Language Learning & Technology
http://llt.msu.edu/issues/june2015/lavolettepoliokahng.pdf
June 2015, Volume 19, Number 2
pp. 50–68

THE ACCURACY OF COMPUTER-ASSISTED FEEDBACK AND STUDENTS' RESPONSES TO IT

Elizabeth Lavolette, Gettysburg College
Charlene Polio, Michigan State University
Jimin Kahng, Northeastern Illinois University

Various researchers in second language acquisition have argued for the effectiveness of immediate rather than delayed feedback. In writing, truly immediate feedback is impractical, but computer-assisted feedback provides a quick way of providing feedback that also reduces the teacher's workload. We explored the accuracy of feedback from Criterion®, a program developed by Educational Testing Service, and students' responses to it. Thirty-two students received feedback from Criterion on four essays throughout a semester, with 16 receiving the feedback immediately and 16 receiving it several days after writing their essays. Results indicated that 75% of the error codes were correct, but that Criterion missed many language errors. Students responded to the correct error codes 73% of the time and responded to more of the codes over the course of the semester, while the condition—delayed versus immediate—did not affect their response rates nor their accuracy on the first drafts. Although we cannot support claims that immediate feedback may be more helpful, we believe that, with proper training, Criterion can help students correct certain aspects of language.

Language(s) Learned in Current Study: English

Keywords: Computer-Assisted Language Learning, Second Language Acquisition, Writing

APA Citation: Lavolette, E., Polio, C., & Kahng, J. (2014). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology*, *19*(2), 50–68. Retrieved from http://llt.msu.edu/issues/june2015/lavolettepoliokahng.pdf

Received: March 20, 2014; **Accepted:** October 26, 2014; **Published:** June 1, 2015

INTRODUCTION

The provision of language feedback on second language learners' writing is one of most widely researched and widely contested issues in the second language learning discipline. The effectiveness of written error correction has been debated since Truscott's (1996) polemic call for the abandonment of written grammar correction. What is not contested is that students want language feedback and that, for teachers, it is time-consuming to give it. Given this information alone, most would agree that if useful feedback can be given by a computer, such feedback could be a positive addition to, not replacement for, teacher feedback. In addition, computer-assisted feedback can be given quickly after writing is produced, something a teacher cannot always do. Thus, this study has two separate yet related purposes. First, we explore the usefulness of computer-assisted language feedback for ESL learners by investigating its accuracy; the first part of the literature review below discusses related research on intelligent computer-assisted language learning (ICALL). Second, we explore what factors affect students' responses to the feedback. Although in our study, we explore four different factors (type of error code, correctness of error code, experience with the software, and feedback timing), it is only feedback timing that has been explicitly addressed in previous research. Thus, in the second part of the literature review we discuss research from second language acquisition (SLA).

ICALL AND SECOND LANGUAGE WRITING

Many ICALL systems have been developed to help language learners improve their language skills, some of which contain grammar checkers that provide feedback on writing. Nagata (2009) developed a system called Robo-Sensei for teaching Japanese, which includes writing practice at the sentence level. Heift (2003, 2010) created E-Tutor for teaching German, which includes opportunities for students to write sentences and essays. Education and testing companies have also developed ICALL systems that provide automated feedback and scores on writing. Vantage Learning uses its IntelliMetric automated essay scoring system in the MyAccess! system and Pearson uses its Intelligent Essay Assessor in the WriteToLearn system. Unfortunately, little data has been reported about how effective any of the above systems are for learning writing in the target languages (but see Chen & Cheng, 2008, for an analysis of students' and instructors' opinions of the effectiveness of MyAccess!).

On the other hand, researchers have examined the accuracy of a few stand-alone intelligent grammar checkers. BonPatron, a grammar, spelling, and expression checker for learners of French, has been reviewed (e.g., O'Regan, 2010), and one evaluation (Burston, 2008) investigated its efficacy. Burston used a corpus of 10 compositions written by learners of French and found that BonPatron detected 88% of the errors, a higher percentage than two other commonly used French grammar/spelling checkers, Antidote (85%) and Microsoft Word (44%). BonPatron also showed low rates of miscorrection and false correction. A more in-depth study of Antidote by Biesemans (2005) (cited by O'Regan, 2010) showed that for a corpus of compositions written by intermediate learners, 60% of errors were identified. O'Regan also examined intelligent grammar checkers for English language learners. Grammar Slammer was found to detect 24% of errors in a learner corpus; Spell Check Plus, 14%; White Smoke Perfect Writing 2009, 40%; and Right Writer 5.0, 16%. These studies show a fairly wide range of error detection accuracy levels among the programs.

Criterion

The most studied ICALL system for language learning is Educational Testing Service's Criterion®. Criterion provides feedback on student essays written in response to prompts. Criterion usually provides indirect feedback by locating an error (without correcting it), and it sometimes also provides suggestions of the correct form. All error corrections are metalinguistic, and most include a brief grammatical explanation (See Figure 1). The feedback is unfocused, covering all errors that Criterion detects. However, Criterion does not detect some errors until others have been corrected. For example, a spelling error may prevent Criterion from detecting a grammar error in the same part of the sentence.

Some studies of Criterion have focused on its accuracy. Han, Chodorow, and Leacock (2006) developed the natural language processing (NLP) system used in Criterion for identifying article errors. According to their tests, Criterion correctly identified 40% of article errors, and of all the errors it identified, 90% of them were, in fact, errors. About five years later, Li, Lee, Lee, Karakaya, and Hegelheimer (2011) looked at the effect of Criterion's feedback on ESL learners' correction of article errors in their writing. They found that Criterion correctly identified about 73% of these errors. In the students' responses to Criterion's feedback, 37% of errors were successfully corrected by following Criterion's suggestions, whereas 31% of errors were ignored, 12% of errors were not corrected successfully, and 20% of errors were eliminated through a change of content. Moving away from articles, Tetreault and Chodorow (2008) examined the accuracy of the NLP system used by Criterion for finding preposition errors. They found that about 80% of the errors identified by Criterion were, in fact, errors, and that the system identified about 12 to 14% of the total preposition errors—which means that the system missed 86 to 88% of the preposition errors.

Other studies have focused on users and learning. Attali (2004) investigated essays submitted to Criterion by L1 English students and showed that their error rates significantly decreased in their resubmissions. Chodorow, Gamon, and Tetreault (2010) showed that second language (L2) English speakers were able

to reduce article error rates in their second and final versions of essays submitted to Criterion. Choi (2010) studied the effects of using Criterion with university English learners in the US and in Korea. The participants wrote essays and revised them based on Criterion's feedback. The students who used Criterion showed greater improvement on holistic scores and accuracy from the first draft to the final revision than the control group. They also got higher final holistic scores, but no statistically significant differences were found.

