the
jaltcalljournal

ISSN 1832-4215
Vol. 11, No.1 Pages 63–76
©2015 JALT CALL SIG

Comparing the pedagogical benefits of both Criterion and teacher feedback on Japanese EFL students' writing

Neil Heffernan
Kurume University
neilhef@gmail.com

Junko Otoshi
Okayama University
Hefferman & Otoshi

This paper reports on a classroom based-inquiry using quantitative methods conducted with Japanese EFL students' writing practice using ETS's Criterion. *The purpose of the study is to examine the actual effects of teachers' feedback on students' writing on Criterion. Twelve university students in Japan participated in this study, while completing three Criterion writing assignments each. Six of the students received feedback from both a teacher and from the Criterion system, while six only received feedback from Criterion. The results of the study demonstrate that while the group who received feedback from a teacher showed improvement in some rhetorical features such as thesis statement and awareness of readers, the group that only received feedback from Criterion did not demonstrate major changes in those areas over their three assignments. From these results we can state that while Criterion can provide useful feedback to EFL learners, a teacher's feedback on early drafts of written work is essential for learners to be able to substantially improve their writing. Finally, some pedagogical implications concerning giving effective feedback while using Criterion will be discussed.*

Keywords: EFL writing; Teacher feedback; Criterion; Autonomous learning

Forum

1. Introduction

When learning a second language, output is deemed to be essential in advancing the process of learning the new language in second language acquisition (Swain, 2005). In the case of second language writing, regular practice is crucial, as it is one of the productive skills. Further, providing feedback to each writing assignment is very time consuming for teachers, even if the outcomes of these efforts are not always satisfactory to learners (e.g., Chandler, 2003; Rob, et al, 1986). Therefore, an alternative approach, which involves students regularly practicing their writing outside of the classroom (Bitchener & Knoch, 2009), all the while receiving effective feedback, is necessary.

In this study, Criterion, an online writing practice tool developed by Educational Testing Service (ETS), was utilized outside of the English as a Foreign Language (EFL) classroom as a self-study tool with the intention of facilitating learners' writing practice, as well as easing the burden on the teachers who were tasked with marking written assignments. According to ETS, the Criterion service motivates English-language learners by giving them frequent writing practice that helps build confidence and improve their English writing skills (ETS, 2013).

With the now ubiquitous use of the Internet evident in all aspects of life, EFL learners can practice and learn the target language (TL) anywhere and at any time as far as online environment is available (Loucky, 2004; Warschauer, 1997). One major advantage of using online systems for both studying and teaching languages is that they are inherently efficient and convenient to use and can lessen the burden for teachers (Long, 2013). At the same time, however, the outcomes of self-study making use of online writing practice tools must be examined by comparing them with that of having actual feedback from a teacher. We are acutely aware that an online writing practice tool like Criterion cannot ultimately replace the feedback from a teacher, despite the aforementioned benefits of using such an online writing practice tool. As Cheville (2004) points out, there is a danger of creating an uneasy reliance on machine-based essay scoring systems such as Criterion, thus undermining the practical knowledge of experienced EFL teachers.

The present study was also in line with previous empirical research which studied, observed and measured phenomena deriving from actual teaching contexts (American Educational Research Association, 2015). Previous studies of this nature have mostly focused on the perceived effectiveness of Automatic Writing Evaluation (AWE) systems by the users themselves and not on whether they can effectively emulate human raters (see Chen & Cheng, 2008; Nielson, 2011). Long's (2013) study was unique because it considered the differences in feedback between Criterion and teachers in order to examine the effectiveness of students' writing. His study analyzed 145 randomly selected papers for both differences in Criterion and human feedback and for improvement over an academic year, focusing on linguistic features such as grammar usage, mechanics and style. Further, Chen and Cheng (2008) analyzed 53 complete surveys on their users' actual experience with an AWE system. Finally, Nielson's (2011) study investigated 326 participants' participation and advances in language learning in two phases over 20 and 26 weeks respectively.

