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

Electronic Journal of Foreign Language Teaching
2014, Vol. 11, Suppl. 1, pp. 149–164

Automated versus Human Scoring: A Case Study in an EFL Context

Shih-Jen Huang
(ufpadata@kuas.edu.tw)
National Kaohsiung University of Applied Sciences, Taiwan ROC

Abstract

One major development of computer technology involving English writing is automated essay scoring (AES). Previous research has investigated different aspects of AES in writing assessment, such as human and automated scoring differences (Bridgeman, Trapani, & Yigal, 2012), and students' essay structure identification (Burstein & Marcus, 2003). This study addresses two research questions. First, how does automated scoring differ from human scoring in EFL writing? Second, what are EFL learners' perceptions of AES and its effectiveness? The instruments involved in this study include an AES system developed by Educational Testing Service (ETS), *Criterion*, and a survey. The findings of the study suggest that the AES and human scoring are weakly correlated. Besides, the study also finds that an AES system such as *Criterion* is subject to deliberate human manipulation and can suffer from insufficient explanatory power of computer-generated feedback. The pedagogical implications and limitations of the study are also discussed.

1 Introduction

Automated essay scoring (AES), also known in the literature as *automatic essay assessment* (Landauer, Laham, & Foltz, 2003), *automatic essay evaluation* (Shermis & Burstein, 2013), *automated writing evaluation* (AWE) (Warschauer & Ware, 2006) or *machine scoring* (Ericsson & Haswell, 2006), is "the ability of computer technology to evaluate and score written prose" (Shermis & Burstein, 2003, p. xiii). An AES system incorporates various computing methods such as natural language processing (Burstein et al., 1998), text categorization (Larkey, 1998), latent semantic analysis (Foltz, Landauer, & Laham, 1999), and other technologies that are far beyond the scope of discussion in the paper. One of the major AES systems is *Project Essay Grade* (PEG), which is based on a calculation of a multiple-regression equation of two types of variables, proxes and trins (Page, 1968). *e-rater* and *Criterion* are both developed by Educational Testing Service (ETS) on the basis of natural language processing (NLP). *Intelligent Essay Assessor* (IEA), developed by Foltz, Landauer and Laham (1999), employs latent semantic analysis. *MyAccess*, owned by Vantage Learning, uses a computing method called *IntelliMetric*. Other minor AES systems, to name a few, are *Bayesian Essay Test Scoring System* (BETSY, http://ericae.net/betsy), *Intelligent Essay Marking Systems* (Ming, Mikhailov, & Kuan, 2000), and *Automark* (Mitchell, Russel, Broomhead, & Aldridge, 2002)

1.1 The rise of AES

The rise of AES may have come from two fronts. The first is the instructional overload. Teaching writing mostly includes class instruction in class and the heavy load of grading and commenting on students' papers after class. It cannot be denied that the after-class grading of students' writing assignments demands a great deal of time and attention. Mason and Grove-Stephenson (2002) estimated that grading students' writing takes up almost 30% of writing teachers' work. "Unfortunately, instructors don't always have sufficient time or resources to effectively grade students' compositions or provide feedback on their reading comprehension skills" (Calfee, 2000, p. 35). By adopting AES systems in the classroom, there is a possibility that writing instructors could be saved from being over-burdened by piles of writing assignments.

The second front is the human factor in the evaluation of writing proficiency. Technology is introduced to writing evaluation for an advantage that human scoring can hardly compete with. The major advantage of AES is consistency, since the criteria of scoring are programmed and executed as such. Meanwhile, human judgment is not a stable cognitive process (Bejar, 2011). It is constantly worn out over a lengthy period of time, distracted by environmental factors, or interrupted by unexpected interferences.

Furthermore, the subjective judgment of human scoring would pose a problem when large-scale high-stakes tests (e.g. Test of English as a Foreign Language [TOEFL] or Graduate Record Exam [GRE]) are involved. To achieve the maximum degree of fairness, writing performance is evaluated according to the intended rating scale and rubrics by test developers and assessed on the basis of agreement between raters. Nonetheless, the variance of human scoring inevitably leads to rating inconsistency in terms of intra-rater and inter-rater reliability (Kondo-Brown, 2002; Shohamy, Gordon, & Kraemer, 1992). As a result, AES is developed to be immune from human cognitive weakness and to improve rating consistency. Shermis, Koch, Page, Keith and Harrington (2002) reported that "automated essay-grading technique (r = .83) achieved statistically significant higher inter-rater reliability than human raters (r = .71) alone on an overall holistic assessment of writing" (p. 16).

1.2 Issues of AES effectiveness

Different issues such as AES performance, feedback, and revision have been discussed in previous studies. Burstein and Chodorow (1999) investigated the performance of *e-rater* on the Test of Written English (TWE). The participants were a group of English non-native speakers with a variety of L1 backgrounds and a group of English native speakers. They were required to write argumentative essays according to the given prompts. The results showed that there were significant differences between the scores of the English native speakers and the non-native speakers ($F(4,1128) = 76.561$, $p < .001$). Moreover, a comparison of the means of *e-rater* scoring and human scoring also indicated a statistically significant difference ($F(1,1128) = 5.469$, $p < .05$). In addition, Attali (2004) studied the relationship between automated feedback and revision on *Criterion* based on 9,275 essays, which were submitted to *Criterion* more than once. 30 specific error types under the categories of organization and development (e.g. thesis, main points, and supporting ideas), grammar (e.g. fragments, run-son sentences, and subject-verb agreement), usage (e.g. missing article, wrong form of word, and preposition error), mechanics (e.g. spelling, missing final punctuation, and missing initial capital letter in a sentence), and style (e.g. repetition of words, inappropriate words or phrases, and passive voice) were investigated. The results revealed that the rate of the error types was significantly reduced. Students were able to "significantly lower the rate of most of the 30 specific error types that were identified by the system and reduced their error rates by about one quarter (with a median effect size of .22)" (p. 17). Moreover, Higgins, Burstein, Marcu, and Gentile (2004) reported that *Criterion* was capable of identifying the relationship of the writing to the essay prompts, relationship of and relevance to discourse elements, and errors in grammar, mechanics, and usages. Other findings about the positive impact of AES on learners'

writing were also reported (Burstein, Chodorow, & Leacock, 2003; Chodorow & Burstein, 2004; Elliot & Mikulua, 2004; Fang, 2010; Wang, Shang, & Briody, 2013).