Instructors and students have reported beneficial effects of Criterion because it can lead to increased writing practice (Myers, 2003), increased motivation for students, and can save teachers' time (Warschauer & Grimes, 2008). However, most of the studies that examined the effects of Criterion and its feedback on students' writing have limitations in that they dealt with only a single essay assignment over a short period of time, such as one week, and improvement was often measured by the scores generated by Criterion or the number of errors that Criterion found (but c.f. Choi, 2010). Difficulties with using computer-generated feedback have also been reported. For example, when implementing MyAccess!, Chen and Cheng (2008) found that one instructor thought the feedback was too vague to be useful.

Our focus in this section has been on the accuracy of ICALL systems. While researchers such as Burston (2008) have examined the accuracy of intelligent grammar checkers for French, grammar checkers for English cannot necessarily be expected to be equally accurate. In fact, O'Regan (2010) found much lower accuracy rates for English, with White Smoke Perfect Writing 2009 showing the highest accuracy: 40%. For Criterion itself, Tetreault and Chodorow (2008) found that Criterion identified 12 to 14% of preposition errors; Han et al. (2006) found that Criterion correctly identified 40% of article errors, while five years later, Li et al. (2011) found that Criterion correctly identified about 73% of article errors. The difference between the Han et al. and Li et al. studies may be due to the different versions of Criterion they used. To our knowledge, no studies have examined the overall accuracy of Criterion's linguistic feedback.

THE ROLE OF IMMEDIATE FEEDBACK

Of the four factors that we explore as possible influences on students' response to the feedback (type of error code, correctness of error code, experience with the software, and feedback timing), only feedback timing has been addressed in the theoretical and empirical literature. In fact, this study was motivated by Polio (2012), who argued that some approaches to SLA suggest that feedback on language needs to be immediate for it to be effective. Further, although there is no research showing that Criterion's immediate feedback is superior to delayed feedback, Educational Testing Service (ETS) stated in 2012 that the immediacy of Criterion's (overall) feedback is one of the keys to its usefulness to students:

> Students get a response to their writing while it is fresh in their minds. They find out *immediately* how their work compares to a standard and what they should do to improve it. The Criterion service also provides an environment for writing and revising that capable and motivated students can use independently. This environment, coupled with the opportunity for *instant feedback*, provides the directed writing practice so beneficial to students (2012, our emphasis).

This statement most likely refers to feedback on all aspects of writing, but according to some theories of second language acquisition, it should apply particularly to feedback on language.

Immediate (as well as *instant*) are terms used by ETS in the quote above, but they are not used consistently in the literature. Here, we define immediate feedback on writing as feedback provided at the *end* of a writing task. Although it is technically possible to provide feedback as soon as a student makes an error (e.g., Aubrey & Shintani, 2014), such feedback is generally impractical. We define delayed feedback as feedback provided at a time later than the end of the writing task. In the current study, the delayed feedback was provided one to three weeks after the task was completed.

Certainly, other factors are related to the effectiveness of language feedback for improving language use in the long term. Indeed, Polio (2012) argued that the learner has to pay attention to the feedback, the feedback may have to be at the right developmental level, and that explicit (including metalinguistic) information may be helpful for writing more accurately. Given these criteria, Criterion should fare well in that it draws learners' attention to errors and provides explicit and metalinguistic feedback. (It does not, however, give feedback aimed at the learners' level, as a teacher might be able to do.) In this limited space, we cannot review all of the research on language feedback. Therefore, we state only that much of the controversy has centered around research design and what constitutes evidence of effective correction (Guénette, 2007; Polio, 2012; Xu, 2009). The immediacy of the feedback is what essentially sets apart computer-assisted feedback from human feedback. Although the topic of immediate feedback has not been well researched in the L2 literature, below, we discuss its possible role in SLA.

Second Language Learning and Teaching

Both SLA theory and three empirical studies suggest that immediate feedback in writing might be beneficial. Polio (2012) claimed that, theoretically, immediate feedback on writing errors should be beneficial to L2 writers according to several approaches in SLA, including the interaction approach, usage-based approaches, and sociocultural theory. She stated that the immediacy of recasts during speaking may contribute to their effectiveness under the interaction approach, which is an advantage that written corrections generally do not have. Ellis (2012) talked about priming as a factor in usage-based learning, meaning that a speaker is more likely to use a structure right after hearing it. In sociocultural-based approaches, feedback is used to help learners scaffold, which suggests that it should be immediate.

Although no empirical studies comparing feedback timing have been conducted within usage-based or sociocultural approaches, two studies related to feedback timing have been conducted from a skill acquisition theory perspective (Evans, Hartshorn, & Strong-Krause, 2011; Hartshorn et al., 2010). Skill acquisition theory originated in psychology and has been applied to a wide variety of learning situations, including language learning (Dekeyser, 2007). This theory suggests that through continued practice, knowledge can be automatized, resulting in faster reaction times, fewer errors, and less required attention. Implicit in this theory is the notion that feedback provided during practice must be timely, although no explicit indication is provided as to how to operationalize this.

Hartshorn et al. (2010) used a technique called "dynamic written corrective feedback", which reduced, to about one day, the time between the error being produced and the feedback being received and used. Advanced ESL students wrote for 10 minutes every day for 15 weeks, and the teachers coded all of the students' errors and returned the assignments the following day. The students corrected the errors, and the process was repeated until no errors remained. Compared to a control group taught using traditional process writing methods, the dynamic written corrective feedback group showed significantly greater improvement in accuracy, with no statistically significant differences between the groups in rhetorical competence ratings, fluency, or complexity. It is not clear whether both groups received the same amount of feedback; however, it is clear is that the treatment group received feedback sooner after writing than did the control group.

Evans et al. (2011) replicated Hartshorn et al.'s (2010) study with undergraduate ESL writers. The procedures in the replication were similar to those in the original study, with the exception that the students in the dynamic feedback group wrote for 10 minutes three or four times per week rather than every day. The results were similar to the original study, with the students who received dynamic feedback significantly outperforming the students in the control group in accuracy, with no significant differences in fluency or complexity. These studies were intended to examine the effect of the timeliness of the feedback in addition to other unique features of the feedback, including the amount of practice and the repeated feedback until the writing contained no errors. Nevertheless, based on skill acquisition theory, these studies argue that timely feedback is an important component of effective feedback.

