

Towards Best ESL Practices for Implementing Automated Writing Evaluation

STEPHANIE LINK

AHMET DURSUN

KADIR KARAKAYA

VOLKER HEGELHEIMER

Iowa State University

ABSTRACT

In the past decade, a new wave of writing assessment tools has evolved as a result of advances in computer technology, highlighting the potential of automated writing evaluation (AWE) tools in the English as a second language (ESL) writing programs and comprising a concerted effort to move from use of AWE tools for testing to teaching. Today's innovative AWE tools offer ESL instructors promising solutions for providing immediate feedback and meeting the demands for better practices in the digital age. However, the outcomes of using AWE are not free of challenges (Chen & Cheng, 2008; Wang, Shang, & Briody, 2012) and necessitate a comprehensive understanding about how to effectively integrate AWE and improve students' writing (Ware, 2011). In situ studies are, therefore, necessary to identify the ways in which teachers are incorporating AWE into their classrooms in order to provide an insight into best practices. This longitudinal qualitative study investigates the practices and perspectives of five university-level writing instructors as they meet challenges and possibilities of using AWE in seven ESL writing courses. Data collection included observations, individual interviews, and delayed focus group interviews. In this paper, we highlight ESL instructors' teaching strategies, perception of the effectiveness, satisfaction, and concerns with the AWE tool (*Criterion*) they employed. We also describe changes in strategies and perceptions after a second semester of using the AWE tool. We conclude with final suggestions for best practices for integration of AWE into the ESL curriculum. Our research intends to expand the current knowledge of AWE by outlining appropriate and effective implementation of AWE technologies.

KEYWORDS

Automated Writing Evaluation, CALL, Technology, ESL Writing Instruction

INTRODUCTION

Major developments in automated writing evaluation (AWE, also known as automated essay scoring¹) have begun to impact the field of second language (L2) writing instruction by enabling immediate feedback on students' writing. The improvements in AWE, furthermore, have motivated researchers and practitioners to explore opportunities with repurposed commercial testing products, which today are making large strides in entering the realm of classroom practices.

Since the first automated essay scoring system was developed in the 1960s for the testing context (Project Essay Grade (PEG); Page, 2003), numerous scoring engines have emerged to evaluate texts on mechanical correctness, stylistic control, organizational structure, grammar usage, word choice, and semantic meaning (e.g. e-rater[®], Intelligent Essay Assessor[™], and IntelliMetric[®]). Today, these automatized scoring engines power a

range of modernized writing evaluation tools for classroom use, such as *Criterion*SM by Educational Testing Service (ETS) (*e-rater*[®]) and MY Access![®] by Vantage Learning (*Intellimetric*[®]), which have stimulated a new era of integrating automated essay assessment and assistance into writing instruction. However, incorporating AWE into the classroom can come with unforeseen challenges (see Chen & Cheng, 2008; Wang, Shang, & Briody, 2012); thus, more research is needed to provide better insight as to how AWE can most effectively be integrated and used in the classroom.

We conducted a case-study of five instructors in seven ESL academic university-level writing classrooms. Observations and individual interviews were completed to provide a comprehensive description of the innovative uses of AWE and the perceptions of instructors on the different uses of AWE inside and outside of writing classes. Focus group interviews after two semesters of AWE use revealed changes in classroom practices and perceptions once instructors became more experienced AWE users.

The New Era of Writing Assessment and Assistance

Since the onset of automated essay scoring in the 1960s (Page, 2003), several essay scoring engines have been developed including *e-rater*[®] by the Educational Testing Service[®] (Burstein et al., 1998) and *Intellimetric*TM of Vantage Learning, the current engine used for evaluating essays on the Graduate Management Admissions Test Analytical Writing Assessment (GMAT AWA; Elliott, 2003; Shermis, Raymat, & Barrera, 2003). Research on both essay scoring engines has consistently shown high correlations between human and computer ratings; for example in validating *e-rater*[®] V2.0, Attali and Burstein (2006) found a true-score correlation of .93 between human and machine scores, indicating that both were assigning similar scores. Although high correlation and agreement may be insufficient conditions for the validity of score use (Chung & Baker, 2003), score users can be more confident than in the past that the engines are effectively evaluating the overall quality of an essay, highlighting the potential of AWE for classroom use.

To enhance AWE tools for classroom purposes, developers have expanded the capabilities of the essay scoring engines to include more detailed feedback and additional writing resources (see Table 1). For example, *Criterion* Online Essay Evaluation (see <http://www.ets.org/criterion/> for an overview of features) relies on the *e-rater*[®] essay scoring engine, which is designed to provide only a holistic score, and *Critique*[®], which includes other forms of individualized feedback and writing assistant resources, such as access to portfolio options, detailed feedback explanation in a writer's handbook, history reports that save first and last drafts of essays, and a spell checker (Burstein, 2003). With the addition of individualized feedback and writing assistant resources, AWE now has the potential not only for guiding students in their development of writing but also for fostering more effective writing instruction.

Table 1
Automated Writing Evaluation Feedback and Resources

Software Engine	Commercial Product	Company	Feedback	Writing Assistant Resources
<i>e-rater</i> [®]	<i>Criterion</i>	Educational Testing Service	<ul style="list-style-type: none"> • Holistic Score • Organization/Development • Grammar • Usage • Mechanics • Style 	<ul style="list-style-type: none"> • Electronic portfolio • 'Writer's handbook' • Student history report • Graphical pre-writing tools • Multilingual feedback
<i>Intellimetric</i>	<i>MY Access!</i>	Vantage Learning	<ul style="list-style-type: none"> • Holistic Score • Focus/Unity • Organization • Development/Elaboration • Sentence Structure • Mechanics/Conventions 	<ul style="list-style-type: none"> • Electronic Portfolio • 'Writing dashboard' • Student history report • Graphical pre-writing tools • Multilingual dictionary • Thesaurus/Translator

Research on AWE Use in the Classroom

In the past decade, research has promoted the overall reliability and validity of automated essay scoring through vendor-sponsored and system-centric evaluation (e.g., Attali & Burstein, 2006; Elliot, 2003; Shermis & Hamner, 2013); recently, calls for ongoing research on the automated scoring engines for high-stakes assessment have been made (Weigle, 2013; Williamson, Xi, & Breyer, 2012). The scope of research must now extend to a thorough investigation of the validity of AWE use in the classroom for promoting students' writing development (Ware, 2011).

One pertinent quantitative study demonstrated the use of *Criterion* in classroom settings throughout the United States (Attali, 2004). The study investigated AWE use by sixth to twelfth graders. To determine whether the AWE feedback was helpful for students, the researcher evaluated the effectiveness of the automated feedback and revision features of *Criterion* by focusing on "the [possible] improvement in feedback from first to last submission of an essay" (p. 3). Findings showed that of the 33,171 submissions of 50 words or more, 23,567 (71%) were submitted only once and 9,604 were submitted more than once, suggesting that "most students did not exploit the revision capabilities of the *Criterion* system" (p. 4). Moreover, students were able to correct errors in subsequent versions of their essays, but the absence of significant differences between thesis statements in the first and final draft indicated that the change was typically in spelling and grammar rather than in organization.

