Language Learning & Technology
http://llt.msu.edu/vol12num2/chencheng/
June 2008, Volume 12, Number 2
pp. 94-112

BEYOND THE DESIGN OF AUTOMATED WRITING EVALUATION: PEDAGOGICAL PRACTICES AND PERCEIVED LEARNING EFFECTIVENESS IN EFL WRITING CLASSES

Chi-Fen Emily Chen and Wei-Yuan Eugene Cheng
National Kaohsiung First University of Science and Technology, Taiwan

Automated writing evaluation (AWE) software is designed to provide instant computer-generated scores for a submitted essay along with diagnostic feedback. Most studies on AWE have been conducted on psychometric evaluations of its validity; however, studies on how effectively AWE is used in writing classes as a pedagogical tool are limited. This study employs a naturalistic classroom-based approach to explore the interaction between how an AWE program, *MY Access!*, was implemented in three different ways in three EFL college writing classes in Taiwan and how students perceived its effectiveness in improving writing. The findings show that, although the implementation of AWE was not in general perceived very positively by the three classes, it was perceived comparatively more favorably when the program was used to facilitate students' early drafting and revising process, followed by human feedback from both the teacher and peers during the later process. This study also reveals that the autonomous use of AWE as a surrogate writing coach with minimal human facilitation caused frustration to students and limited their learning of writing. In addition, teachers' attitudes toward AWE use and their technology-use skills, as well as students' learner characteristics and goals for learning to write, may also play vital roles in determining the effectiveness of AWE. With limitations inherent in the design of AWE technology, language teachers need to be more critically aware that the implementation of AWE requires well thought-out pedagogical designs and thorough considerations for its relevance to the objectives of the learning of writing.

INTRODUCTION

Automated writing evaluation (AWE), also referred to as automated essay scoring (AES)[1], is not a brand-new technology in the twenty-first century; rather, it has been under development since the 1960s. This technology was originally designed to reduce the heavy load of grading a large number of student essays and to save time in the grading process. Early AWE programs, such as *Project Essay Grade*, employed simple style analyses of surface linguistic features of a text to evaluate writing quality (Page, 2003). Since the mid-1990s, the design of AWE programs has been improving rapidly due to the advance of artificial intelligence technology, in particular natural language processing and intelligent language tutoring systems. Newly developed AWE programs, such as *Criterion* with the essay scoring engine "e-rater" by Educational Testing Service and *MY Access!* with the essay scoring engine "Intellimetric" by Vantage Learning, boast the ability to conduct more sophisticated analyses including lexical complexity, syntactic variety, discourse structures, grammatical usage, word choice, and content development. They provide immediate scores along with diagnostic feedback in various aspects of writing and can be used for both formative and summative assessment purposes. In addition, a number of AWE programs are now web-based and equipped with a variety of online writing resources (e.g., thesauri and word banks) and editing features (e.g., grammar, spelling, and style checkers), which make them not only an essay assessment tool but also a writing assistance tool. Students can make use of both AWE's assessment and assistance functions to help them write and revise their essays in a self-regulated learning environment.

Although AWE developers claim that their programs are able to assess and respond to student writing as human readers do (e.g., Attali & Burstein, 2006; Vantage Learning, 2007), critics of AWE express strong skepticism. Voices from the academic community presented in Ericsson and Haswell's (2006) anthology,

for example, question the truth of the industry's publicity for AWE products and the consequences of the implementation of AWE in writing classes. They distrust the ability of computers to "read" texts and evaluate the quality of writing because computers are unable to understand meaning in the way humans do. They also doubt the value of writing to a machine rather than to a real audience, since no genuine, meaningful communication is likely to be carried out between the writer and the machine. Moreover, they worry whether AWE will lead students to focus only on surface features and formulaic patterns without giving sufficient attention to meaning in writing their essays.

The development and the use of AWE, however, is not a simple black-and-white issue; rather, this issue involves a complex mix of factors concerning software design, pedagogical practices, and learning contexts. Given the fact that many AWE programs have already been in use and involve multiple stakeholders, a blanket rejection of these products may not be a viable, practical stand (Whithaus, 2006). What we need are more investigations and discussions of how AWE programs are used in various writing classes "in order to explicate the potential value for teaching and learning as well as the potential harm" (Williamson, 2004, p. 100). A more pressing question, accordingly, is probably not *whether* AWE should be used but *how* this new technology can be used to achieve more desirable learning outcomes while avoiding potential harms that may result from limitations inherent in the technology. The present study, therefore, employed a naturalistic classroom-based approach to explore how an AWE program was implemented in three EFL college writing classes and how student perceptions of the effectiveness of AWE use were affected by different pedagogical practices.