A recent L2 writing study of note is that of Aubrey and Shintani (2014), who used Google Docs to provide feedback to a (synchronous feedback) group of English learners while they were writing. Another (asynchronous feedback) group received feedback a few minutes after completing their essays, and a third group received no feedback. Only hypothetical conditionals were targeted, and the feedback was only the correct form. No significant difference was found between the feedback groups on an immediate posttest. On a delayed posttest, the synchronous feedback group outperformed the no-feedback group, while no difference was found between the asynchronous group and the no-feedback group.

In sum, based on the available empirical evidence, and because of the differences in the studies, it is not clear whether immediate feedback on writing is better than delayed feedback. Hartshorn et al. (2010) and Evans et al. (2011), for example, argued that timely feedback is important. However, it is not clear if other factors, such as the total amount of feedback that each group received, were factors in the dynamic feedback groups' outperforming the control groups in their studies. Aubrey and Shintani (2014), on the other hand, found some evidence that feedback provided during writing may be more effective than feedback provided after a delay. Finally, giving feedback through Criterion allows two advantages over the feedback given in previous studies. The feedback is more timely than that used by Hartshorn et al. and Evans et al. and the feedback is more logistically feasible and more comprehensive than that given by Aubrey and Shintani.

RESEARCH QUESTIONS

Our ultimate goal is to examine whether Criterion helps students improve their language on a new piece of writing. We agree with Truscott (1996) that improvement on a new piece of writing, as opposed to improvement during revision, should be the focus of written error correction research. In addition, we need to determine how well Criterion can give feedback and how students interpret and respond to the feedback during revision. Although Criterion is the most researched natural language processing system available, little is known about how students interact with it or how well it performs overall; previous studies have focused only on articles and prepositions. Thus, our research questions are:

1. How well can Criterion give language feedback?
    a. What types of language errors does Criterion correctly detect?
    b. To what extent does Criterion miss errors?
2. How do type of error, correctness of the error code, timeliness of the feedback, and practice over time affect student response to the feedback?
3. Does immediate feedback result in more grammatically accurate writing?

Questions 1a and 1b were addressed by examining Criterion's feedback. Based on the previous literature, we expected to find that Criterion was fairly accurate at finding article errors (Li et al., 2011) and less accurate at finding preposition errors (Tetreault & Chodorow, 2008), but we did not have any particular expectations for other error types. Question 2 was addressed by examining students' revisions. We expected that the students would improve in their accuracy in responding to the feedback over time as they became more comfortable with the software. We did not have any basis on which to make predictions about students' responses related to the correctness of the code nor of the type of error codes, although some codes, such as subject-verb agreement, seem intuitively easier for students to understand. Based on the emphasis on the timeliness of feedback in SLA (Polio, 2012) and skill-acquisition theory (Evans et al., 2011; Hartshorn et al., 2010), we hypothesized that students might respond more accurately to immediate feedback versus delayed. However, while skill-acquisition theory predicts that sustained and timely practice should result in more accurate grammar, it is not clear whether it predicts a more accurate response to the feedback. Research question 3, therefore, seeks to determine if students in the immediate group improved their accuracy over the course of the treatment. Research question 3 is not, however, the

main focus of our study because we were not able to provide what could be considered sustained feedback as was done by Evans et al. (2011) and Hartshorn et al. (2010).

METHOD

Materials[1]

A screenshot showing feedback when the user's mouse rests on the second highlighted *the* is shown in Figure 1. The usage, grammar, and mechanics feedback were enabled for this study, but the organization and development and style feedback were not enabled because the focus of our study was only on language feedback. In addition, students received a holistic score generated by Criterion for each essay and revision along with the feedback because we thought that a desire to see the scores increase might motivate students to respond to the feedback. The writing prompts were TOEFL essay prompts that were integrated into the application. Those used in the current study are shown in the Appendix. We also asked students for some demographic information and about their comfort level typing in English.