A lack of information about the demographics of the participants and the detailed procedures for AWE implementation in Attali's study does not provide any insight towards a best practices model for AWE utilization. Furthermore, since there was no control group, the research does not provide a convincing argument for or against AWE use. In consideration of this issue, Wang et al. (2012) explored the impact of using one AWE tool, *CorrectEnglish*, in English as a foreign language classrooms in Taiwan by using an experimental and control group. Their study investigated differences in writing accuracy between the two groups and before and after pre- and post-test scores within the experimental group. They also studied the effects of AWE on students' autonomy and interaction. Their findings showed that the experimental group made significantly fewer errors. The same students held positive views towards the effect of AWE on their writing accuracy, and, accordingly, students' awareness of their autonomy was enhanced even though interaction with the tool had its limitations

(e.g. some argue that AWE feedback is too rule-based and too vague). While the study did not examine how students' attitudes may influence teachers' use of AWE, it provides a strong argument for use of AWE to enhance accuracy and autonomy. However, the researchers raised concerns about the effects of AWE on students' attitudes. Furthermore, although the authors provide pedagogical implications, a general limitation in AWE research is the lack of information on how to transfer these suggestions to effective utilization of the tools, which imposes the need to seek additional classroom techniques for fostering adequate revision and writing development.

Instructors' Perception of AWE and Areas of Continued Concern

To improve the approaches to enhancing revision behavior using AWE and, ultimately, improving students' writing, there is a need for a clear view of current classroom practices. Such practices are best explored using classroom observations and instructor interviews since these data collection procedures can provide a means of reflection and discussion that are useful for development of a model for best practices.

One study on instructors' perceptions revealed that AWE simplified classroom management by making writing instruction easier, saving time, creating a more enjoyable teaching experience, and allowing instructors to focus on higher-level concerns of writing instead of mechanics (Grimes & Warschauer, 2010). In their study, most instructors treated automated scoring as useful even though their confidence in the scoring was not strong. Interestingly, teachers in the study preferred to balance AWE use with conventional writing methods (e.g. commenting on individual essays orally or in writing), but the researchers provided a limited account of how and why exactly this was done. That is, the instructors' justification for AWE use and nonuse was unclear.

Filling the gap is Chen and Cheng's (2008) study, which showed that instructors' use of AWE was justified through differences in "the teachers' attitudes towards AWE scores and feedback, their views on the role of human feedback, their conceptions of teaching and learning of writing, and their technology-use skills in working with the AWE program" (p. 103). In their study, they addressed the use of *MY Access!* in an EFL setting of Taiwan. Three instructors participated in the research as they taught three university-level ESL writing courses for third-year English majors. Each of the courses taught academic writing skills, used the same textbook and similar content, lasted 18 weeks, and adopted a process-writing approach. The researchers conducted individual interviews with the instructors to uncover the pedagogical practices with AWE and found three noteworthy differences in: (1) the integration of the tool, (2) the use of automated scores and feedback, and (3) the decisions of when and when not to use the tool.

Chen and Cheng's research provides detailed accounts of students' and instructors' experiences with AWE along with their perceived effectiveness of using AWE, but stops short of exploring precisely how the teachers use AWE due primarily to a lack of observational notes. The researchers also did not examine teachers' attitudes towards AWE after they have become more experienced with integrating and managing the use of AWE. In addition, the context specificity of Chen and Cheng's study makes the transferability of results to the ESL context questionable. Our study deepens and refines the description of AWE use by investigating a university context in the United States; our work also explores the teachers' perspectives when they first used AWE and after two semesters of experience.

THE STUDY

Theoretical Framework

In order to address the various gaps in past research, such as the lack of research in the university-level ESL setting and the limited understanding of how and why instructors integrate AWE into the classroom, this study utilizes an interpretivist perspective to understand the meaning of events, concepts, and experiences of instructors in the classroom context (Creswell, 1998; Merriam, 2002). From this perspective, we are able to

construct meaning within our university context as it is shaped by the ESL instructors' perceptions and understanding of the use of AWE in their classrooms. This study also looks at the relationship between classroom pedagogy and computer-assisted language learning (CALL) with the intention of providing implications for ESL teaching that will enhance the context for integrating language learning and teaching with technology. In this regard, we take a sociocultural perspective that acknowledges the relationship between people and their practices that foster language learning (e.g. Lantolf, 2000; Lantolf & Thorne, 2006). The angle of investigation aligns well with Ortega (2007), who states that "social experience is an object of study rather than a random noise that needs to be eliminated from theory development" (p. 247).

We also intend to highlight best practices (Edge & Richards, 1998; Zemelman, Daniels, & Hyde, 2005), or 'good practice', which we define not as 'the only way' but as "the most current and professionally effective practice in order to benefit students". That is, our research takes a more generative approach proposed by McKeon (1998) "with instructors asking questions, exploring the research, making educated guesses about the models that are most likely to fill the bill, trying those models, and observing the effect those models have on their classes and their practice" (p. 498).

In light of the connection between CALL and classroom pedagogy, AWE may be useful in providing support for language development. We, therefore, investigate the integration of AWE in ESL academic writing classrooms in order to suggest a best practices model for maximizing the potential of AWE for language learning. More specifically, the study was guided by three research questions: (1) How are ESL instructors implementing the AWE tool in the ESL writing classroom? (2) What are instructors' perceptions of their experiences with the AWE tool? And (3) What areas of concern and suggestions do ESL instructors have for future AWE practitioners?

Methodology

We employed a basic interpretive case-study design and relied on classroom observations, individual interviews, and delayed focus group interviews with instructors. This design provided a comprehensive description of the uses of and perceptions towards AWE inside and outside the writing classes.

Context/Classes

This study was carried out at a large Midwestern research university during Spring 2011 and Fall 2012 semesters. Data were collected from instructors who taught two different levels of academic ESL writing classes during Spring 2011: (1) *English 101B: Academic English I* and (2) *English 101C: Academic English II*. A total of seven classes were part of the study (see Table 2 for characteristics of each class).

Table 2
Characteristics of the Seven ESL Writing Classes

	Instructor	Course ID	Number of Students	Online or Face-to-Face
Class 1	Abbi ^a	<i>English 101B</i>	14	Face-to-Face
Class 2	Adele	<i>English 101C</i>	24	Face-to-Face
Class 3	Adele	<i>English 101C</i>	24	Face-to-Face
Class 4	Michael	<i>English 101C</i>	24	Face-to-Face
Class 5	Jason	<i>English 101C</i>	24	Face-to-Face
Class 6	Jason	<i>English 101C</i>	24	Face-to-Face
Class 7	Ellie	<i>English 101C</i>	15	Online

^aNote: all names are pseudonyms

Non-native English speaking students were placed into *English 101B* and *English 101C* based on their scores on the writing section of an in-house English placement test. Their language proficiency was considered high intermediate (students in *English 101B*) to advanced low (students in *English 101C*) based on their scores on standardized English proficiency tests. Students were mainly from China with few from Korea, Malaysia, Indonesia, Thailand, and other countries. *English 101B* students were undergraduate students who demonstrated the need to further develop their academic writing skills, therefore requiring two additional semesters of ESL instruction. The course focuses on advanced English structure as it applies to academic writing features such as choosing topics and developing and organizing material. Successful completion of *English 101B* enables undergraduate students to proceed to the next level of ESL writing instruction, *English 101C*, which is designed to prepare students for first-year composition classes by introducing the process-writing approach, including pre-writing, revising and editing activities.

