

Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction

Jinrong Li ^{a,*}, Stephanie Link ^b, Volker Hegelheimer ^b

^a *Department of Writing and Linguistics, Newton Building 2221A, Georgia Southern University, P.O. Box 8026, Statesboro, GA 30460, United States*

^b *Department of English, Mailbox 206 Ross Hall, Iowa State University, Ames, IA 50011, United States*

Abstract

The development of language processing technologies and statistical methods has enabled modern automated writing evaluation (AWE) systems to provide feedback on language and content in addition to an automated score. However, concerns have been raised with regard to the instructional and assessment value of AWE in writing classrooms. The findings from a few classroom-based studies concerning the impact of AWE on writing instruction and performance are largely inconclusive. Meanwhile, since research provides favorable evidence for the reliability of AWE corrective feedback, and that writing accuracy is both an important and frustrating issue, it is worthwhile to examine more specifically the impact of AWE corrective feedback on writing accuracy. Therefore, the study used mixed-methods to investigate how Criterion[®] affected writing instruction and performance. Results suggested that Criterion[®] has led to increased revisions, and that the corrective feedback from Criterion[®] helped improve accuracy from a rough to a final draft. The potential benefits were also confirmed by the instructors' interviews. The students' perspectives were mixed, but the extent to which the views vary may depend on the students' language proficiency level and their instructors' use and perspectives of AWE.

© 2014 Elsevier Inc. All rights reserved.

Keywords: AWE; ESL writing; Corrective feedback; Mixed-methods research

Introduction

Since the first automated essay scoring (AES) system in 1966 (Page, 2003, p. 43), increasingly sophisticated language processing technologies and statistical methods have enabled newly developed scoring engines, such as e-rater[®],¹ Knowledge Analysis TechnologiesTM,² and IntelliMetric[®]³ to analyze a wide range of text features at lexical, syntactic, semantic, and discourse levels. Powered by these scoring engines, some automated writing evaluation (AWE) tools, such as Criterion[®] and MY Access![®], have been developed to provide formative feedback on

* Corresponding author. Tel.: +1 515 441 3936.

E-mail address: jli@georgiasouthern.edu (J. Li).

¹ <http://www.ets.org/erater/about/>

² <http://www.pearsonassessments.com/pai/ai/Products/NextGeneration/ScoringandReporting.htm>

³ <http://www.vantagelearning.com/products/intellimetric/>

various rhetorical (e.g., organization) and language-related dimensions (e.g., grammar and mechanics) and supplementary resources in addition to automated scores. Although most AES systems have been supported by favorable validity evidence based on the consistency and agreement between the automated system and human raters (Enright & Quinlan, 2010; Keith, 2003; Vantage Learning, 2006; Weigle, 2010), the potential instructional and assessment value of AWE still remains to be examined. Questions have been raised that go beyond the scope of previous psychometric research. For example, researchers believe that high correlation and agreement between AES scores and human ratings is an insufficient condition for the validity of score use (Chung & Baker, 2003, p. 29) and the purpose of such use needs to be taken into account (Chung & Baker, 2003; Weir, 2005). Moreover, critics have expressed concerns over the relevance of the features evaluated by AWE to the true qualities of writing, particularly the social and communicative dimensions (Ericsson, 2006). Recent controversies are increasingly concerned with the use of AWE corrective feedback. The developers and some researchers believe that the automated feedback could help students revise and improve writing while others caution that by diverting students' attention from content development to formal attributes, AWE may lead to negative washback effects on students' writing behavior (CCCC, 2006; Cheville, 2004).

These concerns, together with the lack of adequate evidence of the use of AWE corrective feedback in ESL writing classrooms and its impact on students' writing, prompted the current study. Specifically, the present study explores the role of AWE corrective feedback in an ESL writing curriculum in which one such online writing evaluation service, Criterion[®], has been integrated into the instructional framework of university ESL writing classrooms. We examine the use and the impact of the automated feedback through a naturalistic classroom-based approach. Responding to a recent call for AWE research initiated by Ware (2011), we center our discussion around the need for clearer understanding of “the context in which the system is used, the content of what is written, and the impact on key stakeholders as part of its integration” (p. 773). Next, we review key classroom-based studies that have contributed to AWE research to date. We then outline our theoretical framework, research questions, methodology, and important findings before we end with a discussion of future directions for the field of L2 writing in general and computer-generated feedback in particular.

Previous research on AWE

The narrow scope of psychometric research and the lack of studies that examine the use of AWE as it is systematically integrated into writing instruction have opened up a new research agenda, which was taken up by a few studies exploring the effectiveness of AWE on improving students' standardized writing test scores (Attali, 2004; Vantage Learning, 2007), instructors' and students' use of AWE (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Wang, Shang, & Briody, 2012; Warschauer & Grimes, 2008), and their attitudes toward its use in writing courses (Chen & Cheng, 2008; Lai, 2009).

Different measures have also been used to assess more directly the effectiveness of AWE in helping students revise. Although they seemed to have provided some insights into the effects of AWE, the results were far from conclusive. For example, Attali (2004) used the number of submissions as an indicator of students' use of an AWE system and found that 71% of the essays were submitted only once (p. 4). The number of submissions provides a rough estimate of the frequency of students' use of AWE, but any inferences about how students used AWE or benefited from using it based only on the number of submissions would be inaccurate.

To address the issues concerning the process of using AWE, an increasing number of studies have relied on classroom observations and interviews with teachers and students. Chen and Cheng (2008), for instance, examined the effectiveness of MY Access![®] in assisting writing improvement in three EFL writing classes mostly through interviews with the teachers and the students. By looking into the process, the authors were able to demonstrate a sharp contrast between the three classes in terms of the teacher's use and requirements of using the system and the students' response, and possible connections between the teacher's pedagogical decisions and the students' perception. Meanwhile, the process-oriented approach also allowed the researchers to reveal the complexity in evaluating the effects of AWE. While the study revealed that some students saw the automated feedback as vague and formulaic, others held that it was helpful for identifying and correcting grammatical and mechanical errors, particularly for students with lower English proficiency. Chen and Cheng (2008) cautioned that a quick conclusion about the usefulness of AWE might not be accurate but did not include analysis of writing samples to further the investigation of the issue. The process-oriented approach also unveiled another complexity in evaluating the effects of AWE: Most

students were concerned about the limitations of automated scoring, and reported that it favors lengthiness, overemphasizes the use of transition words, ignores coherence and content development, and discourages unconventional ways of essay writing. Although AWE can provide scores and feedback, and both of them may be helpful for revision, it would probably be more helpful to assess them separately because they may be used for different purposes and in different ways.

Some studies have responded by including analysis of writing samples and/or narrowing down the research focus. In Warschauer and Grimes (2008), for example, both writing samples and interviews were used to examine the effects of Criterion[®] and My Access![®] in four secondary schools in the United States. The analysis of writing samples showed that AWE feedback was more likely to lead to word- or sentence-level revision rather than substantial change of content. Although this may be regarded as a drawback, the idea that AWE may help with grammar and/or mechanical issues is relevant to ESL writing instruction. However, quantitative measures, in addition to detailed descriptions, would help assess the effects of AWE feedback on revision more accurately. Therefore, to contribute to the on-going research on the effects of AWE feedback on revision, the current study focused on the use of AWE corrective feedback among ESL students, documented the context and process of the use of AWE in detail, and incorporated multiple types of data in examining how AWE corrective feedback may affect revision.