Different aspects of AES have been also explored. First, Chen and Cheng (2008) investigated the effectiveness of an AES system as a pedagogical tool in a writing class. They implemented *MyAccess* in three writing classes with a different emphasis on the use of assessment and assistance features. They found that students' attitude towards AES was not very positive partly because of "limitations inherent in the program's assessment and assistance functions" (p. 107). In particular, they advised that human facilitation is important in AES-implemented learning. Second, Lai (2010) conducted comparative research into student peer feedback and AES feedback with the implementation of *MyAccess* in a writing class. The results showed that students preferred peer feedback to AES feedback and that the former provided more help in improving their writing. Third, Wang and Goodman (2012) looked into students' emotional involvement during the AES process. They reported that students might experience the emotions of happiness, sadness, anxiety, and anger while they engaged in writing in an AES environment. Of the four types of emotion, students did not experience any stronger emotion of anxiety than other types of emotion. Pedagogically, it is suggested that "in cases where there is a complex interplay of positive and negative emotions – such as curiosity and sadness in this study – awareness can aid teachers who want to build on the strengths of positive emotions and mitigate the effects of negative emotions" (Wang & Goodman, 2012, p. 29).

Still, there are counter-examples to the effectiveness of AES. For example, Shermis, Burstein and Bliss' (2004) study (as cited in Warschauer & Grimes, 2008) investigated the impact of *Criterion* on a statewide writing assessment in Florida. *Criterion* was used in the experimental group, but the participants in the control group did not get access to *Criterion*. The results showed that there were no significant differences between the experimental and control groups. The participants in the experimental group did not demonstrate a statistically significant improvement in writing. Moreover, Grimes and Warschauer (2006) did not find any statistically significant improvement in writing in their investigation of the effectiveness of *MyAccess*, another AES system developed by Advantage Learning. In addition, Otoshi (2005) found that *Criterion* had difficulty detecting errors specifically related to nouns and articles. *Criterion* also failed to detect the errors related to discourse context, topic content, and idiosyncrasies (Higgins, Burstein, & Attali, 2006). More previous studies have failed to firmly establish the pedagogical effect of AES on learners' improvement of writing (Grimes & Warschauer, 2006; Hyland & Hyland, 2006; Yang, 2004).

1.3 Human and AES rating

In addition to the effectiveness of AES, one major area of investigation is the comparison of human and automated scoring differences (Attali & Burstein, 2006; Bridgeman, Trapani, & Yital, 2012; Page, 1968; Wang & Brown, 2008). The previous studies demonstrated positive correlations between human scoring and different AES systems. Page (1968) reported a correlation coefficient of .77 between *Project Essay Grade* (PEG) and human scoring. Attali and Burstein (2006) compared *e-rater* with human scoring and also reported a very high correlation up to .97. Foltz et al. (1999) compared the scores of over 600 GMAT essays graded by *Intelligent Essay Assessor* (IEA) with the scores of human raters and achieved a correlation of .86, which is almost the same as the inter-rater correlation of ETS human raters.

However, Wang and Brown (2008) reported a different outcome. They found a low correlation of scores between human raters and *IntelliMetric*'s *WritePlacer Plus*: "The correlational analyses, using the nonparametric test Spearman Rank Correlation Coefficient, showed that the overall holistic scores assigned by *IntelliMetric* had no significant correlation with the overall holistic scores assigned by faculty human raters, nor did it bear a significant correlation with the overall scores assigned by NES human raters" (Wang & Brown, 2008, p. 319).

1.4 Research questions

The study seeks to answer the following research questions: First, how does AES differ from human scoring in EFL writing? Second, what are EFL learners' perceptions of AES and its effectiveness?

2 Methodology

2.1 Participants

The participants were 26 English majors in a public technical university in Taiwan. They were taking the second-year English Writing course, which is a required course in the curriculum. There were 5 male and 21 female students. The age ranged from 18 to 20 years old. All of them had graduated from vocational high schools and none of them were English native speakers or possessed near-native proficiency of English. In the first-year writing course, the participants had practiced writing different types of paragraphs and learned some basics of essay writing. Furthermore, they were digital natives who were very familiar with writing on the computer in an Internet environment.

2.2 Data collection: Criterion

The first instrument of data collection was *Criterion*. *Criterion* was initially developed by ETS to assist in the rating of GMAT essays. The two components of *Criterion* are *e-rater* and *Critique* (Burstein, Chodorow, & Leacock, 2003). *e-rater* is AES technology that provides holistic scores, and *Critique* employs Trait Feedback Analysis to supply immediate diagnostic feedback and revision suggestions to students. *Criterion* analyses five traits of essays: grammar, usage, style, mechanics, and organization and development. In addition, *Criterion* provides supporting tools for online writing. One of these is the Outline Organizer, which provides several different outline templates to help students with pre-writing planning; students fill their ideas in the blanks of outline templates and *Criterion* will convert the outline templates into workable outlines in the writing box. Another tool is the Writer's Handbook, which is an online version of a grammar and writing reference book; students do not need to leave the working area of *Criterion* when they are in need of grammar or writing consultation.

A typical writing process is a cycle of composing, submission, feedback, and revision. A student is expected to make a pre-writing plan, which is an optional activity depending on students' individual writing routines, convert the pre-writing plan into a text composition in the working area of *Criterion*, submit a draft to *Criterion* for feedback, and revise the draft according to the feedback. After revision, the draft enters the writing cycle again until satisfaction is achieved.