For the duration of the semester, instructors of *English 101C* classes were asked to follow the same syllabus and assignment sheets for each major paper (four papers in *English 101C*) with equal time spent on each assignment; the arrangement of assignments within the classes helped to ensure that students had the same number of weeks to work on each paper. The *English 101B* syllabus contained three major papers. Beyond the pedagogical and methodological choices in writing instruction, instructors were encouraged to integrate the AWE tool using their own creativity. In addition, all of the instructors had at least one class period per week in a computer lab, except for the fully online course.

Participants

In Spring 2011, a total of five ESL writing instructors in the university’s English department participated in the study. Table 3 highlights the demographic information of each participant. At the time of the study, all the participants were Ph.D. students in an applied linguistics and technology program and rated themselves as proficient users of technology. Of the five instructors, two were Chinese, two Korean, and one Turkish. Both Adele and Jason taught two sections of *English 101C*. Ellie was the only instructor to teach her course online.

**Table 3
Demographic Profile of the ESL Writing Instructors**

	Abbi ^a	Adele	Michael	Jason	Ellie
Course	<i>101B</i>	<i>101C</i>	<i>101C</i>	<i>101C</i>	<i>101C</i> online
L1	Korean	Korean	Chinese	Turkish	Chinese
Years of ESL Teaching Experience	4	10	5.5	5	7
Gender	Female	Female	Male	Male	Female

^aNote: all name are pseudonyms

The AWE Tool

Criterion, a web-based AWE tool developed by Educational Testing Service, was used in the study. The tool was designed to detect errors in native speakers’ English writing, but has been adapted for nonnative English speakers by including level-appropriate prompts. Students can use *Criterion* from the planning stage to the final submission. The tool provides instructors and students with feedback on essays through score reporting and diagnostic feedback. For example, if instructors decide to assign students a prompt from the

range of prompts suggested by *Criterion*, students can receive two types of feedback after their submissions: a holistic score (ranging from 1-6) and trait feedback analysis (i.e. linguistic feedback on grammar, mechanics, usage, style, organization and development). Interpretation of the feedback is also provided in the 'Writer's Handbook' that is accessible through the students' personal account at any time during the writing process. If the instructor chooses not to use one of the *Criterion* prompts, only trait feedback is provided.

Along with the two types of feedback and the 'Writer's Handbook', the software provides other writing assistant resources, including pre-writing graphic organizers, a spell checker, a portfolio option, and the 'Student Quick Access Guide', which provides students with directions on how to utilize the features of the program. The instructors and students can also use the program to leave comments about the essay.

Procedure

Semi-structured interviews with the instructors were conducted at the beginning and end of Spring 2011. Observation data were collected during the same semester. Then, delayed focus group interviews took place at the end of Fall 2011 to gauge general changes in pedagogical approaches and perspectives from the spring to the fall semester once instructors had gained more experience with using and integrating the AWE tool into their classes.

Individual Semi-structured Interviews

The first interview aimed to collect initial attitudes and perceptions, and the later interview sought post-application attitudes and perceptions of instructors towards AWE (see Appendix A for interview questions). These interviews also asked how instructors made use of AWE in their classes. The interviews followed an interview protocol which targeted three sub-sections and a chance for final remarks: (1) instructors' readiness to use *AWE*: to explore instructors' background with respect to use of technology in teaching ESL writing, (2) instructors' classroom use of *AWE*: to elicit information on the use of the software in the classroom, and (3) instructors' perceptions about the use, problems and helpfulness of *AWE*: to aid the researchers in knowing how instructors perceived the overall use of *AWE* in their classes. This third sub-section also examined in what ways feedback, grading, and the software interface affected instructors' teaching. The final sub-section allowed the instructors to provide additional comments and suggestions. The same set of interview questions were used for both interviews, but the first section of questions on teachers' readiness was only asked during the first interview and additional questions based on observational notes were asked during the final interview.

Observations

A total of seven classroom observations were conducted at various times throughout the semester based on instructors' recommendations and availability to see how instructors used the AWE tool in a computer lab class to teach ESL writing at various stages of the paper assignments. Out of the seven observations, three were done in *English 101B* and the other four in each of the face-to-face *English 101C* classes. Observational notes helped formulate questions during the final interview of the semester to facilitate an in-depth discussion of what occurred in class and to collect instructors' justifications for specific AWE activities. Notes also allowed for member checks during interviews to validate interpretations of data (Esterberg, 2002).

Delayed Focus Group Interviews

A focus group instructor interview was conducted one semester after the initial implementation of AWE to gauge whether instructors' use of the software and/or perceptions had changed over the two semesters once instructors became more experienced AWE users. In focus group interviews, "participants get to hear each other's

responses and to make additional responses beyond their own initial responses as they hear what other people have to say" (Patton, 1987, p. 135). Additionally, focus groups can lead to a more natural and relaxed atmosphere for the participants (Marshall & Rossman, 2006). The interview in this study was formatted to help researchers glean detailed information since instructors were able to reflect as a group and share responses freely. The focus group interview included a total of six questions (see Appendix B). A moderator asked each question and instructors answered questions by building on one another's answers.

DATA ANALYSIS

Qualitative analysis was conducted to answer each of the research questions. For the purpose of gathering reports of how *Criterion* was implemented, what the instructors' perceptions of *Criterion* use were, and what concerns and suggestions they may have had, all data from classroom observations, semi-structured interviews, and delayed focus group interviews were first carefully transcribed verbatim. The transcripts were inductively coded by three researchers utilizing grounded theory to identify patterns in the data. Each of the researchers followed procedures outlined by Lincoln and Guba (1985). That is, the researchers first identified the overriding patterns in the data by grouping together similar items. Then the data were divided into smaller and more meaningful units; at this point, themes related to use, perceptions, areas of concern, and suggestions emerged from the data. In order to check the reliability of these analysis procedures, an outside researcher acted as an external auditor to check the codes and provide opinions. The final dataset was discussed through peer debriefing and peer editing within the research group (Esterberg, 2002).

Instructor Implementation of AWE

To answer the first research question about how ESL instructors implement *Criterion*, we analyzed instructors' individual interviews and then evaluated any change in their pedagogical approach by analyzing delayed focus group interviews. Based on our qualitative analysis, ESL instructors' use of *Criterion* can be discussed in terms of in-class use and out-of-class use of AWE.

