

» An Analysis of the Use of *Criterion* in a Writing Classroom in Japan

Junko Otoshi, Ritsumeikan University, Japan

Criterion is a web-based writing assessment system developed by English Testing Service (ETS) which automatically provides learners with feedback as well as a score on essays that are written using the system. This descriptive study examined 28 Japanese adult students' TOEFL writing essays to explore what Criterion can and cannot do with regard to providing feedback on the essays. Criterion's critique function was compared with a human instructor's error feedback focusing on five error categories: verbs, word choice, nouns, articles, and sentence structures. The results revealed that Criterion experienced difficulties in detecting errors in all of the five categories. Through the study, it is suggested that Criterion should be implemented into a classroom with a team of teachers rather than as a standing-alone evaluation tool.

Providing error feedback to students' essays is very complicated issue in second language writing pedagogy. As seen in the critical debate between Truscott (1999) and Ferris (1999), research results and theories over error feedback are mixed. There are researchers who did not find significant effects from instructors' error feedback (e.g., Cohen & Robbins, 1986; Polio et al., 1998), while on the other hand, others (e.g., Robb et al., 1986; Frantzen & Rissell, 1987) have argued for the effectiveness of error feedback, demonstrating reductions in errors in students' essays.

Since the wide-ranging acceptance of process-oriented approaches in second language writing, the issues of writing instruction have been predominantly what and how error feedback should be given to the students (e.g., Celce-Murcia & Larsen-Freeman, 1999). In reality, error feedback is a time-consuming job for writing instructors, particularly in English as a Foreign Language (EFL) contexts. Because EFL classes often involve more than 30 students in class, it is almost impossible to give detailed feedback to each student's essay (Benoit, 2003). Complicating matters further is the fact that, as discussed earlier, teachers are often left with some degree of uncertainty regarding the effectiveness of error correction given that the literature into error feedback has shown mixed results concerning the effects on student's writing.

Despite both theoretical and practical concerns of second language writing pedagogy, many Japanese people learn English writing to prepare for TOEFL Writing Essays (TWE). According to the data of the developers of TOEFL and TWE, English Testing Service (ETS), about 120,000 Japanese have taken TOEFL to apply for universities in

the United States or Canada over the past five years and the TWE has been mandatory since 2000. For those learners, improving the score on the TWE is a major purpose for studying English writing. Currently, TOEFL has been shifted to computer-based formats in Japan and test-takers also need to take the TWE on the computer screen. In computer-based tests, test-takers need to complete an argumentative essay on a given assigned topic in 30 minutes. After 30 minutes, the screen is automatically closed and no revision is allowed. Therefore, time is one of the most significant factors for test-takers. In TWE writing, test-takers need to be concerned with both writing fluency and accuracy in the assigned essay.

As a writing evaluation tool, *Criterion*, a web-based writing assessment tool, has been developed by ETS. *Criterion* gives a holistic score and feedback to students about their essays using natural language processing. Students can instantly receive their score and also can receive feedback of their essay on the screen. *Criterion* comprises two applications: *E-rater* and *Critique*. *E-rater* provides score extracting linguistically based features from an essay and uses a statistical model of how these features are related to overall writing quality to assign a holistic score to the essay. On the other hand, *Critique* is comprised of a suite of programs that evaluate and provide feedback on errors in grammar, usage, and mechanics, identify the essay's discourse structure, and recognize undesirable stylistic features (Burstein et al., 2004).

Furthermore, *Criterion* also has two applications for human instructors: Pop-ups and Instructor's Commentary. Using these two applications, human instructors can freely give the learners feedback both in a holistic and an analytical way. The Pop-ups function is a highlighted box on the essay screen where instructors can usually write a brief correction concerning sentence level errors. Students can confirm this correction by hitting the "I" marks. In the Commentary, instructors can write a short paragraph under each student's essay, where instructors will often make a general comment on the overall quality of the essays.

As can be seen by the functions of *Criterion* above, this writing assessment tool seems systematic, organized, and can be expected to have a positive influence in a process-writing approach, providing students with instant feedback.

However, although this tool has been utilized in writing instruction in Japan for the past three years, the actual use of *Criterion* has not received much attention in the literature. Despite its apparent effectiveness, it must be noted that complicated error feedback issues cannot be resolved easily with a new technology. Therefore, *Criterion*'s assessment should be carefully examined to see where it is has potential limitations in a language learning context.

