http://e-flt.nus.edu.sg/

Electronic Journal of Foreign Language Teaching
2015, Vol. 12, No. 1, pp. 79–100
© Centre for Language Studies
National University of Singapore

Effects of an Automated Writing Evaluation Program: Student Experiences and Perceptions

Pei-ling Wang
(peiling@kuas.edu.tw)
Kaohsiung University of Applied Sciences, Taiwan

Abstract

The issue of whether automated writing evaluation programs facilitate student writing has provoked numerous discussions over the last two decades, but most findings are inconclusive. This study examines the degree of student satisfaction with the functions of *Criterion*®, how the program affects the revision practices of students, and why the program is helpful or unhelpful. The researcher surveyed 53 English major students at a Taiwanese university and examined 530 writing samples from them to discover the strengths and weaknesses of the program. This study used quantitative and qualitative methods to collect the data. The results revealed that many students valued the instant scoring speed (93.8%), the error analysis of usage (75.5%), and the feedback for organization and development (71.4%). However, most students were dissatisfied with the program's scoring rubric (8.2%) and scoring summary (34.7%), the style error analysis (26.5%), and the 'Plan' tool (26.5%). The analysis of error correction rates in students' final drafts confirmed that the feedback for grammar and usage errors was much more useful for student revision than the feedback for mechanics and style errors. The researcher verification showed that the current *Criterion*® tool is limited in its ability to detect errors related to tenses, conjunctions, compound words, word choice, and word order of indirect questions. Another problem of the program is that it may occasionally generate false alarm messages. AWEs have both merits and drawbacks, which may explain why approximately two-thirds of participants believed that the combination of machine scoring with the teacher's explanations was the optimal implementation method for a writing class. Future studies may include more participants and investigate the extent to which these findings can be generalized to students with limited English writing proficiency.

1 Introduction

Today, the computer has become a writing tool and a communication medium for many people. This trend has revolutionized the pedagogical practice of writing teachers over the past decades (Chen & Cheng, 2008). Teachers have applied various electronic writing media such as word processors, e-mail exchanges and bulletin boards to their teaching. Recent advances in automated writing evaluation programs (AWE) have attracted many teachers to implement this new technology for grading and assessing students' writing (Grimes, 2008; Li, Link, & Hegelheimer, 2015).

For students, electronic writing media make writing and editing tasks much quicker and easier. Consequently, students are more willing to revise their essays (Moseley, 2006). For teachers – especially those who believe that writing is a recursive process and that students should continuously rewrite, revise, and edit their writing to improve their compositions – involving students in repetitious practices may seem necessary. However, teachers may become exhausted and lose their passion after concentrating on correcting papers and individually providing specific feedback if they teach a significant number of students. Thanks to the application of AWE, teachers may be more willing to give writing assignments more frequently to students.

Several studies over the past two decades have demonstrated the advantages of AWE programs: the instantaneous feedback regarding revisions (Phillips, 2007); the increased motivation of students (Chou & Chung, 2013; Grimes & Warschauer, 2010; Warschauer & Grimes, 2008); the writing of longer texts with fewer errors (Grimes, 2008); positive changes in student perceptions, including the perception that writing is a recursive process rather than a linear process (Moseley, 2006). However, studies on the impact of AWE programs are inconclusive. Numerous studies on AWE programs (e.g. Grimes, 2008; Li et al., 2015; Moseley, 2006) were conducted in English as a second language (ESL) rather than English as a foreign language (EFL) settings. For example, while Vantage *My Access*® and ETS *Criterion*® are currently the most common AWE applications in Taiwan, studies regarding these two programs in Taiwanese classroom contexts have seldom been conducted. Furthermore, among the limited EFL studies on AWE effectiveness, most studies (e.g. Chen & Cheng, 2008; Chou & Chung, 2013; Lai, 2010; Yang, 2004; Yu & Yeh, 2003) examined student perceptions of *My Access* rather than *Criterion*®. Moreover, Yu and Yeh (2003) only investigated 19 students in their quantitative study. Further studies involving more subjects are necessary.

Thus, this study addresses four major inadequacies of previous studies on AWE. First, the subjects were 53 Taiwanese college students in an EFL setting, a number that is significantly higher than those in previous quantitative studies (e.g. Yu & Yeh, 2003). Second, while previous studies (e.g. Chen, Chiu, & Liao, 2009) on the functions provided by *Criterion*® only evaluated one category of diagnostic feedback (viz. grammar feedback), this study evaluates five categories of diagnostic feedback (viz. grammar, usage, mechanics, style, and organization and development), its scoring functions, and its writing assistance tools. Third, previous studies examined either only student (e.g. Ou, 2011) or researcher (e.g. Chen, Chiu, & Liao, 2009) perceptions of the program functions, but this study used both student evaluations and researcher verification to examine the usefulness of the program. Such a method may enhance the reliability and credibility of this study's findings regarding the contributions and drawbacks of *Criterion*®. Fourth, the subjects wrote essays of three different rhetorical modes (process, cause/effect, and comparison/contrast essays), unlike previous studies (e.g. Frost, 2008; Moseley, 2006; Otoshi, 2005), which investigated only persuasive writing.

This study may assist educators and program designers in understanding students' perceptions of the program, and whether the machine feedback can help improve students' writing skills. In view of these research objectives, the following questions were posed:

1. What are students' perceptions of the functions provided by *Criterion*® and the usefulness of *Criterion*® for learning English writing?
2. To what extent does the program affect the revision practices of students?
3. What is the degree of accuracy of the diagnostic feedback?

To help readers understand the theoretical foundation of this study, the relevant literature on the history of AWE development and the effects of implementing AWEs are reviewed in the following section.

2 Literature review

2.1 Brief history of the development of Criterion®

The electronic essay rater (*e-Rater*®) was developed by Burstein and Kaplan at the Educational Testing Service (ETS) during the 1990s. This system employs natural language processing (NLP) techniques and artificial intelligence techniques to identify linguistic features and syntactical cues from texts (Burstein, 2003).

The *e-Rater*® assumes that the features of good (or bad) essays are similar to those of other well (or poorly) written essays, and adopts a corpus-based approach to identify discourse elements and detect violations of grammar rules, which requires numerous sources of text to examine sample essays. However, unlike other corpus-based AWE systems that normally use well-edited text sources such as newspapers, *e-Rater*® requires unedited text corpora that fit the rhetorical modes of student essays scored by at least two human raters on a 6-point holistic scale to build models (Dikli, 2006).

Rudner and Gagne (2001) showed that the first version of *e-Rater*® uses 60 different features to perform discourse structure analysis and content analysis. The second version of this system (*e-Rater*® V.2) can measure grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage (Attali & Burstein, 2006). Several improvements were made to the system, but *e-Rater* V.2 was still unable to completely consider the context in which the words are used.

Criterion® is the instructional application of the *e-Rater*®. ETS claims that *Criterion*® can assess various writing genres and topics at various levels, including Grades 4 through 12 and undergraduate level. A minimum of 465 essays scored by expert raters are used to modulate the system on a topic. *Criterion*® reflects the overall quality of the writing and can provide feedback for several dimensions of writing (ETS, n.d.). Additionally, this system has an accuracy rate of approximately 97%, compared to that of human judges (Chodorow & Burnstein, 2004).

2.2 Studies on the effects of using the AWE system

2.2.1 Advantages of using AWE programs

AWE programs are normally equipped with word processing features, the Internet, and electronic portfolios. These features have been proven to be beneficial to writing. First, studies (e.g. LinHuang, 2010; Williamson & Pence, 1989) have confirmed that the use of word processing facilitates editing and revising, which makes students more willing to improve their writing, especially the aspects of grammar and spelling. Moreover, student papers written on the computer are longer than those written on paper with pen and pencil (Flinn, 1986). Other studies have shown that students' motivation increased because of computer-assisted writing. For example, Grimes and Warschauer's study (2006) indicates that the U.S. high school students in their study experienced increased motivation for practicing writing when *My Access* and *Criterion*® were used in their writing class. Furthermore, AWE programs can significantly encourage more revisions (Li et al., 2015; Warschauer & Grimes, 2008).

Second, the Internet permits students to access relevant information for brainstorming, consult an online dictionary, conduct online peer reviews, and significantly improve their writing skills (Butler-Pascoe & Wiburg, 2003; Suh, 2002). Several studies indicate that Internet use can provide a non-threatening environment and reduce learner anxiety in writing (Daiute, 1986; Li, 2009; Sullivan & Pratt, 1996).

Third, the use of electronic portfolios saves storage space (Barrett, 2000). Furthermore, it assists by displaying student growth in writing (Herter, 1991; Ware, 2011), sharing the values of the writing-process curriculum, and providing students with opportunities to revise (Hamp-Lyons, 1994). Gottlieb (1995) showed that e-portfolios motivate students toward becoming responsible for their learning and developing a sense of ownership. E-portfolio is a common feature in various AWE programs. Both learners and teachers can see the process of student revisions and student progress, which can assist students in managing their writing (Dikli, 2006; Wang, 2011; Yang, 2004).