In-class Use of Criterion

In terms of in-class use, *Criterion* was used as a toolbox for pre-writing, peer review, editing/revising, and other short in-class activities (Table 4 presents a summary of pedagogical practices in the five writing courses). The first main in-class implementation was to emphasize pre-writing. Four of the five instructors agreed that students did not typically start with planning when they were not told to, but since *Criterion* provides the 'make a plan' feature, planning became an explicit step when integrating the software for a process writing approach. On the other hand, Ellie, who taught the online course, was the only one who did not train students to use the pre-planning feature because of her initial conception of AWE tools as only helping with grammar issues.

Table 4
In-Class Pedagogical Practices with Criterion in Seven ESL Writing Classes

Instructor	Course	In class Use	Grading	Human Feedback	Number of essays
Abbi	English 101B	Pre-writing tool Peer review manager Text editor Revision tool Grammar checker	Not used	Used Microsoft Word to supplement feedback	Diagnostic essay 3 major papers 1 in-class writing
Adele	English 101C (Two sections)	Pre-writing tool	Recorded students pre-writing plan as complete or incomplete	Provided more feedback on logic and organization.	Diagnostic essay 4 major papers Final essay writing
Michael	English 101C	Pre-writing tool Peer review manager Text editor Grammar checker Writing practice tool (free writing) Mechanics workshop Grammar clinic	Used holistic score for midterm diagnostic. Required holistic score of 4 before peer review and 5-6 before final submission	Provided more feedback on structure and content with little on grammar.	Diagnostic essay 4 major papers 5 in-class writing 2 grammar clinics
Jason	English 101C (Two sections)	Pre-writing tool Text editor Peer review manager Grammar checker	Required holistic score of 5-6 before final submission	Provided more feedback on organization	Diagnostic essay 4 major papers 1 in-class writing
Ellie ^a	English 101C online	Text editor Essay submission manager	Graded first and last draft	Provided a balance of feedback in Microsoft Word, but spent more time on focused discussion of language issues.	4 major papers 1 in-class writing

^a in-class writing in Ellie's class was implemented online

The second major in-class use was for peer-review exercises, in which case, the AWE tool assisted in managing the peer review process although there was not a built-in peer review feature. Three of the instructors used the software to do peer-review in class. Michael stated, "...I also asked the students to do the peer review on Criterion by leaving comments on their partner's computer..." A similar activity was observed in Jason's class.

His students opened their paper in *Criterion* and then students moved from one computer to another, leaving comments on their peers' papers using guidelines the instructor provided. Abbi did not originally use peer review in her *English 101B* class, but after discussing the possibilities of the AWE tool with her colleagues and observing one of her colleagues implement a peer review activity, she too decided to try peer review for the fall semester, which she indicated worked fairly well for her purpose: "I think doing peer review on *Criterion* is much more...helpful for [students] because they can check the feedback from *Criterion* together to see whether the partner had some problem to interpret feedback." In other words, students were able to work together to check the AWE feedback and negotiate errors.

Since *English 101B* focuses more on grammar, Abbi was the only instructor to utilize a large portion of in-class time exploring feedback in *Criterion* during the editing and revising stage of writing. The finding suggests that use of the AWE tool may be more meaningful for in-class use when grammar is a large focus of a lesson or class curriculum. Classroom observations illustrated this point; Abbi had set up a procedure in her class where students entered the lab class on 'editing days' and immediately entered their *Criterion* account and/or a word processing software to start editing their paper while Abbi circled the classroom to help with individual students. The word processing software was used because Abbi preferred it over *Criterion* for leaving her instructor's comments since she was still not familiar with giving feedback in *Criterion*. Interestingly, after two semesters of using the tool, Abbi continued to use the word processing software because of limitations in the teacher commenting features within *Criterion*, which we will discuss in a later section. Familiarity with the word processing software also influenced students to work with both programs at the same time because they were still learning the features of *Criterion*.

Out of all the instructors, Michael was the most exploratory with the use of *Criterion*. He remarked:

This semester, I am trying to use *Criterion* as much as I can in class just to explore the functions and to reduce my own workload... For example, in class if I can use *Criterion*'s 'Writer's Handbook', then I don't have to gather so many other examples of errors.

Along with pre-writing and peer review, Michael used *Criterion* prompts for free writing to provide additional practice and to track students' writing. Moreover, he used the 'Writer's Handbook' for workshops on punctuation and introductions to some of the trait feedback (e.g. style, usage, grammar, mechanics, organization and development), and trait feedback for grammar clinics, which taught grammar using metalinguistic language from the AWE tool to not only improve students grammar knowledge but to also enhance the understanding of *Criterion* feedback and the identification of errors in their writing (see Li & Hegelheimer, 2013 for more information on the grammar clinics).

Surprisingly, as Michael was the one to use the AWE tool in the most diverse ways, he was the only one to explicitly express his mixed-opinion towards AWE. He commented that the tool was not so useful for in-class use. That is, the AWE tool is much better if the students can use it after class because the functions are very convenient for them in that students have unlimited access to the tool to resubmit their papers and receive immediate feedback. They can resubmit their papers and receive feedback any time they want. In this way, the toolbox features of the AWE tool may play a role in increasing learner autonomy outside of the class.

Out-of-class Use of *Criterion*

In terms of out-of-class use, *Criterion* acted as an assistant and a grammar checker. As an assistant, the tool helped give feedback on students' papers, as illustrated by Jason's comments:

I feel now I have an assistant in my class. So students can ask him... questions as many times as they want. So that is something that is kind of a

relaxing point for me, so I feel like they have something there to get suggestions from.

After two semesters of using *Criterion* in the classroom, Jason still provided the same comments in the delayed focus group, “[AWE] is kind of an assistant...like a professor has, like a lab, like my TA. I see it in that kind of way.” Adele saw the tool as an aide to help regulate the amount of feedback given on students’ writing and in an appropriate amount of time, which was crucial since she and the other instructors were instructors at the same time as being full-time students, so without *Criterion* Adele did not always have time to provide an ample amount of feedback.

As a grammar checker, *Criterion* dealt with a large portion of students’ grammatical issues, allowing four of the instructors to focus more of their feedback on organization, structure, and content. Ellie, however, seemed to have a slightly divergent perspective when teaching her online class. In fact, she argued that the tool allowed her to spend more time on language related issues because the sometimes vague comments brought a need for developing more online lectures to balance focus on form and content.

Whether or not more time was spent on material development, the tool was still recognized as a general assistant for writing instruction and also a grammar checker for out-of-class needs. Overall, instructors attested that the AWE tool fulfilled significant needs for ESL writing instruction (both in class and out of class). In order to successfully adapt a given practice, instructors’ post-use perceptions can provide beneficial insight into how to change or modify an innovation for successful implementation.