LITERATURE REVIEW

Assessment Validation of AWE: Theory-Based Validity and Context Validity

AWE programs have been promoted by their developers as a cost-effective option of replacing or enhancing human input in assessing and responding to student writing[2]. Due to AWE vendors' relentless promotion coupled with an increasing demand for technology use in educational institutions, more and more teachers and students have used, are using, or are considering using this technology, thus making research on AWE urgently important (Ericsson & Haswell, 2006; Shermis & Burstein, 2003a; Warschauer & Ware, 2006). Most of the research on AWE, however, has been funded by AWE developers, serving to promote commercial vitality and support refinement of their products. Industry-funded AWE studies have been mostly concerned with psychometric issues with a focus on the instrument validity of AWE scoring systems. They generally report high agreement rates between AWE scoring systems and human raters. They also demonstrate that the scores given by AWE systems and those by other measures of the same writing construct are strongly correlated (see Dikli, 2006; Keith, 2003; Phillips, 2007). These findings aim to ensure AWE scoring systems' construct validity and provide evidence that AWE can rate student writing as well as humans do.

Assessment validation, however, is more complex than simply comparing scores from different raters or measures. Chung and Baker (2003) caution that "high reliability or agreement between automated and human scoring is a necessary, but insufficient condition for validity" (p. 29). As Weir (2005) points out, construct validity should not be seen purely as "a matter of the *a posteriori* statistical validation," but it also needs to be viewed as an "*a priori* investigation of what *should* be elicited by the test before its actual administration" (p. 17). Weir stresses the importance of the non-statistical *a priori* validation and suggests that "theory-based validity" and "context validity" are crucial for language assessment. From a socio-cognitive approach, Weir notes that these two types of validity have "a symbiotic relationship" and are influenced by, and in turn influence, the criteria or construct used for marking as part of scoring validity (p. 20). He calls special attention to the role of context in language assessment, as context highlights the social dimension of language use and serves as an essential determinant of communicative language ability.

To examine the theory-based validity of AWE, we need to discuss what writing is from a theoretical perspective. A currently accepted view of writing employs a socio-cognitive model emphasizing writing as a communicative, meaning-making act. Writing requires not only linguistic ability for formal accuracy but, more importantly, meaning negotiation with readers for genuine communicative purposes. Writing thus needs to take into account both internal language processing and contextual factors that affect how texts are composed and read (Flower, 1994; Grabe & Kaplan, 1996; Hyland, 2003). Most AWE programs, however, are theoretically grounded in a cognitive information-processing model, which does not focus on the social and communicative dimensions of writing. They treat texts solely as "code" devoid of sociocultural contexts and "process them as meaningless 'bits' or tiny fragments of the mosaic of meaning" (Ericsson, 2006, p. 36). They "read" student essays against generic forms and preset information, but show no concern for human audiences in real-world contexts.

Even the Conference on College Composition and Communication (CCCC) in the U.S. has expressed disapproval of using AWE programs for any assessment purpose and made a strong criticism: "While they [AWE programs] may promise consistency, they distort the very nature of writing as a complex and context-rich interaction between people. They simplify writing in ways that can mislead writers to focus more on structure and grammar than on what they are saying by using a given structure and style" (CCCC, 2006). CCCC's criticism expresses their concern with not only the theory-based validity of AWE programs but also a washback effect, or a "consequential validity" (in Weir's, 2005, terminology): AWE use may encourage students to write to gain high scores by giving more attention to the surface features that are more easily detected by AWE systems than to the construction of meaning for communicative purposes (Cheville, 2004).

With regard to the context validity, information on how and why AWE is used to assess student writing in educational contexts is often lacking (Chung & Baker, 2003). When the context and the purpose of using AWE are unknown, it is difficult to truly judge the validity of AWE programs. In addition, Keith (2003) points out that most psychometric studies on AWE have been conducted on large-scale standardized tests rather than on classroom writing assessments; hence, the validity of AWE could differ in these two types of contexts. He speculates that the machine-human agreement rate may be lower for classroom assessments, since the content and meaning of student essays is likely to be more valued by classroom teachers.

