English Teaching, Vol. 75, No. 1, Spring 2020, pp. 67-92
DOI: https://doi.org/10.15858/engtea.75.1.202003.67
http://journal.kate.or.kr

The Long-Term Effect of Automated Writing Evaluation Feedback on Writing Development

Young-Ju Lee*

Lee, Young-Ju. (2020). The long-term effect of automated writing evaluation feedback on writing development. *English Teaching*, *75*(1), 67-92.

This is a longitudinal case study using a mixed-methods research design to track how two Korean university students improved their English writing competence over one year with the aid of automated writing evaluation (AWE) program, Criterion. The participants wrote essays outside of class every month for one year, submitting first and later second drafts. The participants completed a TOEIC writing test at the beginning and end of the study; students' reflections on their writing development, obtained through interviews and journal entries, were also examined. A comparison of scores, errors, and quantitative measures of fluency and grammatical complexity indicated writing improvement. Both participants used Criterion feedback effectively to render informed judgments and valid corrections. Essay revision based on Criterion feedback yielded more self-directed learning and greater comfort with writing in content courses. It is suggested that the effect of AWE feedback transfers to long-term improvement. The results point to the potential benefit of AWE use in individual out-of-class writing practices.

Key words: automated writing evaluation feedback, second language writing development, Criterion, longitudinal study

* Author: Young-Ju Lee, Professor, Department of English Language and Literature, Hanbat National University, 125 Dongseo-daero, Yuseong-gu, Daejeon 34158, Korea; Email: yjulee@hanbat.ac.kr

Received 15 January 2020; Reviewed 21 February 2020; Accepted 29 February 2020

1. INTRODUCTION

Providing writing feedback is very important for writing instruction, because students can notice what grammar or lexical errors they tend to make, which will help to improve their writing accuracy. However, provision of manual writing feedback takes a lot of time and effort on the part of writing teachers. In light of this, automated writing evaluation (AWE) programs can be a reasonable option, since they can provide instant feedback to students. AWE programs such as Criterion, My Access, and Summary Street are currently available as classroom instruction tools and the best-known AWE program is Criterion. It has two complementary applications called *e-rater* and *critique* (Burstein, Chodorow, & Leacock, 2003). E-rater is a scoring application that assigns holistic scores on a 6-point scale. Critique analyzes errors in five categories: grammar, organization & development, usage, mechanics, and style (See Appendix A). Based on instant feedback in each area and an overall score, Criterion helps students write and revise essays.

Previous studies (Chodorow, Gamon, & Tetreault, 2010; Lee, 2015; Li, Link, & Hegelheimer, 2015; Moon & Pae, 2011) have focused on short-term effects of Criterion in classroom settings (e.g., 10-16 weeks). Results of previous studies have shown that Criterion feedback led to a decreased number of errors across drafts; however, whether the decreased error rate would transfer to long-term writing development has not yet been investigated.

This longitudinal case study evaluates the long-term effect of AWE over one academic year, using a case study approach and a mixed-methods research design. It investigates English writing development in two Korean students, supported by AWE feedback, over the study period. This is the first longitudinal study of Criterion use in individual out-of-class writing practice with a mixed-methods research design. Two Korean undergraduate students wrote one essay outside the classroom each month and revised their essays after reflecting on Criterion feedback. As an external index for writing development, the study employed a test-retest design: participants took a TOEIC writing test at the beginning and end of the study period. Essay scores and error count as reported by Criterion, and quantitative measures of fluency and grammatical complexity were also analyzed. In addition to quantitative data, students' reflections on their writing development were examined through interviews and journal entries. The following three research questions are addressed:

1. To what extent does students' writing improve over one year using Criterion, as indicated by TOEIC writing scores, essay scores, and error count?
2. How does students' writing proficiency, measured by fluency and grammatical complexity, develop using Criterion over one academic year?

3. How do students perceive their writing trajectory after using Criterion over one year?

2. LITERATURE REVIEW

2.1. English Writing Development

Previous studies have investigated English writing development due to studying in the ESL context, whether it is long-term development (Knoch, Rouhshad, Oon, & Storch, 2015; Li & Schmitt, 2009) or short-term development (Storch & Hill, 2008; Storch, 2009).

Storch (2009) explored the impact of studying in an L2-medium university on developing academic writing after one semester's study, without formal language support. She compared 25 students' essays before the beginning of and toward the end of the semester, using quantitative measures of fluency, accuracy, grammatical complexity, and lexical complexity. Although pre-post comparison of overall scores from 5.6 to 6.2 on a scale of 1 to 9 indicated improvement in writing skills, quantitative measures showed no change in fluency, accuracy, grammatical complexity, and lexical complexity. Storch attributed the lack of improvement in quantitative measures to the study's short time frame and the absence of feedback.

Knoch et al. (2015) investigated change in 31 international undergraduate students' writing proficiency after three years' study in an Australian university, using the institution's ESL proficiency assessment conducted at the beginning and end of their degree programs. Participants were also interviewed concerning their writing practices and their perceptions of their writing development. Results showed that writing proficiency assessment scores, both holistic and analytic, decreased from first to second assessment. Discourse-analytic measures showed that participants' writing improved in terms of fluency, measured by number of words used, but not in accuracy or grammatical or lexical complexity. During interviews, participants expressed that they felt their writing had not improved, because of insufficient opportunities to submit writing assignments and lack of feedback. Based on these previous findings, we know that studying in ESL contexts without the opportunity to revise based on corrective feedback may not automatically induce writing development.