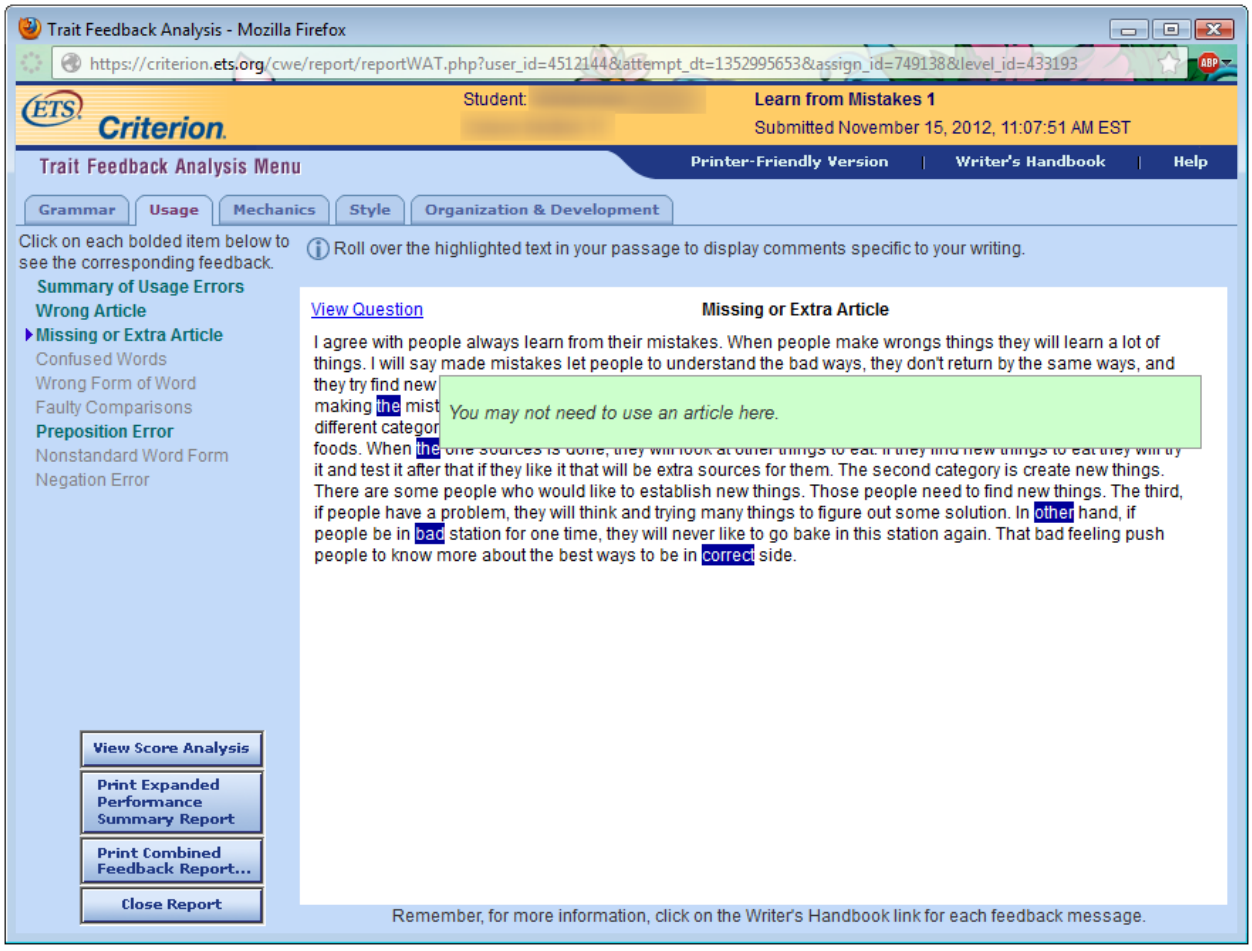

Figure 1. Example of feedback provided by Criterion

Participants

The participants from this study came from five academic writing classes at a U.S. university. Thirty-two students completed all five sessions in the study, and 24 students provided demographic information. The students were admitted or provisionally admitted to the university and were enrolled in a six-credit grammar and composition section. The majority were Chinese-speaking undergraduate students with a mean age of 21. Other native languages included Korean, Arabic, and Italian.

Procedure

All sessions took place in a computer lab. During the first meeting, the students were trained in how to use TextEdit (the text editor included in Mac OS), how to use Criterion, and what types of language feedback to expect from Criterion. Students wrote their essays in a text editor so that they would not submit the essays to Criterion while writing the first draft. This procedure was essentially the same as writing the essay directly in Criterion. In addition, the students were informed that Criterion's feedback was not always accurate because we wanted them to learn to evaluate the feedback. We showed them examples of when Criterion was correct, when it miscoded errors (with particular emphasis on verb tense problems as this was where we found the greatest number of miscodes), and when it noted errors that did not exist. The training lasted about 70 minutes.

The students in both groups wrote essays using TextEdit based on prompts taken from Criterion's bank of TOEFL questions. They were given 40 minutes to write each essay. The immediate group received feedback immediately after writing by copying and pasting their work from TextEdit into Criterion, then clicking a button to submit it. The delayed group received feedback from Criterion in the same way, but one to three weeks after writing. The exact schedule for the delayed feedback depended on when the teachers could schedule their students for the sessions. Immediately after receiving the feedback, both groups revised their essays and resubmitted them. They had 20 minutes to revise and submitted up to two revisions on each essay. Students did not always need the entire 20 minutes, but after piloting the software, we determined that 20 minutes was an appropriate amount of time to allow all students to finish. In some cases, Criterion found no errors in students' essays after the first or second submission, meaning that the students had no feedback to respond to, and these students were not included in this analysis. This procedure was repeated in subsequent sessions on new prompts until all students had written and revised a total of four essays. Only data from Sessions 2 and 5 (the first and final essays) are examined here because we were interested in change over the semester. The order of the essay prompts could not be fully counterbalanced due to logistical issues related to using intact classes. The timing for the training, writing, and revisions is shown in Table 1. We note here that even the delayed group was able to respond immediately to feedback after the first revision, but feedback on the revision was still removed from when the students wrote the original essays.

Table 1. *Training, Writing, and Revision Schedule*

	Immediate feedback group			Delayed feedback group		
Session		Essay topic	Time (min.)		Essay topic	Time (min.)
1	Training		70	Training		70
				Write	B*	40
			3-week interval			
2	Write	B	40	Revise	B	20
	Revise	B	20	Write	C/D	40
			1-week interval			
3	Write	A/C	40	Revise	C/D	20
	Revise	A/C	20	Write	A	40

			3-week interval			
4	Write	C/D	40	Revise	A	20
	Revise	C/D	20	Write	D/C	40
			1-week interval			
5	Write	D/A	40			
	Revise	D/A	20	Revise	D/C	20
	Questionnaire		10	Questionnaire		10

*Note: *The letters refer to the topics in the Appendix.*

Coding the Feedback

Examples of students' writing and the corresponding feedback produced by Criterion are shown in Tables 2 and 3. To answer Research Question 1a, regarding the error types detected, each sentence receiving a Criterion error code was put into a spreadsheet with its error code. A list of all of the error codes that Criterion produced during the current study is shown in Table 4. Each error code was then classified as correct error code, wrong error code, or no error by two of the authors. Examples of each classification are provided in Table 2. The intercoder reliability was 96%. Discrepancies in the two authors' codes tended to come from their disagreement over the grammatical accuracy of a student's sentence or because we were unsure of the student's intended meaning.

Table 2. *Coding the Correctness of Criterion's Feedback*

Error code category	Examples
Correct error code: Criterion correctly identified the problem.	**Student sentence**: Since leader should divides the task to each member. Criterion feedback: Fragment **Student sentence**: If the group has problem or some obstacle all members are going to ask the leader to solve this problem and then he will get more stress. Criterion feedback: Missing article
Wrong error code: Criterion appropriately coded a structure as incorrect but gave it the wrong code.	**Student sentence**: Because these mistakes he learned form which are he used made. Criterion feedback: Missing article **Student sentence**: A critical idea should be came up in some occasion because it is the time to speak out and discuss till the work is satisfied. Criterion feedback: Compound word
No error: Criterion coded a structure as an error but there are none in the sentence or an error exists in the sentence but is unrelated to the underlined structure.	**Student sentence**: The ability to hold the whole group together, the power to require the privilege, and to achieve more publicity are three reasons for being a leader rather than just a member. **Criterion feedback**: Subject-verb agreement **Student sentence**: This shouldn't stop us from looking for guidelines along the way. Criterion feedback: Fragment

Note: The underlining of the words was included as part of the feedback.

To answer Question 1b regarding which errors Criterion was likely to miss, we randomly selected 10 essays and divided them into T-units. A T-unit is a main clause plus its dependent clauses (Hunt, 1965). All T-units were marked as containing an error or not by two authors, who agreed on 94% of the T-units.

We noted the T-units that the coders had determined contained an error but were not flagged by Criterion. This approach was used for two reasons. First, coding individual errors can be difficult (see Polio, 1997) in that depending on how a sentence is corrected, the number and types of errors can vary. Second, we felt that if Criterion identified a sentence as containing an error, the student's attention would be drawn to the sentence, and the student would have an opportunity to find multiple errors or a variety of ways to fix the sentence. Nevertheless, this approach underestimated the number of errors that Criterion missed.