In addition to the above-mentioned advantages, the corrective feedback and the instant scoring of AWE programs are two other benefits perceived by some instructors and students. For example, Chou and Chung (2013) investigated non-English majors' perceptions of the use of *My Access*. Many students felt that the diagnostic feedback helped them notice their individual writing problems. The instant grading process also motivated the students to correct their errors. Chou and Chung concluded that using AWE is helpful for EFL students at a lower level of English proficiency. Moreover, in the study of Li et al. (2015), 18 out of 27 ESL students interviewed were highly satisfied with the corrective feedback of *Criterion*®. Most of the interviewed instructors valued the corrective feedback for grammar and mechanics, although some of them acknowledged the ineffectiveness of the machine feedback for organization and development.

2.2.2 Disadvantages and limitations of using AWE programs

On the other hand, several studies have also pointed to the numerous disadvantages and limitations of AWE tools. Cheng (2006) examined 68 English major college students' perceptions of *My Access*. The results showed that only 55% of the students regarded the program as "slightly helpful" to them for improving their writing skills. Many students were dissatisfied with the system's grading function, because it failed to provide specific feedback on the content and rhetorical aspects of their writings.

Furthermore, there was evidence of student dissatisfaction with the scoring and feedback mechanisms of *My Access* in other studies (e.g. LinHuang, 2010; Yang, 2004; Yu & Yeh, 2003). The studies by Yang (2004) and Yu and Yeh (2003) indicated that most students believed the feedback from *My Access* is useful only for the first revision and that the subsequent similar and repeated feedback was ineffectual.

LinHuang's study (2010), based on a sample of 58 senior students, showed that human raters outperformed *My Access* in identifying errors for student essays. Among 200 student writing samples, human raters detected a mean error rate of 18 compared to 13.2 errors by *My Access* when scoring the same essays. However, *My Access* showed a better correlation rate (0.811) than the judgments of human raters when scoring essays of lower writing proficiency levels, compared to those of higher proficiency levels, which suggests that the program may be more beneficial to students with limited writing proficiency.

Inconsistencies in the scores given by AWEs and human raters were also found in the study by Li, Link, Ma, Yang, & Hegelheimer (2014). In this study, the instructors had only a neutral to low level of trust in the scores from *Criterion*®. Although they had some trust in low scores from *Criterion*®, they did not trust high scores from *Criterion*®. In other words, the instructors believed that if a student gets a low score from the machine, the student's writing quality is most likely problematic. However, if a student gets a high score from *Criterion*®, the instructors may still not give a high grade to this writing.

Chen, Chiu and Liao (2009) showed the limitations of AWE programs by examining the feedback messages provided by *My Access* and *Criterion*® for 269 randomly selected essays. They found that only three set of feedback (viz. for spelling errors, clause errors, and subject-verb agreement) provided by *My Access* reached 20% accuracy; that is, most machine feedback messages from *My Access* were false alarms. In contrast, the majority of grammar feedback messages provided by *Criterion*® had 70% accuracy. Although *Criterion*® had higher accuracy rates and fewer false alarms than *My Access*, Chen, Chiu and Liao concluded that both systems required further improvements in their error feedback mechanisms.

Other shortcomings of AWE systems include financial considerations, technical glitch issues when students log onto the Internet simultaneously, and insufficient technological skill training and familiarity of teachers and students (Busbee, 2001; Lee, 2008). Furthermore, Page (2003) criticized the systems for only performing the processes they were programmed to execute and for their inability to appreciate essays as human raters may. Warschauer and Ware (2006) asserted that computerized scoring systems eliminated the human element from writing assessments and were inadequate in terms of human interaction.

The preliminary findings from previous studies showed that AWE programs have both merits and drawbacks. However, most studies investigated the effect of *My Access* instead of *Criterion*®. Moreover, many of these studies were sponsored by AWE-associated companies (e.g. the developers of *Criterion*®; see Burstein et al., 1998; Chodorow & Burstein, 2004). In other words, most of these evaluations were not provided by the actual users, that is, students; thus, this study examines this uncharted area. Furthermore, previous studies exploring the disadvantages and limitations of AWE use did not analyze students' writing samples to support the researchers' perspectives. Therefore, this study attempts to examine students' writing samples and provide some specific examples of the machine feedback to help readers understand the degree of accuracy of the machine feedback. This study's procedure is described in detail in the following section.

3 Methodology

3.1 Subjects

The subjects comprise two classes of 53 sophomore students majoring in English from a technical university in Southern Taiwan (48 women and 5 men). Their English writing proficiency was intermediate, according to their average pre-test scores on *Criterion*® (mean = 4, SD = 0.62). They had never used AWE before participating in this study. Prior to this study, the English department had just subscribed to *Criterion*® and was concerned about the program's effectiveness; however, budget constraints only permitted the use of 120 accounts and passwords for the program, which enabled four English writing classes (maximum of 30 students per class) to use the program. The researcher was assigned to teach two of the classes by the department, and these two classes were selected as the participants of this study with consideration to their convenience and availability (Creswell, 1994).

3.2 Instruments

The instruments used to collect data for this study include a questionnaire and students' writing samples from the e-portfolio in *Criterion*®. The questionnaire comprises 29 questions, including 15 five-point Likert-type questions, two multiple-choice questions, and 12 open-ended questions (see Appendix A). The questionnaire required students to judge the quality of the three functions provided by *Criterion*®: scoring functions (Q1 to Q3), diagnostic feedback functions (Q4 to Q8), and writing assistance tools (Q9 to Q11). The open-ended questions allowed students to explain their reasons for their evaluation (Q1.1 to Q8.1). Moreover, the questionnaire required students to reflect on their experiences of using *Criterion*® and to offer suggestions for writing teachers (Q12 to Q21). The reliability of the Likert-type items in the questionnaire was measured by computing the Cronbach's α, which shows that the questionnaire is reliable ($\alpha = 0.76$). The questionnaire was not anonymous. However, the researcher assured students that their opinions would only be used for the purpose of the research and would not influence their course scores.

3.3 Application of the Criterion® program in English writing class

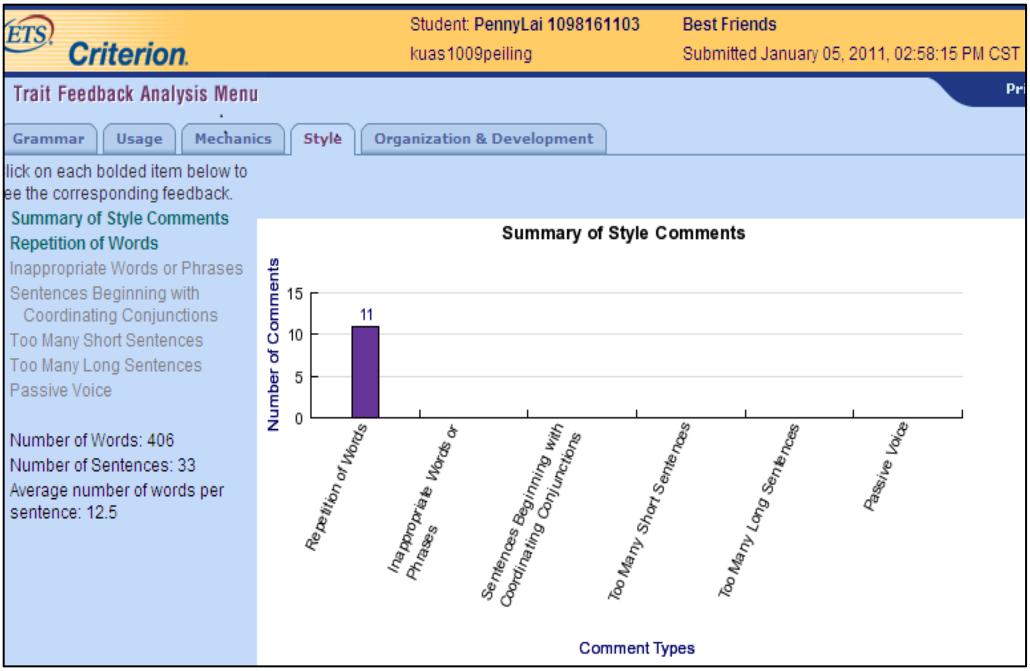
A computer laboratory was arranged to allow each student to access a computer and a *Criterion*® account in the writing classes of this study. The teacher/researcher demonstrated how to log into the *Criterion*® website with an individual student account number and password at the beginning of class. Furthermore, various functions of the software were explained.

When students were writing their drafts, the researcher walked around the laboratory, monitored student progress, and provided assistance whenever students raised their hands. Moreover, the researcher checked student scores and praised students who achieved a score of 6 (the highest score in the system) after all the students had submitted their first drafts. Thereafter, after obtaining the respective students' permission, the researcher displayed their papers to the class as model essays and pointed to the strengths. Students with essay scores below average received instant private tutoring from the researcher and were given suggestions on how to revise their essays (based on the researcher's expertise and machine feedback). Thus, the researcher encouraged communication, negotiation, and elaboration between instructor and learners and among learners. Furthermore, the researcher randomly reviewed student essays and verified their machine's comments. When students were confused or frustrated by the machine's vague advice, the researcher offered clear directions for improvement. She comforted students by stating that their automated scores would make up only 10% of their final grades. Additionally, she encouraged students to identify false alarms and errors that the machine failed to detect. Finally, she advised both classes to discuss the effectiveness of the *Criterion*® program after the students had justified their criticisms.