The appropriate use of lexical phrases (native-like formulaic sequences such as *it should be noted* and *as a result*) is very important in English academic writing (Li & Schmitt, 2009). Li and Schmitt (2009) investigated lexical phrase acquisition in a case study of a Chinese MA student in the United Kingdom over 10 months. Written assignments (8 essays and a dissertation) were analyzed, and interviews were conducted with the participant after submission of each assignment. The participant received feedback on use

of lexical phrases in her essays and reflected on this feedback when writing subsequent essays. Frequency and diversity of lexical phrases did not increase consistently over the year; however, the proportion of lexical phrase tokens considered appropriate by NS judge panels increased from about 40% to 90%. The participant expressed increased confidence in her use of lexical phrases and indicated that she had benefited from the provided academic reading materials, explicit instruction in English for Academic Purposes (EAP) courses, and explicit feedback on use of lexical phrases in essays. Li and Schmitt's study differs from Storch (2009) and Knoch et al. (2015) in that the participant received feedback on her essays, which contributed to appropriate usage of lexical phrases.

In the EFL setting, only a few studies have been conducted on writing development. Kobayashi and Rinnert (2013) investigated a Japanese writer's L1, L2 (English), and L3 (Chinese) writing competence over two-and-a-half years, based on written essays, retrospective stimulated recall of pausing behavior, and interviews. To shed light on English writing development over time, three types of measures—fluency (words and characters per minute), sentence length, and lexical diversity (modified type/token ratio adjusting for text length)—were examined. The participant's fluency, lexical diversity, and sentence length were higher than those of experienced English writers, a comparison cohort group. The participant attributed her increased English fluency to self-initiated writing practice.

The literature shows a lack of research on English writing development under continued provision of feedback. Although previous studies have investigated short-term writing development in ESL context, few studies have tracked long-term writing development in the EFL context using a case study approach. We need more longitudinal case studies of English writing development in EFL settings, especially of the effect of self-initiated writing practice on writing competence over time. The current study addresses this gap.

2.2. Effect of Criterion Feedback

Previous studies on Criterion feedback (Chodorow et al., 2010; Lee, 2015; Li et al., 2015; Moon & Pae, 2011) have mostly investigated its short-term effect in classroom settings and students' views of the feedback's usefulness. Some studies (Dikli & Bleyle, 2014; Li, Link, Ma, Yang, & Hegelheimer, 2014) also examined consistency of either holistic scores or feedback types between instructors and Criterion. Previous studies (Chodorow et al., 2010; Li et al., 2015) have shown that Criterion feedback led to a decreased number of errors across drafts. Most AWE systems have been created and studied with native speakers in mind, and their use in EFL settings is little explored (Weigle, 2011, 2013). This section covers recent studies on Criterion in both EFL and ESL contexts.

Chodorow et al. (2010) investigated the short-term effects of Criterion feedback on student writing and compared the effects of article error correction feedback, employing Lipnevich and Smith's (2009) data, for 268 NS and 71 NNS students in an introductory psychology course in the US. Participants wrote essays in class and revised them based on Criterion feedback. A significant difference was found in the number of article errors between the first and revised essays for NNS students, but no significant difference was found for NS students.

Li et al. (2015) investigated Criterion feedback's impact on revision and writing accuracy in a university ESL writing classroom, as well as instructors' and students' views of Criterion feedback. Criterion was integrated into two academic writing courses in which students submitted four assignments over a semester. The number of draft submissions under each assignment was used to examine writing practices, and Criterion error reports were used to compare accuracy from the first to the final drafts.

The results of Li et al. (2015) showed that the average number of submissions increased from the first assignment to the second but decreased for the last two assignments. The average error rates were significantly lower from first to final drafts of the second assignment. One possible reason for the significant improvement of accuracy in the second assignment was that the instructors required the students to attain a certain minimum score from Criterion before submitting their papers. Students had to use Criterion repeatedly and received iterative feedback as many times as needed to attain the minimum score. Instructors generally perceived Criterion feedback as having potential but had negative reactions to the quality of actual feedback. Most students expressed high satisfaction with feedback on grammar and mechanics but mixed satisfaction with organization feedback.

In the EFL context, Lee (2015) explored 72 Korean university students' perceptions of Criterion feedback using surveys over one semester, hypothesizing that 1) satisfaction with feedback would differ by English proficiency; 2) participants would prefer Criterion feedback to teacher feedback; and 3) grammar and mechanics would be positively related to satisfaction. All three hypotheses were disconfirmed: t-test results showed no significant mean differences in satisfaction with Criterion feedback between low- and intermediate-level students (and no advanced students were included); students viewed teacher feedback as significantly more effective than Criterion feedback; and regression analysis showed that the positive relationships of grammar and mechanics to satisfaction were not significant. Organization & development feedback was positively related to satisfaction (R-squared = 0.128, $p < 0.01$), contrary to previous studies (Li et al., 2015; Moon & Pae, 2011), but Lee did not explore why.

Previous studies have generally focused on the short-term effect of the use of Criterion in classroom settings: that is, no longer than one semester. The literature review has shown a lack of research on the long-term effect of Criterion feedback on writing development,

especially in EFL settings and non-classroom situations, which this study will address. This study examines the development of two Korean undergraduate students' English writing proficiency over one year, using a case study approach and a mixed-methods research design.

3. METHODOLOGY

3.1. Participants

Two Korean undergraduates voluntarily participated in the study. They were enrolled in an academic writing course the semester before the study took place in which Criterion was used. Their profiles are presented in Table 1; pseudonyms are used. Based on TOEIC listening and reading scores, Song had advanced English proficiency and Kim belongs to the intermediate level.

TABLE 1

Profile of Case Study Participants

Name	Gender	Age	Grade	TOEIC Score	English Proficiency
Kim	Female	22	Junior	740	Intermediate
Song	Female	23	Senior	900	Advanced

3.2. Data Collection Procedures

Participants wrote essays outside the classroom every month for one year and submitted 12 first drafts, followed by 12 second drafts after reflection and revisions based on Criterion feedback. They wrote first drafts for about one hour without referring to a dictionary and they did not have access to the Internet. Writing prompts were on topics from the Criterion library: "Good Friend," "Peer Pressure," "Male and Female Roles," "Famous Person to Put on a Postage Stamp," "Beautiful Place," "Favorite TV Show," "Special Place," "Single-Sex Education," "Salary and Satisfaction," "After-School Jobs," "Reasons for Attending University," and "Choosing a Job."

