Criterion: Its effect on L2 writing

Ritsuko Ohta
Keio University

Reference data:

Ohta, R. (2008). Criterion: Its effect on L2 writing.
In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*. Tokyo: JALT.

Some writing instructors are interested in using *Criterion*, an online essay evaluation system, while others question its effectiveness. The present study explores this controversial issue by analyzing data collected from Japanese university students who enrolled in a semester-long TOEFL preparation class. First, to examine learner gains in writing quality (holistic ratings) and fluency (essay length) over time, statistical comparisons were made between the first and last submissions of the students' essays written on similar topics. It was found that students whose TOEFL scores were 500 or above significantly increased in writing quality and fluency, while those with scores below 500 did not. Second, to investigate student attitudes and views toward *Criterion*, a questionnaire was given at the end of the semester. Results showed that many students favored the automated system although they indicated the feedback features need improvement.

ETSが開発したオンラインライティング自動評価ツール(Criterion)の学習効果については教育者の間で賛否が分かれている。本研究では、TOEFL対策コースを１学期間（４ヶ月）履修した日本人学生のデータを分析、検証した。はじめに、英作文能力の伸びを調べるため、学期初めと最後に書かれたエッセイを質（総合的評価）と流暢さ（エッセイの長さ）の２つの尺度で統計的に分析した。その結果、TOEFLスコア５００以上を取得した学生のエッセイは質、流暢さ、両方において有意な増加が見られたが、スコア５００以下の学生のエッセイでは有意差は認められなかった。次に、Criterionを使用した感想を学期末にアンケートにより尋ねたところ、多くの学生がTOEFL対策クラスでの自動評価システム使用に賛意を示した。しかし、Criterionのフィードバック機能についてはさらなる向上が必要なことが示唆された。

A dilemma that many writing instructors are faced with is that they are unable to give writing assignments as frequently as they would wish; students benefit from writing, but this requires a classroom teacher to read and respond to 30 or more essays. As a solution to this problem, the Educational Testing Service (ETS) has developed an automated essay evaluation service known as *Criterion*. By using this online system, students can submit essays on topics assigned by their instructor and immediately receive an overall holistic score on a 6-point scale. In addition to instant scoring, *Criterion* generates diagnostic feedback on grammar, usage, mechanics, style, and organization/development as well as allows instructors to post their own comments within the system, both of which help students revise their essays (ETS, 2007). Some researchers have demonstrated the positive effect of these capabilities on student writing. Burstein, Chodorow, and Leacock (2003) reported that there was approximately 97% agreement on holistic scores between E-raters (*Criterion*'s scoring application) and human raters. Attali (2004) conducted a study on a data set of more than 9,000 essays from six to twelfth grade students in the US to examine the effectiveness of the feedback features of the system. Results showed that the students were able to reduce their error rates and improve their essay organization from the first draft to the revision. In contrast, some researchers have cast doubt on the value of *Criterion*. Sheehan (2001) stated that essay length might be an influential factor for holistic computer scores. Otoshi (2005) questioned the grammar feedback feature of *Criterion* based on her study that found the system could not detect as many errors as human instructors.

Accordingly, this study aims to investigate the impact of *Criterion* on student writing performance, qualitatively and quantitatively analyzing data collected from 43 Japanese university students in a semester-long TOEFL preparation class.

Method

Research questions

The present study addresses the following two research questions:

1. Does *Criterion* help improve the writing skills of Japanese university students at different levels of English proficiency?
2. Do students find it helpful to use *Criterion*?

Class design

The participants in this study were 43 Japanese university students who came from two TOEFL preparation classes each taught in the same way by the researcher for 4 months. They varied in major and grade (from freshman to senior). The mean TOEFL score for the students was 485.65. During the semester, they completed 13 essays (an essay per week, each written within 30 minutes) using *Criterion* at home: 10 first drafts on topics assigned by the instructor and 3 revisions for the first three topics. Only the first and last essay submissions were used for analysis in this study.

Analysis

Research question 1

Although 43 students were enrolled in the two classes in total, 32 students were sampled for this study. Two students who did not submit the last assignment and one student whose TOEFL score was too high (577) were excluded. In order to create two distinct proficiency levels, only students who held a score of 500 or above or below 486 on the TOEFL were chosen for analysis, which included 32 students. A score of 500 was used as a cut-off since many colleges consider this score the minimum level of acceptable performance. A score below 486 was chosen to clearly differentiate between the two proficiency levels based on an ETS report (1995) that in the case of the paper-based TOEFL score, the standard error of measurement is 14. Finally, the 32 students were divided into two groups according to proficiency, as measured by their TOEFL scores: The Upper Level (500 or above) and the Lower Level (below 486). The mean scores for the two groups were significantly different ($t = 10.17$, $p = .000$). Actual scores are included in Table 1.