Theoretical framework

The lack of a comprehensive view of AWE from an empirical perspective leads to an even more integral role of theory in framing research on the role of AWE corrective feedback in ESL writing instruction. Specifically, SLA and L2 response research provide useful lenses to examine the potential benefits of AWE corrective feedback for ESL students (Bitchener & Ferris, 2012). In SLA, corrective feedback is also referred to as negative feedback or negative evidence. Defined as input provided to show the learners unacceptable instances of language use (Long, 1996), corrective feedback is believed by many to be beneficial because it may facilitate the process of noticing, an essential element in L2 acquisition (Gass, 1988, 1990, 1991; Schmidt, 1990, 1994;). Earlier research focused more on oral corrective feedback (Long, Inagaki, & Ortega, 1998; Lyster, 2001) while an increasing number of recent studies have demonstrated the potential benefits of written corrective feedback on L2 acquisition in text-based computer-mediated communication (Lee, 2008; Yilmaz & Yuksel, 2011).

In contrast, L2 response research focuses on written corrective feedback (WCF) and is still far from reaching consensus about its effects. Some have claimed that WCF in fact has no place in L2 writing instruction and should be abandoned (e.g., Truscott, 1996, 2010; Truscott & Hsu, 2008), whereas others have acknowledged its role in L2 writing instruction and moved on to examine ways to provide effective WCF (Ferris, 2010). For example, in a study by Ferris and Roberts (2001), findings showed that explicit feedback both with and without corrections coded by error type were more effective in helping students to self-edit than not providing feedback at all. Furthermore, providing WCF to L2 writers can be frustrating because the teachers may have to repeat similar feedback (Ferris, Brown, Liu, & Stine, 2011). It can also be challenging because coding and labeling student errors by type may not be consistent, raising the “likelihood of teacher error” (Ferris, 2010, p. 199).

Since both language and writing are important in ESL writing classrooms, and improvement in linguistic accuracy may result in better communication of ideas for L2 writers, we believe the more important question is how to provide effective corrective feedback. Previous research suggests that AWE corrective feedback may be particularly helpful for ESL writers, and a comparison of the characteristics of AWE corrective feedback and the essential traits of effective WCF proposed by Hartshorn et al. (2010) further indicates the possible use of AWE corrective feedback in ESL writing classrooms and its potential benefits for ESL writers.

Specifically, Hartshorn et al. (2010) proposed dynamic WCF as an instructional methodology and described it as *individualized, meaningful, timely and constant, and manageable*. They tested the effects of the dynamic WCF on two groups, where the treatment group received dynamic WCF on 10-minute writing assignments nearly every day of class while the control group only received feedback on multiple drafts of four papers during the experimental period. The effects were compared using differences between measures of students’ accuracy, rhetorical competence, fluency, and complexity from a pre-test to a post-test. Specifically, the researchers measured accuracy using the error-free T-unit ratio (EFT/T), and found that the treatment group that received dynamic WCF had significantly higher accuracy scores on the post-test. Since one of the main differences was that the treatment group received more frequent writing

assignments and feedback, and was also required to go through more revisions, it stands to reason that dynamic WCF is most effective probably because it is timely and constant, and manageable.

These features also characterize AWE corrective feedback. To illustrate, being timely and constant requires that students have opportunities to write in every class and can always get feedback (Hartshorn et al., 2010). AWE probably outperforms dynamic WCF in this respect because it provides feedback immediately after students' submission. It also guarantees constant feedback so long as students continue to revise and resubmit. Although this does not mean AWE corrective feedback is as accurate and/or individualized as that from an instructor, the restrictive nature of issues related to grammar and mechanics allows a more active role of AWE corrective feedback.

Moreover, corrective feedback has to be manageable for both instructors and students (Hartshorn et al., 2010). Instructors need to have enough time to provide quality feedback, and students need to have the time and ability to process, learn from, and apply the feedback to their writing. For these purposes, Hartshorn et al. (2010) limited the length of the students' in-class writing to ten minutes. AWE, on the other hand, would lessen the restraint of essay length on the instructor, and allow the students the possibility of working on one particular type of error in a single submission. Both would help make the task more manageable for the instructor and the students.

In addition, AWE corrective feedback is comparable to dynamic WCF in terms of being meaningful. Meaningfulness emphasizes the idea that students are given opportunities to learn why errors occur and how to make corrections (Hartshorn et al., 2010). Criterion[®] can offer similar benefits, because although it distinguishes 35 types of errors under four main categories and provides a descriptive label or a brief explanation for each, it does not offer precise corrections for all errors. The technical limitation mirrors the opportunities for cognitive development provided by coded symbols used in dynamic WCF. However, to help students use AWE corrective feedback in meaningful ways, the instructor's role is essential. Although AWE corrective feedback can be timely and constant, it is the instructor that gives the students guidance with regard to how to use the feedback. This echoes the importance of documenting the details of the context and process of using AWE while studying its impact on writing.

Research questions

While the quality of AWE corrective feedback and its larger impact on writing quality or writing competency need further attention, the current study aims to explore the use of an AWE system by ESL writing instructors and students and the influence of AWE corrective feedback on revision and accuracy. Specifically, the study was carried out to address the following four research questions.

1. How do the instructors use and view Criterion[®] corrective feedback in light of their writing pedagogy?
2. How do the students view the AWE corrective feedback?
3. How does the use of the AWE corrective feedback affect the students' writing practice?
4. How does the use of the AWE corrective feedback affect the change of writing accuracy from the first draft to the final draft?

Method

To examine both the processes and the results of integrating AWE corrective feedback into ESL writing instruction, we adopted a Convergence Triangulation Model (Creswell & Plano Clark, 2007, p. 63), a mixed-methods design where complementary qualitative and quantitative data were collected on the same topic. Qualitative data, including individual interviews with the instructors and students, were collected to examine the instructors' and students' use of Criterion[®], and the perceived benefits and problems of using the corrective feedback from the system. Concurrent with this data collection, quantitative data, including the number of submissions for each paper recorded by the system and AWE error reports, were collected to analyze the influence of AWE corrective feedback on the students' writing practice and the change of accuracy on their paper drafts.

Context and participants

The study was conducted in six face-to-face sections of two academic ESL writing courses (English 101B and English 101C1–5) over a 15-week semester in the spring of 2011 in a major mid-western university in the United

States. The students were placed into the courses based on the results of a 30-minute writing test as part of an institutional English Placement Test. Each class met three hours per week, and one of the contact hours was in a computer lab.

There was one section of the lower-level course (English 101B) that focused on advanced English grammar in academic writing at paragraph level. In this section, three major papers are required (a personal essay, a process essay, a summary and response essay). Successful completion of the course enables students to proceed to the higher-level course (English 101C). The remaining five sections were of the higher-level, which aimed to help students with both content development and language-related issues in writing academic essays. Successful completion of the higher-level course enables students to proceed to first-year composition designed for native speakers.

The five sections of the higher-level course shared the same textbook and a collaboratively created backbone syllabus that included four major papers (a personal essay, a cause-and-effect essay, a compare-and-contrast essay, and an argumentative essay). The topics for the first two essays were chosen from the pre-determined Criterion[®] prompts developed by ETS⁴ that were considered to have a similar level of difficulty, and the topic for the last two papers were different between the sections as the instructors created their own topics⁵ that were not found in Criterion[®] based on a joint consensus of what would fit curricular expectations (see [Appendix A](#) for a description of the writing prompts).

All instructors followed a process writing approach and provided feedback to the students. However, there was variation between the instructors in the amount, content, form, and frequency of their feedback. Moreover, except for a two-hour teleconference training workshop facilitated by an ETS certified trainer in December 2010, no further requirements were given to the instructors with regard to the use of Criterion[®] in their classrooms. Thus, the instructors' requirements of using the AWE system and their students' actual use of it were slightly different. Despite the differences, all the instructors required the students to use the system in revision and submission of the final drafts; at the same time, the instructors all encouraged their students to use Criterion[®] outside of class. In addition, for the first two papers where the topics were the same for English 101C1–5, all 101C instructors met and decided to require their students to achieve a minimum automated score before allowing them to submit their papers for the instructors to grade.