3.3. TOEIC Writing

A test-retest research design was employed to gauge writing development; participants took two TOEIC writing tests one year apart. Participants took the TOEIC writing test twice on the same day and the official TOEIC writing test scores were used in the current

study. This test involves eight tasks and takes approximately one hour. Five tasks require writing a sentence based on a picture, two involve responding to a written request, and one involves writing an opinion essay. The test measures ability to use written English to perform communication tasks typical of daily life and the global workplace.

3.4. Interviews

To track writing development trajectories, the two participants were individually interviewed on three separate occasions at four-month intervals. Interviews were conducted in Korean, extensive notes were taken, and interviews were recorded for transcription. Each interview lasted approximately 30 minutes. Some questions overlapped across interviews. The first interview mainly elicited participants' perceptions of their writing proficiency, the usefulness of Criterion feedback, and their overall experience of writing. During the second interview, participants were asked about the effectiveness of Criterion feedback and how they perceived their writing skills after versus before using Criterion. The third interview encompassed the efficacy of Criterion use and how participants dealt with incorrect feedback, which may or may not lead to errors, and with writing in their courses. The interview questions are presented in Appendix B. The interviews were transcribed, and emerging themes relating to participants' perceived improvement or lack thereof were qualitatively analyzed.

3.5. Journal Entries

Participants wrote one journal entry in English immediately after submitting second drafts (for a total of 24 journal entries, 12 from each participant). Journal entries helped to investigate participants' perceptions of how their writing proficiency developed and how they understood, interpreted, and used Criterion feedback in revisions.

3.6. Data Analysis

Data were collected using the instruments and procedures described above. The essays and participants' scores were examined to compare the quality of first and second drafts across 12 essays. Essays were closely examined in terms of word count, T-units, and clauses to investigate fluency and grammatical complexity.

Qualitative data obtained during the interviews were analyzed. The participants' verbal protocols were also transcribed, analyzed, and summarized. For journal entries, responses were typed as they appeared, including misspellings and grammatical errors. An initial coding scheme was developed by investigating transcripts of the first interviews and

journal entries; subsequent data were coded according to this scheme, using NVivo 11.0. The coding scheme was then refined to combine subcategories into emerging themes.

3.7. Text Analysis of Essays

Quantitative measures of fluency and grammatical complexity were employed, which required all essays to be coded for T-units, clauses, and word count. The analysis of students' first drafts involved independent judgment of T-units and clauses. A second rater with a doctoral degree in English education and many years of teaching experience examined 30% of them. Pair-wise correlations were calculated in order to determine the inter-rater reliability between the researcher and the second rater. The inter-rater reliability for T-units and clauses, respectively, was 0.949 and 0.975, which are highly acceptable figures.

4. RESULTS

4.1. RQ 1: Three Indexes of Writing Development

4.1.1. TOEIC writing scores

As an external index of writing development, TOEIC writing scores before and after using Criterion are compared in Table 2.

TABLE 2

Change in TOEIC Writing Scores (Out of 200)

	Pre-TOEIC Writing Score	Post-TOEIC Writing Score	Increase
Kim	140	150	10
Song	150	190	40

Kim's initial TOEIC writing score of 140 increased by 10 to 150 after one year of using Criterion, and Song's higher initial score of 150 increased dramatically by 40 to 190. Overall, the results indicated improvement in writing skills.

4.1.2. Criterion scores

As a second index of writing development, Criterion essay scores (from 1-6) were examined. First-draft scores were employed, since it was deemed very likely that scores would increase from first to second drafts, after reflection on feedback.

To examine patterns of writing development, essays were also classified into three stages: Time 1 (essays 1-4), Time 2 (essays 5-8), and Time 3 (essays 9-12). Average first-draft scores for all three stages, provided by Criterion, were used to track participants' writing development. As can be seen in Table 3, participants' first-draft essay scores increased from Time 1 to Time 3. Kim's score increased by 0.25 from an average score of 4.25 on the first drafts of essays 1-4 to 4.5 on essays 9-12, while for Song, there was a slightly higher increase of 0.5 from Time 1 to Time 3.

TABLE 3

Change in Average Essay Scores

	Time 1	Time 2	Time 3
Kim	4.25	4.5	4.5
Song	4.5	4.5	5

Overall, essay scores show that Criterion afforded improvement in participants' writing skills over one year.

4.1.3. Error report

As a third index of writing development, error counts for four categories reported by the Criterion program-grammar, usage, mechanics, and style-were examined; organization & development feedback was excluded, as Criterion merely schematically identifies and checks whether a particular sentence constitutes or contains introductory material, thesis statement, main idea, supporting idea, conclusion, or transition. That is, Criterion simply highlights the first sentence in the essay and asks if it is the thesis statement, then highlights the first sentence in each paragraph and asks if it is the main idea, highlights the last sentence in the essay and asks if it is the conclusion, and so on. Therefore, students may tend to ignore organization & development feedback as simplistic and unhelpful.

Instances of Criterion error feedback during Times 1, 2, and 3 were added. Table 4 and Table 5 present changes in average error count for each participant.

TABLE 4

Change in Average Error Count: Kim

	Time 1	Time 2	Time 3
Grammar	2	3	5
Usage	13	13	18
Mechanics	6	2	3
Style	13	1	1
Total	34	19	27

TABLE 5

Change in Average Error Count: Song

	Time 1	Time 2	Time 3
Grammar	3	0	2
Usage	11	14	19
Mechanics	2	4	1
Style	1	0	0
Total	17	18	22

For Kim, the total error count decreased from Time 1 to Time 2 but increased at Time 3 from 19 to 27. The decrease in error count for Kim was striking, especially in the style category, from 13 to 1 from Time 1 to Time 3. For Song, overall error count increased steadily from 17 in Time 1 to 22 in Time 3; in particular, the usage category increased from 11 to 19. The increase can be partly attributed to the fact that Song added new content in her second drafts, as she discussed in her eighth journal entry (section 4.3.2.4). However, her average error count did decrease from Time 1 to Time 3 for grammar and style.