This study deals with *Criterion*'s ability to provide feedback on sentence-level errors in essays written using the system. Specifically, it seeks to determine which sentence errors *Criterion* can detect and which it has difficulties with, and how its error detection differs from that of human instructors. It will be valuable to give specific data of *Criterion* to improve its functionality and to show the best use of the application through micro-level research. As Burstein et al. (2003) argues, *Criterion* is intended to be an aide, not a replacement for teacher writing assessment/error feedback. Therefore, both researchers and practitioners should have critical attitudes toward *Criterion* when they use it. In doing so, it will be clear for us on how to implement *Criterion* in second language writing pedagogy.

The Study

This study was a descriptive study aiming at exploring how *Criterion* feedback differs from the human instructors' feedback. In order to obtain specific data, the following five major error categories were examined considering error feedback issues: verb errors, noun ending errors, article errors, wrong word, and sentence structures. These five categories have been considered as major errors in much of the second language writing literature and the description of each category below was based on Ferris' study (2001), as shown below.

Table 1: Description of Error Categories Used for Feedback and Analysis

Error Type	Description
Verb errors	All errors in verb tense or form, including relevant subject-verb agreement errors.
Noun ending errors	Plural or possessive ending incorrect, omitted, or unnecessary; includes relevant and subject-verb agreement errors.
Article errors	Articles or other determiner incorrect, omitted, or unnecessary
Wrong Word	All specific lexical errors in word choice or word form, including preposition and pronoun errors. Spelling errors only included if the (apparent) misspelling results in an actual English word.
Sentence Structure	Errors in sentence/clause boundaries (run-ons, fragments, comma splices), word order, omitted words or phrases, unnecessary words or phrases, other unidiomatic sentence construction.

This study was conducted using 28 Japanese adult students' TWE essays. They were enrolled in a large private language institution in Tokyo between April in 2003 and March in 2004. Most of them were taking the TWE course to apply for graduate programs in American universities. The average age of the subjects was 27 years old.

In TWE course at the institution, after the students attended classroom sessions, they took TWE in *Criterion* and received the feedback from the human instructors within a week. A native-speaker English instructor and a Japanese instructor cooperated in giving the error feedback. The feedback was mainly direct correction of sentence-level errors; however, a coding system was also used for major five error categories above. The coding system was introduced to the students during the classroom sessions.

In terms of the error feedback procedure, the Japanese instructor gave error feedback using Pop-ups on students' TWE essay screens and the native-speaker instructor checked the Japanese instructor's corrections. Also, the native instructor wrote a general comment concerning the overall quality of the essays in the Commentary box. After having the feedback from the human instructors on the screen, the students took the TWE essay again with a different topic.

Data Analysis

First, in quantitative assessment of student writing, the average number of words, the number of words per sentence, and the score were measured. Regarding the comparison

of error feedback between *Criterion* and human instructors, the Japanese instructor (the author) classified the detected errors into the five error categories. The errors detected by *Criterion* were classified from the *Critique* section of grammar and mechanics. As for the errors detected by human instructors, classification was conducted by examining the Pop-ups.

In this study, error rate was measured. Error counts have been normalized by dividing the number of errors by the number of words and multiplying by a standard, which was set at 200 (see Biber, Conrad, & Reppen, 1988) for both *Criterion* and the human instructors' feedback.

Results and Discussion

Table 2 shows the average number of words, number of words per sentence, and score. As shown in the table, the average number of words per text was 209, the average number of words per sentence was 15.3, and the average score was 3.28 (maximum of 6).

Table 2: Average Number of Words, Numbers of Words per Sentences, and Score (Mean/SDs)*

Word Count	Number of Words per Sentence	Score
209.6 (61.01)	15.3 (3.9)	3.28 (0.71)

*N=28

Table 3 shows the error rate marked by *Criterion* and human instructors. As explained earlier, error counts have been normalized by dividing the number of errors by number of words and multiplying by the standard of 200. As shown in Table 3, there is a large difference between *Criterion* and the human instructors in detecting the major five surface errors. In all five categories, the human instructors detected more than *Criterion* did. Of particular interest is that *Criterion* did not detect noun errors at all in this study.