During the 18 weeks of instruction, three rhetorical modes of writing were taught: process, cause/effect, and comparison/contrast. Thereafter, students were requested to write five essays (1 process, 2 cause/effect, and 2 comparison/contrast essays) using *Criterion*® and submit three drafts for each essay. All students had at least 10 minutes of individual tutoring for their first draft of the second essay. The second essay was selected for this task to ensure that the students had already understood all of the functions provided by *Criterion*® before they wrote the other three essays.

3.4 The diagnostic feedback of Criterion®

Criterion® has five categories of diagnostic feedback messages: summary of grammar errors, usage errors, mechanics errors, style comments, and organization and development analysis. Students could access these messages by using the "View Trait Feedback Analysis" tool (see Fig. 1), which displayed the feedback and informed writers of the errors in their essays. When students moved the mouse over the highlighted texts in their passages, they could view the locations of their errors and the advice for correcting them (see Fig. 2).



Student: PennyLai 1098161103
kuas1009peiling

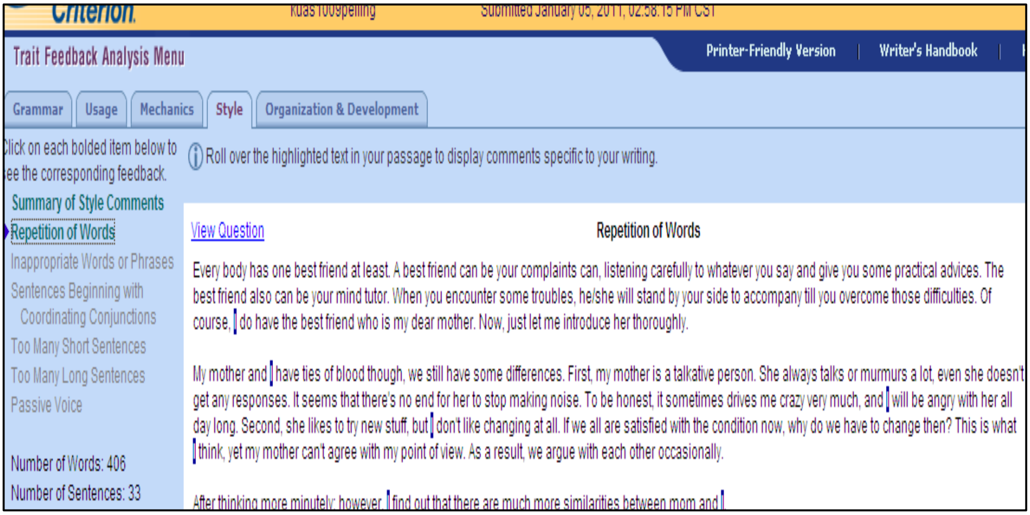
Best Friends
Submitted January 05, 2011, 02:58:15 PM CST

lick on each bolded item below to ee the corresponding feedback.

Number of Words: 406
Number of Sentences: 33
Average number of words per sentence: 12.5

ETS
Criterion
Trait Feedback Analysis Menu
Pri
Grammar
Usage
Mechanics
Style
Organization & Development
Summary of Style Comments
Repetition of Words
Inappropriate Words or Phrases
Sentences Beginning with Coordinating Conjunctions
Too Many Short Sentences
Too Many Long Sentences
Passive Voice
Summary of Style Comments
Number of Comments
15
10
5
0
11
Repetition of Words
Inappropriate Words or Phrases
Sentences Beginning with Coordinating Conjunctions
Too Many Short Sentences
Too Many Long Sentences
Passive Voice
Comment Types

Fig. 1. Example of style comments by *Criterion*® on trait feedback analysis


Criterion.
Kuas 100spelling
Submitted January 05, 2011, 02:58:15 PM CST
Trait Feedback Analysis Menu
Printer-Friendly Version | Writer's Handbook |
Grammar
Usage
Mechanics
Style
Organization & Development
Click on each bolded item below to see the corresponding feedback.
Roll over the highlighted text in your passage to display comments specific to your writing.
Summary of Style Comments
Repetition of Words
Inappropriate Words or Phrases
Sentences Beginning with Coordinating Conjunctions
Too Many Short Sentences
Too Many Long Sentences
Passive Voice
Number of Words: 406
Number of Sentences: 33
View Question
Repetition of Words

Every body has one best friend at least. A best friend can be your complaints can, listening carefully to whatever you say and give you some practical advices. The best friend also can be your mind tutor. When you encounter some troubles, he/she will stand by your side to accompany till you overcome those difficulties. Of course, I do have the best friend who is my dear mother. Now, just let me introduce her thoroughly.

My mother and I have ties of blood though, we still have some differences. First, my mother is a talkative person. She always talks or murmurs a lot, even she doesn't get any responses. It seems that there's no end for her to stop making noise. To be honest, it sometimes drives me crazy very much, and I will be angry with her all day long. Second, she likes to try new stuff, but I don't like changing at all. If we all are satisfied with the condition now, why do we have to change then? This is what I think, yet my mother can't agree with my point of view. As a result, we argue with each other occasionally.

After thinking more minutely; however, I find out that there are much more similarities between mom and I

Fig. 2. Example of highlighted errors “Repetition of words-- I” and comments by *Criterion*®

3.5 Procedure for data collection

Data were collected in the Fall semester of the 2010 school year. Before students participated in the study and before their writing samples were examined, a consent form was given to students. Students completed five essays and wrote three drafts for each essay during this study. Since *Criterion*® only saved the students’ first and final drafts, a total of 530 writing samples were collected.

All the student essays were stored in the electronic portfolios of *Criterion*®. The total number of words, the scores, and the machine error analyses for each student’s first and last submissions were recorded. Students completed the questionnaire in the last week of the semester. A total of 49 students of 53 responded to the questionnaire; thus, the questionnaire return rate was 92%.

3.6 Data analysis

The researcher employed the following methods to analyze the collected data. First, descriptive statistics were used to indicate the students’ perceptions of the functions and the use of the program. Second, descriptive statistics were applied to present the revision rates of errors in students’ writing from their first to final submissions. The researcher analyzed the students’ writing samples from portfolios and during the individual 10-minunte tutoring to examine the degree of accuracy of the automated diagnostic feedback and to evaluate whether the detected errors were actual errors or false alarms. Third, student responses to the open-ended questionnaire were coded using the content analysis technique, which “involves the simultaneous coding of raw data and the construction of categories that capture relevant characteristics of the document’s content” (Merriam, 1998, p. 160). For example, for the question “Do you know any strategies to obtain higher scores in *Criterion*®?”, the researcher read the collected data and generated two themes and categories: strategies for surface revision (i.e. rewording and correcting grammar errors) and strategies for deeper revision (i.e. focusing more on content, style, and organization development). Thereafter, the researcher searched for instances of the patterns or themes that emerged from the data, and assigned student responses to one of the two categories. Furthermore, other plausible explanations for these data and the linkages among them were searched. Finally, the data were interpreted and summarized.

4 Results and discussion

4.1 Students' perceptions of the functions provided by Criterion®

Criterion® provided three functions: scoring, diagnostic feedback, and writing assistance tools. Table 1 presents students' perceptions of these functions. First, *Criterion*'s scoring function provided a single score on a scale of 1 to 6, and a holistic score summary that informs students of the strengths and weaknesses in their essays. Student perceptions of these functions were as follows: 46 students (93.8%) agreed that the program offered instantaneous scoring, but only 17 students (34.7%) liked the function of holistic score summary. Several students indicated that the summary component of *Criterion*® was fixed and not sufficiently informative. For example, feedback such as "The essay text does not resemble other essays written about the topic" may only inform students that the essay is off-topic, but students did not receive advice on how to revise their papers. In contrast, only four students (8.2%) agreed that *Criterion*'s scoring rubric was objective, and believed that the program was able to effectively assess their English writing ability.