Nielson (2011) notes that while there are many CALL systems that cater to autonomous language learning, very few are complete language learning solutions in terms of teaching and testing in the classroom. Keith (2003) states that most of the studies on AWE systems were carried out on large-scaled standardized tests rather than actual classroom writing. Chen and Cheng (2008) suggest that AWE systems may actually encourage language learners to focus more on surface elements such as grammar rather than communicating their actual meaning when writing. As a result, the chief complaint of AWE systems is that they seemingly eliminate the human element of writing: the essential notion that writing is an interaction between people and not between a person and a machine. With such caution in mind, this classroom-based inquiry was unique in that we examined the benefits of AWE (i.e., Criterion) and human feedback on students' writing, with careful attention paid to the specific interaction between AWE and EFL writers.

The current study explores how the outcomes of writing practice outside the classroom using Criterion are different between two groups: students who receive feedback from both a teacher and Criterion, and students who only receive feedback from the Criterion system. It is hypothesized that the group which receives feedback from a teacher and Criterion will improve their writing ability over time to a greater extent than the group only receiving automatic feedback. Nevertheless, this research aims to test this hypothesis and further find out which features of writing components will be affected by teachers' feedback. Based on the results of the study, pedagogical implications will be discussed focusing on teachers' guidance for successful self-study.

2. Criterion's feedback and teacher feedback on rhetorical features

Essays submitted on Criterion receive a score instantly by an automated rating system, called *e-rater.* E-rater evaluates learners' essays based on what it predicts human raters would holistically give an essay based on several criteria (Enright & Quinlan, 2010). In particular, the e-rater system evaluates the essays with a holistic scoring ranging from one to six, which takes into account the following dimensions: grammar, usage, mechanics, style, organization and development, lexical complexity, topic-specific vocabulary usage (Quinla, Higgins, & Wolff, 2009).

In addition to scoring essays, Criterion actually has a built-in function called *View Feedback Analysis* that can be used to provide feedback to users. The learners who use the system can receive online feedback through *View Trait Feedback Analysis* according to each of the dimensions listed above. The dimensions taken into account in Criterion's e-rater system highly correspond with other widely used rubrics, such as the *ESL Composition Profile* (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981). Therefore, Criterion's e-rater seems to evaluate the same constructs as other rubrics commonly used in the EFL classroom (Becker, 2010).

While it seems to value all the components of writing, the Criterion service does not grade essay content (Lim & Kahng, 2012). This might be understandable since many automated programs are theoretically grounded in a cognitive process model of the human brain, all the while ignoring the social and interactional domains (Ware & Warschauer, 2006). The advantages of e-rater are many: it returns a score to a user within seconds; it is completely objective (as scores are distributed based a statistical analysis); and assigns extremely reliable scores that can be compared to human raters' scores (Attali, 2007; Attali & Bernstein, 2006). However, the main detriment to the system is that it cannot evaluate an essay's argumentation, logic or coherence (Heilman & Tetreault, 2012). Thus, judging the quality of the content of an essay is left to the subjectivity of a human rater.

Essentially, the traits of *Organization and Development* in Criterion focus on "the number and length of text units, and not organization and development as a human rater might interpret the terms" (Deane, 2013, p.14). In Criterion, *Organization and Development* in e-rater's *View Trait Feedback Analysis* feature is classified into the following further detailed elements: (A) *introductory material and* (B) *thesis statement (topic relationship and technical quality)*, (C) *main ideas,* (D) *supporting ideas,* (E) *conclusion* and (F) *transitional words and phrases*. E-rater is known to be able to assess these elements reliably and with a validity equal to a human raters' scoring (Weigle, 2010).

Due to Criterion's use of *Organization and Development*, comparing and contrasting the above detailed components with the existing literature concerning coherence in EFL writing is necessary. In order to do so, the authors decided to utilize Lee's classifications of coherence for second language writing classrooms (2002), as her work makes the complicated constructs of coherence clear for classroom use. Lee operationally defines each element of coherence with instructional purposes in writing classrooms. To this end, the authors found her classifications informative in providing feedback in the EFL writing classroom, especially regarding the coherence of students' writing. Lee's classifications, which were outlined in her seminal 2002 paper, are specifically designed for instructing the complicated concept of coherence in writing, and have been used as a benchmark in evaluating second language writing ever since. She states that coherence is "traditionally described as the relationships that link the ideas in a text to create meaning for the readers" (p.135). Thus, coherence is an area in which the organization and development of an essay are dealt with. As such, we sought to gain insight into whether there is a difference between Criterion and Lee's taxonomies by comparing them to each other. A comparison of Criterion's concepts of *Organization and Development* and Lee's concepts of Coherence are presented in Table 1 below.