2.3 Data collection: Survey

The second instrument of data collection was a 5-point Likert scale survey to elicit the participants' responses to AES. To construct the survey, a pilot study was conducted to elicit the initial responses to the implementation of *Criterion* in class.

The pilot study was composed of five open-ended questions. They were:

1. What is the strength of *Criterion*?
2. What is the weakness of *Criterion*?
3. How does *Criterion's* feedback help you revise the essays?
4. To what extent does *Criterion's* scoring reflect the quality of your essays?
5. Do you recommend that the writing class continue to use *Criterion* in the next semester?

The questions were used for class discussion and the participants were required to write down their responses. The responses were then used as the basis to produce survey statements.

The survey was divided into four sections. The purpose of the first section was to understand how the participants used *Criterion*. The second section (Feedback and interaction) was used to find out how the participants responded to the diagnostic feedback from *Criterion*. The third section (Assessing AES performance) sought to determine how the participants evaluated the assessment performance of *Criterion* as an AES system for online writing. The last section asked the participants to give an overall evaluation of *Criterion*.

The preliminary version of the survey was reviewed by a colleague at another university to achieve expert validity. It was also discussed in class to check the participants' comprehension of each statement. Clarifications about the wording and phrasing of the statements in the survey were offered and revisions were made accordingly. An intra-class reliability test was conducted. The Cronbach alpha was .876, which indicated a rather satisfactory reliability of the survey. In the end, a survey of 20 statements plus one open-ended question was produced.

2.4 Procedures

The implementation of *Criterion* spanned through the spring semester, 2013. In the first week, the participants were told that an AES system would be used as part of the writing course. They were required to complete at least four essays by the end of the semester, depending on their writing performance and progress. In the second week, the participants were given *Criterion* accounts and passwords as well as instructions on how to use *Criterion*. In the third week, the participants started to write the first essay in class. The in-class writing aimed at orienting the participants towards the AES program. The participants were encouraged to explore the outline templates and writers' handbook, and to experiment with submitting essays and reading diagnostic feedback and analysis. The number of online submissions for revision was unlimited. A brief class discussion was held immediately after the practice in class. The participants were required to achieve a minimum score of 5 out of 6 on a holistic grading scale. To get the benchmarked grade, they had to repeatedly revise their *Criterion*-rated essays according to *Criterion* feedback. The average number of essay submissions for revision is 7.62. Lastly, although in-class *Criterion* writing classes were scheduled, the participants did not always finish writing in the two-hour period. Since the submission-revision loop is an important process of learning, the participants were encouraged to keep revising after receiving feedback and thus were allowed to complete the unfinished parts after class. However, they were explicitly advised that plagiarism was strictly forbidden. They were not allowed to copy-and-paste articles from any sources. To further secure the integrity of essays, random selected segments in each essay were checked by means of a Google search. The main purpose of this was to check whether there was any identical match between the selected segments and the results yielded by the Google search. No identical match was found. The four topics of the essays are listed in Table 1.

At the end of the semester, 103 *Criterion* essays on the four topics were collected. Two raters were invited to score the 103 essays. One rater, designated as "Human Rater 1", was a Ph.D. candidate in TESOL and the other, "Human Rater 2", a lecturer with a master's degree in TESOL. Both raters were experienced instructors in teaching college-level writing for years. Furthermore, the two raters had never implemented any AES systems in their writing classes. The rationale to exclude raters who had used AES systems (such as *Criterion*) in writing classes is that they would have known how it tended to score students' essays. The likelihood that they could have been subconsciously influenced to score the participants' essays as an AES system would could not be completely ruled out. The raters were given an official *Criterion* scoring guide (see Appendix 1). The *Criterion* scoring guide (http://www.ets.org/Media/Products/Criterion/topics/co-1s.htm) is a reference that specifies the writing quality and features that correspond to the *Criterion* scores from 1 to 6 on the grading scale. It is included in *Criterion* as a reference for students. The inter-rater reliability was 0.705, which was an acceptable value for the scoring consistency of the two raters. In addition to the collection of the essays, the participants completed the survey of 20 statements in the final week of the semester (see Appendix 2). One open-ended question was also included in the survey to ask the participants to report their thoughts and reflections regarding dif-

ferent aspects of AES and *Criterion*. 24 copies of the survey were collected because two participants did not show up in class.

Table 1. The prompts of essay topics

Topics	Prompts
A+ professor	What makes a professor great? Prominence in his or her field? A hot new book? Good student reviews every semester? What standards should be used to assess the quality of college faculty members? Support your position with reasons and examples from your own experience, observations, or reading.
Billionaire dropouts	A number of high-profile businesspeople are college dropouts who abandoned college to focus on pursuing their dreams. With such success stories in the high-tech and entertainment fields, it is easy to understand the temptation some students feel to drop out of college to pursue their entrepreneurial dreams. If a friend were thinking of dropping out of college to start a business, how would you decide whether or not to encourage that decision? Support your position with reasons and examples from your own experience, observations, or reading.
Defining a generation	Every generation has something distinctive about it. One generation may be more politically active, another more self-centered, while yet another more pessimistic. Identify a significant characteristic of your own generation, and explain why you think that this characteristic is good or bad. Support your point of view with examples from your own experience, reading, or observation.
Gap year	At least one major United States university officially recommends that high school students take a year off — a so-called "gap year" — before starting college. The gap year idea is gaining popularity. Supporters say it helps students mature and focus on their goals. Detractors say taking a year off from school will get students off track and that many will never go to college if they don't go right away. Do you think taking a gap year is a good idea? Why or why not? Support your point of view with examples from your own experience, reading, or observation.