Overall, the two participants' proficiency in different aspects of writing improved. Kim's general writing proficiency improved, especially in style, while Song's proficiency improved in grammar and style.

4.2. RQ 2: Quantitative Measures of Fluency and Grammatical Complexity

Three measures of writing fluency were examined as indicators of proficiency development, adopting the approach of Storch (2009): the total number of words, the number of T-units, and the length of T-units. A T-unit in this study is defined as one independent clause and any dependent clause connected to it. Also, as an index of grammatical complexity, the ratio of clauses to T-units, also known as a subordination ratio, was examined. It generally increased in a linear relationship to the proficiency level, regardless of the task (Wolfe-Quintero, Inagaki, & Kim, 1998). The subordination ratio is designed to measure grammatical complexity: the higher the number of clauses per T-unit, the higher the complexity of writing (Wolfe-Quintero et al., 1998). For example, Casanave's (1994) analysis of L2 learners' journal entries showed 1.07 clauses per T-unit for beginner-level students and 2.17 for advanced-level students.

The two participants' progress on fluency and grammatical complexity measures from Time 1 to Time 3 was presented in Table 6. Figures 1, 2, and 3 show progress on three fluency measures: words, T-units, and T-unit length, respectively.

TABLE 6

Fluency and Grammatical Complexity Measures

	Time 1	Time 2	Time 3
Words			
Kim	323	348	353.5
Song	406.5	376	434.25
T-units			
Kim	33	31	30
Song	33.25	27.75	30
T-unit Length			
Kim	11.25	11.47	11.7
Song	12.4	13.8	14.5
Ratio of Clauses to T-units (C/T)			
Kim	1.32	1.45	1.62
Song	1.42	1.62	1.76

FIGURE 1

Change in Words

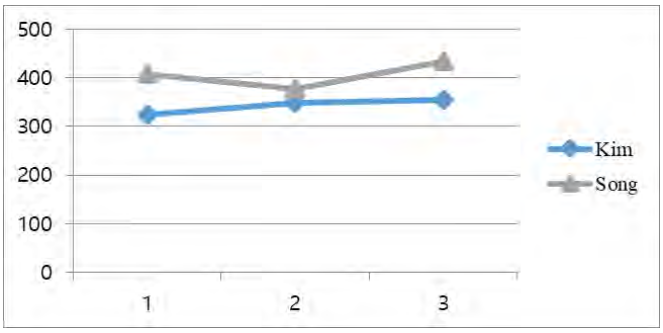


FIGURE 2

Change in T-units

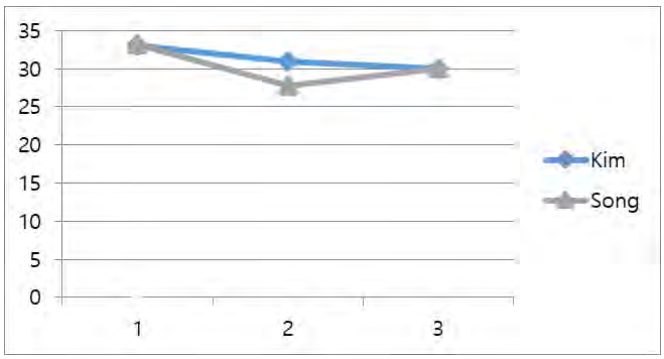
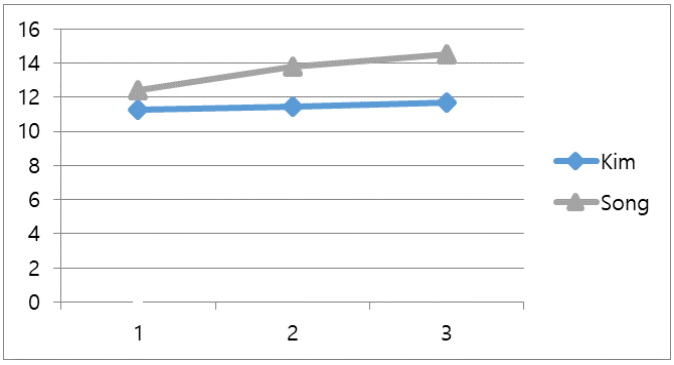


FIGURE 3

Change in T-unit Length

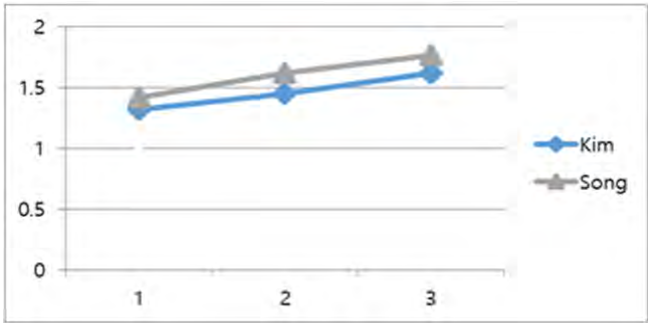


Overall, the participants' fluency measures increased from Time 1 to Time 3. In the case of words, it increased similarly for both participants from Time 1 to Time 3: from 323 to 353.5 for Kim and from 406.5 to 434.25 for Song. Regarding T-units, it decreased from 33 to 30 for Kim and from 33.25 to 30 for Song. As for T-unit length, it increased for both participants: for Kim, from 11.25 to 11.7, and for Song, from 12.4 to 14.5. That is, in the first four essays, Kim wrote an average of 323 words with an average T-unit length of 11, whereas in the last four essays she wrote an average of 353 words with an average T-unit length of 12. For Song, in the first four essays, she wrote an average of 407 words with an average T-unit length of 12, whereas in the last four essays she wrote an average of 434 words with an average T-unit length of 15.