When discussing coherence, Lee (2002) delineates six essential components for second language writing instruction. These six components are: (1) *purpose, audience and context of situation*; (2) *macrostructure;* (3) *information distribution and topical development*; (4) *propositional development and modification;* (5) *cohesion*; and (6) *metadiscourse* (p. 140).

Introductory material and *thesis statement (topic relationship and technical quality*) in Criterion's e-rater seem to be similar to Lee's *purpose, audience* and *context of situation* since the *introductory material* and *thesis statement* in an essay should involve these specific elements. As for Criterion's *main ideas* and *supporting ideas*, they are comparable to *propositional development and modification*, which deal with how topics in texts are illustrated and developed. *Transitional words and phrases* in Criterion's classifications can be closely linked to *cohesion* in Lee's category. That is, Lee refers to "reference, substitution and conjunctions" as being important parts of cohesive writing (p.140).

Although Lee's subcategories of coherence contain *information distribution and topic development,* which deal with how a topic is developed and distributed in texts, Criterion deals with this matter within the scope of *main ideas* and *supporting ideas*. Further, *metadiscourse,* which is concerned with a readers' point of view, while paying attention to attitude markers and topicalizers, are not found in Criterion's *Organization and Development*. This is one example of how Criterion cannot effectively evaluate a writer's content while taking a readers' point of view into effect. It is here that we can see one possible advantage of human raters.

Therefore, it is necessary to know how accurately Criterion has evaluated these coherence components with a specific rubric which can be used as a supplement by human raters. In doing so, teachers who are using Criterion can give feedback paying closer attention to these areas. After comparing and contrasting Criterion's *Organization and Development* with Lee's categories – as shown Table 1 – the following four dimensions were considered to be the most important elements for the researchers in this study to analyze the students' essays: thesis statement, topical development, propositional development, and awareness of readers.

Table 1: Comparison of organization and development in Criterion and Lee's six topics of coherence

Organization and development in Criterion	**Lee's Six Topics of Coherence**
(A) Introductory material and (B) Thesis statement	(1) Purpose, audience and context of situation
(A) Introductory material and (E) Conclusion	(2) Macrostructure
(C) Main ideas and (D) Supporting ideas	(3) Information distribution and topical development
(C) Main ideas and (D) Supporting ideas	(4) Propositional development and modification
(F) Transitional words and phrases.	(5) Cohesion
Not applicable	(6) Metadiscourse

3. Purpose of the study

The current empirical research was designed to explore how Criterion as a self-learning tool can be utilized outside of the EFL classroom using quantitative methods. In accordance with one key characteristic of empirical research, the following questions – which were derived from the actual teaching context – were poised to guide this study.

1. How do specific features of students' writing improve over time following i) automatic and teacher feedback and ii) only automatic feedback?
2. What is the correlation between human and automatic scoring in L2 writing?
3. Which rhetorical features are sensitive to a teacher's feedback in terms of writing quality?

4. Method

This study was conducted in the spring semester of the 2012 academic year in Japan with twelve students in three different national universities. This study took the form of an experiment group (Group 1) and a control group (Group 2), and was specifically designed to investigate the differences of a teacher's feedback on samples of writing on the Test of English as a Foreign Language (TOEFL) Criterion e-rater system. Since this research project dealt with Japanese learners of English, a written consent form was collected from all participant students in the study.

The students in Group 1 came from a TOEFL Institutional Testing Program (ITP) preparation class in which one of the authors taught at the time the research was conducted. Due to a lack of class time, which was mostly dedicated to instructing learners exclusively for the reading and grammar parts of the TOEFL, the instructor decided to implement the use of Criterion outside of the class as a self-study writing tool.

Another six students (Group 2) were recruited from different national universities as a comparison group by taking into account their English proficiency and motivation, which were considered equivalent to the experimental group students in Group 1. In particular, the average Test of English for International Communication (TOEIC) score of Group 1 was 639 and the average in Group 2 was 661. The students were mostly in their first- year and second year when the study was conducted. The majors of the students in both groups varied from Science to Humanities. All instructions on exactly how to carry out the Criterion writing tasks were distributed to the student participants in Japanese via an instruction sheet before the study began. Figure 1 depicts the procedure of Criterion writing tasks in this study.