Table 1. Students' TOEFL scores (N = 32)

Proficiency Level	Mean	SD
Upper Level (N = 11)	511.18	9.94
Lower Level (N = 21)	467.14	12.34

Accordingly, a total of 64 essays were analyzed for the present study: 32 first drafts on Topic 1 and 32 on Topic 10. The two topics were selected since they were similar in subject matter and prompt. The prompts were:

> Topic 1: Why do you think people attend college or university?
>
> Topic 10: Why do some students study abroad?

The two sets of data were compared statistically in terms of quality and fluency at each proficiency level. In order to measure change in writing quality over time, 6-point scaled holistic ratings (with higher being better) automatically generated by *Criterion* were used. *Criterion* evaluates a student's essay by comparing its linguistic features with those of the human-scored essays stored in the system's database (Burstein, et al., 2003). The fluency of student writing was measured by the number of words per essay, as suggested by Reid (1990).

Research question 2

To investigate student attitudes and views about *Criterion*, a questionnaire was conducted at the end of the semester. The questionnaire comprised six items asking the students their general opinions about using *Criterion*, the benefits of the system, their assessments of its feedback features, and the number of assignments given. The students were asked to answer all the questions but did not have to provide any personal information. The rationale for making the questionnaire anonymous was that it would encourage more honest, accurate responses. A disadvantage was that it would not permit analysis of differences in responses between the two proficiency groups. A total of 41 students completed the questionnaire since two students arrived late and did not participate in the survey.

Results and discussion

Research question 1: Does Criterion help improve the writing skills of Japanese university students at different levels of English proficiency?

Paired t-tests were used to assess improvement in the quality and fluency of student writing from the first to the last essay assignment. Table 2 shows the results for quality. The mean holistic scores for the Upper Level significantly increased between the first and last assignment. One intriguing finding is that the holistic score for Topic 10 was higher than 4. Because essays rated above 4 are considered good, the result might make a case for the positive influence of *Criterion* on the quality of student writing, at least with those whose TOEFL scores are over 500. On the other hand, the Lower

Level did not make any statistically significant progress in their writing quality. This may be because students at lower proficiency levels need more time to improve the quality of their writing or they misinterpret or do not understand *Criterion*'s written English feedback. In sum, this study seems to indicate that students with a TOEFL score of 500 or above are likely to enhance the quality of their writing using *Criterion*, whereas those holding a score of below 486 are not.

Table 2. Holistic scores (Quality)

Proficiency Level	Topic 1	Topic 10	t (df)	p
	Mean (SD)	Mean (SD)		
Upper Level	3.82 (.60)	4.45 (.52)	- 4.18 (10)	**
Lower Level	3.10 (.54)	3.38 (.92)	- 1.45 (20)	n.s.

**$p < .01$

The results for fluency are presented in Table 3. The Upper Level students wrote significantly more words on Topic 10. The average increase of approximately 70 words per essay should be encouraging, particularly for Japanese college or university students, many of whom have difficulty producing organized essays under timed conditions. The Lower Level students also increased their essay length, although the results were not significant. What should be noted here is that the *p* value was close to .05 (*p* = .057), whereas the standard deviation for Topic 10 was large. This means there was a great deal of variation in improvement in fluency among the students of lower English proficiency. What the results in Table 3 suggest is that the frequent practice opportunities offered by *Criterion* can help students write longer essays but that how much they can improve their writing fluency varies depending on their level of English proficiency. One could argue that mere practice without the system would develop the fluency of L2 student writing as well, and I would not preclude the possibility. However, in reality, it would be a daunting task for a classroom instructor to read and give feedback to the first drafts of 40 students' essays every week in parallel with checking their revisions and grading their final drafts, all of which can be instantly done by *Criterion*.

Table 3. Total number of words (Fluency)

Proficiency Level	Topic 1	Topic 10	t (df)	p
	Mean (SD)	Mean (SD)		
Upper Level	203. 64 (65.06)	271.27 (62.72)	- 6.55 (10)	***
Lower Level	155. 67 (41.70)	179.95 (70.15)	- 2.02 (20)	n.s.