Figure 4 shows the grammatical complexity measure from Time 1 to Time 3. The ratio of clauses to T-units increased for both participants from Time 1 to Time 3: for Kim, from 1.32 to 1.62; and for Song, from 1.42 to 1.76.

FIGURE 4

Change in Ratio of Clauses to T-units



Figures 5 and 6 offer a glimpse into how writing fluency and grammatical complexity developed from Time 1 to Time 3. Figure 1 displays an excerpt of Kim's first draft of essay 1, and Figure 2, of her essay 12. Essay 1 had a word count of 311 with an average T-unit length of 8.9, for a ratio of clauses to T-units of 1.14, while essay 12 had a word count of 349 and an average T-unit length of 11.3, so that the ratio of clauses to T-units was 1.45.

FIGURE 5

Excerpt of First Draft of Essay 1 (Kim)

Many people have friends. They often spend time or have a talk with friends. However, they sometimes feel lonely although they are with friends. The reason is these friends aren't good friends. Then, what is it good friends? I introduce three things about them.
First, good friends don't put relation of gain and loss. For example, just friends call to you for borrowing money or ask you for help to solve their assignments. Moreover, if they don't get something when I need to help, they won't give me a hand. However, good friends are different. They help you anytime and don't be concerned to gain something to you.

FIGURE 6

Excerpt of First Draft of Essay 12 (Kim)

Nowadays, people not only consider salaries, but also a variety of parts when they choose a job. The reason is they think that wages are not everything, so am I. Thus, if two different companies offered me similar jobs at similar salaries, I choose the company that meets the three parts I consider. The three parts are vacations, flexible company system and return-to-work. Now, I explain them in detail and why I think them.
First, it is the company that gives employees a lot of vacations. Holidays of most companies are only two or three days and workers sometimes don't go on vacations because of their boss. So, their efficiency is gradually declining and their tiredness is increasing. If the corporation make them to go on vacation and guarantees 5~7 days as holidays, they will work harder and the satisfaction of the company will be increasing.

Overall, the results of quantitative measures of fluency and grammatical complexity indicated improved writing proficiency for both participants over one year.

4.3. RQ 3: Participants' Perceptions of Writing Development

From close analysis of journal entries and interview data, five themes arose in the participants' perceptions of writing development: usefulness of Criterion feedback, self-directed learning, improved writing proficiency, comfort with writing in content courses, and overall confidence in writing.

4.3.1. Intermediate writer: Kim

The intermediate writer, Kim, had been learning English since the third year of elementary school, but she had not received substantial writing instruction in English until

entering university. Because of the multiple-choice format of the college entrance exam, Korean students receive limited English writing instruction in junior high and high school and have little chance to learn about English academic writing conventions until university.

At the time of the study, Kim was a junior majoring in English language and literature, and her institution's department offered some English-medium courses where students had to write short-term papers in English as course assignments. Initially, Kim expressed some difficulties completing writing assignments.

Kim received a TOEIC listening and reading score of 740. She was a very hardworking and conscientious student, who self-assessed her writing proficiency as intermediate and expressed concern about her lack of vocabulary in her first journal entry, as follows.

> My writing proficiency is intermediate. The reason is that I don't know many words well, so I often have to find a dictionary when I write an essay. Also, I don't write long sentence well because I'm confused how to write. (1st journal entry)

In her third journal entry, Kim mentioned that Criterion feedback, especially on grammar and style, was useful. She tried to write long sentences in response to "short sentences" feedback under the category of style. As she mentioned in her first journal entry, she usually did not write long sentences; short sentences feedback thus must have been useful for her. In contrast, she found that the repetition of words feedback was the least useful, because Criterion only indicated repetition of pronouns.

> I think grammar feedback is useful, so I could recognize my mistake and reduce it. The help of the Criterion made to develop me. After I continually received feedback of short sentence, I tried to write long sentences. So, I didn't receive feedback of short sentence in this third essay. I think repetition of words feedback is the least useful. Of course, to repeat same words isn't good. However, it comments only repetition of pronouns in this essay. So, I didn't change anything. (3rd journal entry)

Kim's negative perception of repetition of words feedback was again illustrated during the first interview.

> *The least useful feedback is repetition of words. The Criterion program points out how many times I used pronouns such as I, you, they, etc. I just ignore this feedback.*

In her seventh journal entry, Kim expressed that she had gotten into the habit of studying examples of sentences using provided words in a dictionary, which she found very helpful. The following journal entry shows her perception of self-directed learning.

> These days when I wrote an English writing, I always used internet dictionary. By using the dictionary, I was helped from examples of sentences as well as just looked up the word which I didn't know. I think that checking examples of sentences often is more helpful than just looking up the word that I don't know. (7th journal entry)

Kim also perceived that using Criterion had led her to refer to grammar books to find out if feedback was correct. Similar to the advanced writer, Song, Kim said receiving incorrect feedback was helpful because it made her search for the right answer in grammar books.

> The Criterion program gave two wrong feedbacks to me. I think not only wrong feedbacks give confusion to me. Also, it makes me find some grammar books. So, I think the wrong feedbacks have merits and demerits. (9th journal entry)

During the second interview, Kim said her writing proficiency had improved over the last six months and she had become able to write essays more quickly. The following interview response shows her perception of improved writing proficiency.

> *I believe my writing has improved because I look up the words in a dictionary less than before. The time for writing assigned essays has decreased from 80 minutes to 50 minutes. I had a chance to read the first essay. I was a little shocked because there were so many errors there. I guess I didn't know about writing much at that time.*

Kim perceived that she had greatly benefited from Criterion feedback and that this perceived benefit had transferred to greater comfort with writing in two content courses, *Current English* and *History of English Language*. In her sixth journal entry, she mentioned being able to compose exam answers easily and to write the term paper in a timely manner.