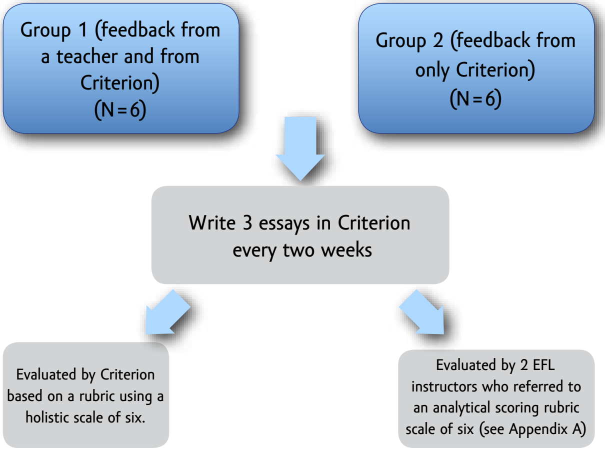
Group 1 (feedback from a teacher and from Criterion) (N=6)
Group 2 (feedback from only Criterion) (N=6)
Write 3 essays in Criterion every two weeks
Evaluated by Criterion based on a rubric using a holistic scale of six.
Evaluated by 2 EFL instructors who referred to an analytical scoring rubric scale of six (see Appendix A)

Figure 1. The procedure of the writing task evaluation

As shown above, students in both groups wrote essays in response to three writing tasks that were provided by the Criterion prompt pool, totaling 36 essays in all (the three prompts used in this study are in Appendix B). The students were asked to write an essay on the Criterion website. No time limit was placed on the students, but they were all encouraged to respond to the teachers' feedback soon after receiving it in order to improve on their second and third drafts. The feedback to the participants in Group 1 was given two times: after their first and second drafts. The teacher provided feedback with the use of the comment functions in Criterion itself, while she also used the pop-up feedback function built into Criterion in order to make specific corrections on vocabulary and grammar directly in the text of an essay. The feedback comments focused on the following four rhetorical features: thesis statement, topical development, propositional development, and awareness of readers, referring to the operational definitions of coherence by Lee (2002). For instance, in the comment section, the teacher wrote comments such as "Explain the usual classroom situation in Japan before making your opinions" when she noticed that the writer did not consider the readers' understanding.

As for the pop-up function, she sometimes corrected ungrammatical sentences directly by guessing what the writer was trying to say. However, the focus of her feedback was on the coherence of the students' writing so that the grammatical corrections in the pop-up functions were limited to local errors such as tense and subject-verb disagreement. Group 2 were also required to submit three drafts of each essay; the main difference being they did not receive any feedback from the teacher. However, they were strongly encouraged to check the e-rater's *View Trait Feedback Analysis* feedback and scores provided by Criterion.

All thirty six essays were evaluated by Criterion and two experienced EFL instructors, one of them being an author of this study. The two EFL instructors marked all thirty six essays using the six-point scale that focuses on the rhetorical features discussed earlier. The instructors did not know to which students the essays belonged. The inter-rater reliability between the two instructors' scores was confirmed to have a value of .78. Their scores were summated and a mean score for each essay was then calculated.

5. Analysis procedures

The data were analyzed quantitatively making use of the two evaluations: evaluated by e-rater and the two EFL instructors. First, descriptive statistics for the holistic scores evaluated by both Criterion and the two researchers were examined in order to see the changes over the three writing tasks. Next, the scores evaluated by the two instructors focusing on the four rhetorical features mentioned above were used to identify which rhetorical feature changed over the three writing tasks, and these were compared with the scores evaluated by Criterion.

6. Results

Table 2 summarizes the descriptive statistics of the results of the students' three writing tasks evaluated by both Criterion and the researchers on a six-point scale. Please refer to Appendix C for a complete table of the raw scores of all of the participants in this study. A correlational analysis reveals that there are statistically significant correlations between Criterion and the two researchers' evaluations, with a value of .46 at the .01 level.