Another important validation issue for AWE is the credibility of the scoring systems. A number of studies found that writers can easily fool these systems. For instance, an essay that is lengthy or contains certain lexico-grammatical features preferred by the scoring systems can receive a good score, even though the content is less than adequate (Herrington & Moran, 2001; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002; Ware, 2005). Students can thus devise means of beating such systems, rather than making a real effort to improve their writing. Moreover, since AWE systems process an essay as a series of codes, they fail to recognize either inventive or illogical writing (Cheville, 2004), nor can they recognize nuances such as sarcasm, idioms, and clichés used in student essays (Herrington & Moran, 2001). When meaning and content are more emphasized than form, the fairness of AWE scoring is often called into question.

Pedagogical Foundation of AWE: Formative Learning and Learner Autonomy

To enhance their pedagogical value, several AWE programs, such as *MY Access!* and *Criterion*, have been developed for not only summative but also formative assessment purposes by providing scores and diagnostic feedback on various rhetorical and formal aspects of writing for every essay draft submitted to their scoring systems. Students can then use the computer-generated assessment results and diagnostic advice to help them revise their writing as many times as they need. The instructional efficacy of AWE, as Shermis and Burstein (2003b) suggest, increases when its use moves from that of summative

evaluation to a more formative role. Though AWE scoring systems' validity remains contended, their diagnostic feedback function seems pedagogically appealing for formative learning.

Formative assessment is used to facilitate learning rather than measure learning, as it focuses on the gap between present performance and the desired goal, thus helping students to identify areas of strengths and weaknesses in gaining directions for improvement (Black & Wiliam, 1998). For second language (L2) writing, formative feedback, as Hyland (2003) suggests, is particularly crucial in improving and consolidating learners' writing skills. It serves as an in-process support that helps learners develop strategies for revising their writing. Formative feedback can therefore support process-writing approaches that emphasize the need for multiple drafting through a scaffold of prompts, explanations, and suggestions. Although formative feedback is a central aspect of writing instruction, Hyland and Hyland (2006) point out that research has not been unequivocally positive about its role in L2 writing development since many pedagogical issues regarding feedback remain only partially addressed. The form, the focus, the quality, the means of delivery, the need for, and the purpose of feedback can all affect the usefulness of feedback in improving writing. These issues are vital not only for human feedback but for automated feedback as well.

Studies on AWE for formative learning, however, have not been able to demonstrate that automated feedback is of much help during students' revising process. The most frequently reported reason is that automated feedback provides formulaic comments and generic suggestions for all the submitted revisions. Thus, students may find it of limited use. Moreover, since such feedback is predetermined and unable to provide context-sensitive responses involving rich negotiation of meaning, it is useful only for the revision of formal aspects of writing but not of content development (Cheville, 2004; Grimes & Warschauer, 2006; Yang, 2004; Yeh & Yu, 2004). Additionally, Yang's (2004) study reveals that more advanced language learners appeared to show less favorable reactions toward the AWE feedback. Learners' language proficiency may constitute another variable affecting the value of such feedback.

AWE programs, like many other CALL tutors, are designed to foster learner autonomy by performing error diagnosis of learner input, generating individualized feedback, and offering self-access resources such as dictionaries, thesauri, editing tools, and student portfolios. In theory, such programs can provide opportunities for students to direct their own learning, independent of a teacher, to improve their writing through constant feedback and assistance features in a self-regulated learning environment. However, whether students can develop more autonomy in revising their writing through computer-generated feedback and making use of the self-help writing and editing tools available to them is uncertain. This may lead to questions of student attitudes toward and motivation for the use of AWE (Ware, 2005). Additionally, Beatty (2003) cautions that CALL tutors often "follow a lock-step scope and sequence," thus giving learners "only limited opportunities to organize their own learning or tailor it to their special needs" (p. 10). Such a problem may also occur when AWE is used.