The participants of the study included four ESL writing instructors⁶ and 70 non-native English-speaking students. We first recruited the instructors by contacting them individually to explain the purpose and nature of the project. Four out of six instructors, Abbi, Adele, Michael, and Jason,⁷ agreed to participate and signed the informed consent form. They were all doctoral students in applied linguistics and each had over four years of ESL teaching experience. Meanwhile, although they had not used AWE in teaching prior to the study, they all indicated that they were proficient in using technology in language teaching and were generally not concerned about possible technical difficulties.

Out of the 141 students that received the informed consent form, 70 students, or 50%, agreed to participate. [Table 1](#) shows the L1 background of all the participants, which was largely representative of the L1 background of the students enrolled in the writing courses. Considering the increasing number of Chinese students on U.S. campuses, this sample may also represent more broadly international students who are placed into ESL writing courses that directly precede first-year composition in major mid-western universities in the United States; however, for universities with more diverse ESL populations, results should be interpreted with caution. Among the 70 student participants, 27 volunteered for interviews, and a team of three researchers interviewed the volunteers twice throughout the semester.

The AWE system

Criterion[®], a web-based service developed by ETS, provides an “initial evaluation of student skill levels, allowing instructors to benchmark writing...adjust instruction and track progress” (ETS, 2011). The system provides corrective feedback that covers 35 subcategories of language issues under the four main categories of grammar, usage, mechanics, and style⁸ (see [Appendix B](#) for a list of the categories). It highlights errors and provides brief suggestions

⁴ Educational Testing Service.

⁵ This was largely because the instructors planned to assign different readings. Although Criterion[®] can provide corrective feedback for writing samples on the instructor-created topics, it is not able to grade them.

⁶ Some instructors taught two sections while others taught one. See [Table 2](#) for details.

⁷ Pseudonyms were used.

⁸ There may be differences between these categories and those included in the newer versions of the system.

Table 1

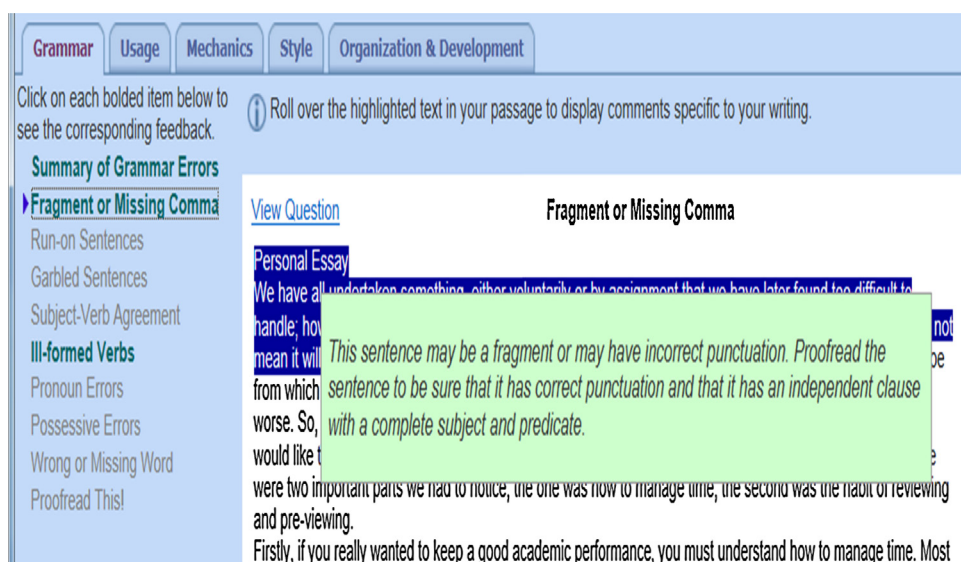
L1 background of student participants.

L1	Mandarin	Korean	Malay	Thai	Arabic	Spanish	Indonesian
Number	58	3	3	1	2	2	1
Percentage	83%	4%	4%	1%	3%	3%	1%

for correction. Fig. 1 illustrates how errors are highlighted and explanations and/or suggestions for corrections are provided in the system using the example of a fragment or missing comma error. Detailed explanations for grammar and language use can also be found in the *Writer's Handbook* accessible through students' online Criterion[®] accounts. The version of the system used for this study only saved the first and the final draft; thus our analyses focused on the changes between these two drafts. While students varied in terms of the number of submissions for each paper, the final drafts were treated equally as the revised version because these were what the students submitted for the instructors to grade.

Data collected

We conducted one-on-one semi-structured interviews with the four instructors and the 27 student volunteers to understand the instructors' and the students' use and views of the AWE corrective feedback (see Appendices C and D for the interview protocols). Concurrently, we collected quantitative data from the AWE system that included the number of submissions for each paper and the number of errors identified and classified by Criterion[®] in the students' first and final drafts. The number of submissions for each paper was used to measure the amount of writing practice. Since the students were informed that the AWE feedback could help them improve different aspects of writing and get a better score from the instructors, and that the instructors did not have a specific requirement for the frequency of using the program, the number of submissions was viewed as an indication of the students' efforts in revision (Attali, 2004). The Criterion[®] error reports on the first and the final drafts were used to compare changes of accuracy in writing. Although writing accuracy can be assessed using several measures (Wolfe-Quintero, Inagaki, & Kim, 1998), we chose to use the normed error counts recorded by the AWE system mostly because it allowed us to track the changes of the specific errors that have been identified and highlighted by the system, and thus to examine its influence more directly. In addition, considering the possible influence of complexity on the measure of accuracy between

Fig. 1. Screenshot of corrective feedback from Criterion[®].

different drafts of the same paper may be small, we chose to use the total number of words instead of the number of clauses (Bardovi-Harlig & Bofman, 1989).

Elicitation procedures

Two 30- to 60-minute semi-structured interviews were conducted with each of the four instructors and two 30-minute semi-structured interviews were conducted with each of the 27 students. The instructors and the students were first interviewed during the week after the first two paper assignments were completed and were interviewed again in the week after the last two paper assignments were completed. Conducting two separate interviews provided a more in-depth description of instructors' experiences with AWE as they became more experienced with using the software. A similar set of main questions was used in the two interviews but different follow-up questions were added. The interviews were held in private offices and were recorded by a digital voice recorder and transcribed verbatim.

The number of submissions and the error counts were automatically recorded by Criterion[®], and the reports were retrieved in accordance with institutional review board policies. The number of submissions and error counts on the two drafts were recorded on an Excel spreadsheet. The essays were formatted and saved as MS Word documents, and the word counts generated by MS Word were entered in the Excel spreadsheet.

Data analysis

The first research question about the instructors' uses and views of the corrective feedback from Criterion[®] was addressed by coding the instructors' interviews. The authors first developed a tentative coding scheme while reading through a sample of interview transcripts. Six researchers coded the rest of the data for emerging themes independently using Nvivo 9.0 or traditional methods. The authors had several meetings during the iterative process of refining the coding scheme and identifying emerging themes and sub-categories in order to provide a layer of peer debriefing and to improve the trustworthiness of the qualitative analysis (Denzin & Lincoln, 1994). One of the authors summarized the emerging themes and their subcategories with representative quotes, which were peer reviewed as a final step in the qualitative process to validate the findings. The second research question about the students' views of Criterion[®] corrective feedback was addressed by coding the students' interviews. We followed a similar procedure as we did for the instructors' interviews.