Table 1. Students' perceptions of the functions of the *Criterion*®

Questions	Agree		Unsure		Disagree		Mean	*SD*
	N	%	N	%	N	%		
1. I think its scoring rubric is objective.	4	8.2	29	59.2	16	32.6	2.73	.63
2. I think its holistic summary is useful.	17	34.7	20	40.8	12	24.5	3.10	.77
3. I think its scoring speed is satisfying.	46	93.8	3	6.1	0	0	4.20	.57
4. I think its error analysis of grammar is useful.	30	61.2	10	20.4	9	18.4	3.53	.93
5. I think its error analysis of usage is useful.	37	75.5	10	20.4	2	4.1	3.80	.64
6. I think its error analysis of mechanics is useful.	28	57.2	17	34.7	4	8.2	3.53	.71
7. I think its error analysis of style is useful.	13	26.5	23	46.9	13	26.6	2.94	.92
8. I think its error analysis of organization development is useful.	35	71.4	12	24.5	2	4.1	3.82	.75
9. I think the function of 'Question Statement' in *Criterion*® is helpful.	31	63.2	14	28.6	4	8.2	3.57	.67
10. I think the function of 'Plan' in *Criterion*® is helpful.	13	26.5	20	40.8	16	32.6	2.90	.91
11. I think the 'Sample essays' provided by *Criterion*® is helpful.	22	44.9	17	34.7	10	20.4	3.25	.88

Note: 1. Agree % included the percentages of "Totally agree" and "Agree"
2. Disagree % included the percentages of "Totally disagree" and "Disagree"

Furthermore, students' dissatisfaction with the scoring objectivity surfaced in their answers to the open-ended question "Do you know how to obtain higher scores in *Criterion*®?" Among the 35 students answering this question, 24 students (68.6%) indicated that the longer an essay was, the better the score would be. Four students (11.4%) stated that a good essay required at least five paragraphs. A student (2.9%) informed that the probability of obtaining a score of 6 increases if the draft has more than 500 words. Thus, this finding echoes Chen and Cheng's study on *My Access* (2008), indicating that the AWE favored lengthiness.

The second function of *Criterion*®, the diagnostic feedback, offered five category feedback reports: grammar, usage, mechanical, style, and organization and development errors. The program highlights errors and offers advice on how to correct them. Numerous participating students believed that the error analysis of usage (37 students, 75.5%) and the feedback for organization and development (35 students, 71.4%) were beneficial. More than half of the students appreciated the error analysis of grammar (30 students, 61.2%) and of mechanics (28 students, 57.2%). However, the error analysis of the style function was valued by only 13 students (26.5%). Numerous students complained that the machine indicated that they repeated some words too frequently (e.g. "I") in their essays, but the machine did not teach them how to improve this shortcoming.

The feedback on organization and development is a special function in *Criterion*®, which uses five colors to identify the five elements of a good essay: the introduction in blue, thesis statement in red, topic sentence of the supporting paragraph in light green, supporting sentences in dark green, and the conclusion in yellow. Thus, a good essay should be developed in the order of blue, red, light green, dark green, and yellow. When student writings were not marked in these five colors, or the order of the colors was chaotic, then students would know that their writing was not well-organized. With this useful feature, the majority of the participants felt that the feedback on organization and development was beneficial to them.

However, the feedback on organization and development still had some drawbacks. Of the 11 students who answered the open-ended question "How do you feel about the function of organization and development?", 4 students (36.3%) stated that they learned to "trick" the machine after several essay submissions. They noticed that the program can only identify if an essay had these five elements: introduction, thesis statement, supporting paragraph 1, supporting paragraph 2, and conclusion. Nevertheless, the machine was unable to judge whether they were semantically appropriate to the particular context. Therefore, students would receive higher scores if they developed their essay in the "ideal" format, even if they provided information that deviated from the topic.

The error analysis of the mechanics function (i.e. spelling and punctuation) was reported as useful by only half of the subjects (28 students, 57.2%), possibly because numerous word processors are already equipped with this function. Moreover, 8 of the 11 students (72.7%) answering the open-ended question "How do you feel about the function of mechanics?" stated that *Criterion*® was unable to recognize several new words such as Facebook and MP4. Another drawback was that the machine failed to identify place/human names, proper nouns, and contractions. Therefore, the program would evaluate these words as incorrect spellings.

The third function of *Criterion*®, writing assistance tools, includes Question Statement, Sample Essay, and Plan. Students' perceptions of these tools were as follows. Thirty-one students (63.2%) considered the Question Statement tool to be beneficial, 22 students (44.9%) judged the Sample Essay tool to be useful, and only 13 students (26.5%) agreed that the Plan tool was valuable. However, relatively few students created a plan prior to writing an essay; that is, many students did not use this tool. The *Criterion*® program designers may consider making this function mandatory for students before writing an essay.

Students' perceptions of the machine functions could also be found in their responses to the following two open-ended questions. Forty-five students answered the question "What are the best parts of the *Criterion*® program?" The greatest advantages in the students' opinion were that the program facilitates grammar correction (21 students, 46.7%), reminds them of spelling mistakes (10 students, 22.2%), identifies organization and development (eight students, 17.8%), provides instant scoring (four students, 8.9%), and forces students to revise repeatedly (one student, 2.2%). As Pei-Fan (all names stated here are pseudonyms) stated, "It's like magic. I sent and got the machine feedback in seconds. Since its response was so instant, I knew where I made mistakes right away, and then I was more willing to revise my writing again and again. It indeed pointed out some errors which I had ignored." The machine feedback also increased students' self-confidence. Hsiu-Hsiu wrote, "No matter how bad my essay is, *Criterion*® always gives me at least a score of three, which makes me feel confident. Besides, I think it's good that the machine helped me revise the first draft and then I sent the second draft to my teacher. In this way, I was sure that when the teacher evaluated my paper, my writing would not be too lousy."

Forty students responded to the question "What are the worst parts of the program?" Twenty-two students (55%) did not trust the scoring mechanism. They doubted that the machine truly understood the content of an essay. As Wei-Lun pointed out, "Could it really understand my ideas? It seemed that I would get a higher score as long as I wrote longer!" Nine students (22.5%) complained that the system's feedback was not precise. Five students (12.5%) indicated that the machine gave false alarms regarding the spelling of proper nouns. Ya-Han said, "It doesn't know MP4 and people's names! Whenever I used these nouns, it said I was wrong! It really drove me crazy!" Four students (10%) alleged that the machine was incapable of detecting all mistakes and that it occasionally provided strange suggestions. For example, Chia-Hsin wrote, "I could not totally count on the machine, because my teacher would still find some errors in my paper after I had revised my paper based on the machine feedback. Besides, the machine comments sometimes were very ridiculous; for example, I once wrote 'encouragement to people,' but *Criterion*® suggested to me 'encouragement two people.' Wasn't it funny?"

4.2 Student perceptions of the usefulness of Criterion® for learning English writing

Although 35 students (71.4%) believed that their English writing ability improved after using *Criterion*® (item 12), the students' answers to Item 16 (I think *Criterion*® assisted me in the following aspects) show that *Criterion*® failed to nurture their English writing abilities, except for grammar and organization. More specifically, while roughly half the participants agreed that *Criterion*® helped them make progress in the areas of grammar (25 students, 51%) and organization (22 students, 44.9%), fewer believed that the program assisted them in the following aspects:, logic development (17 students, 34.7%), sentence coherence (16 students, 32.7%), generating ideas (12 students, 24.5%), punctuation (11 students, 22.4%), and vocabulary (11 students, 22.4%).

Surprisingly, only 13 students (26.5%) were willing to use *Criterion*® again in the future, although the majority (37 students, 75.5%) believed that *Criterion*® was user-friendly. Moreover, less than one-third of the students (11 students, 22.4%) were satisfied with the *Criterion*® program. The number of students with a neutral attitude was high (more than 40%). Around half of the students (23 students, 46.9%) doubted that the program was satisfactory, and 20 students (40.8%) were uncertain about using *Criterion*® again (See Table 2).

Table 2. Student perceptions of the usefulness of the *Criterion*® for learning English

Questions	Agree		Unsure		Disagree		Mean	*SD*
	N	%	N	%	N	%		
12. I think my English writing ability has I proved after using *Criterion*®	35	71.4	6	12.2	8	16.3	3.61	.83
13. I think *Criterion*® is user-friendly.	37	85.5	9	18.4	3	6.1	3.73	.86
14. I am willing to use *Criterion*® again in the future if I have the chance.	13	26.5	20	40.8	16	32.7	2.94	.77
15. Generally speaking, I am satisfied with *Criterion*®.	11	22.4	23	46.9	15	30.6	2.92	.73

Note: 1. Agree % included the percentage of "Totally agree" and "Agree"

2. Disagree % included the percentage of "Totally disagree" and "Disagree"

Students' opinions of the optimal method for using *Criterion*® in writing classes (Item 17) indicated that all surveyed students (53 students, 100%) disapproved of their essays being evaluated only by *Criterion*®. Thirty-seven participants (69.8%) believed that they required teacher tutoring as a follow-up to the feedback generated by *Criterion*®. They indicated that *Criterion*® can assist in checking the organization of an essay, but that the teacher should be responsible for assessing content quality; that is, the teacher's explanation should supplement the AWE tools in class. Moreover, 16 students (30.2%) preferred that teachers did not use computerized scoring programs at all. They indicated that teacher scoring was significantly more objective.