Students' Responses to Feedback

To address Research Question 2, we coded the students' responses to the feedback. This was sometimes difficult given the multiple error codes per sentence and the number of judgment calls required, including whether the initial Criterion code was correct, whether the revision made by the student was related to the Criterion error code, and whether the revision corrected or improved the sentence. Thus, we chose a simple, but perhaps less informative, classification system based on whether any revision made by the student was in any way related to the Criterion code, meaning we did not judge the correctness of the revision. This was not only because it would have been difficult to get acceptable reliability but also because our goal was to determine if the students understood the feedback, rather than if they knew the correct form. This coding system may have overestimated how well the students understood the feedback because of our assumption that a related change meant that the feedback was understood. An additional category was added for a deleted structure or sentence. Table 3 shows these codes. In the first example, the student made an article change in response to another error code but did not fix the run-on sentence, so this was coded as no change/unrelated change. For the analysis of the revisions, unlike the analysis of the accuracy of Criterion's codes, we examined only the first revision. Coding for revision can sometimes be difficult because students may make major changes in their essays or completely change the structure of the sentence. In this study, however, students made only surface-level changes (as defined by Faigley & Witte, 1981), which simplified the coding process. This is likely due to the nature of the feedback itself, which indicated only surface-level errors.

Table 3. *Coding Students' Revisions.*

Revision category	Examples
No change/unrelated change	**Original sentence**: And there was one time I got hurt from my mistake, I forgot to drop my trash away from table. Error code: Run-on **Revision**: And there was one time I got hurt from my mistake, I forgot to drop my trash away from the table.
Related revision	**Original sentence**: To sum up, the member of the team will decide how group work, but not how group will work. **Error code**: Missing article **Revision**: To sum up, the member of the team will decide how group work, but not how the group will work.
Deletion	**Original sentence**: Without food, people would past away. **Error code**: Confused word **Revision**: Without food, people would past away.

Using the coded student responses, we performed a 2 x 2 mixed ANOVA, using the number of related revisions (for the correct Criterion codes) in the students' first and final essays as a within-subject factor and feedback timing group (immediate or delayed) as a between-subject factor.

Finally, we also wanted to check the possibility that Criterion got better at identifying students' errors over the course of the study. Therefore, we performed a repeated-measures ANOVA using the percentage of correct error codes produced by Criterion on the first and final essays.

Students' Change in Accuracy Over Time

To determine whether there was an effect for feedback timing on the students' grammar, we tallied the number of error-free T-units in the first drafts written in the second and last sessions.[2] As noted above, in answering research question 1b, we coded all T-units as containing an error or not with an inter-coder reliability of 94%. The results of this analysis were used in a mixed ANOVA to check for group and time effects on students' accuracy in their first drafts.

RESULTS

Question 1a: What Types of Language Errors Does Criterion Correctly Detect?

We examined four essays, one first draft and its first revision from the students' first and final essays, from each of the 32 students, resulting in 1540 error codes produced by Criterion. We determined that 1159 of the error codes (75%) were correct, 208 (14%) correctly identified an error but miscoded it, and 173 (11%) were for structures that were already correct. The full results are shown in Table 4, listed in order of frequency with the first fifteen codes appearing 20 times or more. Note that the codes *proofread* and *garbled* were always correct because they could have referred to any error in the sentence. Criterion did very well at identifying capitalization, missing comma, wrong word, and ill-formed verb errors; these codes were correct over 85% of the time. The codes for run-on sentence, wrong article, and, surprisingly, spelling errors, were incorrect or miscoded over 50% of the time. (Criterion interpreted plural noncount nouns as spelling errors.) Below is an example of each of these miscodings. (The underlined words indicate the errors highlighted by Criterion.)

(1) Coded as run-on

> As we enter into new stages in our lives, the advices we receive from them is very helpful because they have already had bad similar experiences.

(2) Coded as wrong article

> The member in a group is much better than leader for two reasons

(3) Coded as spelling error

> As we enter into new stages in our lives, the advices we receive from them is very helpful because they have already had bad similar experiences.

Table 4. *Correctness of Criterion's Error Coding*

Criterion code	Correct error code		No error		Wrong error code		Total	
Missing article	272	(78%)	20	(6%)	56	(16%)	348	(23%)
Capitalization	157	(97%)	2	(1%)	3	(2%)	162	(11%)
Fragment	81	(60%)	16	(12%)	38	(28%)	135	(9%)
SV agreement	85	(79%)	3	(3%)	20	(19%)	108	(7%)
Missing comma	86	(85%)	12	(12%)	3	(3%)	101	(7%)
Preposition error	72	(74%)	24	(25%)	1	(1%)	97	(6%)
Run-on	45	(49%)	17	(18%)	30	(33%)	92	(6%)
Extra article	49	(58%)	32	(38%)	3	(4%)	84	(5%)

Ill-formed verb	74	(95%)	3	(4%)	1	(1%)	78	(5%)
Wrong article	33	(43%)	15	(19%)	29	(38%)	77	(5%)
Spelling	29	(48%)	19	(31%)	13	(21%)	61	(4%)
Confused word/wrong missing	50	(89%)	3	(5%)	3	(5%)	56	(4%)
Proofread	42	(100%)	0	(0%)	0	(0%)	42	(3%)
Compound words	36	(97%)	0	(0%)	1	(3%)	37	(2%)
Garbled	20	(100%)	0	(0%)	0	(0%)	20	(1%)
Possessive error	8	(62%)	1	(8%)	4	(31%)	13	(1%)
Faulty comparison	8	(100%)	0	(0%)	0	(0%)	8	(1%)
Hyphen	6	(100%)	0	(0%)	0	(0%)	6	(0.4%)
Pronoun error	1	(17%)	2	(33%)	3	(50%)	6	(0.4%)
Missing question mark	0	(0%)	4	(100%)	0	(0%)	4	(0.3%)
Missing or extra article	2	(100%)	0	(0%)	0	(0%)	2	(0.1%)
Negation error	1	(100%)	0	(0%)	0	(0%)	1	(0.1%)
Nonstandard word form	1	(100%)	0	(0%)	0	(0%)	1	(0.1%)
Wrong word form	1	(100%	0	(0%)	0	(0%)	1	(0.1%)
Total	1159	(75%)	173	(11%)	208	(14%)	1540	

A few other issues stand out. Criterion was much better at identifying missing articles (78% correct) than wrong articles (43% correct), which is not surprising given that most singular count nouns need an article and that the context is less important than when determining which article to use.[3] Regarding the preposition error codes, 25% of them misidentified what were already correct sentences. This is similar to Tetreault and Chodorow's (2008) finding that 20% of the preposition error codes produced by Criterion were for already correct structures. In addition, structures that Criterion coded as extra article were already correct 38% of the time. Examples (4) and (5) below show error codes given to correct structures.