Research questions three and four were addressed by analyses of the quantitative data using SPSS 20.0. To answer the third question of the extent to which Criterion[®] corrective feedback may affect students' writing practice, we looked at the total number of submissions under each major paper for each individual student and the percentage of students who submitted three or more times, and the change of the numbers across the four papers throughout the semester. To answer the fourth research question about the change of accuracy, we calculated the average number of errors identified by Criterion[®] on each draft for each section, normed the raw numbers to a standard scale of errors per hundred words, and then examined the following aspects: First, we obtained descriptive statistics about the class average change of error rates in total from the first draft to the final draft for all four papers. After confirming that the data met the assumptions for comparing mean differences, we conducted within-subjects *t*-tests (Agresti & Finlay, 2009) to examine the direction and magnitude of the change of the error rates in total. Since the goal was to look at the improvement from a first draft to a final draft of the same paper, differences in writing topics that may cause variations of language use were not considered as a confounding factor in the within-subjects *t*-tests.

Results and discussion

RQ 1. How do the instructors use and view Criterion[®] corrective feedback in light of their writing pedagogy?

To answer the first research question, we examined the instructors' interviews. Seven major themes emerged concerning the instructors' use and views of the system: expectation, use, activity, suggestion, teacher feedback, teacher evaluation, and teacher observation. The seven major themes highlight the perspectives of the instructors in terms of their expectations of the AWE system, their experiences using it with various types of activities, suggestions on using it, the teachers' change of approach to providing feedback, their evaluation of the system, and their observation of their students' response to the use of it. The following discussion focuses on the use of Criterion[®] in providing feedback and the instructors' evaluation of the feedback. This decision came from an

iterative and reflexive process that began as we started the interviews. When we first talked to the instructors, we found that most of their comments evolved around the use of the AWE corrective feedback even though they were given open-ended questions to talk about any aspect of the use of AWE. Both the use and the evaluation of the AWE corrective feedback were frequently commented on by all the instructors. We then made an effort to learn more details about how they used and felt about using the corrective feedback (for a more detailed description of the instructors' strategies of using Criterion[®] and sample activities, see [Link, Dursun, Karakaya, & Hegelheimer, 2014](#)).

All four instructors commented on how, after using AWE, they began to focus more on aspects of student writing other than grammar and mechanics, as illustrated by Jason's comments below.

Yes. Finally I don't have to concentrate on their grammar mistakes because this is their responsibility. Can't get feedback from every mistake they have so now I have put my focus on organization, and I skip grammar if I see it as their responsibility because they have a guide there. (Jason, Interview 2)

Like Jason, the other three instructors who commented on the change of their response strategy also expressed positive attitudes towards how Criterion[®] allowed them to focus on content development and organization of student writing.

At the same time, the teachers' evaluations of Criterion[®] also center on its corrective feedback. Both the positive and negative impact of the corrective feedback was frequently discussed, with all four instructors commenting on some aspects of its positive values, and three instructors commenting on some aspects of its weaknesses. These comments address different aspects of the AWE corrective feedback and are therefore not contradictory. Specifically, positive comments focus on how the AWE corrective feedback would allow the students to be more aware of grammatical errors in their writing, to learn some meta-linguistic terms, to separate feedback on content from that on form to make revision less overwhelming, and to make the instructors feel less obligated to provide a large amount of feedback on grammar. Adele's quote below is an example of how all the instructors felt about using Criterion[®] in providing feedback to their students.

I feel less guilty when I made comments on the students' paper because I know they will get a certain amount of feedback on their grammar from the program. (Adele, Interview 1)

Some instructors also believe that simple reminders and/or the identification of the problems would be enough to help the students see some of the recurring problems and make corrections. For that purpose, AWE corrective feedback would be adequate, and Jason, in his comments on the AWE system, even compared Criterion[®] to teaching assistants in providing corrective feedback.

The negative comments about the corrective feedback, on the other hand, focus on the quality of the feedback and its long-term effects on error reduction. Among others, the instructors acknowledged that Criterion[®] cannot identify all grammatical errors and sometimes its feedback was inaccurate. They also expressed the concern that such corrective feedback may not help reduce errors in the long run. Similar concerns about the long-term effects of WCF on accuracy have been expressed by some researchers ([Truscott & Hsu, 2008](#)). However, the concern about the long-term effects of the AWE corrective feedback partly came from dissatisfaction of the quality of the feedback. A few instructors mentioned that without teacher intervention, some of the corrective feedback could be confusing as it is not always accurate or consistent.

In comparison with the instructors' evaluation of the corrective feedback from Criterion[®], their evaluation of its feedback on content and organization is mostly negative. All the instructors commented on the ineffectiveness of content feedback.

...but the content feedback or structure feedback is not helpful at all. And many students will say this: it is just a tougher style of writing, and it's not creative or they had to follow the patterns. It's not that good. (Michael, Interview 1)

Michael believed that the AWE feedback on organization was more like a template and may restrain students' creative thoughts. Other instructors also believed that the same content feedback for all students and all types of writing assignments did not provide meaningful help for content development or reorganization, and instead promoted the use of a fixed pattern of writing. The results suggest that the larger impact of AWE on overall writing quality and writing competence would be an important topic for future research.

Another recurring theme centered on the instructors' evaluation of the problems caused by the use of the automated score. Based on [Chen and Cheng \(2008\)](#), automated scoring seemed to be an important feature that can promote learning. In their comparisons of three instructors' use of an automated writing evaluation program in EFL writing classes, one instructor required the students to achieve a minimum automated score of four out of six before they could turn their papers in for the instructor to grade. This requirement seemed to have motivated the students to use the program more frequently and the students in that class also had more positive attitudes toward the use of the program at the end of the study (p. 103). However, the teacher participants in our study reported that the same requirement of a minimum automated score of four out of six before formal submission caused some problems. The biggest problem was the inconsistency of the automated scoring and the teachers' grades, which caused the students' and the teachers' frustration. Jason's comments below serve as an example.

Grading was something I was not very happy about in Criterion because you could get a 6 although your content was not very well organized. But they could play with the grammar a little bit. Articles, words, for example, the number of words were some things that were effective in terms of getting high scores. But when students got their feedback from me after their first submission, most of them were kind of confused saying we got a score of 6 but this feedback. So I don't know, either at that point they hate Criterion or they hate me. I don't know which one but one of them happened probably because we were not consistent in terms of the grades. (Jason, Interview 1)

Like Jason, all the other instructors commented that using automated scoring caused some problems: Students were frustrated either because they could not achieve the minimum score, or that the scoring of the program and the teacher was not consistent. A possible reason is that the pre-determined prompts may indeed have different expectations that were not clear to the teachers at the time the topics were chosen. Therefore, although the scoring features could be used as an incentive to motivate students to use AWE corrective feedback more frequently, it would be necessary for teachers to fully understand the requirements and the grading rubric for a writing prompt before it is used as a topic for a writing assignment (see [Li, Link, Ma, Yang, & Hegelheimer, 2014](#)).

RQ 2. How do the students view the AWE corrective feedback?

To answer the second research question, we examined the emerging themes from the students' interviews. We discuss the students' experiences and responses in terms of positive and negative views of Criterion[®] as they relate to students' level of satisfaction and confidence in using the feedback, and the challenges of using it. Overall, 18 out of 27 student interviewees expressed a high level of satisfaction, and only two students were not satisfied because sometimes the program can make mistakes. All 18 students who expressed positive responses commented on the helpfulness of the corrective feedback. Specifically, students frequently commented on how AWE corrective feedback can be a helpful resource for them to resolve language-related issues in writing and can motivate them to use other tools in order to make a correction. The following quote is an example.

... I like the grammar parts. This program find out a lot of grammar mistakes, like spelling or commas, or something like that. (Dawson, Interview 1)

The students' positive reaction to the use of AWE corrective feedback seems to be closely connected to their instructors' perspectives and pedagogy. As shown above, when commenting on the benefits of the AWE system, all four instructors focused on the value of its corrective feedback. Moreover, some instructors were interested in AWE corrective feedback and planned to use it in teaching prior to the start of the study. For example, Dawson's instructor, Adele, mentioned, among others, the possibility of using AWE corrective feedback when asked about what motivated her to participate in the study (Adele, Interview 1).