Unlike the participants in previous studies who showed more trust in AWE programs (e.g. Chou & Chung, 2013; Li et al., 2014; Li et al., 2015), this study indicates that the participants lacked confidence in the programs, even though they were willing to have the technology integrated into a writing class. The following five reasons may explain why students did not trust the programs. First, the participants were second-year English majors, who have mastered fundamental linguistic forms and writing skills. They may consider content development to be more important than form, and prefer more freedom when constructing their essay, which may account for the reason why many students perceived the AWE program's assistance to be inadequate. In contrast, the students in the study of Chou and Chung (2013) and the study of Li et al. (2014) were non-English majors. Since AWE programs could help these low to intermediate English proficiency level students revise surface-level errors, most students were very satisfied with the programs. Second, student antipathy to machine-only scoring may be explained by Confucius' philosophy, which emphasizes the expertise of the teacher. Previous studies (Nelson & Carson, 1998; Wei, 1995) revealed that Asian students had a tendency to trust the teacher's feedback the most, because they regarded the teacher as an authority figure, which may explain why the majority of participants trusted the teacher's ability to compensate the inadequacy of machine feedback. Third, unlike the study of Li et al. (2015), which investigated students' perceptions through the method of interviewing, this study used a questionnaire survey to understand students' perspectives. This allows students to express their opinions without restraint. Given the fact that 83% of the participants in the study of Li et al. (2015) were Chinese students, it is easy to understand why the researchers found that students' positive attitudes toward the use of AWE seemed to be influenced by their instructors' attitudes and pedagogy. To show respect for their teachers, Chinese students tend not to orally criticize the teacher or the instruments used in the classroom, especially when interviewed by unknown people (Nelson & Carson, 1998). Fourth, the fairness of the automated scoring was inadequate. Several students in this study complained that the machine favored longer essays, and ignored coherence issues and illogical ideas. Other students stated that the program restricted their creativity in content development, because if they did not follow the conventional style of writing, they would receive lower scores. Fifth, the suggestions generated by the machine were not sufficiently clear. Students indicated that the AWE program may be beneficial for detecting fundamental form errors during the preliminary revision process, but they were unable to improve themselves without human communication and concrete teacher comments during the following revision stage.

4.3 The effect of diagnostic feedback on student revision

To understand and verify the effect of diagnostic feedback on student revision, the researcher downloaded the student essays (n = 530) and manually recorded the number of errors in grammar, usage, mechanics, and style in each student's first and final submissions of their essays. In this way, the researcher could calculate the revision rates in students' final submissions (see Tables 3, 5, 7, and 9). Next, the researcher examined the degree of accuracy of the machine messages from random student writing samples on the computer (n = 30) and from student essays reviewed during the tutoring time (n = 53). This examination verified whether the machine feedback was helpful for student revision, which helped to explain why some revision rates for errors were high but for other aspects were low. Systematic manual analysis was conducted for the four types of feedback messages (i.e. grammar, usage, mechanics, and style), and the researcher discovered that certain feedback from the program was beneficial to students, but others did not promote or even hinder student revision, because these feedback messages were incorrect. In fact, the machine sometimes gave invalid warnings when there were no errors in student writings. Occasionally, the machine did not detect some errors and therefore students did not notice these writing mistakes (see Tables 4, 6, 8 & 10).

4.3.1 Feedback for grammatical error

Table 3 shows that the grammar feedback assisted students in correcting approximately 70% of their grammatical errors. More specifically, it helped correct 76% of fragment or missing comma errors, 60% of run-on sentence errors, 100% of garbled sentence errors, 84% of subject-verb agreement errors, 75% of ill-formed verb errors, 78% of possessive errors, and 57% of wrong or missing word errors. However, it could not help revise pronoun errors.

Furthermore, false alarms for sentence fragments occurred when interrogative sentences such as "How about…" or exclamatory sentences such as "No way" were written. Additionally, the system gave false alarms for ill-formed verbs when the subject consisted of "something plus an adjective" in interrogative sentences such as "Will *something bad* become nice?" A sentence such as "You can ask the clerk give (to give) you" was misjudged as "subject-verb agreement error," when it is actually an issue of "ask + object + to infinitive."

Moreover, *Criterion*® failed to detect several grammatical errors such as verb forms, conjunctions, parts of speech, run-on sentences, noun clauses, word order of indirect questions, and tenses. For example, a sentence such as, "It lets us know what can we do (we can do)" was not detected by the system (see more examples in Table 4).

Table 3. The revision of grammatical errors

Essay	Errors in the 1st submission							
	A	B	C	D	E	F	G	H
1	142	27	0	17	16	0	1	9
2	74	49	3	34	27	0	2	0
3	44	38	1	30	20	0	4	2
4	64	57	0	23	17	1	1	11
5	30	43	2	36	22	0	1	1
Total	354	214	6	140	102	1	9	23
Essay	Errors in the final submission							
	A	B	C	D	E	F	G	H
1	102	18	0	6	4	0	1	6
2	29	21	0	6	10	0	0	0
3	17	13	0	2	6	0	0	1
4	18	25	0	3	0	1	1	3
5	10	13	0	6	6	0	0	0
Total	86	90	0	23	26	1	2	10
Revision Rate	76%	60%	100%	84%	75%	0	78%	57%

A= fragment or missing comma B= run-on sentences C= garbled sentences
D= subject-verb agreement E= ill-formed verbs F= pronoun errors
G= possessive errors H= wrong or missing word

Table 4. Examples of false alarms and undetected errors for grammatical errors

Category	Error type	Examples
False alarms	Fragment	1. "No way!", she shouted. 2. "What's the movie's name?" "No rules" 3. How about other people's situations?
	Verb forms (pronoun + adj.)	4. Will something bad become nice?
	Subject-verb agreement	5. You can ask the clerk give you.
Undetected errors	Ill-formed verbs	1. because they don't (*aren't*) afraid of being caught. 2. A lot of people continued came (*coming*) in. 3. I am amazing (*amazed*) by an egg. 4. because some rules are not allow *(allowed)*. 5. I felt that it *(is)* just like the end of the world. 6. Choice (*Choosing to*) stay in the beautiful dream is a good way.
	Conjunction, part of speech (word choice)	7. They can enter another's house and using *(use)* any furniture, *(and)* electricity *(electrical)* appliance.
	Run-on sentence	8. Even (*though*) the one who you beat has died, you won't get a penalty.
	Noun clause (determiner pronoun)	9. The police didn't come to handle accidents due to *(the fact that)* the world had no rules.
	Indirect question (word order)	10. It lets us know what can we do *(we can do)*.
	Tense	11. Without thinking, I knew that person who takes (*had taken*) it away.

4.3.2 Feedback for usage errors

Table 5 shows that students revised approximately 70% of usage errors for their final drafts. More specifically, it helped correct 82% of wrong articles, 66% of missing or extra articles, 76% of confusing words, 82% of wrong forms of word, 71% of preposition errors, 100% of nonstandard word forms, and 50% of negation errors.

Nevertheless, it was unhelpful for the revision of faulty comparisons. In addition, the system incorrectly advised students to use "cannot" instead of "can't." Furthermore, it was too strict with the usage of articles and prepositions (see Examples 2 and 3 in Table 6). Moreover, the system did not identify that a countable noun required either an article or a plural form (e.g. ".. cause traffic jams or even car accident(s)").

Table 5. The revision of usage errors

Essay	Errors in the 1st submission							
	A	B	C	D	E	F	G	H
1	15	71	19	4	1	12	1	0
2	14	87	17	1	0	15	0	1
3	15	81	22	0	1	24	0	1
4	24	62	13	1	0	3	0	0
5	21	139	16	5	0	28	0	0
Total	89	440	87	11	2	79	1	2
Essay	Errors in the final submission							
	A	B	C	D	E	F	G	H
1	5	44	7	0	0	4	0	0
2	2	21	5	0	1	1	0	1
3	5	24	5	1	0	12	0	0
4	2	34	0	1	0	4	0	0
5	2	28	4	0	1	2	0	0
Total	16	151	21	2	2	23	0	1
Revision Rate	82%	66%	76%	82%	0	71%	100%	50%

A= wrong article B= missing or extra article C= confusing words D= wrong form of word
E= faulty comparisons F= preposition errors G= nonstandard word form H= negation error

Table 6. Examples of false alarms and undetected errors for usage errors

Category	Error type	Examples
False alarms	Negation	1. But I can't.
	Missing or extra article	2. It seemed like everyone on the streets looked weird.
	Preposition error (collocations)	3. The law is a very important part in our life.
Undetected errors	Countable noun	1. cause traffic jams or even car accident(*s*).

4.3.3 Feedback for mechanical errors

Table 7 shows that the feedback was less beneficial for revising mechanical errors. More specifically, it was effective in revising the following errors: missing apostrophes (100%), fused words (100%), duplicates (100%), compound words (75%), and hyphen errors (67%). However, it was less effective in revising errors such as missing final punctuation (0%), missing commas (17%), missing capitalization of proper nouns (44%), missing question marks (44%), and spelling (47%).