(4) Preposition error: Most of the citizens in this country do not have to worry about their futures.

(5) Extra article: After I gathered the information from everyone, though it is a hard work, I shared the feedback to members in the group which greatly help them to understand each other.

Criterion does not appear to code for verb tense errors and instead misinterprets many as subject-verb agreement errors, as in Example (6). As with some of the wrong article errors, Criterion does not account for the context, thus providing the incorrect error code.

(6) Subject-verb agreement: After all this pain and difficulties he lead the army to win the war.

Question 1b: To What Extent Does Criterion Miss Errors?

The randomly selected essays included 266 T-units, 206 of which contained at least one student error. However, Criterion identified only 111 as containing an error and thus missed 95 (46%) of these T-units that contained an error. Thus, while Criterion was good at coding errors, it missed nearly half of the T-units that contained at least one error, as in Examples (7) and (8).

(7) It cannot avoid that it will have some controversy between members, even to the leaders.

(8) So, being a leader must be brave and full of responsibilities.

In these 95 T-units, there was a wide variety of errors, so it is difficult to describe exactly what Criterion was missing, but lexical errors, as in Example (9), where the student used the word *character* instead of *characteristics*, seemed to be particularly problematic for Criterion.

(9) There are several characters that a good leader should have, such as good leader status, a clear mind and good management.

Even though Criterion did not code for verb tense errors, only two of the missed T-units contained verb tense errors, probably because Criterion detected some other error in the T-unit or miscoded the verb tense errors. Criterion missed seven T-units containing preposition errors, which was surprising given that it tended to overcode preposition errors. Criterion also missed nine T-units containing article errors. Previous studies of Criterion's accuracy examined articles (Han et al., 2006; Li et al., 2011) and prepositions (Tetreault & Chodorow, 2008), but it is difficult to directly compare these results with ours because these studies did not use T-units.

Question 2: What Factors Affect Students' Revisions?

To answer this question, we examined the influence of the following: the type of error code (i.e., for which structure); the accuracy of the error code (correct, no error, or wrong error); time (session 1 vs. 4); and the timing of the feedback (delayed vs. immediate).

Before looking at the possible influences, we examined the individual response rates of the students to correct error codes, and we combined no response and unrelated response into one category, unrelated. Only one student deleted any of the sentences with correct error codes. The number of related changes that the students made ranged from 16 to 96% with a mean of 73% and a standard deviation of 18%, showing that some students were outliers and simply did not respond to much of the feedback.

In Table 5, we show how the students responded to error codes for which there were over 20 correct occurrences. For the missing comma, preposition, and spelling error codes, the students made changes related to the error only about 50% of the time. In contrast, for the ill-formed verb, proofread, and subject-verb agreement error codes, the students made related changes about 85% of the time. What is puzzling is that students ignored capitalization 41% percent of the time but ignored ill-formed verb only 15% of the time. It seems that the former is more straightforward to understand and respond to than the latter.

Table 5. *Students' Responses to Criterion's Most Common Error Codes (20 Occurrences or More)*

	Type of change							
Criterion code	Related		Unrelated		Deleted		Total	
Capitalization	59	(59%)	41	(41%)	0	(0%)	100	(10%)
Compound words	16	(62%)	10	(38%)	0	(0%)	26	(3%)
Confused word	28	(76%	8	(22%)	1	(3%)	37	(4%)
Extra article	50	(77%)	15	(23%)	0	(0%)	65	(7%)
Fragment	61	(73%)	21	(25%)	2	(2%)	84	(8%)
Ill-formed verb	47	(85%)	8	(15%)	0	(0%)	55	(6%)
Missing article	174	(73%)	66	(28%)	0	(0%)	240	(24%)
Missing comma	34	(55%)	28	(45%)	0	(0%)	62	(6%)
Preposition error	32	(51%)	31	(49%)	0	(0%)	63	(6%)
Proofread	24	(86%)	4	(14%)	0	(0%)	28	(3%)

Run-on	47	(78%)	13	(22%)	0	(0%)	60	(6%)
Spelling	15	(47%)	17	(53%)	0	(0%)	32	(3%)
Subject-verb agreement	83	(87%)	12	(13%)	0	(0%)	95	(10%)
Wrong article	36	(68%)	17	(32%)	0	(0%)	53	(5%)
Total	706	(71%)	291	(29%)	3	(0.3%)	1000	(100%)

Table 6 shows how students responded differentially to the codes that were correct, used the wrong error code, or coded a correct structure. We see here that students indeed ignored the error codes that Criterion produced for already correct structures more often than they ignored other error codes. Nevertheless, about half the time they tried to change structures that were already correct. This may be because the students were not always sure if what they had initially written was correct.

Table 6. *Students' Responses to Criterion's Error Codes by Correctness of Codes*

	Type of change							
Rater code	Related		Not related		Deleted		Total	
OK	587	(73%)	219	(27%)	2	(0.2%)	808	(76%)
Wrong error code	100	(72%)	38	(28%)	0	(0%)	138	(13%)
No error	63	(56%)	48	(43%)	1	(0.9%)	112	(11%)
Total	750	(71%)	305	(29%)	3	(0.3%)	1058	(100%)

Next, we examined the number of related revisions (in response to correct Criterion codes) in the students' first and final essays to determine if they improved at responding over the course of the study. We also looked at a possible feedback timing effect between the groups that received delayed and immediate feedback. Table 7 shows the combined data from all students, and there appears to be an effect for time but not group. To determine if the difference was significant, we calculated the percentage of related revisions for each participant on the first and final essays, then performed a mixed ANOVA. It indicated a significant effect for time, $F(1, 30) = 11.14$, $p = .002$, and an effect size suggesting that time accounted for 27% of the variance error ($\eta^2_{partial} = .27$). (See Brown, 2008, for a discussion of $\eta^2_{partial}$.) There was no significant effect for group, $F(1, 30) = 0.09$, $p = .77$, $\eta^2_{partial} = .003$, and no interaction effect, $F(1, 30) = 1.01$, $p = .32$, $\eta^2_{partial} = .03$. Thus, it appears that the students improved at responding to Criterion's codes and finding related errors. Another possibility is that Criterion improved at identifying students' errors because the students' language skills improved; in other words, Criterion did better because later essays contained fewer errors. Thus, we checked the percentage of correct error codes from the first and final essays and found no significant increase, $F(1, 30) = 0.88$, $p = .36$, $\eta^2_{partial} = .03$, meaning that Criterion was not more accurate on the final essay in terms of identifying the students' errors, but that the students did get better at responding to them. In addition, the timing of the feedback did not play a role in the students' responses to the error codes.