3 Results and discussion

3.1 RQ1: How does AES differ from human scoring in EFL writing?

3.1.1 Quantitative results

Two sets of paired-samples t-tests were conducted to compare AES and human scoring. In the first set, there was a significant difference between the scores of Human Rater 1 (M=3.63, SD=0.078) and AES (M=4.65, SD=0.074): t(102)=-10.557, p=0.000. In the second set, there was also a significant difference between the scores of Human Rater 2 (M=3.57, SD=0.082) and AES (M=4.65, SD=0.074): t(102)=-11.194, p=0.000. The results suggest that the scores given by automated scoring and human scoring did differ. Specifically, the average scores of the human raters (M=3.63 for Human Rater 1 and M=3.57 for Human Rater 2) is lower than the average scores of AES (M=4.65). Furthermore, a correlation test was conducted and found that there was a very weak positive correlation between AES and Human Rater 1 (r=0.193, n=103, p=0.050). However, there was a statistically significant positive correlation between AES and Human Rater 2 (r=0.244, n=103, p=0.013).

3.1.2 What was measured

Although the comparison of automated scoring and human scoring yielded statistic differences and weak correlations, what is missing is what exactly was measured in automated scoring and human scoring. AES relies on linguistic, mathematic, or algorithmic models to calculate the sur-

face grammatical features and semantically relevant vocabulary and assess the writing quality of essays. However, there is evidence that human raters examine other aspects of writing in addition to grammar and words when they evaluate essays. In a study that compared scores given by English native speakers and non-native speakers, Shi (2001) reported that there was no statistically significant differences between the two groups of raters. Yet, he found that, based on the raters' self-reports, the raters focused on different sets of linguistic and discourse features in the process of rating, even though they gave the same scores. More emphasis was given to contents and language use by English native speakers, but non-English native speakers would pay more attention to essay organization and length. Similarly, Kondo-Brown (2002) showed that individual raters displayed differences of emphasis in their ratings, even though a high correlation of inter-rater scoring was achieved. She concluded that "… it may not easily eliminate rater characteristics specific to them as individuals. In other words, judgments of trained teacher raters can be self-consistent and overlapping in some ways, but at the same time, they may be idiosyncratic in other ways" (p. 25). In short, while AES systems tend to assess the surface linguistic and organizational features, human rating would evaluate essays with a different emphasis on linguistic and discourse aspects.

3.2 RQ2: What are EFL learners' perceptions of AES and its effectiveness?

3.2.1. Quantitative results

In the first section (see Table 2), the participants did not show a preference to use the writing support provided by *Criterion*, although they were told that *Criterion* provided online writing aids in the second week of the writing class. They did not make full use of the Outline Organizer (Q1, M=2.71, SD=0.91). Over a quarter of the participants (29.17%) did not use the Outline Organizer. Also, the Writer's Handbook was not particularly favored by the participants (Q2, M=3.00, SD=0.83). 58.33% of the participants did not show any particular interest in using the Writer's Handbook provided by *Criterion* as a reference. In addition, 66.67% of the participants agreed that the prompts of essay topics were clear enough to avoid writing deviations from the topics (Q3, M=3.83, SD=0.70). 62.5% of the participants would more or less write in the way *Criterion* expected them to write to get a higher score (Q4, M=3.54, SD=0.66). It indicates a possible washback effect on the participants, which will be discussed later.

Table 2. Section 1 – General use of *Criterion*

Statement	5 (%)	4 (%)	3 (%)	2 (%)	1 (%)	Mean	SD
1. I used the Outline Organizer provided by *Criterion* to help me organize essays.	4.17	29.17	37.50	25.00	4.17	2.71	0.91
2. I used the Writer's Handbook provided by *Criterion* to help me improve English.	4.17	16.67	58.33	16.67	4.17	3.00	0.83
3. The description of essay prompts was clear enough for me to know what the topic asks of.	16.67	50.00	33.33	0.00	0.00	3.83	0.70
4. I tended to write essays in the way *Criterion* expects me to do to get a higher score.	0.00	62.50	29.17	8.33	0.00	3.54	0.66

Note. 5= Strongly Agree, 4= Agree, 3= Neutral, 2= Disagree, 1= Strongly Disagree

The second section (see Table 3), "Feedback and interaction," was used to find out how the participants perceived the diagnostic feedback from *Criterion*. The participants demonstrated a moderately positive response to the diagnostic feedback and thought that the feedback would be useful in improving their grammar (Q5, M=3.79, SD=0.66), usages (Q6, M=3.63, SD=0.65), mechanics (Q7, M=3.63, SD=0.71), style (Q8, M=3.38, SD=0.71), and organization and development (Q9, S=3.42, SD=0.72). However, the confidence in *Criterion's* feedback decreased from grammar (M=3.79), usages (M=3.63), mechanics (M=3.63), organization and development (M=3.42), to

style (M=3.38). In addition, the participants thought that the description of diagnostic feedback was just clear enough to be understood for further revision (Q10, M=3.50, SD=0.88). Over half of the participants (Q11, M=3.63, SD=0.71) would regard the submission of essays and *Criterion's* immediate feedback as an interaction.

Table 3. Section 2 – Feedback and interaction

Statement	5 (%)	4 (%)	3 (%)	2 (%)	1 (%)	Mean	SD
5. *Criterion's* feedback was useful to improve the grammar of essays.	8.33	66.67	20.83	4.17	0.00	3.79	0.66
6. *Criterion's* feedback was useful to improve the usage of essays.	4.17	58.33	33.33	4.17	0.00	3.63	0.65
7. *Criterion's* feedback was useful to improve the mechanics of essays.	8.33	50.00	37.50	4.17	0.00	3.63	0.71
8. *Criterion's* feedback was useful to improve the style of essays.	0.00	50.00	37.50	12.50	0.00	3.38	0.71
9. *Criterion's* feedback was useful to improve the organization and development of essays.	0.00	54.17	33.33	12.50	0.00	3.42	0.72
10. *Criterion's* feedback could be clearly understood for revision.	8.33	45.83	37.50	4.17	4.17	3.50	0.88
11. I considered the submission of essays and *Criterion's* immediate feedback as an interaction between Criterion and me.	8.33	50.00	37.50	4.17	0.00	3.63	0.71