However, the students' comments on the usefulness of the feedback on organization were mixed. For example, a few students felt feedback on organization helped, as illustrated by the following quote.

Interviewer: What kind of feedback was helpful the most to you?

Andrew: I think it's the last one that says part of my essay in colors.

Interviewer: The organization?

Andrew: Yeah. It tell me to develop the organization because I may miss some part when I write it.

Interviewer: So you like that kind of feedback the most?

Andrew: Yeah. (Andrew, Interview 1)

A possible reason why these students found the feedback on organization helpful is that they needed some more guidance on the basic patterns of English essays, and Criterion[®] offers reminders by highlighting the sentences that should be the thesis statement or the topic sentences and thus the students could review and possibly revise. Since such feedback functions more like a reminder, it was not mentioned in the instructors' comments on the potential benefits of the AWE system or in their description of how the AWE system was integrated into their teaching. However, for students with limited experience with English academic writing, the reminder may be helpful. It is also possible that students like Andrew may only have limited English proficiency and therefore structural feedback with visual representation may be easier to understand than corrective feedback that contains many English meta-linguistic terms.

For most students who need advice on more sophisticated organization strategies or individualized response, such feedback might not be perceived as meaningful. For example, one student made the following comments.

... feedback like about my essay itself, I cannot give opinions like that, because you know, it's a machine, how to say, well, anyway, it's not that clearly. ... (Dawson, Interview 1)

Here, Dawson was commenting on AWE feedback on the essay "itself," after talking about the feedback on grammar, and he felt that AWE was not able to provide adequate feedback for that purpose (i.e., about how to present one's opinion) because it was a machine. Other students expressed a similar opinion by contrasting Criterion[®] feedback with that from the instructor. The following quote is an example.

I like that Criterion offers feedback about minor grammar errors while the feedback from the instructor is about the organization, contents, ideas of the essay. (Jennifer, Interview 2)

Both Dawson and Jennifer were taught by Adele, and their views of the AWE system seem to echo Adele's perspective. In describing how she introduced the AWE system to the class, Adele emphasized that "this is just software, just a machine" (Adele, Interview 1). The same language was seen in a subsequent interview with Dawson when he commented that "It's a machine, how to say, well, anyway, it's not that clearly" (Dawson, Interview 1). Furthermore, Adele's use of the AWE system may also have influenced her students' views on different aspects of the system. In the second interview with Adele, she described how she used AWE corrective feedback to supplement her own feedback for student papers.

I didn't count on the feedback on organization from Criterion in the first place. So I didn't expect my students to get much information from the Criterion feedback. So when I gave them feedback, I paid much more attention to organization and maybe the logic in their paper, maybe the way they present their ideas in their paper. (Adele, Interview 2)

Clearly, Adele focused on providing feedback on content development and organization when she responded to student papers while requiring students to use AWE corrective feedback for issues related to grammar and mechanics. This strategy had strong influence on students' view of how the AWE system can help them during revision. Quite a few students expressed similar views that they can use AWE feedback to correct basic grammar errors, but need instructor's help on larger issues such as organization and rhetorical strategies.

Although not all the students were satisfied with the corrective feedback from the system, it seemed clear that their views on the usefulness of its feedback on content were less positive. When comparing the students' response to how the instructors viewed and used the system, it is not difficult to see that the instructors' focus on content when giving feedback while relegating corrective feedback to the AWE system probably had an impact on the students' perspectives. In the following section, we further examine how the teachers' and students' use and views may be linked to the change of accuracy on the students' writing samples.

Moreover, the students expressed mixed opinions about the use of a holistic score as a minimum requirement before submission to the instructor. Some felt that the score provided a target and thus motivated them to use the AWE corrective feedback, while others held the view that holistic scores were not reasonable because machines cannot score an essay. Still others felt frustrated when a minimum score requirement could not be achieved even after all the identified language-related issues had been addressed. Like many other limitations of AWE programs, this weakness seems to present an opportunity to teach students that revision involves more than correcting errors. The extent to which the requirement of a minimum score may influence the number of submissions is examined in our discussion of the findings for the third research question, but the potential opportunities for teaching students to revise remains a question for future research.

RQ 3. How does the use of the AWE corrective feedback affect the students' writing practice?

Among the 70 students, six missed one or two assignments and were removed from the analysis. Therefore, data from 64 students were included in the final analyses. To answer the third research question, we looked at the number of submissions under each assignment to make inferences about the amount of writing practice that the students were engaged in, assuming the higher the number of submissions, the more writing practice the students had.⁹ Since all the instructors required at least two drafts as evidence of process documents, we believe that the percentage of students who submitted three or more times for one paper assignment can be regarded as an indicator of the students' use of AWE in revision.

Table 2 shows the average number of submissions, standard deviations of the number of submissions, and the percentage of students who submitted three or more times by class and by paper assignments. Aggregate numbers for the higher-level classes were presented at the bottom of the table. Several patterns emerged from the data. First, the average number of submissions increased from paper 1 to 2, and then went down for papers 3 and 4 for all sections, except for Jason's sections. This result may be connected to the requirement that paper 1 and 2 used pre-determined Criterion[®] prompts and that the students needed to achieve a minimum automated holistic score before submitting to the instructors. This finding may suggest that requiring students to reach a benchmark score may be a clearer goal or motivation for students to consult the AWE corrective feedback and revise their papers accordingly.

Second, Adele's sections had relatively higher submission numbers for all papers, and Jason's sections had relatively lower average submission numbers for all four papers. The differences between sections taught by different instructors were also observed in the percentage of students whose number of submissions was equal to or more than three. Again, Adele's sections had the highest percentages (85% and 100% for paper 1 and 2 in one section, and 100% and 100% in the other section), and Jason's sections had the lowest percentages (none of the students submitted three or more times for either paper 1 or 2). Although the design of the study does not allow us to establish strong causal connections between the instructors' use of the AWE system and the students' revision behavior, we noted some connections in Adele's and Jason's classes. Jason also valued the AWE system mostly for its corrective feedback, but his use of it was less consistent from the student's perspective: Jason created various topics for the students to practice writing and revision on the AWE system in addition to asking them to use it for the revision of the major paper assignments. In contrast, Adele's students learned from the very beginning that they would use the AWE system mostly for its corrective feedback and that they were expected to use the system consistently throughout the semester. The instructor's view may have shaped some of the students' views, and more importantly, the clear and consistent goal of using AWE corrective feedback may be one of the motivating factors for Adele's students to use the AWE system during revision. The extent to which such motivation would encourage students to focus more on language-related issues instead of content development, and how this would then affect the quality of revision deserve further research.

More importantly, the aggregate result for the higher-level course showed that overall, slightly over half of the students were actively using the AWE system in revision. The lowest percentage was 44% for paper 1, and it went up to 58% for paper 2 and paper 3, and 72% for paper 4. This indicates that Criterion[®] encouraged over half of the students to have three or more submissions, or to be engaged in more writing practice, which contradicts with Attali's (2004) findings that students did not exploit Criterion[®] feedback to work on their revision based on the fact that about 71% of the essays were submitted only once, and only a very small proportion of the essays went through multiple revision drafts.

Such differences can be caused by many factors. For example, Attali (2004) examined younger students (6th through 12th grade), which may indicate that adult nonnative English speakers may find more use for AWE feedback. Therefore, further research looking at both populations is necessary to explore the influence of such variables on the amount of revision in ESL classes. Moreover, differences in teaching style, the way AWE is integrated into the curriculum, and the consistency of the use of AWE can be essential, and thus future research could look into the connections between the instructors' approaches and the students' use and perspectives (Chen & Cheng, 2008), specifically with regard to AWE corrective feedback.