Table 7. *Average (SD) Percentage of Changes Related to Criterion's Error Codes for Essays 1 and 4, by Feedback Group*

Feedback group	Essay 1	Essay 4
Immediate	66% (21%)	85% (15%)
Delayed	69% (28%)	79% (17%)

Question 3: Does Immediate Feedback Result in More Grammatically Accurate Writing?

We performed a 2 x 2 mixed ANOVA of the percentage of error-free T-units, using change over time as the within-subject factor (from the first to fourth session) and immediate or delayed feedback as the between-subjects factor (Table 8). No significant main or interaction effects were found: main effect of time, $F(1, 30) = 0.88$, $p = .36$, $\eta^2_{partial} = .028$ (small effect size); main effect of feedback timing, $F(1, 30) = 0.009$, $p = .92$, $\eta^2_{partial} < .001$ (very small effect size); interaction between time and feedback timing, $F(1, 30) = 0.13$, $p = .72$, $\eta^2_{partial} = .004$ (small effect size).

Table 8. *Average (SD) Percentage of Error-free T-units for Essays 1 and 4, by Feedback Group*

Feedback group	Essay 1	Essay 4
Immediate	76% (16%)	74% (12%)
Delayed	78% (12%)	73% (18%)

DISCUSSION

Summary of Findings

The motivation for our study was to explore a method for providing relatively immediate feedback—a factor highlighted by Hartshorn et al. (2010) and Polio (2012). Toward this goal, we needed to fully understand how well an NLP program could give language feedback and how students responded it to. We wanted to find out what factors influenced student responses to the feedback and if the timing of the feedback played a role. To summarize the findings, Criterion's codes were correct 75% of the time, but 11% of the time, the codes were on structures that were already correct, and 14% of the time, the wrong error codes were given. In addition, Criterion missed at least 46% of the errors. With regard to how students responded to the codes, we found much individual variation, more responsiveness to the feedback over time, and greater responsiveness to correct error codes than to incorrect ones. Despite the immediacy of Criterion's feedback as the impetus for the study, we found no significant difference in the student responses between the immediate and delayed feedback groups, with the caveat that what we call immediate feedback was provided up to 40 minutes after the error was produced. Furthermore, we found no difference between the two groups in accuracy over time on their initial drafts.

Implications Related to Feedback Timing and Research

SLA theory suggests that immediate feedback is more beneficial than delayed, and Criterion is partially marketed with the claim that immediate feedback is beneficial, but this study did not confirm that relatively immediate feedback results in better language revisions. Of course, a long-term effect is possible, but we suspect that none exists for two reasons. First, the feedback provided in this study may not have been immediate enough. The only study to date that isolated the variable of feedback timing on L2 writing and found an effect was Aubrey and Shintani (2014), and the more effective feedback was provided while the participants were still writing their essays. In addition, writing is different from speaking in that students can go back and see what they have written when getting the feedback. Limitations on cognitive resources are not a factor because the writing serves as an extension of working memory (Hayes & Chenoweth, 2006). Given conflicting results in research outside of SLA (Azevedo & Bernard, 1995; Hattie & Timperley, 2007), we suspect that the issue of feedback timing is much more complex than we had anticipated.

Implications for Computer-assisted Feedback in Teaching

We believe that used properly, that is, with student training, Criterion can reduce a teacher's workload. The training aspect is essential not only because of the fair amount of incorrect feedback that Criterion gave but also in light of our finding that the students responded to more of the feedback over time. As

Hubbard (2013) argued, learner training may help students use technology effectively, sooner than they would without training.

In our study, students revised only for language and generally did not add or delete sentences. This is not surprising given that we suppressed the other feedback functions, but it remains an open question as to whether students will revise at a more global level if they also get feedback on organization and development. Teachers perhaps could give feedback on global factors of writing, and then use Criterion for language feedback after the students have revised, thus freeing up teacher time.

Another matter that teachers need to consider is Criterion's capacity for students to resubmit papers after making corrections. Criterion can often catch missed errors upon resubmission after the student makes corrections on a related part of a sentence. Perhaps due to the nature of its parser, Criterion seems, probably unintentionally, to withhold some feedback, providing new feedback on each submission. On one hand, this could be beneficial; Shih (1998) reported that students wanted individual grammar feedback but sometimes felt overwhelmed by the amount of it. On the other hand, numerous submissions could cause too much reliance on the program, again highlighting the need for training.

Regarding the incorrect error codes, for advanced students with metalinguistic knowledge, the incorrect codes might still be highly beneficial if they promote noticing. Criterion can help students find errors, but it makes them think because the program does not do all the work. That is, it requires them to consider the error codes and evaluate them in context to determine whether the error codes are correct. Criterion is correct often enough that students will see it as beneficial, but also incorrect so often that students cannot rely solely on the feedback. We see it as a tool for teaching and encouraging independent editing skills. Ferris (1995a, 1995b) and Shih (1998) have provided descriptions of procedures for helping students learn to self-edit, and computer-assisted feedback could be included in such procedures. Many have noted that students need to engage with the feedback in some way for it to be effective (e.g., Sachs & Polio, 2007). Suzuki (2012), for example, had students write about the feedback that they received. For students who can tolerate ambiguity, computer-assisted feedback could actually increase engagement with the feedback because of the need to evaluate it. On the negative side, we suspect that students will miss many errors using Criterion because they will focus on only the parts of the essay that Criterion has underlined.

Limitations and Future Research

This study examined only accuracy in writing, so we do not know what effect the feedback may have had on other aspects of the students' writing (e.g., linguistic complexity) during revision or over the semester. For example, if students are focused on accuracy, they may neglect other aspects of writing. Given that we did not find group differences with regard to accuracy, it is unlikely that either group was overly focused on accuracy. Furthermore, because of logistical difficulties, the study was not designed to test the effects of immediate sustained feedback.

More information about students' attitudes toward the program would be helpful, particularly how they felt about the sometimes ambiguous or incorrect feedback. When students receive teacher feedback, they are not taught to evaluate it for its accuracy, and the necessity to evaluate feedback adds a new dimension to processing the feedback. This processing of the feedback could be studied more closely by collecting think-aloud data as students read and responded to the feedback.[4]

We conclude by saying, however, that research with less advanced students should be considered cautiously. Given some of the problems with the feedback and the use of metalanguage, not all students will be able handle the feedback in English, nor will Criterion be as accurate in providing feedback as it was with the advanced students in the current study.