***$p < .001$

Research question 2: Do students find it helpful to use Criterion?

The data collected from the questionnaire (N = 41) indicates that the students responded well to *Criterion*. Many students (88%) said they enjoyed using the program, while the majority of students (92%) suggested the instructor should keep using *Criterion* in TOEFL preparation classes. The top five responses to the question "What are some benefits of using *Criterion?*" are as follows:

- Can get immediate feedback on weaknesses (66%)
- Can prepare for TOEFL writing (63%)
- Can receive instant scoring (61%)
- Can become accustomed to writing in English (54%)
- Can use at home (51%)

It may be safe to conclude from the results that the students considered *Criterion* a useful preparation tool for TOEFL essays mainly because of its immediacy (instant scoring and feedback). Chandler (2003) notes this 'immediacy' appears crucial for improving student writing:

> Perhaps when ESL students can see their errors corrected soon after writing, they internalize the correct form better. Perhaps the greater cognitive effort expended in making their own corrections is offset by the additional delay in knowing whether their own hypothesized correction was in fact accurate. (p. 291)

Given that most instructors would have to spend at least one week checking students' first drafts, both instructors and students may benefit from using *Criterion*.

On the other hand, there seems to still be room for improvement in the feedback features of the system. As presented in Table 4, the students were not satisfied with the feedback on style or organization/development, all of which are related to content. The results seem plausible since *Criterion* does not detect flaws in logic nor does it give detailed comments on how to develop ideas. This implies instructors need to provide constructive feedback on these areas.

Table 4. Was each of the feedback features useful? (1: strongly disagree, 5: strongly agree)

Means (SDs) N = 41				
Grammar	Usage	Mechanics	Style	Organization/Development
3.61 (1.00)	3.44 (.98)	3.54 (.84)	3.24 (.94)	3.27 (.98)

Conclusion

Research question 1 asked about the longitudinal impact of *Criterion* on student writing. Analysis revealed that students with scores of 500 or above on the TOEFL at the start of the course *qualitatively* and *quantitatively* made significant progress in their writing, while those with a score below 486 at the start of the course did not show any significant improvement in either quality or fluency, although the result of the latter was not discouraging (an average increase of 24 words per essay, p = .057). The data collected from the questionnaire indicated that in general the students found it helpful to use *Criterion* despite the fact that they expected more explicit feedback on style, organization, and development (Research question 2). With the limit that the present study was conducted on a small scale, the following conclusions and pedagogical implications can be drawn:

1. Although students are positive about and motivated by using *Criterion*, not all of them can benefit from the system in the same way.

2. The use of *Criterion* is likely to yield long-term growth in students' writing quality and fluency if their TOEFL scores are 500 or above. However, even those students may require teacher feedback on the content of their essays.

3. Many practice opportunities offered by the automated system might help some students at lower proficiency levels to write longer essays, but they would not necessarily lead to a significant improvement in the overall quality of writing. At this level of L2 proficiency, instructors should guide students through the revising and editing steps of the writing process; they can post their own comments in Japanese within the system or communicate face-to-face so students can clarify the meaning of feedback given by *Criterion*.

As long as instructors understand the shortcomings of the system, *Criterion* can be an effective writing aid for students and a supportive instructional tool for teachers.

Ritsuko Ohta teaches at several universities. Her research interests include L2 writing assessment, writing development, and CALL.

References

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

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion^SM online essay evaluation: An application for automated evaluation of student essays. In J. Reid & R. Hill (Eds.), *Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence* (pp. 3-10). Menlo Park, CA: AAAI Press.

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.

Educational Testing Service [ETS]. (1995). *The official guide: TOEFL sample test* (6th ed.). Princeton, NJ: Author.

Educational Testing Service. (2007). *Criterion: User Manual*. [Online] Available: <criterion1.ets.org/cwe/News/Criterion%20User%20Manual%207.2% 206304.pdf>.

Otoshi, J. (2005). An analysis of the use of Criterion in a TOEFL writing program. *Ritsumeikan Gengo Bunka Kenkyu*, *16*, 305-313.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191-210). Cambridge, UK: Cambridge University Press.

Sheehan, K. (2001). *Discrepancies in human and computer generated essay scores for TOEFL CBT essays*. Unpublished manuscript.