Instructors’ Reflections of AWE Tool Use

To provide a meaningful interpretation of instructors’ use of an AWE tool, it was also important to see how instructors perceive the utility of *Criterion*, which is addressed in our second research question, “What are instructors’ perceptions of their experiences with the AWE tool?” In other words, we asked participants to reflect on the areas in which they think the tool is effective and/or ineffective as well as areas of satisfaction. A summary of findings is shown in Table 5.

Table 5
Summary of Instructors’ Perceptions towards Criterion

AWE is effective for:	(1) increasing learners’ metalinguistic ability (2) reducing the workload of the instructors (3) giving grammar feedback
AWE is ineffective for:	(1) providing enough necessary feedback (2) providing high-quality feedback (3) giving reliable holistic scores to students’ essays
Factors contributing to level of satisfaction:	(1) low satisfaction with holistic score reports (2) mixed satisfaction with Criterion feedback (3) high satisfaction with ability to promote student autonomy and motivation

EFFECTIVENESS OF CRITERION

In Table 5, three main areas of effectiveness were identified: (1) increasing learners’ metalinguistic ability, (2) reducing the workload of the instructors, and (3) giving grammar feedback. Instructors reported that AWE feedback fostered students’ metalinguistic ability. That is, *Criterion* allowed the learners to increase their knowledge of language through the

exposure to the linguistic terms they needed to understand in order to correct their mistakes based on trait feedback. Ellie stated that AWE contributed to her students' metalinguistic development, "*In fact, I probably spend more time on language issues for this semester because after using Criterion, students became more aware of the metalinguistic terms used in Criterion in the feedback.*" The other instructors expressed similar observations and underlined this benefit of fostering students' metalinguistic ability to reach course objectives.

The use of *Criterion* resulted in a reduced workload for the instructors. Since this tool was able to give instant feedback to the learners and play the role of an assistant, instructors felt that they were able to return their students' papers earlier compared to their previous experiences where they did not have such a classroom tool: "*...I think my time grading papers was reduced...I feel like I can grade more papers on a given day.*" (Ellie). That is, Ellie thought *Criterion* provided a sufficient amount of grammar feedback; thus, she focused less on the grammar mistakes when providing feedback, reducing the time she spent on grading papers.

Previously we mentioned that *Criterion* is able to detect numerous grammar features (e.g., run-on or garbled sentences, subject-verb agreement, ill-formed verbs, pronoun errors, wrong or missing word). The earlier findings from this study suggest that *Criterion* was effective in terms of providing the grammar feedback to the students so that instructors could focus on other issues like organizational problems. "*I think it's a good supplement, like it takes part of my job of correcting grammar errors...and I can focus on the organization*" (Michael). However, when it comes to ineffectiveness, it is noteworthy that instructors pointed to three main areas closely associated with feedback generation. The first point was that the tool is missing feedback for certain content-related characteristics such as coherence and cohesion, which are crucial features for promoting students' writing development. Jason noted, "*When I look at the Criterion feedback I feel if I don't give feedback to my students they will miss a lot.*"

The second ineffective feature of *Criterion* is the inaccuracy of feedback. All of the instructors thought that the AWE tool did not always provide the correct feedback which resulted in confusion, especially when detecting the missing or extra article or identifying how to reduce the number of repetitive words. Abbi indicated that "*Criterion does not detect all grammar errors in students' writing, so students can be misled by the wrong feedback from Criterion and that's why students got confused in my classes.*" For Ellie, the incorrect feedback created a situation where students wanted to defend themselves:

If there is incorrect or inaccurate feedback, they want to know why. They want to argue about every little thing because I think they don't really understand that grammar is not an absolute thing...sometimes there are exceptions and sometimes it depends on context.

While Adele had a similar experience, she felt that the ineffectiveness of *Criterion* could be turned into a learning experience by showing students how to approach the AWE feedback from an analytical perspective in order to improve grammatical knowledge and to judge or evaluate the AWE feedback. In other words, the instructors make it evident that *Criterion* is not flawless, but Adele's situation suggests that when a tool has its limitations, students can be made aware of the issues so that the technology can be adapted to fulfill the needs of the students in different contexts of learning. As highlighted by Hubbard (2004), learner training plays a major role for efficient adaptation of the technology. Since different technological environments offer various challenges, learners can be equipped with necessary training.

Finally, instructors thought *Criterion* did not provide reliable scores on students' essays. In other words, instructors mentioned that when assigning a score to essays, the tool misled students either by giving students a high score where actually the essay was below expectations or giving a low score when the essay was carrying the necessary features according to the instructors' criteria (see Li, Link, Ma, Yang, & Hegelheimer, 2013).

As a result of *Criterion's* inconsistent scoring system, teachers had mixed feelings as to the satisfaction with the tool.

Satisfaction

Teacher satisfaction with a tool plays a key role in its classroom integration. Therefore, we asked the instructors during their interviews, "In general, how satisfied are you with the AWE tool?" Table 6 shows that the average rating was 4.5 on a six-point Likert scale, where 1 means strongly dissatisfied and 6 means strongly satisfied.

Table 6
Instructors' satisfaction with AWE

Instructor	Rating	Explanation
Abbi	4.5	"4.5 because I believe that it helps students a lot because they can realize their grammatical errors by using <i>Criterion</i> , but from the instructors perspective, sometimes it took more time because I have to look at different stages of students' writing: brainstorming, the first draft, the second draft, and the third one"
Adele	4.5	
Michael	4.5	
Jason	5	I'm not very strongly satisfied but as I mentioned just before it is a good program but it can be definitely better. Because now we don't really have content feedback and that is something we work on a lot in class because we are following a process approach and content organization is more important for us. But sometimes <i>Criterion</i> ...does not give that feedback, and students sometimes get confused because they get a higher score.
Ellie	4.5	I would say between 4 and 5... in a way, it makes my job a little more difficult, so I have to give [students] more encouragement because it's... more frustrating to them and I have to explain things more clearly so that they can understand what's going on....But still, it provides more opportunities for students to just notice areas that are potentially problematic.

Indeed, instructors' overall satisfaction was not stable during the research period. That is, their satisfaction increased from Spring 2012 to Fall 2012 after they had used *Criterion* for a second semester. Abbi was able to comment on this during the fall focus group interview:

At the beginning of [Spring 2012] I was a little overwhelmed because I had to get used to using and understanding different types of interface and features of *Criterion*. However, as time passed by, I became familiar with using it. So, these days, I don't have difficulties in using *Criterion* as an instructor.

It is worth noting, however, that Abbi entered the teaching context as a proficient user of technology and indicated that she was not worried about exploring the use of technology in the classroom.

Accuracy of holistic scores, issues with feedback, motivation, and learner autonomy surfaced in the results to help explain the instructors' levels of satisfaction. First, instructors

saw Criterion's holistic scores as one of the main reasons of lower satisfaction because they caused students to think that the program was more objective than the instructors. Ellie commented on this issue: "Yeah, a student wrote in his journal that Criterion is more objective than the instructors because the instructor would give higher grades to his pets." Moreover, instructors shared views about limitations of comprehensive and consistent AWE feedback, which raised questions regarding the reliability of the program. Michael sheds light on this issue: "Even when they got zero errors based on Criterion feedback, there were still a lot of mistakes." Furthermore, Michael mentioned "Getting a high score from Criterion does not mean you are good, but if you get a low score in Criterion, it means you are problematic," thus, the inconsistency became a problem for many of the instructors.