Table 3: Errors Marked by *Criterion* and Human Instructors (Means/SDs)*

Error Type	Criterion	Human Instructors
Verbs	0.47/ (0.64)	0.84/ (0.86)
Nouns	0.00 /(0.6)	0.94/ (1.21)
Articles	0.07/ (0.27)	2.00/ (1.95)
Word Choice	0.11/ (0.32)	2.32/ (1.15)
Sentence Structure	0.32/ (0.50)	6.31/ (2.90)

*N=28

In Table 4 through Table 8, each category was subdivided by classifying the error types. First, Table 4 shows the error types of verbs detected by *Criterion* and the human instruc-

tors. As seen in the table, while both *Criterion* and the human instructors detected subject-verb agreement, *Criterion* did not detect tense errors at all. Because the tense form changes depending on the context (Celce-Murcia & Larsen-Freeman, 1999), it is anticipated that it is difficult for *Criterion* to detect contextual verb tense errors.

Table 4: Error Types of Verbs Detected by Criterion and Human Instructors

Type of Verb Error	Criterion	Human Instructors
Subject-verb agreement	9	6
Ill-formed	3	9
Tense	0	10
Total	12	25

Table 5 shows the noun error detected by *Criterion* and the human instructors. No noun errors were detected by *Criterion*. On the other hand, the human instructors detected 24 cases which were all incorrect use of singular and plurals. Noun errors are considered surface errors; however, *Criterion* did not detect this type of error in this study.

Table 5: Error Types of Nouns Detected by Criterion and Human Instructors

Type of Noun Error	Criterion	Human Instructors
Incorrect use of singular and plurals	0	24
Total	0	24

Table 5 shows the article errors detected by *Criterion* and the human instructors. The two cases detected by *Criterion* were incorrect use of “these”. On the other hand, the human instructors detected various types of article errors by Japanese writers. It is well known that the command of articles is very difficult for Japanese English writers (e.g. Leki, 1991). Since the human instructors have acknowledged this point, it is assumed that they carefully detected the article errors which were made by Japanese learners.

Also, it is noted that Japanese instructor sometimes failed to detect article errors in the error feedback process. The native instructor could not help correcting both article errors by the students and the Japanese instructor. Since articles are echoed with writing fluency of discourse (Celce-Murcia & Larsen-Freeman, 1999), this is one category where a native speaker’s assistance is needed in error feedback.

Table 6: Error Types of Articles Detected by Criterion and Human Instructors

Type of Article Error	Criterion	Human Instructors
Incorrect Use	2	13
Omitted	0	35
Unnecessary	0	6
Total	2	54

In Table 6, wrong word errors detected by *Criterion* and human instructors are shown. The human instructors detected various types of errors in this category, too. In this category, *Criterion* detected some spelling, confused words, and word choice errors; however, it did not detect preposition and word form errors at all. Although they are both considered surface errors, *Criterion* did not have a function to detect them. On the other hand, the human instructors detected many word choice errors. In this study, the human instructors not only pointed out wrong words of the essay but also even rewrote more appropriate words considering the discourse context. The wrong word category of error is one of the major categories for which human instructors can contribute to error feedback.

Table 7: Error Types of Wrong Word Detected by Criterion and Human Instructors

Types of Word Error	Criterion	Human Instructors
Spelling	2	12
Confused words	1	0
Preposition	0	14
Word choice	2	27
Pronoun	0	9
Word form	0	4
Total	5	66

Table 8 shows the sentence structure errors detected by *Criterion* and the human instructors. The human instructors detected this error category most. In this category, it is noted that there were 75 cases which the human instructors considered as unclear types of errors. These cases involved a mix of variety sentence errors and unidiomatic expressions of English. Both the Japanese and the native English instructor tended to rewrite the sentences when they could guess the meaning from the context. In other words, human instructors were trying to understand what the writers wanted to say from the context.

Therefore, human instructors detected 75 cases and rewrote the sentences as well as giving error feedback.

Table 8: Error Types of Sentence Structure Detected by Criterion and Human Instructors

Error Types of Sentence Structure	Criterion	Human Instructors
Fragment	9	23
Unnecessary words and phrases	0	30
Omitted words	0	38
Word order	0	9
Run-on	0	1
Others	0	75
Total	9	176

Conclusion

This study was designed to explore what *Criterion* can do and what it cannot do when compared with human instructors' error feedback, focusing on major five error categories. As shown in the tables above, the human instructors detected errors much more than *Criterion* did. In particular, *Criterion* had difficulty in detecting errors with regard to nouns and articles. Although these errors are considered as surface errors, these errors are not rule-governed errors and non-native English writers need more time to control those errors in improving comprehensive English proficiency (Truscott, 1999; Reid, 1997). In the same way, the other three error categories also involve the issue of rule-governed and non-rule-governed errors (Celce-Murcia & Larsen-Freeman, 1999).

