Title of Project: Exploring Automated Corrective Feedback in EFL Writing Classroom Context: A Case Study of Criterion

Researcher: Giang Hoang University of Melbourne lgiang8380@gmail.com

Dissertation Supervisors: Associate Prof. Ute Knoch University of Melbourne

Final Report

Motivation for the Research
Automated writing evaluation (AWE) systems are increasingly used in classroom settings to provide formative feedback to learners. This study emerged from a concern that the increasingly prominent presence of automated writing instructional programs in numerous teaching and learning contexts is not accompanied by adequate understanding of their impact on students’ learning to write in a second language. While corporate funded researchers are in a better position to access big data to study how (well) a system scores student essays, there is still a scarcity of research evidence about the impact of automated feedback on accuracy development or writing/revision practices and a lack of longitudinal studies into learners’ engagement with automated feedback. Therefore, the current study was conducted to seek insights into learners’ actual interaction with automated form-focused feedback through learners’ perspectives, which produced implications for both service providers and end-users of automated writing instructional programs.

Research Questions
1. What is the nature of Criterion automated corrective feedback (ACF)?
2. How accurate is Criterion’s automated corrective feedback?
3. How do EFL students engage with and respond to Criterion ACF?
4. How does EFL students’ accuracy in L2 writing change after a semester’s access to and use of Criterion ACF?
5. What are students’ perceptions of Criterion ACF?

Research Methodology
The study adopted a pre-post quasi-experimental design to examine the effectiveness of ACF as a formative assessment tool for L2 learners’ writing development. Participants included 81 English majors in a teacher-training degree in the English Department of a university in Hue, central Vietnam. The students, sub-divided into the experimental and control groups, were undertaking a 15-week writing course during which both groups had three practice sessions to compose essays to three writing prompts. The control group wrote their essays on paper and submitted them to the instructor for feedback. The experimental group, however, had
access to the automated writing instructional program called *Criterion* during their practice sessions, so they were able to revise their drafts in response to its feedback before submitting revised drafts to the teacher for feedback.

Beside test essays, other data included the following: (1) first and revised drafts from *Criterion* practice sessions; (2) recorded think-aloud protocols (TAP), which were conducted with 14 students as they revised essays using *Criterion* corrective feedback; (3) follow-up interviews with four students, and (4) end-of-term focus group interviews. Error analysis was conducted on the test scripts of the control and experimental groups to investigate any changes in accuracy over time. Two accuracy measures were used: (1) a holistic measure based on Foster and Wigglesworth’s (2016) weighted clause ratio approach, and (2) analytic measures of accuracy in using specific grammatical morphemes following an obligatory occasion analysis. Findings from the quantitative pre- and posttest analyses were triangulated with the qualitative data, showing how the students engaged with the feedback from *Criterion* and implemented it in their revision practices. Students’ revisions were analyzed for the type of changes made in response to *Criterion* corrective feedback. Analyses of TAP recordings employed Schmidt’s (1993) key concept of “noticing” to shed light on students’ cognitive engagement with the automated feedback and to clarify revision practices found. Further triangulation came from students’ perceptions of automated feedback in TAP follow-up and focus group interviews that were thematically analyzed.

**Summary of Findings**

*Criterion* was found to perform differently across different error types. Overall, the automated corrective feedback was satisfactorily accurate with 21 error types exceeding the 80% precision threshold. Findings on students’ engagement with and subsequent revisions using *Criterion* ACF revealed more perfunctory than substantive processing of the feedback. It was also found that unhelpful feedback tended to initiate more extensive engagement among the learners. Following their engagement, students performed mostly local textual operations to revise their texts and were modestly successful at correcting the errors *Criterion* pointed out to them, with 56.2% success rate and 7.1% retention of their original texts in response to false positives. Generally, the surface-level nature of the feedback and students’ superficial engagement with it, along with simple and straightforward textual operations suggest a strong connection between the nature of the feedback, the types of errors addressed, and the way in which students responded to the feedback.

Over the time frame for the study, there was no significant treatment effect of the use of *Criterion* ACF on experimental students’ accuracy across the five morphemes for which they received most error tags from *Criterion*, including article, third person singular, copula, plural, and comma usage. Their overall accuracy did not improve significantly either.

The qualitative analyses of the focus group interviews and stimulated recalls revealed students’ overall positive perceptions about *Criterion* ACF. Their satisfaction with the feedback was thanks to *Criterion*’s immediacy, systematicity, and comprehensiveness in generating feedback. Students also discussed issues that caused concerns about *Criterion* ACF, including the lack of elaboration in metalinguistic explanations and absence of feedback on more complex grammatical issues. The findings also pointed out learners’ varied expectations about how explicit the feedback should be and their evaluations of teacher compared to automated corrective feedback.
Overall, the validation of Criterion ACF as a learning and assessment tool in the EFL writing classroom reveals a mixture of support for its use and rebuttal evidence against its use. Criterion was able to address EFL learners’ needs for surface-level errors, but it still lacked coverage of some major issues in the students’ L2 writing. It can be praised for facilitating revising, as well as self-regulatory writing strategies, and triggering noticing among the students; however, Criterion’s approach to feedback generation was not pedagogically based, resulting in a lack of meaningful engagement with the feedback. Overall, despite students’ positive feelings about Criterion ACF, reservations about its value remain due to learners’ middling revision success rates and the absence of significant intervention or retention effects of the use of Criterion ACF on their accuracy gains over the period of study.

Implications
The findings extend our understanding about students’ engagement and use of the automated corrective feedback by adding a missing piece in the research on EFL learners’ use of Criterion ACF. Qualitative data from think-aloud protocols and stimulated recalls unveiled the underlying mechanisms that moderate learners’ final revising decisions and somewhat invalidate certain findings in earlier research which assumed that the absence of any change to a flagged error indicates the learners’ non-use of the feedback.

The study’s main pedagogical implications relate to formative feedback practices in the classroom, including the need to supplement Criterion automated feedback with teacher feedback to support L2 writing instruction and classroom-based assessment. Specifically, writing instructors can provide oral feedback sessions during meeting hours so that learners can bring up clarification questions after they have engaged with Criterion ACF. Furthermore, strategy training sessions can be embedded in classes during which learners learn cognitive/metacognitive strategies for revisions or share reference sources they find most helpful to seek further information for the error codes received. The instructor should also be open to the current shortcomings of Criterion ACF.

From developers’ perspectives, Criterion corrective feedback should be designed to be more adaptable to focus learners’ attention on the relevant issues for their developmental stage. Solutions include making Criterion feedback more flexible in terms of toggling between its error types or feedback options to maximize the benefits for a larger number of learners. In addition, more lenient and flexible treatment of writing mechanics issues should be considered to accurately reflect the dynamic nature and changing realities of writing.
References

Activity 4: Complexity in oral vs. written language. (n.d.). Retrieved from https://carla.umn.edu/learnerlanguage/spn/comp/activity4.html?fbclid=IwAR1LSZQAsiNWCVo2YgO8hSbMcDH1vw-hrPzwaasxXcnrj-W4kHgtHvX7MZk


Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). Analysis of discourse features and verification of scoring levels for independent and
integrated prototype written tasks for the new TOEFL [Monograph]. *ETS TOEFL Monograph Series, MS-30*. Princeton, NJ: ETS.


the Annual Meeting of the American Educational Research Association (AERA), San Francisco, CA.