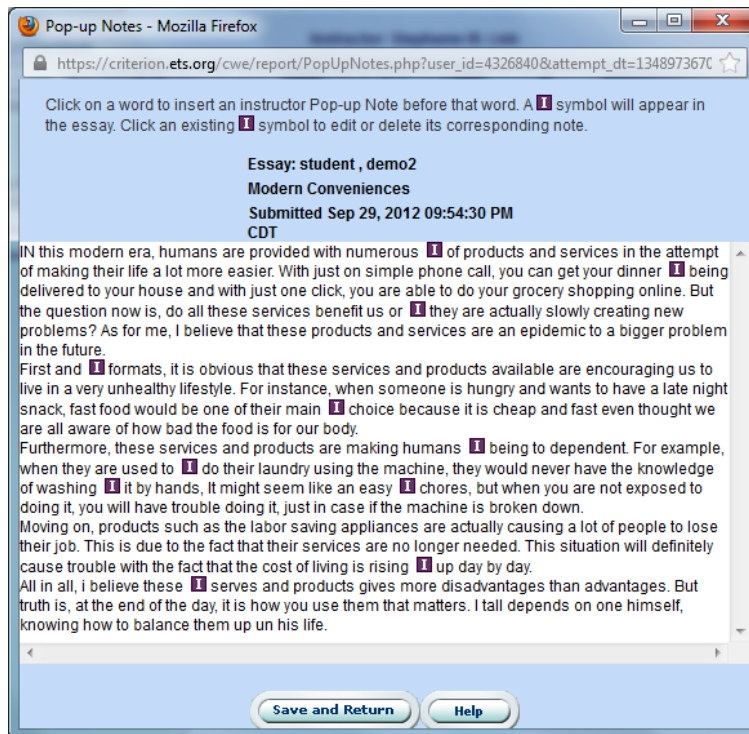
The second factor, issues with feedback, elicited mixed satisfaction with *Criterion*. The instructors compensated for the inaccurate and inadequate AWE feedback by commenting more on organization, but at the same time, instructors appreciated the AWE feedback since it acted as an assistant and seemed to reduce errors. Because the tool provided grammar feedback and hypothetically helped learners develop metalinguistic ability, instructors allotted more time and effort to other aspects such as essay development.

Finally, the five instructors were highly satisfied with the AWE tool because of its ability to promote student autonomy and motivation as it can be seen from Abbi's commentary: "Students had no problems with using Criterion, and they can review their essays by themselves with it so I think that's why using Criterion is helpful...it let them revise and check their grammar by themselves." Additional comments about the success of the tool for enhancing autonomy and motivation demonstrate the potential of AWE for developing students' writing; however, instructors' concerns about the use of AWE are still worthy of further investigation.

ADDITIONAL AREAS OF CONCERN AND INSTRUCTORS' SUGGESTIONS

In response to the third research question focusing on areas of concern and suggestions for future AWE practitioners, the results indicated two additional issues: technical limitations and pedagogical concerns. In the first place, instructors stated that the *Criterion* interface created unnecessary challenges in that they could not do everything they wanted to with the available functions. Abbi noted that feedback on drafts was not stored because only the first and last drafts are archived in *Criterion*. Interestingly, Adele also found the commenting feature (see Figure 1) disappointing, "I was planning to give comments in Criterion, but when I heard that once they read the comments, the comments were removed... instead of giving comments there, I decided to give comments on Microsoft Word." Two instructors also observed the small font size that tool uses, which makes reading feedback difficult. These limitations show the potential of AWE for in-class use, but user interface limitations with the current version of this particular AWE tool points out areas for future development and implications for other AWE developers.

Figure 1
Instructors' in-text comments on a student's draft in *Criterion*.



The second category of concern was related to pedagogy. There were two major aspects pointed out by the instructors: (1) affective factors and (2) instructor training. The first aspect that instructors expressed was difficulty dealing with affective factors. That is, instructors stated that they were put in an uncomfortable position when students questioned the reliability of instructor-given grades due to discrepancy between the grades and the AWE holistic scores, causing frustration on both sides. Ellie touches on this issue: *"It is more frustrating because...I felt it was really difficult to tell [my student] why I gave him a C-, but he already got a 5 from Criterion."* Abbi suggests that instructors should explain each different type of AWE feedback and provide examples because students cannot understand some of the errors, such as fragments and run-on sentences. This comes with understanding the metalanguage yourself, as Ellie stated, *"You would be surprised sometimes that your explanation might not be the same as what the system actually means. And so when that happens it's going to be a disaster for students."*

Additionally, as highlighted by Hubbard and Levy (2006), instructors raised a concern regarding instructor training. Most of the instructors stated that they were not fully aware of the potential of *Criterion* due to the lack of a comprehensive and cyclical training. Michael, the most exploratory instructor in the group, said, *"I noticed that there are still some functions that we did not explore."* As a suggestion, both Abbi and Adele advised playing around with the software to better answer students' questions. Adele stated, *"If the instructor does not know how to use the program...how do I expect the students to use the program."*

Instructors also made a few more suggestions, starting with thoughts about how to use AWE in the classroom. Ellie and Michael insisted that instructors strive to make good use of the tool. Although it is not necessary to make use of all the AWE features, Michael stated, *"You paid for this service...you have to make full use of it."* Additional research on specific tasks and classroom practices that lead to effective writing development and the effects of instructor training could provide insight as to how experience with the tool can foster more creativity in technology use. Indeed, the more people strive to understand and

benefit from an innovation, the more rapid its rate of successful adoption is likely to be (Rogers, 2003). Likewise, Rogers (2003) states that the success of an innovation also depends on how well it evolves to meet the needs of more and more demanding individuals and how these individuals become a part of the redevelopment of the innovation.

DISCUSSION

This study aimed to examine the practices and perspectives of university-level ESL writing instructors as they met challenges and possibilities of using a specific AWE tool, *Criterion*. Highlighted specifically were instructors' teaching strategies, perceptions of the effectiveness, satisfaction, and concerns with the software. Any changes in strategies and perceptions after a second semester of using the tool when the instructors were more experienced with the tool were also discussed. Several major findings are discussed in the following section.

In the classes investigated, it appears that AWE has many uses in the ESL classroom. As a general toolbox, *Criterion* offered opportunities for instructors to become more exploratory when designing classroom activities. As an assistant, *Criterion* allowed students to gain autonomy and motivation with some signs that the tool may also foster metalinguistic skills, and as a grammar checker, *Criterion* helped most instructors to spend less time on students' language issues and more time on organization and meaning, which fits the nature of a process-based approach to writing.