Table 2: Descriptive statistics of the participant students' writing tasks

	Criterion's Evaluation Mean (SD)			**Researchers' Evaluation Mean (SD)**		
Writing #	**1**	**2**	**3**	**1**	**2**	**3**
Group 1 (feedback from a teacher)	3.8 (.75)	4.6 (.81)	4.5 (.83)	3.6 (.18)	4.6 (.63)	4.5 (.92)
Group 2 (no feedback from a teacher)	4.1 (.75)	4.3 (.51)	4.1 (.75)	3.9 (.76)	3.6 (.40)	4.4 (.77)

As indicated in the table, while both evaluations showed a higher mean score for Group 2 in Writing Task 1, Group 1 showed a higher mean score than Group 2 in Writing Task 2 and 3 (henceforth referred to as Writing 1, 2 and 3, respectively).

In order to determine which rhetorical features were sensitive to the teacher's feedback, the mean scores obtained from the descriptive statistics were compared between Group 1 and Group 2. Table 3 shows the descriptive statistics of the twelve participant students' writing scores evaluated by the two EFL instructors focusing on the four rhetorical features.

Table 3: Descriptive statistics of the scores of participant students' writing tasks on rhetorical features

Writing	Thesis statement			Topical development			Propositional development			Awareness of readers		
task #	1	2	3	1	2	3	1	2	3	1	2	3
Group 1	3.7 (.85)	4.9 (.66)	4.8 (1.1)	4.3 (.88)	4.6 (.77)	4.5 (.90)	3.6 (.88)	4.1 (.83)	4.0 (.66)	3.5 (.92)	4.8 (.71)	4.6 (.98)
Group 2	4.0 (.95)	3.8 (.71)	4.3 (1.4)	3.6 (1.4)	3.8 (1.0)	4.0 (1.6)	3.3 (1.3)	3.4 (.51)	3.6 (1.9)	3.7 (1.3)	3.5 (.67)	3.7 (1.8)

As shown in the table above, it was found that Group 1 showed a higher improvement in all the four rhetorical features than Group 2, and especially between Writing 1 and Writing 2. Taking a closer look at the table, it was noticed that while Group 1 indicated great improvement in *thesis statement* and *awareness of readers* between Writing 1 and Writing 2, Group 2 did not show any improvement in these features.

7. Discussion and pedagogical implications

From these results, it was found that Group 1 (the group that received feedback from a teacher) scored higher than Group 2 (the group that received no feedback from the teacher) on the essays scored by both Criterion and the two instructors in this study. Further, we can see some differences between two of the rhetorical features used in evaluating the learners' essays. Thus, it is possible to state that the group that received feedback from Criterion and a teacher improved their writing ability over time to a greater extent than the group who only received automatic feedback.

As previously mentioned, participants in both groups were given an instruction sheet in Japanese about how to use the Criterion system. However, we must state that there is no way to determine whether the students in either group actually used the system for feedback correctly or even used it at all. Even though we strongly encouraged these participants to refer to the feedback on the system regarding their writing, we cannot be sure if they did. Further research should endeavor to interview participants to glean how much they actually used the Criterion system to gain feedback in order to subsequently improve their writing.

As far as RQ1 is concerned, it was found that, as a whole, and when used regularly over a period of time, Criterion as a self-learning writing practice tool can be consistent with evaluations by human raters. Even though Criterion does not evaluate the content of writing, it can evaluate the same constructs of content from the results of the instructors' evaluation in this study. Therefore, Criterion's evaluations were deemed to be consistent with the teachers' analytical scoring focusing on the rhetorical features prevalent in Criterion. The students who received only feedback from Criterion did not demonstrate as much improvement on their writing as the students who received feedback from both Criterion and a teacher.

Regarding RQ2, despite the reliability of scoring in Criterion, using it without a teacher's feedback did not support overall improvement in the three writing assignments in this study. The primary reason for this is that the participants regarded the feedback from the teacher on Criterion as direct feedback from the teacher herself. That is, even though the teacher did not provide any instructions in class concerning Criterion use, the students who participated in this study recognized that she was the one who gave feedback on their writing tasks in Criterion. This lends credence to the importance of the *awareness of readers* category in the coherence rubric in this study. That is, as with most situational uses of language, users are most successful when they have an awareness of the context and social dimension of how the language should be used (Weir, 2005). Thus, we can state that there is, to some degree, a correlation between human and automatic scoring.