NOTES:

1. ETS provided access to Criterion for this study.
2. There are a variety of ways to measure accuracy. Error-free T-units is a widely used measure, albeit one that may not fully capture accuracy. Polio and Shea (2014) compared a wide variety of accuracy measures including error-free clauses and number of errors, but they were not able to determine that the more fine-grained measures were a more valid measure of accuracy.
3. Criterion uses the term *article* to refer to any determiner.
4. In a separate study, we collected think-aloud data from a few participants. The quality of the data was not good enough to report here, but with students who were able to think aloud, we were able to understand how they interpreted the feedback. For example, some of the students challenged the feedback.

APPENDIX. Writing Prompts

A. Experience or Books

It has been said, "Not everything that is learned is contained in books." Compare and contrast knowledge gained from experience with knowledge gained from books. In your opinion, which source is more important? Why?

B. Group Member or Leader

Do you agree or disagree with the following statement? It is better to be a member of a group than to be the leader of a group. Use specific reasons and examples to support your answer.

C. Money on Technology

Some people think that governments should spend as much money as possible on developing or buying computer technology. Other people disagree and think that this money should be spent on more basic needs. Which one of these opinions do you agree with? Use specific reasons and details to support your answer.

D. Learn From Mistakes

Do you agree or disagree with the following statement? People always learn from their mistakes. Use specific reasons and details to support your answer.

ACKNOWLEDGEMENTS

We thank ETS for providing access to Criterion for use during this study. We also thank the teachers and students who participated in the research. We appreciate the helpful comments of the anonymous reviewers who helped us improve this article.

ABOUT THE AUTHORS

Elizabeth (Betsy) Lavolette is Director of the Language Resource Center at Gettysburg College. She holds a PhD in Second Language Studies from Michigan State University. Her research interests include computer-assisted language learning, assessment, and feedback.

E-mail: elavolet@gettysburg.edu

Charlene Polio is a Professor in the Department of Linguistics & Germanic, Slavic, Asian, & African Languages at Michigan State University. She is the co-editor of the *Modern Language Journal* and has published several articles and book chapters on L2 writing from an SLA perspective.

E-mail: polio@msu.edu

Jimin Kahng is an assistant professor in TESOL (Teaching English to Speakers of Other Languages) at Northeastern Illinois University. Her areas of research include second language acquisition and education, ICALL, language assessment, and development of second language fluency.

E-mail: J-Kahng@neiu.edu

REFERENCES

Attali, Y. (2004). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education, San Diego, CA.

Aubrey, S. C., & Shintani, N. (2014). *The effects of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment*. Paper presented at the meeting of the American Association of Applied Linguistics, Portland, OR.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111–127.

Brown, J. D. (2008). Effect size and eta squared. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(2), 38–43.

Burston, J. (2008). BonPatron: An online spelling, grammar, and expression checker. *CALICO Journal, 25*(2), 337–347.

Chen, C.F. E., & Cheng, W.Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94–112. Retrieved from http://llt.msu.edu/vol12num2/chencheng.pdf

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.

Choi, J. (2010). *The impact of automated essay scoring (AES) for improving English language learner's essay writing*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3437732).

Dekeyser, R. M. (2007). Skill acquisition theory. In B. Vanpatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Mahwah, NJ: Erlbaum.

Educational Testing Service. (2012). *Frequently asked questions about the Criterion® online writing evaluation service*. Retrieved from http://www.ets.org/criterion/ell/about/faq/

Ellis, N. (2012). Frequency-based accounts of second language acquisition. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 193–210). New York, NY:

Routledge.

Evans, N. W., Hartshorn, K. J., & Strong-Krause, D. (2011). The efficacy of dynamic written corrective feedback for university-matriculated ESL learners. *System, 39*(2), 229–239.

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*(4), 400–414.

Ferris, D. (1995a). Can advanced ESL students become effective self-editors? *CATESOL Journal, 8*(1), 41–62.

Ferris, D. (1995b). Teaching ESL students to become independent self-editors. *TESOL Journal, 4*(4), 18–22.

Guénette, D. (2007). Is feedback pedagogically correct? Research design issues in studies of feedback on writing. *Journal of Second Language Writing, 16*, 40–53.

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

Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly, 44*(1), 84–109.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hayes, J. R., & Chenoweth, N. A. (2006). Is working memory involved in the transcribing and editing of texts? *Written Communication, 23*(2), 135–149.

Heift, T. (2003). Multiple learner errors and meaningful feedback: A challenge for ICALL systems. *CALICO Journal, 20*(3), 533–548.

Heift, T. (2010). Developing an intelligent language tutor. *CALICO Journal, 27*(3), 443–459.

Hubbard, P. (2013). Making a case for learner training in technology enhanced language learning environments. *CALICO Journal, 30*(2), 163–178.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. *NCTE Research Report No. 3*. Champaign, IL: National Council of Teachers of English. Retrieved from http://files.eric.ed.gov/fulltext/ED113735.pdf

Li, J., Lee, H.W., Lee, J.Y., Karakaya, K., & Hegelheimer, V. (2011). *The influence of using Criterion® on students' error correction in writing*. Paper presented at the Second Language Research Forum, Ames, IA.

Myers, M. (2003). What can computers and AES contribute to a K–12 writing program? In M. D. Shermis & J. Burstein (Eds.). *Automated essay scoring: A cross-disciplinary perspective* (pp. 3–20). Mahwah, NJ: Lawrence Erlbaum.

Nagata, N. (2009). Robo-Sensei's NLP-based error detection and feedback generation. *CALICO Journal, 26*(3), 562–579.

O'Regan, B. (2010). From spell, grammar and style checkers to writing aids for English and French as a foreign language: Challenges and opportunities. *Revue Française de Linguistique Appliquée, 15*(2), 67–84.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101–143.

Polio, C. (2012). The relevance of second language acquisition theory to the written error correction debate. *Journal of Second Language Writing, 21*(4), 375–389.

Polio, C. & Shea, M. (2014). Another look at measures of accuracy in L2 writing. *Journal of Second Language Writing, 23*, 10–27.

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.

Shih, M. (1998). ESL writers' grammar editing strategies. *College ESL, 8*(2), 64–86.

Suzuki, W. (2012). Written languaging, direct correction, and second language writing revision. *Language Learning, 62*(4), 1110–1133.

Tetreault, J. R., & Chodorow, M. (2008). *The ups and downs of preposition error detection in ESL writing*. Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08, 1, 865–872.

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

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

Xu, C. (2009). Overgeneralization from a narrow focus: A response to Ellis et al. (2008) and Bitchener (2008). *Journal of Second Language Writing, 18*, 270–275.