Note. 5= Strongly Agree, 4= Agree, 3= Neutral, 2= Disagree, 1= Strongly Disagree

The third section (see Table 4), "Assessing AES performance," sought to determine how the participants evaluate the assessment performance of *Criterion* as an AES system. In general, *Criterion* can indicate grammatical errors (Q12, S=3.75, SD=0.74), usage errors (Q13, M=3.71, SD=0.75), mechanics errors (Q14, M=3.58, SD=0.88), style errors (Q=15, M=3.46, SD=0.72), and organizational errors (Q16, M=3.46, SD=0.78). Additionally, when asked whether the *Criterion* scoring truthfully reflected the writing quality of their essays (Q17) and whether *Criterion* scored essays as expected (Q18), the participants showed relatively lower confidence in the assessment performance of *Criterion*. (Q17, M=3.04, SD=0.75; Q18, M=3.00, SD=0.78). Over half of the participants (Q17, 58.33%) maintained a neutral attitude towards the assessment capability of the *Criterion* scoring. Only a quarter of the participants (25%) thought that the *Criterion* scoring truthfully reflected the writing quality of their essays.

Table 4. Section 3 – Assessing AES performance

Statement	5 (%)	4 (%)	3 (%)	2 (%)	1 (%)	Mean	SD
12. *Criterion* can satisfactorily indicate grammatical errors of essays.	12.50	54.17	29.17	4.17	0.00	3.75	0.74
13. *Criterion* can satisfactorily indicate usage errors of essays.	12.50	50.00	33.33	4.17	0.00	3.71	0.75
14. *Criterion* can satisfactorily indicate mechanics errors of essays.	16.67	33.33	41.67	8.33	0.00	3.58	0.88
15. *Criterion* can satisfactorily indicate style errors of essays.	4.17	45.83	41.67	8.33	0.00	3.46	0.72
16. *Criterion* can satisfactorily indicate organization and development errors of essays.	4.17	50.00	33.33	12.50	0.00	3.46	0.78
17. *Criterion* scoring truthfully reflected the writing quality of my essays.	0.00	25.00	58.33	12.50	4.17	3.04	0.75
18. *Criterion* scored my essays as I had expected.	0.00	29.17	41.67	29.17	0.00	3.00	0.78

Note. 5= Strongly Agree, 4= Agree, 3= Neutral, 2= Disagree, 1= Strongly Disagree

The last section (see Table 5) asked the participants to give an overall evaluation of *Criterion*. *Criterion* was used as a good learning tool of English writing (Q19, M=3.29, SD=0.86) and the participants would recommend the future implementation of AES in the writing class (Q20, M=3.29, SD=0.91).

Table 5. Section 4 – Overall evaluation

Statement	5 (%)	4 (%)	3 (%)	2 (%)	1 (%)	Mean	SD
19. I used *Criterion* as a good learning tool of English writing.	0.00	45.83	45.83	0.00	8.33	3.29	0.86
20. I recommended that *Criterion* be implemented in future writing classes.	0.00	50.00	37.50	4.17	8.33	3.29	0.91

Note. 5= Strongly Agree, 4= Agree, 3= Neutral, 2= Disagree, 1= Strongly Disagree

3.2.2 Qualitative responses

3.2.2.1 Feedback

The participants expressed a variety of reflections and attitudes towards *Criterion* in their responses to the open-ended question of the survey. Two major threads emerged in the participants' responses. The responses of selected participants quoted below were grammatically revised for clarity, but the propositions of the sentences remained intact.

One major thread concerned the computer-generated feedback. The participants' responses in the survey reveal several points of weakness in *Criterion*. First, while the participants welcomed the immediate feedback after the online submission, the participants were not satisfied by the quality of the computer-generated feedback. In terms of accuracy, the corrective feedback did not successfully indicate common errors, as evidenced in Chen, Chiu, and Liao (2009). They analyzed the grammar feedback of 269 student essays from two AES systems (*Criterion* and *MyAccess*) and found that *Criterion* could not mark common grammatical errors regarding word order, modals, tenses, collocations, conjuncts, choice of words, and pronouns. Similarly, Crusan (2010) also reported that the feedback of another AES system, *Intelligent Essay Assessor* (IEA), was "vague and unhelpful" (p. 165). Besides, the participants reported that the quality of feedback was unsatisfactory. The automated feedback consisted of pre-determined formulaic messages. The feedback could indicate an error, but it could not specifically show the participants how the error should be corrected:

> I don't think the feedback of *Criterion* is good because I got almost the same feedback in different essays. (S20)

> The feedback given by *Criterion* is not as specific and complete as the feedback given by teachers. This made students confused and even frustrated. Some of my classmates spent so much time revising their essays to achieve the score the teacher required. They did not receive specific suggestions from *Criterion* to know the problems in their essays. They didn't have directions to improve their writing. (S07)

> *Criterion* gives me some suggestions about my essay, but some of the suggestions are not clear enough. For example, 'preposition errors'? (S01)

> It [*Criterion*] gives you a general direction to revise your essay. It will indicate the weakness of your essay. If your essay does not provide examples to support your arguments, it will tell you to provide examples without explanation. Sometimes the feedback is so ambiguous that it is hard for my revision to meet *Criterion's* standard. (S11)

Moreover, the participants also expected pedagogical intervention from human instructors. It is not surprising that the participants preferred human feedback. One reason is that, in response to the rigid computer-generated feedback, human feedback is detailed and specific, and they *understand* what is required to improve writing. More importantly, key words such as "negotiate" and "discuss" in the participants' responses reflect the essential process of negotiation in student writing. From the perspective of Bakhtinian dialogic perspective, creating the text "cannot be operationalized as the acquisition of the set of static conventions shaping meaning in texts but as a dynamic negotiation that involves the writer in the process of moving with and against given resources, adopting, bending, and diverting available textual patterns and resources to attain his/her communicative ends" (Donahue, 2005, p. 144). As presented in the previous paragraph, the computer-generated feedback fails to a large extent to be the "given sources" for the participants to achieve their communicative ends for writing improvement. Hence, the participants expected more and further teacher involvement in the writing process:

> I prefer the feedback from teachers because I can negotiate with teachers about my opinions that the computer cannot recognize. (S19)

> I think that the teacher can pick up mistakes more correctly, especially contents. (S02)

> The feedback from teachers helps me more. The suggestions given from *Criterion* are too formulaic, but those from teachers are easier to understand. (S15)

> … because I can explain my thought to my teacher and let him know why I write an essay in this way. And I can discuss with him. (S11)

3.2.2.2 Manipulation

Although the participants were not computer experts, they learned several programming fallacies to manipulate AES scoring. First, the participants noted that the scoring capability is not properly balanced between form and content. *Criterion*, like other AES systems, does not really "read" essays. If a participant wrote creatively, *Criterion* might have difficulty identifying and recognizing the participant's creative use of language and might then give a low score:

> *Criterion* doesn't recognize new words or celebrities' names. For example, in the essay Billionaires Dropouts, it [*Criterion*] doesn't know the name of the Facebook founder, Mark Elliot Zuckerberg. Moreover, the word "gapper" is used to call those who take a gap year, but this word would be marked as a wrong word. (S11)

> It [*Criterion*] shows me most of my writing problems, but I wonder if it can really understand what I want to tell readers. Maybe my writing ability does not reach *Criterion's* standards, but, on the other hand, maybe it's because *Criterion* doesn't understand what I want to express. (S25)

Moreover, surface grammatical features are valued more by AES systems. As some participants mentioned, they "cheated" the AES system by typing more transitional keywords or phrases such as "first," "however," or "in conclusion," even though the sentences were not as logically linked as the transitional words indicated.

Furthermore, the participants discovered some predictable scoring factors. For example, essay length is one of the predictable rating factors (Attali & Burstein, 2006). Since contracting propositional content is not the strength of an AES system such as *Criterion*, adding a few additional sentences can manipulate the AES scoring:

> *Criterion* has some blind spots. For example, I can write one or two stupid sentences in a paragraph without being detected as long as my whole concept is not out of topic. (S06)

> For example, I had revised my essays several times and still got 4 out of 6. But it's strange that I finally scored 5 just because I added one sentence in the first paragraph. (S01)

The participants' empirical hands-on "cheating" strategies were also reported by Powers, Burstein, Chodorow, Fowles and Kukich (2001). They invited staff from ETS, researchers in ESL, computational linguistics or artificial intelligence, and opponents of automatic essay scoring to compose and submit essays that could, in their word, "trick" *e-rater* into giving a score higher or lower than what the essays deserved. The invited essay writers used the following "cheating" strategies to receive higher or lower scores from *e-rater*. One strategy was that the invited essay writers deliberately produced lengthier essays and repeated canned texts or even a paragraph. Their accounts of "cheating" strategies are excerpted below from Powers et al. (2001, p. 35):

> The paragraph that is repeated 37 times to form this response probably deserves a 2. It contains the core of a decent argument, possibly the beginning of a second argument, and some relevant references, but the discussion is disjointed and haphazard, with the references and transition words badly misdeployed. *E-rater* might give too high a rating to the single paragraph because of the high incidence of relevant vocabulary and transition words.

> … just a rambling pastiche of canned text (often repeated for emphasis), possibly relevant content words, and high-falutin function words.

Another strategy was that the invited essay writers wrote many transitional phrases and prompt-related content and function words (Powers et al., 2001, p. 36):

> It uses many transitional phrases and subjunctive auxiliary verbs; it has fairly complex and varied sentence structure; and it employs content vocabulary closely related to the subject matter of the prompt. All of these features seem to be overvalued by *e-rater* (even though length per se is not supposed to be a factor).

3.2.2.3 Washback effect

Washback effect is the influence of testing on teaching and learning (Cheville, 2004; Hillocks, 2002). Although AES in a classroom setting does not directly involve high-stakes testing such as TWE (Test of Written English, ETS), it is closely related to writing assessment, because AES still rates the participants' essays with a grade point on a 6-point scale. As a result, it inevitably leads to a possible washback effect on the way the participants write and compose essays. In other words, the participants, intentionally or unintentionally, would write in accordance with the parameters established by AES systems.

> No. I don't want to write essays to the computer anymore. Our writing is way too much like our classmates'. Our essays appeared to be very similar. The first paragraph should be an introduction, the second paragraph should support your main idea, the third paragraph should be another support, and the last paragraph needs to sum up. If I want to make some differences in my essay, what grade will I get? Maybe 3 or 4 or whatever, but it won't be a good grade. (S22)

> Finally, I understand one thing. We just revised for the system in order to get a good grade. We don't revise to correct mistakes. The computer is not always right. (S19)

> It (*Criterion*) will limit creativity, I think. When I wrote my first essay on *Criterion*, I got a 4. But I wrote that essay in a very happy mood and thought a lot about that topic. I also gave reasons to explain every point that I mentioned in that essay. I don't know what went wrong and how I should revise my writing. (S24)

However, what type of washback effect was there on the participants? On one hand, the participants' responses apparently did not seem to indicate positive washback effect because of the in-

creasing similarities among the participants' writing, the more frequent use of AES-conditioned transitional phrases or prompt-related content words, the bounded creativity of writing, and the confusing purpose of writing. On the other hand, it was observed that the participants were getting familiar with the basic components of an essay such as the thesis statement, topic sentences, logical transitional phrases for textual development, and the conventional deployment of 5-paragraph essays.