> I think this program makes me improve writing skill. For example, when I took midterm exams, the exam of *Current English* subject was to write my thinking, feeling and summary after reading an article. Some students felt it is

> difficult, but it was not difficult about it because of this program. So, I could save time and write grammatically correct sentences. Also, when I was doing an assignment of *History of English Language*, I could write faster before. So, I feel a sense of accomplishment. (6th journal entry)

Kim mentioned frankly in her eighth journal entry that she was disappointed at her low score on the revised essay. She was surprised that her score did not increase despite adding new sentences and making changes based on Criterion feedback. She acknowledged that her writing had improved, her overall confidence in writing had increased, and the essay score was not the only standard of writing proficiency; however, the unexpectedly low score made her feel irritated.

> However, against my expectation, my score was low. So, I thought these problems are my wrong grammar and short length. However, although I revised grammar and added some sentences, the revised essay was same score. I don't know why score didn't change. I guess sentences are simple and this content isn't persuasive. As I use Criterion program, my writing skill is better than before. So, I feel more confident than before. However, when I got low grade, I mind this score. Of course, my writing skill isn't evaluated only score, but I am bothered about it. (8th journal entry)

During her last interview, she said her confidence in writing had increased as the result of regular essay writing practice with Criterion feedback over the year.

> *I guess I became more confident than before due to writing opportunities with Criterion feedback. My fear of writing has disappeared significantly, although not completely.*

4.3.2. Advanced writer: Song

Song had substantial previous writing instruction in English from studying in New Zealand at ages 11-14, and had no difficulty writing English essays. Her command of English was very high and, in the first journal entry, she self-assessed herself as an advanced writer, as follows:

> I think my English writing skill is in a high grade for my age, because I can write about the things I think about. For example, if I think about a topic in my head, I can write about the topic in English right away, without translating my thoughts in English from Korean. (1st journal entry)

With her TOEIC listening and reading score of 900, Song was able to work as a teaching assistant at a private institute, where she answered questions on reading and grammar from students taking TOEIC prep courses. She was devoted to developing and improving her academic writing in English.

Song was particularly enthusiastic about participating in the study and actively shared her thoughts with the researcher. During the first interview, Song described her perception of Criterion feedback as follows:

> *I think Criterion feedback is very detailed and specific. The program points out which part of the essay is wrong by highlighting in colors. For example, it does not tell that your grammar in this sentence is wrong. Instead, the program tells you that the subject-verb agreement in the grammar category is wrong.*

It is interesting that Song perceived Criterion feedback as specific rather than general. Song also expressed that immediately after she participated in the study, she became motivated to write and revise with the help of Criterion feedback.

> *Before I used the Criterion program, I wrote essays based on my instinct. I thought I was a very competent writer but now I realize that I make many grammar mistakes. These days I check out my grammar constantly while writing essays.*

In her second journal entry, Song again expressed the utility of grammar feedback, as follows. She made common grammar errors such as missing or extra articles and she was able to correct them in the second drafts with the help of Criterion feedback.

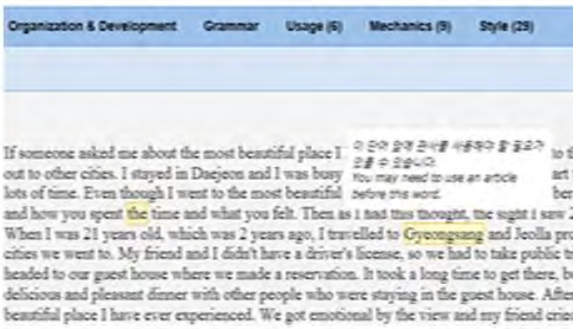
> The most helpful feedback I received is responses about my grammar mistakes. I made a few mistakes such as missing articles from the sentences and adding an extra comma. For example, I wrote a sentence called "the conformity makes students more diligent." When the program pointed out this sentence has an extra article, I couldn't see what my problem was. However, when I looked up in the dictionary, I found out that the word *conformity* doesn't go with the article *the*. (2nd journal entry)

Blind acceptance of corrections suggested by Criterion without reflection does not promote self-directed learning (Chodorow et al., 2010). For instance, Song received incorrect feedback on the sixth essay, as seen in Figure 7. Detailed explanations of the

feedback are provided in various languages, including Korean, and in Figure 7, explanations for corrections are provided in Korean.

FIGURE 7

Screenshot of Incorrect Feedback (Song)



Song received missing or extra article feedback on the proper noun phrase *Gyeongsang and Jeolla provinces*. As seen in her sixth journal entry, she did not blindly accept the feedback, but looked the point up in an online dictionary and did not make changes in response to the incorrect feedback, which was a good response, as the feedback was incorrect.

> I agreed with almost all the errors the Criterion program pointed out, but there was one suggestion that I was not sure about. I wrote 'Gyeongsang and Jeolla provinces' without the article 'the', but the program said that there was a missing article. But, when I looked up in the internet, it said that provinces are not followed by an article. I was not sure about this matter, but I decided to go along with what I have written at the first time. (6th journal entry)

During the last interview, Song said she benefited from the incorrect feedback because she had to search resources such as dictionaries, websites, and grammar books to verify that Criterion *did* make mistakes, which led to self-directed learning. Thus, Song considered Criterion to be like a game she wanted to win, similar to participants in Scharber, Dexter, and Riedel (2008).

> *Whenever I received the incorrect feedback, I said to myself: Who's right this time? Me or the Criterion program? After I found out that the Criterion program did make mistakes, I felt a bit of accomplishment. It was like I won*

> *AlphaGo.*[1] *Searching for the right answer using various resources led to self-directed learning, which also contributed to an increased TOEIC writing score.*

Song benefited from grammar feedback and perceived her writing as having improved greatly as a result, as can be seen from her third journal entry.

> I think my grammar skills have improved a lot since I used this Criterion program. I thought I was good at grammar, but I realized that I was making many small mistakes. So nowadays, I revise my sentences one more time carefully on my own before I submit my first draft. (3rd journal entry)

When asked in the second interview which aspect of her writing she felt had improved most over the last eight months, Song said it was grammar.