⁹ A follow-up study among the students who used Criterion[®] in Fall 2011 has shown increased writing practice through multiple draft submission as one of the major benefits. For example, one student mentioned that "...although it's easy to make an appointment with the writing help center, because we need to correct our articles for several times, we can't make appointments so frequently. It is really waste of time. In this situation, the Criterion system is pretty necessary for us" (Students' Reflection, 2011).

Table 2
Number of submissions.

Section	Instructor	Paper 1		Paper 2		Paper 3		Paper 4	
		N (SD)	P	N (SD)	P	N (SD)	P	N (SD)	P
101B	Abbi	3 (1)	75%	5 (3)	75%	3 (1)	50%	N/A	N/A
101C1	Adele	6 (4)	85%	13 (5)	100%	7 (6)	92%	6 (4)	100%
101C2	Adele	7 (4)	100%	12 (8)	100%	6 (5)	67%	6 (4)	87%
101C3	Michael	4 (4)	33%	15 (34)	93%	4 (3)	53%	3 (1)	53%
101C4	Jason	2 (0)	0%	2 (0)	0%	3 (2)	80%	3 (1)	40%
101C5	Jason	2 (0)	0%	2 (0)	0%	1 (0)	0%	4 (2)	79%
Aggregate (101C)		6 (4)	44%	12 (17)	58%	7 (7)	58%	6 (3)	72%

Notes: N, number of submissions; SD, standard deviation; P, percentage of students that had submitted 3 or more times.

RQ 4. How does the use of the AWE corrective feedback affect the change of writing accuracy from the first draft to the final draft?

To answer the last research question, we looked at the change of the average number of errors. Although the AWE system recorded the number of errors in the categories of grammar, usage, mechanics, and style, we did not examine the change of errors in these categories separately because we were more interested in the overall accuracy. Additionally, there is no strong consensus with regard to the categorization of errors and problems may exist in the classification system used by Criterion[®]. Table 3 summarizes the change of error rates from the first to the final draft for the three papers submitted by the participants in the lower-level course, and the change of error rates for the four papers submitted by participants in the higher-level course. Based on the results, the reduction of the error rates was observed in all three papers in 101B, with the reduction in paper 1 being the greatest. Specifically, the reduction of errors in paper 1 was 10.27, indicating that in every 100 words, the number of errors decreased by about 10 from the first draft to the final draft.¹⁰ Abbi, the instructor of 101B, indicated that since the focus of the course was on grammar and writing, the AWE corrective feedback was particularly helpful, and she also spent considerable amount of time in class helping her students understand some of the corrective feedback. For students at this level, such extra help probably gave them more confidence and thus motivation in using the AWE corrective feedback.

Table 3 also summarizes the change of error rates for all the sections in the high-level course, except for paper 3 in Section 101C5 because of missing data. All the available difference values in Table 3 showed the reduction of the error rates, except for paper 4 in section 101C4. Several sections had quite noticeable decrease of error rates for paper 2 or 3. For example, the decrease of error rates in paper 3 for 101C4 was 17.13, and that in paper 2 for 101C5 was 12.35, indicating that for every 100 words, the number of errors on the final draft decreased by about 17 and 12, respectively. However, the standard deviations (24.95 and 11.39) were also high, suggesting large differences between individual students. The results also suggest that the possible impact of AWE on error correction became clearer from paper 1 to paper 2 and 3, and then faded from paper 3 to 4. However, there were also differences between the sections. Specifically, the decrease of error rates for all four papers in 101C1 and 101C3 was relatively consistent. In 101C1 the range was from 3.70 to 5.34, and in 101C3, the range was from 1.30 to 3.65. The difference scores and their standard deviations in 101C1 were fairly consistent across four papers, indicating most of the students had some improvement in accuracy. Since Adele's sections (101C1 and 101C2) had the highest number of submissions, the result suggests that some connections may exist between the instructor's pedagogy, the frequency of students' use of AWE corrective feedback in revision, and the improvement in accuracy in student papers. However, the design of the study did not allow us to strictly compare such connections between different classes, but this would be an important topic for future research.

Within-subjects *t*-tests by class were then conducted to further examine whether the difference values were statistically significant. Section 101B ($n = 4$) and 101C4 ($n = 5$) were excluded from the analyses because of the small number of students. Also, the test was not run for paper 3 from 101C5 because of missing data. The results showed that

¹⁰ Although error reduction cannot be conclusively attributed to the use of Criterion[®] due to the classroom-based design of the study and the possibility of students engaging in additional revision, it is worth noting that the use of the system did not show detrimental effect.

Table 3
Change of mean number of errors.

	Dif. in Paper 1	Dif. in Paper 2	Dif. in Paper 3	Dif. in Paper 4
101B	10.27 (8.55)	8.24 (8.43)	3.94 (5.33)	N/A
101C1	3.70 (7.53)	4.38 (4.14)	5.34 (6.23)	4.89 (5.11)
101C2	1.42 (3.74)	4.97 (4.74)	9.70 (14.09)	0.57 (4.22)
101C3	1.30 (2.45)	3.65 (3.41)	2.55 (2.79)	2.25 (3.01)
101C4	5.82 (7.30)	6.90 (12.10)	17.13 (24.95)	−1.18 (4.79)
101C5	2.45 (6.18)	12.35 (11.39)	N/A	1.61 (8.52)

Notes: Dif. = the average number of errors on the 1st draft – the average number of errors on the final draft. Negative differences indicate increase of error rates. The values in the parentheses are standard deviations.

Table 4
Within-subjects *t*-tests by class.

Section	Paper	df	<i>t</i>	Sig.	Cohen's <i>d</i>
101C1	Paper 2	13	3.66	.004	1.474
101C2	Paper 2	15	4.06	.001	.0821
101C3	Paper 2	15	5.49	.000	1.482
	Paper 4	15	3.06	.009	0.620
101C5	Paper 2	14	4.40	.001	1.019

the average error rates on the final draft were significantly lower than that on the first draft of paper 2 for all the sections, while statistical significance was not found in any other papers or sections except for paper 4 in 101C3. Table 4 displays a summary of the findings.

Within-subjects *t*-tests of the aggregate data showed significant improvement of accuracy in paper 1 ($t(66) = 3.673$, $p < .01$, $d = 0.368$), paper 2 ($t(66) = 6.912$, $p < .05$, $d = 0.763$), paper 3 ($t(48) = 3.285$, $p < .05$, $d = 0.469$), and paper 4 ($t(63) = 2.807$, $p < .05$, $d = 0.393$). Although the aggregate data showed that the reduction of the error rates on the first three papers was statistically significant, we believe the results by class were more accurate given the different approaches taken by the instructors. On one hand, the results seem to confirm that AWE corrective feedback may help students identify and correct errors during revision (Chodorow, Gamon, & Tetreault, 2010). However, the results were largely influenced by how the software was used. Specifically, one possible reason for the significant results in paper 2 was that the instructors all required the students to get a minimum automated score before they could submit their papers for the instructors to grade and comment. It was likely that the students had a clearer goal in revising paper 2 and thus made better use of the corrective feedback. Additionally, the error rates may also be affected by the complexity of student writing. According to Skehan (2009), complexity and accuracy are two distinctive aspects of performance that compete for language users' attention. Thus, it is possible that changes in accuracy may be constrained by changes in complexity. In addition, since the study focused on the effects of AWE corrective feedback on revision, the change of accuracy was measured by the differences between multiple drafts of the same paper instead of multiple papers. The long-term effects of AWE corrective feedback on ESL writers' ability to learn to write accurately deserve more research in the future.