Instructors' perceptions of the effectiveness of the AWE tool were characterized by evident benefits for both the students and the teachers (e.g. students learned metalinguistic skills and instructors were able to reduce their workload). Additionally, while some researchers have shown that the AWE feedback is sometimes vague and formulaic, suggesting that "the use of AWE as a surrogate writing coach without human feedback may frustrate students..." (Chen & Cheng, 2008, p. 108), teachers in this study found that the inconsistent feedback created learning opportunities for students. Thus, giving grammar feedback was seen as an effective AWE component.

Grammar feedback, however, was not the only concern with the AWE tool; lower levels of satisfaction were also influenced by misleading holistic scores. Nevertheless, instructors perceived the tool as a motivational factor for students even though others have shown the opposite (Wang et al., 2012). With some concerns about the quality of the AWE feedback, it is important to acknowledge that the more familiar instructors are with the program the more satisfied they are likely to be, assuming they receive full training, ideas for implementation, and peer support. Along with these needs, additional suggestions for moving towards a best practices model for integrating AWE into the university-level ESL classroom are provided based on the instructors' reflections and past models of best practices (Edge & Richards, 1998; Zemelman et al., 2005):

1. Instructors in the study found that a lack of familiarity with the AWE tool impeded on the use of all its features. Thus, it is recommended that future AWE practitioners seek full training in how to use the AWE tool of interest from both a student and teacher standpoint. If training is not available, instructors should take time to explore the tool's features independently to prepare for unforeseen concerns.
2. After much discussion with colleagues, several of the instructors in this study were able to implement more creative teaching techniques using the AWE tool. This shows that raising questions between colleagues and even observing each other can help to decipher solutions to classroom issues and gain insight into creative AWE writing activities.
3. It was evident that accuracy rates for *Criterion* were not 100%. Students and learners need to avoid trusting AWE tools blindly, and instructors may help the cause by observing students' problems with the technology and trying to create learning opportunities to help students overcome difficulties with the tool.

4. Instructors in face-to-face versus online classes in this study found different ways to utilize the AWE tool. As a suggestion, instructors should find ways to adapt features of the technology to work best for each classroom context in order to make good use of AWE.
5. Out-of-class use of the AWE tool demonstrated a level of convenience that instructors in this study suggested could heighten learner autonomy. Promoting learner autonomy can, therefore, be established by helping students use AWE as a toolbox for the writing process.

Perhaps the most important suggestion is to not let frustration with AWE technology hinder the continued use of the tool in the classroom. As teachers in this study demonstrated, satisfaction can be heightened when experience increases.

In light of all the findings, it is probable that AWE, used alone, would not be a best-practice model for teaching ESL writing. Used as a supplement to or an extension of human feedback from peers or instructors, future AWE tools may be capable of filling the gaps that current tools are not yet advanced enough to fulfill. As Grimes and Warschauer (2010) state:

Mindful use of AWE can help motivate students to write and revise, increase writing practice, and allow instructors to focus on higher-level concerns instead of writing mechanics. However, those benefits require sensible instructors who integrate AWE into a broader writing program emphasizing authentic communication, and who can help students recognize and compensate for the limitations of software that appears more intelligent at first than on deeper inspections. Thus, like many educational technologies, it is unlikely to improve ineffective teaching, but it can help good instructors be more effective (p. 34).

Therefore, effective implementation of AWE is strongly influenced by how instructors overcome the drawbacks of technology to maximize its full potential in the classroom. Though successful implementation may be affected by many factors, such as students' socio-economic status and familiarity with technology (Grimes & Warschauer, 2010), the strongest predictor is instructors' beliefs about writing pedagogy (Chen & Cheng, 2008; Warschauer & Grimes, 2008). That is, if an instructor does not explore the tool or seek guidance in order to understand the capabilities of the technology, it will be difficult to implement it effectively and to ensure adequate revision and student development.

CONCLUSION

Although the use of *Criterion* seems to produce favorable results, this study has two factors that are worth addressing. First, the participants in this study were all considered "proficient users of technology." Although they received limited training, their familiarity with technology may have eased the integration of the AWE tool in the writing classroom. Second, the instructors may have felt confined to certain expectations because they shared the same syllabus. That is, they did not have the freedom to determine their own assignments, which may have impacted the amount of flexibility the instructors felt they had in developing their AWE in-class activities.

This study has also shown that AWE tools can be integrated into the ESL writing curriculum, but the amount of success largely depends on instructors' willingness to apply the concept of best practices by asking questions, exploring, and observing the effects of technology while overcoming difficulties and by adapting features of the technology that work best for a given context. Although AWE research is increasing in popularity, still more classroom-based research is needed. Thus, future research should focus on the effect of training on the use of AWE tools and of AWE integration into the status quo of a writing curriculum.

ACKNOWLEDGEMENTS

We would like to thank our Criterion Research Group for providing comments throughout our drafting process and for being a base of support.

NOTE

¹ The term automated essay scoring (AES) is most often used in general reference to the historical development of this software, which was originally developed to score essays for testing. Automated writing evaluation (AWE) includes numerical scores and other forms of feedback for classroom purposes (Warschauer & Ware, 2006).

REFERENCES

- Attali, Y. (2004). *Exploring the feedback and revision features of Criterion*. Archives of the Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/erater_NCME_2004_Attali_B.pdf
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, 4(3), 1-21.
- Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-131). Mahwah, NJ: Lawrence Erlbaum Associates.
- Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris M. D. (1998). Automated scoring using a hybrid feature identification technique. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics* (pp. 206-210). East Stroudsburg, PA: Association for Computational Linguistics.
- 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. Retrieved from <http://llt.msu.edu/vol12num2/chencheng.pdf>
- 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, NJ: Lawrence Erlbaum Associates.
- Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.
- Edge, J., & Richards, K. (1998). Why best practice is not good enough. *TESOL Quarterly*, 32(3), 569-576. doi: [10.2307/3588127](https://doi.org/10.2307/3588127)
- Educational Testing Service. (2011). *About the e-rater® scoring engine*. Retrieved from <http://www.ets.org/erater/about/>
- Elliott, S. (2003). Intellimetric™: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates.
- Esterberg, K. G. (2002). *Qualitative methods in social research*. New York, NY: McGraw-Hill.
- Grimes, D., & Waschauer, M. (2006). Learning with laptops: A multi-method case study. *Journal of Educational Computing Research (JERIC)*, 38(3), 305-332. doi: [10.2190/EC.38.3.d](https://doi.org/10.2190/EC.38.3.d)
- 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), 4-43. Retrieved from <http://files.eric.ed.gov/fulltext/EJ882522.pdf>