Finally, when we look at RQ3, this study reveals that rhetorical features are generally not improved upon in Criterion without feedback from a teacher. While Group 2 did not show a noticeable improvement over the three writing tasks in any of the rhetorical features, Group 1 showed great improvement in *thesis statement* and *awareness of readers* over the three writing tasks. Even though the students who participated in this study were informed of the online feedback evaluation of *Organization and Development* in Criterion from an information sheet distributed to them at the beginning of the study, they might have failed to read it carefully. It is also possible that the students did not comprehend Criterion's feedback very well. From this we can state that clear explanations, and even instruction on the use of the feedback functions in Criterion are necessary. It is here that EFL teachers can explicitly guide their learners through the features of Criterion, for, as Chen and Cheng (2008) argue, without such guidance, the effectiveness of the feedback given by any type of AWE may not serve its intended purpose.

Further to this point, a teacher's involvement is still necessary even if Criterion is used as a self-learning tool. As explained earlier, Criterion has a comment function which requires a teacher to make comments on students' writing. We believe that by making use of this function, teachers can focus on clarifying comments on the *thesis statement* and *awareness of readers* features. When teachers focus on these specific areas, especially after the first writing task, students will be well aware of what is necessary to compose quality writing. They will, in turn, be able to demonstrate improvements as they work toward their final drafts, thus improving the overall quality of their writing (Chodorow, Gamon & Tetreault, 2010). In addition, direct instruction from a teacher plays a key role in this discussion. Similar to what Beatty (2003) argues, an over-reliance on feedback from an AWE alone will not lead to great improvements on successive drafts of writing. So, we believe that the use of Criterion must be accompanied by constructive feedback by knowledgeable EFL teachers.

Lastly, because of the improvement shown by Group 1 in this study, the authors believe that Criterion might actually have encouraged the learners in this group to practice their writing outside of the classroom. Although we cannot categorically state this to be true, as it was beyond the scope of this study to investigate this factor, future work in this area can include interviews with the participants in order to determine how much out-of-class writing practice actually occurred at this stage in the writing process. This rise in self-study outside of the classroom should be encouraged by teachers; however, the role of the teacher in providing feedback to this out-of-class work is still important.

It is here that the methods employed in this study are reflected in the results. First, the teacher providing feedback to the learners played an integral role in guiding the students through the use of Criterion and navigating the feedback given on the system. Further, the use of participants who were highly motivated to study English on their own time was deemed to be essential to the comparison between the teacher feedback group and the non-teacher feedback group. This argument augurs well for motivated language learners who have a clear reason to be studying English: if they can pursue their goals with the assistance of their teachers, they will be more likely to be content in what and how they are learning (LaGuardia, Ryan, Couchman, & Deci, 2000; Ryan & Deci, 2000). Seeing as some differences were found after the teacher's feedback between Writing 1 and 2 in regard to two of the rhetorical features used in this study, we believe that the participants in Group 1 were motivated by both the use of Criterion and also by the teachers' feedback. This is significant because it underlies the impact of direct feedback from a teacher on students' writing; often seen as the key to learners improving their writing from one draft to another (Ferris et al., 1997; Hyland, 2003; Hyland & Hyland, 2006).

Finally, we should note that there are some limitations to the current study. First, the number of participants was relatively small, as was the number of Criterion prompts used. Future studies could utilize a larger sample size and a greater variety of prompts. Second, in addition to the point raised above regarding whether the students in Group 2 actually used Criterion to gain feedback, the same could be said for subsequent drafts of work in Group 1. However, since one of the authors of this study provided guidance to Group 1 through the three drafts and strongly encouraged them in class to refer to the feedback on Criterion, including adding her own feedback to the Criterion system, we are relatively certain the participants in this group used both her feedback and the feedback on the system itself to improve subsequent drafts of their writing.

As a result of these limitations, we can suggest some larger follow-up studies that we intend to undertake. First, we would like to focus on whether students can improve their writing in successive drafts of the same essay by using the feedback provided by Criterion before they get feedback from a teacher. Examining scores over successive drafts would provide evidence as to whether students pay attention to features identified by Criterion before they get teacher feedback, and might help show whether the teacher feedback is efficacious in itself or if it can help teach students how to take advantage of the automated feedback in Criterion. A further idea that should be pursued is to give one group of students feedback only, while giving a second group of students focused instruction on how to use the Criterion system for feedback purposes. The differences between these two groups would highlight the effectiveness of Criterion. Finally, we would also like to conduct interviews and distribute surveys to student participants to glean how much they actually used the Criterion system when making revisions to their papers.