> *Due to the effect of schooling in New Zealand, I acquired grammar naturally. I didn't think about grammar constantly when I wrote. Now I realize that I didn't pay attention to basic grammar such as subject-verb agreement. The Criterion program gave me feedback on it. It made me attend to subject-verb agreement for my future writing.*

She perceived her improved writing proficiency as being manifested in her decreasing use of reference tools to look words up, as seen in the fifth journal entry.

> In writing the 5th journal, I can see that I am now writing my own thoughts better in an organized way. I used to look up in the internet or find the usage of words in the dictionary, but I use less computer now. (5th journal entry)

During her second interview, Song mentioned that she had voluntarily written a term paper in English for a course on *British Novels*, and that she felt quite confident about her writing. The following interview response shows a higher level of comfort with writing in content courses.

> *I took a course on British Novels from Professor P. There was a term paper requiring a critical review in five pages in Korean or in English, if we wished. I felt comfortable with writing it in English because all the resources were in English, and I did it.*

[1] A computer program that plays the board game *Go.*

In addition, she mentioned that she was able to write assignments with ease for her *English Writing II* course.

> *In the past, it took some time for me to figure out what I should write about. These days, it takes less time and I can write essays more smoothly. I have a writing assignment for English Writing II and I have to write biweekly diaries. The Criterion program takes credit for it because my grammar has improved.*

In her eighth journal entry, she expressed that her writing proficiency as well as her thinking skills had developed, and that as a result, she was able to try new expressions and phrases in her essays. As indicated in Table 5 in section 4.1.3, her increased total error count from 17 to 22 from Time 1 to Time 3 might be partly explained by her new writing habit of adding new phrases in revised versions.

> Not only my English writing skills but also my ability to think and expressiveness have improved a lot. The best feature that the program is now providing for me is my challenging attitude. Through these writing activities, I constantly try to use new expressions and phrases. I feel that I am continuously upgrading my writing and thinking skills. (8th journal entry)

Figure 8 illustrates the excerpt of Song's essay 10 written on the topic of after-school jobs. Text additions are highlighted in bold in the second draft paragraph. In the second draft, the first sentence is added at the beginning of the introductory paragraph and some new phrases and clauses are also added at the following sentence.

FIGURE 8

Excerpt of Essay 10 (Song)

<The first draft>
These days, many students are working for several reasons such as working for their living expenses or making career. Also, some parents don't let their children work, but other parents allow to work. As far as I am concerned, I think taking after-school job is beneficial for students. Here are several reasons about it.

<The second draft>
Every year, so many graduates are coming out to society in Korea. A lot of non graduates are working for several reasons such as working for their living expenses or making career **or saving before entering society**. On this part, some parents don't let their children work, but other parents allow children to work. As far as I am concerned, I think taking after-school job is beneficial for students. Here are several reasons for it.

During the last interview, Song mentioned that her confidence in her writing had increased because she had received a TOEIC writing score of 190, an increase of 40 points, even though she was unfamiliar with the test's format. She was surprised by the increase and ascribed it to the writing practice afforded by Criterion.

> *Due to the increased TOEIC writing test score from 150 to 190 out of 200, I got more confident about writing than I used to be. Writing one essay with the help of the Criterion program per month over one year helped me build confidence in my writing. I did not prepare for the TOEIC writing test. For instance, I have not learned about how to write business email messages, which is the second task on the TOEIC writing test.*

5. DISCUSSION

This longitudinal case study investigated the EFL writing development of two Korean university students supported by Criterion feedback over one year, using a mixed-methods research model. The study employed a test-retest design that required participants to take a TOEIC writing test at the beginning and again at the end of the study period. The participants wrote essays outside the classroom every month for one year, and submitted 12 first drafts and, later on, after reflecting on Criterion feedback, 12 second drafts. They wrote a reflective journal entry immediately after the submission of the second draft of each essay and were also interviewed three times each.

The first research question investigated the extent to which students' writing improved, as indicated by three indices: TOEIC writing test scores, Criterion essay scores, and error count. The results indicated an overall improvement in writing proficiency, as shown by the findings in all these areas. First, participants' TOEIC writing scores increased by 10-40 out of a total score of 200 after using Criterion. Second, essay scores provided by Criterion also increased from Time 1 (from essay 1 to essay 4) to Time 3 (from essay 9 to essay 12). Third, the total error count decreased from Time 1 to Time 3, except for the advanced writer, Song. Therefore, different aspects of writing developed differently among these two participants. For Kim, there was a dramatic error decrease from Time 1 to Time 3 in style, and only in grammar and style for Song.

The second research question investigated the extent to which students' writing proficiency developed, as indicated by fluency and grammatical complexity. The results showed that quantitative measures of both fluency and grammatical complexity increased from Time 1 to Time 3. For Kim and Song, fluency measures (i.e., word count) increased to a similar degree from 323 to 353.5 for Kim and from 406.5 to 434.25 for Song.

Regarding grammatical complexity, the ratio of clauses to T-units increased by similar, substantial amounts for both participants: for Kim, from 1.32 to 1.62; and for Song, from 1.42 to 1.76.

The third research question investigated the way students perceived their writing trajectory after using Criterion for one year. The results for the two participants' perceptions of writing development, gathered through the analysis of their journal entries and interview data, can be summarized as follows. First, both participants perceived grammar feedback as helpful in revising the essays. The advanced writer, Song, perceived grammar feedback as sufficiently specific. Second, in order to understand and address their Criterion feedback in writing second drafts, both participants made reference to grammar books and the internet. Third, both participants also acknowledged that their writing proficiency had improved over the year, as manifested, for example, by less frequent dictionary use to look up words and phrases. Kim perceived that her essay composing time had decreased as her writing practices evolved. Fourth, they were able to write term papers and exam answers with greater comfort in their content courses. Fifth, their overall confidence in their writing also increased as a result of regular writing practice, although Kim was disappointed at her low scores on Criterion evaluations.