- Hubbard, P. (2004). Learner training for effective use of CALL. In S. Fotos & C. Browne (Eds.), *New perspectives on CALL for second language classrooms* (pp. 45-69). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hubbard, P., & Levy, M. (2006) *Teacher education in CALL*. Philadelphia, PA: John Benjamins.
- Lantolf, J. (2000). *Sociocultural theory and second language learning*. Oxford: Oxford University Press.
- Lantolf, J., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford: Oxford University Press.
- Li, Z., & Hegelheimer, V. (2013). Mobile-assisted grammar exercises: Effects on self-editing in L2 writing. *Language Learning and Technology*, 17(3), 135-156. Retrieved from <http://ilt.msu.edu/issues/october2013/lihegelheimer.pdf>
- Li, Z., Link, S., Ma, H., Yang, H. & Hegelheimer, V. (2013). The role of automated writing evaluation holistic scores in the ESL classroom. Manuscript submitted for publication.
- Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.
- Marshall, C., & Rossman, G. B. (2006). *Designing qualitative research* (4th ed.). Thousand Oaks, CA: Sage.
- McKeon, D. (1998). Best practice: Hype or hope?. *TESOL Quarterly*, 32(3), 493-501. [doi: 10.2307/3588119](https://doi.org/10.2307/3588119)
- Merriam, S. B. (2002). *Qualitative research in practice: Examples for discussion and analysis*. San Francisco, CA: Jossey-Bass.
- Ortega, L. (2007). Second language acquisition explained? SLA across nine contemporary theories. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 225-250). Mahwah, NJ: Lawrence Erlbaum Associates.
- 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, NJ: Lawrence Erlbaum Associates.
- Patton, M. Q. (1987). *How to use qualitative methods in evaluation*. Newbury Park, CA: Sage.
- Rogers, M. E. (2003). *Diffusion of innovations* (5th ed.). New York City, NY: Free Press.
- Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313-346). New York, NY: Routledge.
- Shermis, M. D., Raymat, M. V., & Barrera, F. (2003, April). *Assessing writing through the curriculum with automated essay scoring*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago: IL.
- Vantage Learning. (2011). IntelliMetric®. Retrieved from <http://www.vantagelearning.com/products/intellimetric/>
- Wang, Y., Shang, H., & 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(3), 234-257. [doi: 10.1080/09588221.2012.655300](https://doi.org/10.1080/09588221.2012.655300)
- Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly*, 45(4), 769-774. [doi: 10.5054/tq.2011.272525](https://doi.org/10.5054/tq.2011.272525)
- Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, 3, 22-36. [doi: 10.1080/15544800701771580](https://doi.org/10.1080/15544800701771580)
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 1-24. [doi:10.1191/1362168806lr190oa](https://doi.org/10.1191/1362168806lr190oa)

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

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2-13. doi: [10.1111/j.1745-3992.2011.00223.x](https://doi.org/10.1111/j.1745-3992.2011.00223.x)

Zemelman, S., Daniels, H., & Hyde, A. (2005). *Best practices: Today's standards for teaching and learning in America's schools*. Portsmouth, NH: Heinemann.

APPENDIX A. INDIVIDUAL SEMI-STRUCTURED INTERVIEW QUESTIONS

1. Instructors' readiness to use Criterion

Q1: 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?

Q2: How would you describe your technology expertise in general? (i.e., highly proficient, proficient, average, low proficiency, poor)

Q3: What types of software/hardware/online resources have you used for teaching? What target language group have you taught with technology and for which skills? How long?

Q4: Other than Criterion, have you ever used any writing software as an instructor or student?

2. Instructors' use of Criterion

Q5: How did you introduce the use of Criterion to your learners?

Q6: How did you make use of Criterion in the classroom? (i.e., activities, stages of use, supplementary Criterion resources used)

3. Instructors' perceptions about the use/problems/helpfulness

Q7: 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?

Q8: Was Criterion helpful to you as an instructor? In what ways? (i.e., interface, features, feedback, grading, affective factors, ease of use)

Q9: Did you encounter any problems with using Criterion? (i.e., interface, features, feedback, grading, affective factors, ease of use)

Q10: In general, how satisfied are you with Criterion?

Q11: Do you think Criterion was helpful to your students? In what ways? (i.e., interface, features, feedback, grading, affective factors, ease of use)

Q12: Did the students encounter any problems with using Criterion? (i.e., interface, features, feedback, grading, affective factors, ease of use)

Q13: In general, how satisfied do you think your students are with Criterion?

4. Additional Comments

Q14: 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?

Q15: Do you have any additional comments? (i.e., workload, feedback, grading)

Q16: Any follow up questions that arise from observations.

APPENDIX B. DELAYED FOCUS GROUP INTERVIEW GUIDES

Introduction:

Hi, my name is _____ and I will be helping to lead this discussion today. We will be here about an hour.

Guidelines:

Here are some guidelines for you to know about:

Taping: Please notice that the conversation will be taped and so please speak in a loud voice and speak one at a time.

Names: I would also like you to know that for research purposes, the information you provide during the interview will not be linked to other types of data through your names.

Acknowledge:

I want to thank you each for being here. Your time is very valuable and your opinions are important. Let's get started by having you introduce yourself to the group and tell us:

1. Your name and the class where you used Criterion this semester.
2. What is it that makes you decide to use Criterion in your class?
3. What did you use Criterion for?
4. How did you feel about using Criterion in your teaching of writing?
5. What was your students' reaction to the use of Criterion in the course?
6. What aspect of Criterion do you think is the most helpful or problematic in an ESL writing class or for an ESL writing student?

AUTHORS' BIODATA

Stephanie Link is a PhD student in Applied Linguistics and Technology at Iowa State University. Her research interests include the use of CALL technologies for second language writing instruction and assessment. Specifically, her work centers around complexity/chaos theory and systemic functional approaches to the development and validation of automated writing evaluation for assessing students' writing development. She also focuses on genre analysis of research articles and academic writing instruction.

Ahmet Dursun is a doctoral student in the Applied Linguistics and Technology program at Iowa State University. He received a bachelor's degree in English Language Teaching from Marmara University, Istanbul in 2006 and a master's degree in Teaching English to Speakers of Other Languages from Minnesota State University in 2009. He has presented his work at professional conferences such as the European Association for Language Testing and Assessment (EALTA), Teachers of English to Speakers of Other Languages (TESOL), the American Association for Applied Linguistics (AAAL), the Computer-Assisted Language Instruction Consortium (CALICO), and Technology for Second Language Learning (TSLL). He also works as an English instructor in Iowa State's ESL academic writing program.

Kadir Karakaya is a PhD student in Applied Linguistics and Technology doctoral program at Iowa State University. His research explores the applications of technology for enhancing students' writing skills, individual differences within automated writing evaluation (AWE) contexts, and assessment in distance learning environments. He is also interested in corpus-based investigations of academic writing.

Volker Hegelheimer, Professor of English in the Program in Applied Linguistics & Technology, teaches courses on technology in language teaching and research, language assessment, and research methodology. His research interests include applications of technology in language learning and language testing. In his more recent projects, he has focused on issues surrounding the implementation and use of automated writing evaluation tools in writing classes. 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 has contributed to several edited volumes on Computer-Assisted Language Learning (CALL) and Language Testing.

AUTHORS' ADDRESSES

smcross@iastate.edu

adursun@iastate.edu

karakaya@iastate.edu

volkerh@iastate.edu