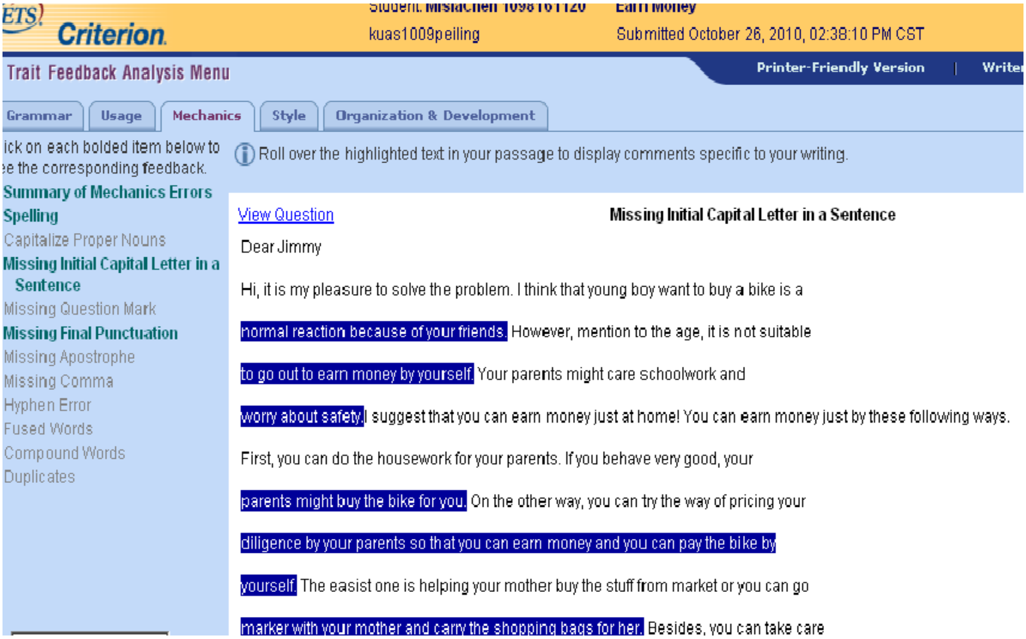
In fact, numerous feedback messages were incorrect, which may explain why the revision rates for several mechanical errors in students' final essays were less than 20%. For example, Figures 3 and 4 show that the system incorrectly asked students to use an initial capital letter for all first words in each line and to add punctuation marks for all final words in each line. Thereafter, it produced false alarm messages for spelling (e.g. people's names and proper names) and compound words (e.g. can not). Furthermore, it identified a statement beginning with a question word as a question; thus, it suggested that students use a question mark instead of a period at the end of the sentence. However, a sentence such as "What's more …" is actually a narrative sentence (see Table 8).

Several mechanical errors were untreated. For example, a sentence such as "People will increase their speed on the road or stress (street)" is grammatically acceptable but semantically incorrect (see more examples in Table 8). The system also misidentified an indirect question as a direct question; for example, the system did not indicate the mistake of using a question mark in a sentence such as “I can’t imagine if the world doesn’t have any rule what will happen?”

Table 7. The revision of mechanical errors

Essay	Errors in the 1st submission										
	A	B	C	D	E	F	G	H	I	J	K
1	67	1	9	6	0	0	1	0	0	14	0
2	61	1	22	2	16	0	1	1	0	21	1
3	56	7	7	3	2	0	4	0	0	18	0
4	45	0	8	2	0	0	0	0	1	18	2
5	157	0	4	3	0	0	0	2	2	10	0
Total	389	9	50	16	18	0	6	3	3	81	3
Essay	Errors in the final submission										
	A	B	C	D	E	F	G	H	I	J	K
1	29	2	38	3	39	0	1*	1	0	7	0
2	17	0	3	1	0	0	1*	0	0	7	0
3	22	3	3	2	0	0	3	0	0	2	0
4	19	0	0	1	0	0	0	0	0	2	0
5	120	0	0	1	0	0	0	0	0	2	0
Total	207	5	44	9	39	0	5	1	0	20	0
Revised Rate	47 %	44 %	12 %	44 %	0%	100 %	17 %	67 %	100 %	75 %	100 %

A= spelling B= missing capitalization of proper nouns C= missing initial capital letter in a sentence
D = missing question mark E = missing final punctuation F= missing apostrophe
G = missing comma H= hyphen error I= fused words J= compound words K= duplicates


ETS
Criterion
Student: MisiaChen 1098161120
Earn Money
kuas1009peiling
Submitted October 26, 2010, 02:38:10 PM CST
Trait Feedback Analysis Menu
Printer-Friendly Version | Write
Grammar
Usage
Mechanics
Style
Organization & Development
ick on each bolded item below to
ee the corresponding feedback.
Summary of Mechanics Errors
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
View Question

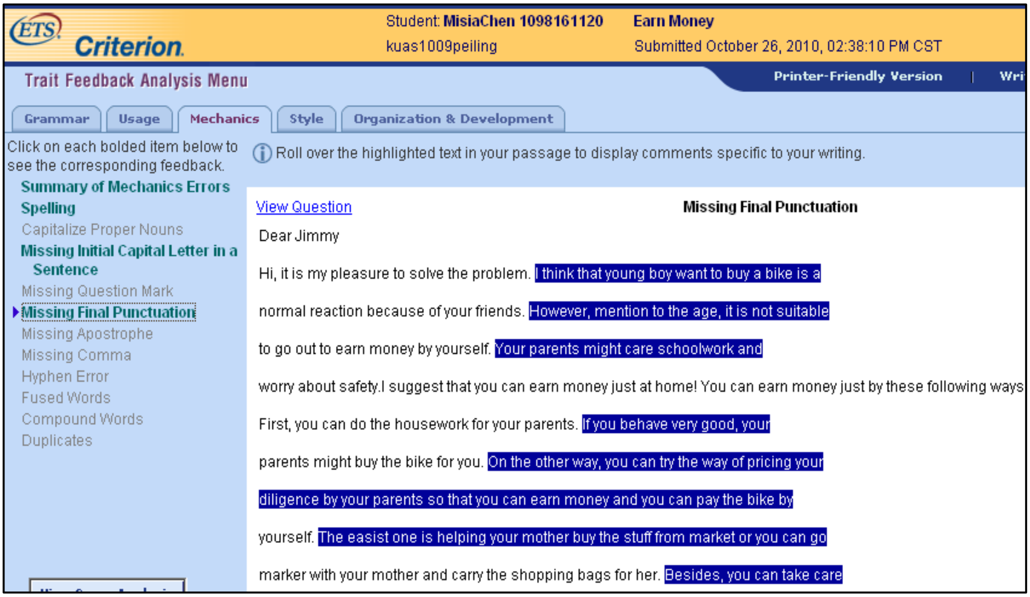
Roll over the highlighted text in your passage to display comments specific to your writing.

Missing Initial Capital Letter in a Sentence

Dear Jimmy

Hi, it is my pleasure to solve the problem. I think that young boy want to buy a bike is a normal reaction because of your friends. However, mention to the age, it is not suitable to go out to earn money by yourself. Your parents might care schoolwork and worry about safety.I suggest that you can earn money just at home! You can earn money just by these following ways. First, you can do the housework for your parents. If you behave very good, your parents might buy the bike for you. On the other way, you can try the way of pricing your diligence by your parents so that you can earn money and you can pay the bike by yourself. The easist one is helping your mother buy the stuff from market or you can go marker with your mother and carry the shopping bags for her. Besides, you can take care

Fig. 3. False alarm for missing initial capital letter in a sentence



Student: MisiaChen 1098161120 Earn Money
kuas1009peiling Submitted October 26, 2010, 02:38:10 PM CST

Trait Feedback Analysis Menu

Click on each bolded item below to see the corresponding feedback.

Roll over the highlighted text in your passage to display comments specific to your writing.

Summary of Mechanics Errors
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

Missing Final Punctuation

Dear Jimmy

Hi, it is my pleasure to solve the problem. I think that young boy want to buy a bike is a normal reaction because of your friends. However, mention to the age, it is not suitable to go out to earn money by yourself. Your parents might care schoolwork and worry about safety.I suggest that you can earn money just at home! You can earn money just by these following ways First, you can do the housework for your parents. If you behave very good, your parents might buy the bike for you. On the other way, you can try the way of pricing your diligence by your parents so that you can earn money and you can pay the bike by yourself. The easist one is helping your mother buy the stuff from market or you can go marker with your mother and carry the shopping bags for her. Besides, you can take care

Fig. 4. False alarm for missing final punctuation

Table 8. Examples of false alarms & undetected errors for mechanical errors

Category	Error type	Examples
False alarms	Spelling (proper nouns, human names)	1. I could not open up my Facebook and my blog. 2. Just like the famous Chinese philosopher, Lau-Tzu,
	Compound words	3. I can not imagine
	Missing question mark	4. What's more, students do not have to attend schools.
	Missing initial capital letter in a sentence	5. .. etc.
Undetected errors	Spelling (word choice)	1. He drove very fast sine (since) that he was afraid to be blamed by his boss. 2. Students were heating (*hitting*) teachers. 3. There will be a big mass (*mess*).
	Question mark	4. I can't imagine if the world doesn't have any rule what will happen? (.)

4.3.4 Feedback for style errors

Table 9 shows that the feedback was helpful for inappropriate words or phrases (100%), sentences beginning with coordinating conjunctions (89%), and too many long sentences (100%); however, the feedback for repetition of words (7%), too many short sentences (30%), and passive voice (22%) was not beneficial to students. The revision rates of these three errors were only 30% or lower. Moreover, the system failed to identify several semantically inappropriate words (see Table 10).

