opinion essay	**Money and Success** *Do you agree or disagree with the following statement?* "Only people who earn a lot of money are successful." *Use specific reasons and examples to support your answer.*
3	Advantage/ Disadvantage essay	**Change Job or Not** *Some people prefer to change jobs or professions during their careers. Others choose to stay in the same job or profession. Discuss the advantages of each choice. Which do you prefer? Use reasons and examples to explain your choice.*

Giang Thi Linh Hoang is a lecturer in TESOL in the English Department, University of Foreign Languages and International Studies, Hue University. With a PhD in Applied Linguistics from the University of Melbourne, her research focuses on language assessment, automated essay evaluation, and in particular teacher and automated feedback in EFL contexts.

Neomy Storch was an Associate Professor in ESL and Applied Linguistics at the University of Melbourne. She is currently an Honorary Fellow at the University. Her research continues to focus on issues related to language pedagogy, and in particular the nature of pair interaction in classroom contexts and the development of academic writing.