Conclusion

In this study, we explored the role of one commercially available AWE system, Criterion[®], in ESL writing classrooms in an effort to examine the use of its corrective feedback by the instructors and the students, and how the use of the corrective feedback may affect students' writing practice and performance. The following five major findings emerged from our analyses of the interviews investigating the use and views of the AWE system and the quantitative data concerning the students' writing practice and the change of accuracy between different drafts and different papers: (1) despite their different approaches to the use of the AWE system, all the instructors focused primarily on the potential use of Criterion[®] corrective feedback; (2) the instructors' views on the potential of AWE

corrective feedback in helping students with grammar and mechanics were largely positive although they were not completely satisfied with the quality of the feedback; (3) in comparison to their views on the AWE feedback on content, the students' views of the AWE corrective feedback were more positive, but they seemed to be connected to the instructors' views and use of the AWE system; (4) AWE corrective feedback seemed to encourage students to engage in writing practices, but other instructional methods, such as setting a minimum score as a prerequisite for submission, may also be important in the process; (5) AWE corrective feedback seemed to be helpful for ESL students to improve linguistic accuracy, and such impact may be influenced by error types and instructional approaches. These findings also suggest that how students may use and benefit from AWE, and resolve possible issues during the process, depend to a large extent on instructors' pedagogy and perspectives.

While the findings of this study certainly shed light on the role of AWE for providing corrective feedback, the potential of AWE in the ESL classroom has yet to be fully elucidated. In this study, the use of AWE prompted students to submit multiple drafts under each paper assignment (i.e., about 38% of the students had only one or two submissions), in contrast to [Attali's \(2004\)](#) study where only a small proportion of students actually submitted more than once. The differences between the results in these two studies suggest that adult nonnative English speakers may find more use for AWE, emphasizing the notion that an AWE tool directed towards the needs of nonnative English speakers is highly desirable and is a timely area of future AWE development. Moreover, instructors' strategies for implementing AWE in the classroom may also have a great impact on how students use AWE for engaging in writing practice, thus more descriptive research would highlight the impact of AWE integration in writing classrooms.

Instructors' implementation of AWE seems to also have an influence on students' linguistic accuracy and their abilities to identify and correct their own errors over time, as indicated by data from both writing samples and the instructors' and the students' interviews. However, the limitations in the study's lack of controlled variables (e.g., instructors' provision of feedback) and control group make it difficult to attribute the findings to AWE corrective feedback alone, especially considering the possibility that repeated writing in itself may be sufficient to lower error rates. Knowing that AWE corrective feedback may not be the only cause for improved accuracy is a reason for continued research. Yet what is known is that students are able to process the AWE feedback to an extent that does show linguistic improvement. Since the effectiveness of WCF relies on students' abilities to manage the feedback ([Hartshorn et al., 2010](#)), students need full support to process the AWE feedback as well. This support may be in the form of in-depth AWE training and tutorials and/or classroom activities that invite frequent and meaningful student–instructor interaction. With added training and interaction with instructors, AWE has a better chance of producing feedback that is meaningful, a characteristic of dynamic WCF, giving students more opportunities to learn why errors occur and how to make corrections and enhance cognitive development. At the same time, the study relies only on the normed error counts recorded by the AWE program, and may not provide a comprehensive picture of changes of writing accuracy. Therefore, future research could use multiple accuracy measures and look into how AWE corrective feedback can be used to facilitate students' cognitive and metalinguistic development and its connection to autonomous learning.

In addition, while the findings show that the role of the instructor may be integral to students' effective use of AWE, it is unclear what instructor strategies and styles were the most successful in improving students' writing using AWE. In fact, the instructors' use of AWE corrective feedback seemed to vary according to their writing pedagogy and practices of responding to students' essays, and questions concerning the specific relationships between the instructor, the use of AWE, and changes shown in writing performance can be addressed by future experimental research.

In this study, we were interested in exploring instructors' use of an AWE system in ESL writing classrooms, and the possible impact of the system, and particularly its corrective feedback, on students' writing. The data from the current study was based on the first time use of the system. Thus, the instructors and the students might not have been aware of its full potential, or integrated it effectively with their writing pedagogy. Therefore, further investigation with experienced instructors will provide more insights into the pedagogical possibilities and the associated rationales. It would also be informative to examine specifically how individual students interact with the AWE system in revision to understand the benefits and challenges for students in more details through stimulated recalls.

Overall, AWE does have a role in ESL instruction. Perhaps the exact role may only be decided by individual instructors' pedagogical philosophies and in-class practices, but nevertheless, its potential in L2 writing is not withstanding. Though quality concerns continue to surface, the timely and constant feedback that AWE provides should not go without recognition. Criticisms of the lack of focus on content and organization in AWE feedback is surely warranted at this state in AWE development; this is to show that the role of human WCF continues to have its

place in L2 writing instruction. Nevertheless, AWE capabilities continue to progress and will hopefully be able to provide not only timely, constant, and manageable feedback but also meaningful feedback that will maximize students' opportunities to improve their quality of writing.

Appendix A. description of writing prompts used in English 101C

Paper I: Personal Essay (500–600 words)	
Option One: Coping with Hostility	Option Two: Difficult Undertaking
This prompt requires students to describe and explain how they have experienced and coped with a hostile situation.	This prompt requires students to describe a difficult undertaking, how they became involved in it, and why it was difficult.
Paper II: Cause and Effect Essay (600–700 words)	
Option One: Virtual Friends	Option Two: Breaking Traditions
This prompt requires students to write an essay explaining the possible causes or effects of the development of communication technologies.	This prompt requires students to write an essay explaining the possible reasons or effects of breaking a tradition.
Paper III: Compare and Contrast (800–900 words)	
Option One: Cultural Comparison	Option Two: Educational Comparison
This prompt requires students to examine and discuss the similarities and differences between the American culture and a culture that they are familiar with.	This prompt requires students to examine and discuss the similarities and differences between the American educational system and that in their own home country.
Paper IV: Persuasive Essay (700–900 words)	
Variable topics on language, cultural customs, identity, or work	

Appendix B. List of categories and subcategories of Criterion[®] feedback

Feedback categories	Feedback subcategories
Grammar	Fragments
	Run-on Sentences
	Garbled Sentences
	Subject–verb Agreement
	Ill-formed Verbs
	Pronoun Errors
	Possessive Errors
Usage	Wrong or Missing Word
	Determiner Noun Agreement
	Wrong Article
	Missing or Extra Article
	Confused Word
	Wrong Form of Word
	Faulty Comparisons
	Preposition Error
	Nonstandard Word Form
	Negation Error
	Wrong Part of Speech

Appendix B (Continued)

Feedback categories	Feedback subcategories
Mechanics	Spelling Capitalize Proper Noun Missing Initial Capital Letter in a Sentence Missing Question Mark Missing Final Punctuation Missing Apostrophe Missing Comma Hyphen Error Fused Words Compound Words Duplicates Extra Comma
Style	Repetition of Words Inappropriate Words or Phrases Sentences Beginning with Coordinating Conjunction Short Sentences Long Sentences Passive Voice

Appendix C. Instructors' interview protocol

*****Questions for first interview only*****

- What motivated you to become a part of this research project? What were your expectations in terms of potential for Criterion[®] and the ways it would impact your teaching?
- How would you describe your technology expertise in general?
- What types of technology have you used for teaching? What target language group have you taught with technology and for which skills? How long?
- Other than Criterion[®], have you ever used any writing software as a teacher or student?
- How did you introduce the use of Criterion[®] to your learners?