The potential washback effect should be taken into pedagogical consideration when an AES system is implemented in a writing class. First, to avoid a negative washback situation in which students only write what AES systems want them to write to get a high score, constant teacher monitoring during the implementation of AES in a writing class is required. As previously noted, AES systems are more form-focused and human feedback is preferred to them. Therefore, it is suggested that the social gap of meaning negotiation be filled by additional feedback provided by teachers or by holding teacher-student writing conferences for individual coaching. Second, teachers have to clearly define the role of AES, depending on the writing approach that teachers intend to adopt in a writing class. For example, if the writing class is process-oriented, AES would be a pedagogical tool to induce the positive washback effect (e.g, in the form of constant AES feedback and opportunities for revisions) on students' writing, because they would be guided to learn the expected form of an essay.

4 Conclusion

The findings of the study can be summarized as below. First, AES and human scoring differed in scoring results and the correlation between them is weak. AES tended to score higher than human raters. Second, in general, the participants held a positive attitude toward the use of AES, although an AES system could be susceptible to users' deliberate manipulation of text input.

It is generally agreed that AES systems cannot replace human instruction (Chen & Cheng, 2008; Warschauer & Ware, 2006), and some pedagogical implications are noted. To begin, it is suggested that teachers pay further attention to the social and communicative aspects of writing, when an AES system such as *Criterion* is used in a writing class. Chen and Cheng (2008) argued that writing demands more than linguistic accuracy. It is a meaning negotiation process with communicative purposes. However, AES systems fail to fill the social gap of meaning negotiation, because most AES systems "are theoretically grounded in a cognitive information-processing model, which does not focus on the social and communicative dimensions of writing" (Chen & Cheng, 2008, p. 96). Take a writing class within the process approach, for example. The process approach proposes the recursive stages of prewriting, drafting, revising, and editing (Tribble, 1996). Of these stages, the feedback-revision cycle is of paramount importance, but is the weakest link between students and AES systems. As a result, bridging the social gap of meaning negotiation is particularly important to EFL learners. According to the participants' responses, they preferred human feedback through interactions by means of discussion, clarification, and clear guidance. It should also be noted that EFL learners learn to write in a language that they are learning at the same time; it is therefore double the linguistic effort to complete a writing task. Although the immediate computer-generated feedback provides superficial grammatical and textual clues that can be applied to different writing cases, the specific details for revisions should be addressed through more social engagement of meaning negotiation for EFL learners.

In addition, a washback effect on students' writing was observed. It was possible that students attempted to "please" AES systems and thus paid more attention to form than content in order to achieve a higher grade. What is achieved, in addition to a good grade, is a textual product in compliance with the textbook standard of a good essay at the cost of students' creative expression, because AES systems are not capable of identifying it.

Limitations of the study need to be addressed. First, contrary to previous studies whose purpose was to investigate the effectiveness of AES systems in large-scale high-stakes testing, the number of the participants in this study is small by the nature of case studies. Furthermore, due to

the limited availability of human raters, only two human raters were involved in the study. The number of human raters was low, thereby limiting the generalizability of this study's results.

References

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

Attali, Y., & Burstein, J. (2006). AES with e-rater V.2. *The Journal of Technology, Learning, and Assessment, 4*(3), 1–31.

Bejar, I. I. (2011). A validity-based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 319–341.

Bridgeman, B., Trapani, C., & Yigal, A. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education 25*(1), 27–40.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays*. Princeton, NJ: Educational Testing Service.

Burstein, J., & Chodorow, M. (1999). *AES for nonnative English speakers*. Paper presented at the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD.

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion 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* (pp. 3–10). Menlo Park, CA: AAAI Press..

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

Calfee, R. (2000). To grade or not to grade. *IEEE Intelligent Systems*, *15*(5), 35–37.

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 & Technology*, *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.

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

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays.* Princeton, NJ: Educational Testing Service.

Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor, MI: University of Michigan Press.

Donahue, C. (2005). Student writing as negotiation. In T. Kostouli (Ed.), *Writing in context(s): Textual practices and learning Processes in sociocultural settings* (pp. 137–163). New York: Springer.

Elliot, S., & Mikulua, C. (2004). *The impact of MyAccess use on student writing performance: A technology overview and four studies*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Ericsson, P., & Haswell, R. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences.* Logan, UT: Utah State University Press.

Fang, Y. (2010). Perceptions of the computer-assisted writing program among EFL College Learners. *Educational Technology & Society*, *13*(3), 246–256.

Foltz, P. W., Landauer, T. K., & Laham, D. (1999). AES: applications to educational technology. In B. Collis & R. Oliver (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 1999* (pp. 939–944). Chesapeake, VA: AACE.

Grimes, D., & Warschauer, M. (2006). *AES in the classroom*. Paper presented at the American Educational Research Association, San Francisco, CA.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*, 145–159.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). *Evaluating multiple aspects of coherence in students essays*. Retrieved from: http://www.ets.org/media/research/pdf/erater_higgins_dis_coh.pdf

Hillocks, G. (2002). *The testing trap: How state writing assessments control learning.* New York: Teachers College Press.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3–31.

Lai, Y. H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology, 41*(0), 432–454.

Landauer, T. K., Laham, D. L., & Foltz, P. W. (2003). Automatic essay assessment. *Assessment in Education, 10*(3), 295–308.

Larkey, L. (1998). *Automatic essay grading using text categorization techniques*. Paper presented at 21st International Conference of the Association for Computing Machinery-Special Interest Group on Information Retrieval (ACM-SIGIR), Melbourne, Australia. Retrieved from http://ciir.cs.umass.edu/pubfiles/ir-121.pdf.

Mason, O., & Grove-Stephenson, I. (2002). Automated free text marking with paperless school. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference* (pp. 216–222). Loughborough: Loughborough University.

Ming, P. Y., Mikhailov, A. A., & Kuan, T. L. (2000). Intelligent essay marking system. In C. Cheers (Ed.), *Learners together*. Singapore: Ngee Ann Polytechnic. Retrieved from http://ipdweb.np.edu.sg/lt/feb00/intelligent_essay_marking.pdf

Mitchell, T., Russel, T., Broomhead, P., & Aldridge N. (2002). Towards robust computerized marking of free-text responses. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference* (pp. 233–249). Loughboroug: Loughborouh University.