8. Conclusion

The use of Criterion as an e-learning and improvement tool has been well documented in the literature (i.e. Attali, 2007; Attali & Bernstein, 2006; Chen & Cheng, 2008). However, while Criterion has generally been regarded to be as valid as human raters in evaluating EFL learners' written work, there are some limitations to its use. For instance, as argued by Ware and Warschauer (2006), e-feedback might have underestimated the learning that takes place on the social and interactional plane. Seeing as EFL students cannot learn a second language in a vacuum, and must take into account the social and contextual aspects of the language (Weir, 2005), any attempt to effectively teach second language writing must include these elements. This is one area in which Criterion is lacking, as the participants who only received feedback from Criterion did not show much improvement in the four rhetorical features that were the focus of this study.

From this discussion we can glean some pertinent results from this research. While EFL learners can rely on Criterion to provide useful feedback to them, a teachers' feedback on early drafts of written work is essential for learners to be able to substantially improve their writing – and especially in terms of the rhetorical features dealt with here. In this study, we focused on an amalgamation both Criterion's *Organization and Development* and Lee's six categories of coherence (2002) to form four rhetorical features that are deemed to be essential in teaching and learning second language writing. The largest improvement was seen in *thesis statement* and *awareness of readers*, lending credence to the notion that instruction from a teacher is of the utmost importance when learning to write. Essentially, taking a teacher's feedback into account when writing successive drafts of essays can lead to great improvements in writing (Hyland, 2003; Hyland & Hyland, 2006). This is especially important for learners preparing to take the TOEFL iBT, as decisions on this test can greatly affect their future academic lives.

This study can contribute to the existing literature in a few ways. First, the human factor in evaluating our learners' written work cannot be overstated. While Criterion can assist students in improving successive drafts of their written work (Chen & Cheng, 2008), a teachers' guidance is especially important when formulating a thesis statement and attempting to put together an essay that takes into account the readers of that essay. Thus, e-learning tools such as Criterion are undoubtedly of use for EFL learners, but, as Hyland (2003) points out, they cannot replace proper instruction from actual teachers.

Finally, since Criterion can effectively gauge the *thesis statement, topical development* and *propositional development* in an essay, it is evident that the *awareness of readers* feature is especially important because of the content and context involved in this feature. That is, since Criterion cannot effectively judge the content of an essay, it is left to teachers to guide their learners during writing instruction on how best to write for a specific audience. This is how we, as EFL teachers, can be of great assistance to our learners, and is also one fundamental area in which Criterion falls short.

References

American Educational Research Association (2015). Standards for reporting on empirical social science research in AERA Publications. Retrieved from http://www.aera.net/Publications/StandardsforResearchConduct/tabid/15746/Default.aspx.

Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL essays. *ETS Research Report No. RR-07-21.* Princeton, NJ: Educational Testing Services.

Attali, Y., & Bernstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning and Assessment, 4*(3), 1–30.

Beatty, K. (2003). Teaching and researching computer-assisted language learning. New York, NY: Longman.

Becker, A. (2010). Examining rubrics used to measure writing performance in US intensive English programs. *The CATESOL Journal, 22*(1), 113–130.

Bitchener, J., & Knoch, U. (2009). The relative effectiveness of different types of direct written corrective feedback. *System, 37*, 322–329.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267–296.

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

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal, 93*(4), 47–52.

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.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7–24.

Education Testing Service: Criterion. (2013). Retrieved from http://www.ets.org/s/criterion/pdf/9286_CriterionBrochure.pdf

Enright, M., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing, 27*(3), 317–334.

Erlam, R., Ellis, R., & Batstone, R. (2013). Oral corrective feedback on L2 writing: Two approaches compared. *System 41*(2), 257–268.

Ferris, D., Pezone, S., Tade, C., & Tinti, S. (1997). Teacher commentary on student writing: descriptions and implications. *Journal of Second Language Writing, 6*, 155–182.

Heilman, M., & Tetreault, J. (2012). *Using automated scoring to analyze student writing*. Paper presented at Georgetown University Roundtable on Languages and Linguistics (GURT) 2012, Washington, DC.