The results have shown that the opportunity to revise essays with the help of Criterion feedback pushed the learners to pay attention to grammar, form, and usage, helping them to convey their intended meanings and refine their linguistic expression. The most important element of successful writing development is the provision of feedback on writing (Ferris, 2003). Two case study participants did not accept Criterion's suggested feedback and corrections blindly; instead, they were generally able to distinguish correct from incorrect feedback and make informed judgments. The development in their EFL writing over one year can be attributed to sustained writing practice, along with continued provision of Criterion feedback.

There are several limitations to the present study. First, the results might have been different if different writing prompts had been used for the 12 essays because topic familiarity could influence students' writing, which may limit the generalizability of the results. Second, one year of Criterion use might be insufficient to fully investigate the long-term effect of Criterion feedback on writing development. Therefore, future studies should examine the effects of Criterion use beyond one year. Third, there might be the maturation effect because taking English major courses can lead to increased English proficiency, hence writing development.

In conclusion, the opportunity to produce essays and also to receive AWE feedback on essays over one year could be possible reasons for improvement in writing proficiency. This study thus offers valuable insights into our understanding of EFL writing

development and should contribute to examination of how Korean college students' writing development evolves.

Applicable Levels: Tertiary

REFERENCES

Burstein, J., Chodorow, M., & Leacock, C. (2003). *Criterion*^{SM} Online essay evaluation: An application for automated evaluation of student essays. In J. Riedl & R. Hill (Eds.), *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico* (pp. 3-10). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Casanave, C. P. (1994) Language development in students' journals. *Journal of Second Language Writing, 3*(3), 179-201.

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.

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1-17.

Ferris, D. (2003). *Response to student writing: Implications for second language students*. Mahwah, NJ: Lawrence Erlbaum.

Knoch, U., Rouhshad, A., Oon, S., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing, 28*, 39-52.

Kobayashi, H., & Rinnert, C. (2013). L1/L2/L3 writing development: Longitudinal case study of a Japanese multicompetent writer. *Journal of Second Language Writing, 22*(1), 4-33.

Lee, E.-H. (2015). Korean EFL college writers' perceptions of Criterion. *English Language and Literature Teaching, 21*(1), 199-216.

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1-18.

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*(1), 66-78.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18*(2), 85-102.

Lipnevich, A., & Smith, J. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied, 15*(4), 319-333.

Moon, Y.-I., & Pae, J.-K. (2011). Short-term effects of automated writing feedback and users' evaluation of Criterion. *Korean Journal of Applied Linguistics, 27*(4), 125-150.

Scharber, C., Dexter, S., & Riedel, E. (2008). Students' experiences with an automated essay scorer. *Journal of Technology, Learning, and Assessment, 7*(1), 1-44.

Storch, N. (2009). The impact of studying in a second language medium university on the development of L2 writing. *Journal of Second Language Writing, 18*(2), 103-118.

Storch, N., & Hill, K. (2008). What happens to international students' English after one semester at university? *Australian Review of Applied Linguistics, 31*(1), 1-17.

Weigle, S. C. (2011). *Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability*. (Research Report No. RR-11-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333- 8504.2011.tb02260.x

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*(1), 85-99.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawaii Press.

APPENDIX A

List of Five Traits and Error Categories

Traits	Error Categories
Grammar	Fragment Run-on sentences Garbled sentences Subject-verb agreement Ill-formed verbs Pronoun errors Possessive errors Wrong or missing word Proofread this!
Mechanics	Spelling Capitalize proper nouns Missing initial capital letter in a sentence Missing question mark Missing final punctuation Missing apostrophe Missing comma Hyphen error Fused words Compound words Duplicates Extra comma
Usage	Wrong article Missing or extra article Determiner-noun agreement Confused words Wrong form of word Faulty comparisons Preposition error Nonstandard word form Negation error Wrong part of speech
Style	Repetition of words Inappropriate words or phrases Sentences beginning with coordinating conjunctions Short sentences Long sentences Passive voice
Organization & Development	Introductory material Thesis statement Main ideas Transitional words and phrases Supporting ideas Conclusion

APPENDIX B

Interview Questions

I. First Interview

1. How long have you been studying English?
2. How often do you usually write in English?
___Never ___Almost never ___A few times a month ___At least once every day
3. What do you focus on when you are writing essays? (e.g., content, grammar, organization, vocabulary, etc.)
4. What types of writing did you have to do in the English department? Did you experience difficulties with writing assignments? Did you get help with writing assignments?
5. How do you assess your writing proficiency?
6. Do you like writing with Criterion?
7. Was Criterion feedback easy or difficult to understand?

II. Second Interview

1. Do you believe that your writing proficiency has developed with the help of Criterion?
2. Do you believe that the increased Criterion score indicates improved essay quality?
3. Do you believe that writing practice with Criterion leads to self-directed learning?
4. Does writing practice with Criterion help you manage the essay assignment or the essay exam in the English department?
5. What is your own perception of the usefulness of Criterion feedback? What is the most useful feedback you received? What is the least useful feedback you received?
6. Do you become motivated to write and revise more when you are using Criterion?

III. Third Interview

1. What is your TOEIC writing score?
2. Do you believe that your writing proficiency has developed with the help of Criterion? Do you think your English has improved over time? If so, which aspects (e.g., grammar, vocabulary, organization, etc.) have improved?
3. What factors helped or hindered improvement?
4. Does writing practice with Criterion help you to lose the fear of writing and to gain confidence in writing?
5. Do you believe that writing practice with Criterion leads to self-directed learning?
6. Does writing practice with Criterion help you manage the essay assignment or the essay exam in the English department?
7. What is your own perception of the usefulness of Criterion feedback? What is the most useful feedback you received? What is the least useful feedback you received?
8. Have you received incorrect feedback (i.e., Criterion-induced errors)? If so, how did you deal with it?