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

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education, 14*, 210–225.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring.* Princeton, NJ: Educational Testing Service.

Shermis, M. D., & Burstein, J. (2003). *AES: A cross disciplinary perspective*. NJ: Lawrence Erlbaum Associate.

Shermis, M. D., & Burstein, J. (2013). *The handbook of AES: Current applications and new directions.* NJ: Lawrence Erlbaum Associate.

Shermis, M. D., Burstein, J., & Bliss, L. (2004). *The impact of AES on higher stakes writing assessment*. Paper presented at the Annual Meetings of American Education Research Association and the National Council on Measurement in Education Conference, San Diego, CA.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*(1), 5–18.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*, 27–33.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*, 303–325.

Tribble, C. (1996). *Writing.* Oxford: Oxford University Press.

Wang, J., & Brown, M. S. (2008). AES versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education*, *8*(4), 310–325.

Wang, M. J., & Goodman, D. (2012). Automated writing evaluation: Students' perceptions and emotional involvement. *English Teaching and Learning, 36*(3), 1–37.

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, *26*(3), 234–257.

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

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

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

Appendices

Appendix 1

The *Criterion* scoring guide

Score of 6:	You have put together a convincing argument. Here are some of the strengths evident in your writing. Your essay: • Looks at the topic from a number of angles and responds to all aspects of what you were asked to do. • Responds thoughtfully and insightfully to the issues in the topic. • Develops with a superior structure and apt reasons or examples (each one adding significantly to the reader's understanding of your view). • Uses sentence styles and language that have impact and energy and keep the reader with you. • Demonstrates that you know the mechanics of correct sentence structure, and American English usage virtually free of errors.
Score of 5:	You have solid writing skills and something interesting to say. Your essay: • Responds more effectively to some parts of the topic or task than to other parts • Shows some depth and complexity in your thinking. • Organizes and develops your ideas with reasons and examples that are appropriate. • Uses the range of language and syntax available to you. • Uses grammar, mechanics, or sentence structure with hardly any error.
Score of 4:	Your writing is good, but you need to know how to be more persuasive and more skillful at communicating your ideas. Your essay: • Slights some parts of the task. • Treats the topic simplistically or repetitively. • Is organized adequately, but you need more fully to support your position with discussion, reasons, or examples. • Shows that you can say what you mean, but you could use language more precisely or vigorously. • Demonstrates control in terms of grammar, usage, or sentence structure, but you may have some errors.
Score of 3:	Your writing is a mix of strengths and weaknesses. Working to improve your writing will definitely earn you more satisfactory results because your writing shows promise. In one or more of the following areas, your essay needs improvement. Your essay: • Neglects or misinterprets important parts of the topic or task. • Lacks focus or is simplistic or confused in interpretation. • Is not organized or developed carefully from point to point. • Provides examples without explanation, or generalizations without completely supporting them. • Uses mostly simple sentences or language that does not serve your meaning. • Demonstrates errors in grammar, usage, or sentence structure.
Score of 2:	You have work to do to improve your writing skills. You probably have not addressed the topic or communicated your ideas effectively. Your writing may be difficult to understand. In one or more of the following areas, your essay: • Misunderstands the topic or neglects important parts of the task. • Does not coherently focus or communicate your ideas, • Is organized very weakly or doesn't develop ideas enough. • Generalizes and does not provide examples or support to make your points clear. • Uses sentences and vocabulary without control, which sometimes confuses rather than clarifies your meaning.
Score of 1:	You have much work to do in order to improve your writing skills. You are not writing with complete understanding of the task, or you do not have much of a sense of what you need to do to write better. You need advice from a writing instructor and lots of practice. In one or more of the following areas, your essay: • Misunderstands the topic or doesn't show that you comprehend the task fully. • Lacks focus, logic, or coherence. • Is undeveloped; there is no elaboration of your position. • Lacks support that is relevant. • Shows poor choices in language, mechanics, usage, or sentence structure which make your writing confusing.

Source: http://www.ets.org/Media/Products/Criterion/topics/co-1s.htm

Appendix 2

The survey

Part 1.
Criterion has been implemented in the writing class for this semester. We want to know what you think of the use of *Criterion* in the writing class. Please indicate to what extent you agree with the following statements.
5= Strongly Agree, 4= Agree, 3= Neutral, 2= Disagree, 1= Strongly Disagree

Statement	5	4	3	2	1
1. I used the Outline Organizer provided by *Criterion* to help me organize essays.					
2. I used the Writer's Handbook provided by *Criterion* to help me improve English.					
3. The description of essay prompts was clear enough for me to know what the topic asked of me.					
4. I tended to write essays in the way *Criterion* expects me to write in order to get a higher score.					
5. *Criterion's* feedback was useful to improve the grammar of essays.					
6. *Criterion's* feedback was useful to improve the usage of essays.					
7. *Criterion's* feedback was useful to improve the mechanics of essays.					
8. *Criterion's* feedback was useful to improve the style of essays.					
9. *Criterion's* feedback was useful to improve the organization and development of essays.					
10. *Criterion's* feedback could be clearly understood for revision.					
11. I considered the submission of essays and *Criterion's* immediate feedback as an interaction between *Criterion* and me.					
12. *Criterion* can satisfactorily indicate grammatical errors of essays.					
13. *Criterion* can satisfactorily indicate usage errors of essays.					
14. *Criterion* can satisfactorily indicate mechanics errors of essays.					
15. *Criterion* can satisfactorily indicate style errors of essays.					
16. *Criterion* can satisfactorily indicate organization and development errors of essays.					
17. *Criterion* scoring truthfully reflected the writing quality of my essays.					
18. *Criterion* scored my essays as I had expected.					
19. I used *Criterion* as a good learning tool of English writing.					
20. I recommended that *Criterion* be implemented in future writing classes.					

Part 2.
Please write down your reflections or thoughts about the use of *Criterion* in the writing class.