- How did you make use of Criterion[®] in the classroom?
- Do you think Criterion[®] had an impact on your approach to writing teaching? If so, can you make direct comparisons between your previous writing teaching methodology and your current one?
- Was Criterion[®] helpful to you as a teacher? In what ways?
- Did you encounter any problems with using Criterion[®]?
- In general, how satisfied are you with Criterion[®]?
- Do you think Criterion[®] was helpful to your students? In what ways?
- Did the students encounter any problems with using Criterion[®]?
- In general, how satisfied do you think your students are with Criterion[®]?
- You mentioned your expectations regarding the use of Criterion[®] in your teaching. Did Criterion[®] meet your expectations in your writing class? If so, in what ways? If not, why not?
- What kinds of modifications would you want to make if you were going to use Criterion[®] next semester or again?
- Are there some students that seem to utilize the program better than others? Please explain.
- If another instructor was interested in using Criterion[®] in the classroom, what are some recommendations that you would give them?
- What are some of the your final reactions to the semester using Criterion[®]?
- Any additional comments?

Appendix D. Students' interview protocol

- Were you satisfied with Criterion[®] feedback?
- What kind of feedback was helpful the most to you?
- How did you use the feedback in terms of revising?
- What kind of strategies did you use to achieve the score of 6?
- Was it easy to correct the highlighted errors in Criterion[®] based on the feedback?
- Are you confident in using Criterion[®]?
- Was it easier to find/identify errors by yourself after using Criterion[®]?
- What kind of errors do you usually make in writing?
- Can you identify your writing weakness from the feedback in Criterion[®]?

- Did you proofread/revise or correct any of your mistakes by yourself before you submitted the paper to Criterion[®] and get to see what was wrong with your paper? If 'yes', Why?/If 'no', why not?
- Do you see any differences between teacher's feedback to your paper and Criterion's feedback? If 'yes', please explain the differences between them.
- Have you encountered any difficulties in using Criterion[®]?
- While you have been using Criterion[®] in your class, what was the most impressive/interesting points while using Criterion?
- Which part do you think that you have made progress while using Criterion[®]? (e.g., grammar/organization/wording/spelling)
- What problems/difficulties did you have with Criterion[®]?
- Which part of Criterion[®] did you dislike the most?
- What activities did you like the most while using Criterion[®] in class? (e.g., free-writing/make a plan/peer review/revision)
- Do you have any comments or suggestions for future use of Criterion[®] in 101C based on your experience during this semester?

References

- Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). New Jersey: Pearson Prentice Hall.
- Attali, Y. (2004). Exploring the feedback and revision features of Criterion. *Journal of Second Language Writing*, 14, 191–205.
- Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11, 17–34.
- Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. New York: Routledge.
- CCCC. (2006). *Writing assessment: A position statement*. Retrieved from <http://www.ncte.org/cccc/resources/positions/writingassessment>
- Chen, C.-F.E., & Cheng, W.-Y.E. (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.
- Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *The English Journal*, 93(4), 47–52. <http://dx.doi.org/10.2307/4128980>
- Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, 27, 419–436.
- Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23–40). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: SAGE Publications.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (1994). *The Sage handbook of qualitative research*. Thousand Oaks, CA: Sage Publishing Company.
- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater[®] scoring. *Language Testing*, 27, 317–334.
- Ericsson, P. F. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 28–37). Logan: Utah State University Press.
- ETS. (2011). *About the e-rater[®] Scoring Engine*. Retrieved from <http://www.ets.org/erater/about/>
- Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition*, 32, 181–201.
- Ferris, D. R., Brown, J., Liu, H. S., & Stine, M. E. A. (2011). Responding to L2 students in college writing classes: Teacher perspectives. *TESOL Quarterly*, 45, 207–234. <http://dx.doi.org/10.5054/tq.2011.247706>
- Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, 10, 161–184.
- Gass, S. M. (1988). Second language vocabulary acquisition. *Annual Review of Applied Linguistics*, 9, 92–106.
- Gass, S. M. (1990). Second and foreign language learning: Same, different or none of the above? In B. VanPatten & J. Lee (Eds.), *Second language acquisition* (pp. 34–44). Clevedon, UK: Multilingual Matters.
- Gass, S. M. (1991). Grammar instruction, selective attention and learning. In R. Phillipson, E. Kellerman, L. Selinker, M. Sharwood Smith, & M. Swain (Eds.), *Foreign/second language pedagogy research* (pp. 134–141). Clevedon, UK: Multilingual Matters.
- 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.
- Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly*, 44, 84–109. <http://dx.doi.org/10.5054/tq.2010.213781>
- Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–167). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Lai, Y. H. (2009). Which do students prefer to evaluate their essays: Peers or computer program? *British Journal of Educational Technology*, 41, 432–454.
- Lee, L. (2008). Focus-on-form through collaborative scaffolding in expert-to-novice online interaction. *Language Learning & Technology*, 12(3), 53–72.
- 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.
- Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards best practices for implementing automated writing evaluation. *CALICO Journal* 31(3).
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & B. K. Bahtia (Eds.), *Handbook of second language acquisition* (pp. 413–468). New York: Academic Press.

- Long, M. H., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *Modern Language Journal*, 82, 357–371.
- Lyster, R. (2001). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, 51, 265–301.
- Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158.
- Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 165–210). London: Academic Press.
- Skehan, P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30, 510–532.
- Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46, 327–369.
- Truscott, J. (2010). Some thoughts on Anthony Bruton's critique of the correction debate. *System*, 38, 329–335.
- Truscott, J., & Hsu, A. Y.-P. (2008). Error correction, revision, and learning. *Journal of Second Language Writing*, 17, 292–305.
- Vantage Learning. (2006). *Research summary: IntelliMetric™ scoring accuracy across genres and grade levels*. Retrieved from http://www.vantagelearning.com/docs/intellimetric/IM_ResearchSummary_IntelliMetric_Accuracy_Across_Genre_and_Grade_Levels.pdf
- Vantage Learning. (2007). *MY Access! Efficacy Report*. Retrieved from <http://www.vantagelearning.com/learning-center/research/#MyAccess>
- Wang, Y.-J., Shang, H.-F., & Briody, P. (2012). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, 26, 234–257.
- Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly*, 45, 769–774.
- Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, 3, 22–36.
- Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, 27, 335–353.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave MacMillan.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Honolulu, HI: University of Hawaii Press.
- Yilmaz, Y., & Yuksel, D. (2011). Effects of communication mode and salience on recasts: A first exposure study. *Language Teaching Research*, 15, 457–477.

Jinrong Li is an Assistant Professor in the Department of Writing and Linguistics at Georgia Southern University. Before that, she taught at Iowa State University and Boston University. Her research interests include computer-mediated communication and second language learning, L2 writing instruction and assessment, and research methods in applied linguistics. She has presented at Second Language Research Forum, AAAL, and Technology for Second Language Learning Conference, and her work has been published in *CALICO Journal* and *International Journal of Computer-Assisted Language Learning & Teaching (IJCALT)*.

Stephanie Link received her M.A. in TESOL from Winona State University and Tamkang University in Taiwan and is a Ph.D. candidate in the Applied Linguistics and Technology program at Iowa State University (ISU). Currently, she is a research assistant for the evolving software Research Writing Tutor for native and nonnative students' academic writing and an instructor for the ISU Graduate College and English Department. Her research interests are in computer-assisted language learning, L2 classroom assessment, and L2 writing. Stephanie's work has been presented at a range of national and international conferences, and her publications can be found in *System* and *CALICO Journal*.

Volker Hegelheimer, Professor, teaches courses on technology in language teaching and research, language assessment, and research methodology. His research interests include applications of the WWW and emerging technologies in language learning and language testing. He has presented his research and held academic workshops at numerous national and international conferences. His publications have appeared in journals such as *Language Learning & Technology*, *Language Testing*, *System*, *Computer-Assisted Language Learning*, *ReCALL*, *CALICO Journal*, and he contributed to several edited volumes on Computer-Assisted Language Learning (CALL).