Hyland, K. (2003). *Second language writing*. New York: Cambridge University Press.

Hyland, K. & Hyland, F. (2006). Contexts and issues in feedback on L2 writing. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp.1–19). Cambridge, UK: Cambridge University Press.

Jacobs, H. L., Zingraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*, Rowley, MA: Newbury House.

Keith, T. Z. (2003). Validity and automatic essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–167). Mahwah, NJ: Lawrence Erlbaum.

La Guardia, J. G., Ryan, R. M., Couchman, C. E., & Deci, E. L. (2000). Within-person variation in security of attachment: A self-determination theory perspective on attachment, need fulfillment and well-being. *Journal of Personality and Social Psychology*, *79*(3), 367–384.

Lee, I. (2002). Teaching coherence to ESL students: a classroom inquiry. *Journal of Second Language Writing*, *11*, 135–159.

Lim, H., & Kahng, J. (2012). Review of Criterion. *Language Learning and Technology, 16*(2), 38–45.

Long, R. (2013). A review of ETS's Criterion online writing program student compositions. *The Language Teacher*, *37*(3), 11–18.

Loucky, J. P. (2004). Maximizing vocabulary development using an online semantic keyboard approach. *Calling Japan, 12*(1), 7–20.

Nielson, K. (2011). Self-study with language learning software in the workplace: What happens? *Language Learning & Technology, 15*(3), 110–129.

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the Construct-Coverage of the e-rater® Scoring Engine. *ETS Research report January 2009*. Retrieved from http://www.ets.org/research/contact.html

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly, 20,* 83–93.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development and well-being. *American Psychologist,* 55, 68–78.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–483). Mahwah, NJ: Lawrence Erlbaum Associates.

Ware, P. D., & Warschauer, M. (2006). Electronic feedback and second language learning. 105–122. In K. Hyland & F. Hyland (Eds.). *Feedback in second language writing: Contexts and issues* (pp. 105–122). Cambridge, UK: Cambridge University Press.

Warshchauer, M. (1997). Comparing face-to-face and electronic discussion in the second language classroom. *CALICO Journal, 13*(2&3), 7–25.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335–353.

Weir, C. J. (2005). Language testing and validation: An evidence-based approach. New York: Palgrave MacMillan.

Appendix A

Rubric used by the researchers (Score 1–6)

Rhetoric features	**Score**
Thesis statement (Is the thesis clearly stated?)	
Topical development (Is the thesis developed well with reasons/examples?)	
Propositional development (Is the proposition logically developed/described?)	
Awareness of readers (Does this essay explain the context of situation well?)	

1	2	3	4	5	6
Poor	Fair		Good		Excellent

Appendix B

Writing prompts

Prompt 1: Why study abroad?

Many students choose to attend schools or universities outside their home countries. Why do some students study abroad? Use specific reasons and details to explain your answer.

Prompt 2: 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.

Prompt 3: 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?

Appendix C

The results of the scores evaluated by Criterion and the two researchers

Student #	Group #	Criterion Writing 1	Criterion Writing 2	Criterion Writing 3	Researcher A Writing 1	Researcher A Writing 2	Researcher A Writing 3	Researcher B Writing 1	Researcher B Writing 2	Researcher B Writing 3
1	1	4	5	4	3.5	4.25	3.75	4.25	4	5
2	1	4	5	5	5.5	5.25	5.25	4.25	5.75	5.75
3	1	5	6	6	3	4.75	5.5	4.25	5.5	4.75
4	1	3	4	4	3.5	4	4.25	3.75	3.5	4.5
5	1	4	4	4	3.25	4.75	4.5	4	4.75	4.5
6	1	3	4	4	3.5	4.5	3.75	3.25	4.75	2.75
7	2	4	5	4	3.75	3	4	2.75	4.5	4.75
8	2	3	4	4	3.25	3.5	5.25	2.5	4.5	5.5
9	2	4	4	3	4.5	3.25	3.25	4.75	4.5	3.5
10	2	4	4	4	4.25	3	3.75	3.75	2.75	4.25
11	2	5	5	5	4.75	4	5.5	5	3.75	5.75
12	2	5	4	5	4.25	3.25	3.25	3.75	4.25	4