Table 9. The revision of style errors

Essay	Errors in the 1st submission					
	A	B	C	D	E	F
1	1731	0	19	50	0	1
2	1069	2	19	111	0	5
3	1339	0	19	76	0	21
4	1265	0	12	117	0	0
5	1027	0	7	90	0	22
Total	6431	2	76	444	0	49
Essay	Errors in the final submission					
	A	B	C	D	E	F
1	1665	0	4	73	0	0
2	838	0	0	58	0	5
3	1538	0	0	40	0	16
4	1066	0	0	81	0	2
5	871	0	4	60	0	15
Total	5978	0	8	312	0	38
Revised Rate	7%	100%	89%	30%	100%	22%

A= repetition of words B= inappropriate words or phrases
C= sentences beginning with coordinating conjunctions D = too many short sentences
E = too many long sentences F= passive voice

Table 10. Examples of false alarms & undetected errors for style errors

Category	Error type	Examples
False alarms	Repetition of words	1. Poor people must be very happy after they got so much money to do what they want to do. 2. When you open your eyes in the morning, you may begin to think what you have to do and where you should go.
Undetected errors	Inappropriate words or phrases	1. After knowing how terrible the world will be, you may have a sense of security *(insecurity)*.

5 Conclusion

5.1 Summary of the findings

This study examined students' perceptions of the three functions provided by *Criterion*® (i.e. scoring, diagnostic feedback, and writing assistance tools) and the use of *Criterion*® for learning English writing. In addition, the researcher calculated the revision rates for errors in students' writing from their first to final submissions, and conducted a thorough analysis of the diagnostic feedback messages to verify whether the program was helpful for student revision. The three main results were as follows. First, the results indicate that the majority of students appreciated the instant scoring speed and the error analysis of usage; however, many students were dissatisfied with the scoring rubric, the error analysis of style, and the writing assistance tool 'Plan'. Furthermore, most students preferred the combination of machine scoring with the teacher's explanations for a writing class. As participating students indicated, the biggest benefit of the program was that they were able to submit 15 drafts in 18 weeks, and they believed their English writing ability improved after using *Criterion*®; however, they did not attribute the improvement to the *Criterion*® program itself. The real reason for students' improvements may have resulted from the fact that they did extensive drafting and correction, and received both instant machine and teacher feedback several times. Therefore, one contribution of this study is that it supports the findings of previous studies (Dyson & Freedman, 1990; Flower & Hayes, 1981) which suggest that writing multiple drafts of a single essay is a necessary but insufficient condition for writing improvement, if the writer does not receive clear advice on how to correct their errors and how to consolidate their writing skills. Another contribution of this study is its discovery that the integration of teacher-student tutoring can effectively compensate the weaknesses of *Criterion*®. In this study, all students were privately tutored and advised of methods to improve their writing, and student complaints regarding the drawbacks of the program such as vague comments or incorrect diagnostic messages were noted. The teacher demonstrated her interest in her students by providing this type of guidance and counseling, which reassured students that her assistance was always available when they were unable to understand the machine's advice.

Second, based on the revision rates of errors in students' writing, the researcher confirmed student perceptions that some diagnostic feedback messages from *Criterion*® were useful for student revision, while others were not. For example, the messages were effective for the correction of grammatical errors (e.g. subject-verb agreement) and usage errors (e.g. non-standard word forms), because the revision rates for these issues were above 80%. However, the program feedback regarding various mechanical errors (e.g. missing final punctuation) and style errors (e.g. repetition of words, and too many short sentences) was not as beneficial to students. The revision rates for these items were all below 30%. Thus, these findings are consistent with results from previous studies (Burstein & Marcu, 2003; Chen, Chiu, & Liao, 2009; Grimes, 2008; Ware, 2011), which showed that AWE programs were more capable of detecting surface errors such as spelling or grammatical errors in students' writings.

Third, after examining the accuracy of the diagnostic feedback messages, the researcher found that some diagnostic feedback was incorrect, and some was confusing. For instance, the mechanical feedback suggested that students use a question mark at the end of a narrative sentence. Next, a feedback message like "You have repeated these words several times in your essay" was not very helpful. Students need to know how to substitute these words. Furthermore, several common errors in tenses, determiner pronouns, and collocations have not been included in the error list of *Criterion*®; therefore, these errors were not treated by the machine.

5.2 Limitations of the study and recommendations for future research

Since this study involved only 53 English major students at one institution, the results cannot be generalized to other populations. Further studies may consider increasing the number of participating students and investigating the effect of AWEs on basic-level learners. Moreover, future studies may examine the strategies teachers can implement for integrating the automated feedback as a revision aid with other learning activities such as peer-review activities. The collection of more forms of qualitative data such as classroom observations and interviews can be helpful in assessing the use of an AWE program. Another limitation of this study is that the questionnaire was not anonymous, which may have affected the participants' responses. Therefore, it would be better for future research to use an anonymous questionnaire to ensure the reliability of the survey.

5.3 Implications of the study

The findings of this study lead to two implications for program designers. First, designers are advised to include model essays that demonstrate strategies for revising style errors (e.g. repeated words). Second, they may wish to address the issue of false alarms and undetected errors related to capitalization, punctuation, modals, verb forms, conjunctions, compound words, pronouns, word choice, and the word order of indirect questions.

Several pedagogical implications can be drawn from this study. First, teachers must be aware of the limitations of current AWE systems and their counterproductive effects on students' beliefs regarding well-written essays. Several drawbacks of the AWE such as the preference of long essays according to the five-paragraph formula, and scoring based on superficial features of writing can mislead students toward writing a verbose but pointless essay. In this study, while all students gained a minimum score of three on a scale of 1 to 6, they seemed more concerned with improving the content of their writing rather than their scores; otherwise, they may not have been so frustrated when the machine was unable to provide advice on more complex revisions (i.e. structural development and meaning of a text). Perhaps one reason for the students' willingness to emphasize quality in their writing rather than writing scores is that they were assured during the first class that the automated scores would only comprise 10% of their final grade. Another possible reason is that these students may have already had strong intrinsic motivation to "master" writing skills rather than "perform" well in line with the scoring mechanism, because they were English majors. Therefore, teachers are recommended not to use the automated scores as the single measurement of student writing performance, and to be cautious of the potential development of students' false concepts regarding well-written essays when implementing AWE systems for students who may not be intrinsically motivated.

The second pedagogical implication of this study is that teachers must consider the English proficiency of students and the strategies for consoling students when they receive low scores or wrong advice. The participants in this study did not trust the machine feedback all the time, because they had adequate grammar knowledge to judge the correctness of these messages. Although some researchers (e.g. Cheng, 2006; Moseley, 2006) believe an AWE program is more appropriate for basic-level learners due to the fact that the program generally handles surface corrections, teachers of beginners must be informed that their students are the most vulnerable to confusing machine messages. In addition, since an AWE program values a five-paragraph essay, beginners may frequently receive low grades from the machine, if they are unable to write multi-paragraph essays with those elements. Therefore, when teachers use *Criterion*® for basic-level writing classes, they are advised to remind students that automated scores are only for reference and should provide consultations to clarify machine feedback.

Third, regarding the strategy of teacher-student conferencing, before teachers hold individual tutoring with students, teachers can request students to continue submitting their drafts to the AWE program until they reach a minimum satisfactory score. Another strategy entails the teacher giving priority to students with essay scores below the class average, because they may have more writing problems and require extra encouragement.

The final pedagogical implication is that teachers are advised to consider the social and communicative dimensions of writing when incorporating AWE into their teaching (Vygotsky, 1978). As this study pointed out, many students were discouraged, because their expressions were not "understood" by the machine, and because it failed to provide insight by "reading" their intentions. Students' comments revealed their desire to have meaningful communications with the machine. However, current AWE programs are not yet capable of these complex processes. Considering a single writing teacher cannot give instant feedback to the whole class, peers may become alternative human readers for genuine communicative purposes. For example, writing teachers can conduct a follow-up peer feedback activity after students have already received their score and feedback from the AWE program. In an ideal situation, students may submit their interim drafts to peers and resubmit their final drafts to the teacher for further teacher guidance or individual consultation on how to revise their essays. Teachers can focus their attention on students' writing problems of coherence and content development, because the machine or peers can handle spelling and grammar errors.

In conclusion, the use of technology alone cannot guarantee its effectiveness in improving student writing. Issues regarding how the program is implemented deserve more attention by language instructors. An AWE system can be a useful tool for improving students' writing skills, if teachers understand how to mediate between students and the machine. For example, teachers can use the beneficial aspects of AWE such as easy editing and immediate feedback to motivate students to write multiple drafts. Furthermore, they can use the program to perform a preliminary check of an early draft before returning it to students. Thereafter, teachers can retrieve student essays from their e-portfolio, interpret what students intend to write, and verify whether they have developed their ideas in later drafts. This study recommends that human feedback is required to compensate for the limitations of the machine's instructions and protect students from false messages. Thus, teacher involvement and/or peer feedback are especially important when the learning goal is to demonstrate the writer's creativity to real audiences.

Acknowledgements

A part of this paper was presented in the 2011 Symposium on Second Language Writing. June 9-11, 2011. Taipei, Taiwan. The author would like to express her gratitude to the audience for their inspiring comments on an earlier draft of this paper.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650/

Barrett, H. (2000). *Create your own electronic portfolio. Learning and leading with technology*. Retrieved from http://electronicportfolio.org/balance/index.html.

Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Kukick, K., Wolff, S., Lu., C., & Chodorow, M. (1998). *Computer analysis of essays*. Retrieved from http:// www.ets.org/research/dload/ncmefinal.pdf.

Burstein, J., & Marcu, D. (2003). A machine learning approach for identification of thesis and conclusion statements in student essays. *Computers and the Humanities, 37*(4), 455–467.

Busbee, E. (2001). The computer and the Internet: Are they really designed to play a major role in English teaching? *English Teaching, 56*(1), 201–205.

Butler-Pascoe, M. E., & Wiburg, K. (2003). *Technology and teaching English language learners*. Boston: Pearson Education, Inc.

Chen, C. F., & Cheng, W. Y. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Teaching, 12*(2), 94–112.

Chen, H. J., Chiu, T. L., & Liao, P. (2009). Analyzing the grammar feedback of two automated writing evaluation systems: *My Access* and *Criterion*. *English Teaching and Learning, 33*(2), 1–43.

Cheng, W. Y. (2006). *The use of a web-based writing program in college English writing classes in Taiwan – a case study of MyAccess* (Unpublished master's thesis). National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan.

Chodorow, M., & Burnstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (TOEFL Research Report No. RR-73, ETS RR-04-04). Princeton, NJ: ETS. Retrieved from https://www.ets.org/Media/Research/pdf/RR-04-04.pdf

Chou, H. N., & Chung, C. M. (2013). EFL college students' writing performance and perceptions of the use of an automated writing evaluation system. *Journal of English Education, 2*(1), 93–126.

Creswell, J. W. (1994). *Research design*. Thousand Oaks, CA: SAGE Publications.

Daiute, C. (1986). Physical and cognitive factors in revising: Insights from studies with computers. *Research in the Teaching of English, 20*, 141–159.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1), 1–36.

Dyson, A. H., & Freedman, S. W. (1990). *On teaching writing: A review of the literature*. Berkeley, CA: National Center for the Study of Writing.

Educational Testing Service (ETS). *Criterion® online writing evaluation service*. Retrieved from http://www.ets.org/*Criterion*

Flinn, J. Z. (1986). *The role of instruction in revising with computers: Forming a construct for good writing*. St. Louis, MO: University of Missouri.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387.

Frost, K. L. (2008). *The effects of automated essay scoring as a high school classroom intervention* (Unpublished doctoral dissertation). University of Nevada, Las Vegas, NV.

Grimes, D. C. (2008). *Middle school use of automated writing evaluation: A multi-site case study* (Unpublished doctoral dissertation). University of California, Irvine, CA.

Grimes, D., & Warschauer, M. (2006). *Automated essay scoring in the classroom*. Paper presented at the American Educational Research Association (AERA) Annual Conference, San Francisco, CA.

Grimes, D., &Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 4–44.

Gottlieb, M. (1995). Nurturing student learning through portfolios. *TESOL Journal, 5*(l), 12–14.

Hamp-Lyons, L. (1994). Interweaving assessment and instruction in college ESL writing classes. *College ESL, 4*(1), 43–55.

Herter, R. J. (1991). Writing portfolios: Alternatives to testing. *English Journal, 80*, 90–91.

Lai, Y. H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology, 41*(3), 432–454.

Lee, S. (2008). Exploring the potential of a web-based writing instruction program and AES: An empirical study using My Access. *Multimedia-Assisted Language Learning, 11*(2), 103–125.

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

Li, W. L. (2009).*The effects of integrating blogging into peer feedback revision on English writing performance and attitude of vocational high school students in Taiwan* (Unpublished master's thesis). National Taiwan Normal University, Taipei, Taiwan.

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*, 66–78.

LinHuang, S. H. (2010).*The exploitation of e-writing in an EFL classroom: Potential and challenges* (Unpublished master's thesis). I-Shou University, Kaohsiung, Taiwan.

Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco, CA: Jossey-Bass Publishers.

Moseley, M. H. (2006).*Creating recursive writers in middle school: the effect of a writing program on student revision practices* (Unpublished doctoral dissertation). Capella University, Minneapolis, MN.

Nelson, G. L., & Carson, J. G. (1998). ESL students' perceptions of effectiveness in peer response groups. *Journal of Second Language Writing, 7*, 113–131.

Otoshi, J. (2005). An analysis of the use of *Criterion* in a writing classroom in Japan. *The JALT CALL Journal, 1*(1), 30–38.

Ou, Y. S. (2011). *An explorative study on the use of the automated writing evaluation system- Applying Rogers' diffusion of innovations theory* (Unpublished master's thesis). National Kaohsiung Normal University, Kaohsiung, Taiwan.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Phillips, S. M. (2007). *Automated essay scoring: A literature review. Kelowna: TASA Institute, Society for the Advancement of Excellent in Education*. Retrieved from http://www.maxbell.org/sites/default/files/036.pdf

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation, 7*(26). Retrieved from http://pareonline.net/getvn.asp?v=7&n=26

Suh, J. S. (2002). Effectiveness of CALL writing instruction: The voices of Korean EFL learners. *Foreign Language Annals, 35*(6), 669–679.

Sullivan, N., & Pratt, E. (1996). A comparative study of two ESL writing environments: A computer-assisted classroom and a traditional oral classroom. *System, 24*(4), 491–501.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wang, Y. J. (2011). *Exploring the effect of using automated writing evaluation in Taiwanese EFL students' writing* (Unpublished master's thesis). I-Shou University, Kaohsiung, Taiwan.

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly, 45*(4), 769–774.

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

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research 10*(2), 1–24.

Wei, C. L. (1995). Sweet burdens: Perceptions of foreign-language majors about peer reviews. *Journal of Taichung Evening School* (National Chung Hsing University), *1*, 129–148.

Williamson, M. M., & Pence, P. (1989). Word processing and student writers. In B. K. Britton & S. M. Glynn (Eds.), *Computer writing environments: Theory, research, and design.* (pp. 93–127). Hillsdale, NJ: Lawrence Erlbaum.

Yang, N. D. (2004). Using My Access in EFL writing. *In Proceedings of the 2004 International Conference and Workshop on TEFL & Applied Linguistics* (pp. 550–564). Taipei, Taiwan: Ming Chuan University.

Yu, Y. T., & Yeh, Y. L. (2003). Computerized feedback and bilingual concordance for EFL college students' writing. In *Proceedings of the 2003 International Conference on English Teaching and Learning in the Republic of China* (pp. 35–48). Taipei, Taiwan: Crane.

Appendix A

Questionnaire
(5= Strongly agree, 4= Agree, 3= Unsure, 2= Disagree, 1= Strongly disagree)
I. My Satisfaction with the Functions of Criterion®

1. I think its scoring rubric is objective.	5	4	3	2	1
1.1 Please give reasons or examples.					
2. I think its holistic summary is useful.	5	4	3	2	1
2.1 Please give reasons or examples.					
3. I think its scoring speed is satisfying.	5	4	3	2	1
3.1 Please give reasons or examples.					
4. I think its error analysis of grammar is useful. (e.g., subject-verb agreement, ill-formed verb)	5	4	3	2	1
4.1 Please give reasons or examples.					
5. I think its error analysis of usage is useful. (e.g., article, preposition, word choice)	5	4	3	2	1
5.1 Please give reasons or examples.					
6. I think its error analysis of mechanics is useful (e.g., spelling, punctuation, capitalization)	5	4	3	2	1
6.1 Please give reasons or examples.					
7. I think its error analysis of style is useful (e.g., repeated words, long/short sentences, passive sentences)	5	4	3	2	1
7.1 Please give reasons or examples.					
8. I think its error analysis of organization development is useful	5	4	3	2	1
8.1 Please give reasons or examples.					
9. I think the function of 'Question Statement' in *Criterion*® is helpful.	5	4	3	2	1
10. I think the function of 'Plan' in *Criterion*® is helpful.	5	4	3	2	1
11. I think the 'Sample essays' provided by *Criterion*® is helpful.	5	4	3	2	1

I. My Perceptions of the Usefulness of *Criterion*® in the Writing Class

12. I think my English writing ability has improved after using *Criterion*®.	5	4	3	2	1
13. I think *Criterion*® is user-friendly.	5	4	3	2	1
14. I am willing to use *Criterion*® again in the future if I have the chance.	5	4	3	2	1
15. Generally speaking, I am satisfied with *Criterion*®.	5	4	3	2	1

16. I think *Criterion*® assisted me in the following aspects:
□vocabulary □grammar □logic development □organization
□sentence coherence □generating ideas □punctuation

17. In your opinion, what is the optimal method for employing *Criterion*® in writing classes?
□ Essays being evaluated only by *Criterion*®
□ Essays being evaluated only by teachers
□ Teacher tutoring as a follow-up to the feedback generated by *Criterion*®

18. Do you know any strategies to obtain higher scores in *Criterion*®?

19. What are the best parts of the *Criterion*® program?

20. What are the worst parts of the *Criterion*® program?

21. What are your suggestions for teachers regarding the implementing *Criterion*® in writing classes?