Additionally, *Criterion* in this study did not detect errors which relate to the context. *Criterion* was especially unable to detect errors in wrong word and sentence structure which on the other hand, the human instructors easily detected errors. These error categories are closely related to the discourse context. *Criterion* could not make best use of its *Critique* function in those categories, either.

From such findings in this study, it is difficult to say that *Criterion* can be used as a stand-alone error feedback tool. Based on the findings of this study, we need to consider how *Criterion* should be implemented into the practical settings. In this study, the subjects were advanced writers who aimed to apply for graduate programs in the U.S or Canada. It is assumed that they need to achieve a certain score of the TWE. In such a situation, the learners might not be too concerned about *Criterion's* critique function; they will just check the total score. Therefore, *Criterion* can be used as an evaluation tool. However, this aspect must be explored more by conducting learners' attitude survey toward the use of *Criterion*. It might be a good idea to investigate which function of *Criterion* learners are concerned

about. In doing so, researchers can discover how learners actually make use of *Criterion* in order to improve their writing score.

In regard to the cooperation with the human instructors, human instructors are still required to check surface errors as well as global errors because of the above mentioned complicated characteristics of errors. As many writing instructors have experienced, they are sometimes providing the students with right words/sentences hoping such feedback will eventually help learners' writing growth in a long run. As shown in Table 8, both the Japanese and the native English speaker instructor in this study rewrote several sentences in the Pop-ups. Even though such model sentences are not considered to be effective unless learners can understand them, human instructors cannot help introducing those sentences. The human instructors hope that showing the better sentences will eventually help the learners in the long run.

As Burstein *et al.* (2004) suggested, *Criterion* can never become a replacement of human instructor's feedback; rather, it should be considered as a guide to help the human instructors provide feedback based on the holistic score given by *Criterion*. In many English classrooms in Japan, writing instruction has attracted the least attention among the four language skills because it involves time-consuming work and less productive results in a short time. *Criterion* has the potential to alleviate some of the load on the teacher in this regard, as well as affording students opportunities for writing both inside and outside the classroom. In saying this, however, while various technical support tools for language instruction such as *Criterion* is expected in this century, both researchers and instructors should have a critical attitude towards using them through careful examination of the tools and their potential strengths and weaknesses.

References

- Benoit, R. (2003). 自由作文とCriterion におけるエッセイライティング指導. [Free writing and teaching essay writing through Criterion] TOEFL Mail Magazine. Retrieved July 2004.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterionism: Online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*.
- Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book* (2nd ed.). Heinle & Heinle Publishers.
- Cohen, A.D., & Robinson, M. (1976). Toward assessing interlanguage performance: The relationship between selected errors, learners' characteristics, and learners' expectations. *Language Learning*, 26, 45-66.
- Ferris, D. (2004). The "Grammar Correction" debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime....?). *Journal of Second Language Writing*, 13 (1), 49-62.
- Ferris, D., Roberts, B., (2001). Error feedback in L2 writing classes: how explicit does it need to be? *Journal of Second Language Writing*, 10 (3), 161-184.

- Ferris, D., (1999). The case for grammar correction in L2 writing classes: a response to Truscott (1996). *Journal of Second Language Writing*, 8, 1-11.
- Franzen, D., & Rissel, D. (1987). Learners self-correction of writing compositions: What does it show us? In B. VanPatten, T.R. Dvorak, & J.F. Lee (Eds.), *Foreign language learning: A research perspective* (pp. 92-107). Cambridge: Newbury House.
- Leki, I. (1991). The preference of ESL students for error correction in college-level writing classes. *Foreign Language Annals*, 24, 203-218.
- Polio, C., Fleck, C. & Leder, N. (1998). "If only I had more time": ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing*, 7, 43-68.
- Reid, J. (1997). Responding to ESL student language problems : Error analysis and revision plans. In J.M. Reid & P. Byrd (Eds.), *Grammar in the composition classroom* (pp. 118-137). Boston: Heinle & Heinle.
- Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly*, 20, 83-93.
- Truscott, J. (1999). The case for "The case against grammar correction in L2 writing classes": A response to Ferris. *Journal of Second Language Writing* 8, 111-122.