Vendor-funded studies on AWE programs have demonstrated significant improvement on standardized writing tests (e.g., Attali, 2004; Elliot, Darlington, & Mikulas, 2004; Vantage Learning, 2007). Although these results are encouraging, Warschauer and Ware (2006) criticize many of these studies for being methodologically unsound and outcome-based. Accordingly, a major problem of this type of research is that "it leaves the educational process involved as a black box" (p. 14). These studies seem to attribute the observed writing improvement to the AWE software itself but ignore the importance of learning and teaching processes. Warschauer and Ware thus suggest that research on AWE should investigate "the interaction between use and outcome" (p. 10), for it can provide a more contextualized understanding of the actual use of AWE and its effectiveness in improving writing.

One recent classroom-based AWE study on the interaction between use and outcome is particularly worth noting. Grimes and Warschauer (2006) investigated how *MY Access!* and *Criterion* were implemented in U.S. high school writing classes. They found two main benefits of using AWE: increased motivation to

practice writing for students and easier classroom management for teachers. More importantly, their study revealed three paradoxes of using AWE. First, teachers' positive views of AWE did not contribute to more frequent use of the programs in class, as teachers needed class time for grammar drills and preparation for state tests. Second, while teachers often disagreed with the automated scores, they viewed AWE positively because, for students, the speed of responses was a strong motivator to practice writing, and, for teachers, the automated scores allowed them to offload the blame for grades onto a machine. Third, teachers valued AWE for revision, but scheduled little time for it. Students thus made minimal use of automated feedback to revise their writing except to correct spelling, punctuation, and grammatical errors. Their revisions were generally superficial and had little improvement in content. In addition, the use of these two programs did not significantly improve students' scores on standardized writing tests. The authors caution that AWE can be misused to reinforce artificial, mechanistic, and formulaic writing disconnected from communication in real-world contexts.

Based on the studies reviewed here, it can be concluded that the validity of AWE scoring systems has not been thoroughly established and the usefulness of automated feedback remains uncertain in any generalized sense. AWE programs, even those designed for formative learning and emphasizing learner autonomy, do not seem to improve student writing significantly in either form or meaning. Therefore, AWE programs are often suggested to be used as a supplement to writing instruction rather than as a replacement of writing teachers (Shermis & Burstein, 2003b; Ware, 2005; Warschauer & Ware, 2006). Yet, how AWE can be used as an effective supplement in the writing class and how different learning contexts and pedagogical designs might affect the effectiveness of AWE warrants further investigation. The present study addresses these issues by exploring the interaction between different pedagogical practices with an AWE program in three EFL writing classes and student perceptions of learning outcomes. The purpose is to reveal how different learning/teaching processes affect the perceived value of AWE in improving students' writing.

METHODOLOGY

The Context

This study is a naturalistic classroom-based inquiry that was conducted in three EFL college writing classroom contexts in a university in Taiwan. The three writing classes, for third-year English majors, were taught by three instructors who were all experienced EFL writing teachers. They shared some common features in their writing instruction: 1) the three writing courses were required of third-year Taiwanese college students majoring in English; 2) their course objectives all aimed to develop students' academic writing skills; 3) the three instructors used the same textbook and taught similar content; 4) each class ran for 18 weeks and met three hours per week; and 5) they adopted a similar process-writing approach, including model essay reading activities followed by language exercises and pre-writing, drafting and revising activities.

An AWE program, *MY Access!* (Version 5.0) (Vantage Learning, 2005), was implemented in the three writing classes for one semester on a trial basis. The main purpose for the AWE implementation was to facilitate students' writing development and to reduce the writing instructors' workload. Before the writing courses started, the three instructors received a one-hour training workshop given by a *MY Access!* consultant. The workshop introduced how each feature of the program worked; however, it did not provide hands-on practice or instructional guidelines. Consequently, the three instructors had to spend extra time working with the program to familiarize themselves with its features and to develop their own pedagogical ideas for the AWE implementation. They had total autonomy to design writing activities with *MY Access!* as they saw fit for their respective classes. No predetermined decision on how to incorporate the program with their writing instruction was made by the institution or the researchers.

Participants

The three writing classes varied slightly in size: there were 26 students in Class A, 19 in Class B, and 23 in Class C. All the students were Taiwanese third-year college students majoring in English. They had formally studied English for eight years: six years in junior and senior high school and two years in college. Their English language proficiency was approximately at the upper-intermediate level. They were taking the required junior year EFL academic writing course and also had taken fundamental academic writing courses in their freshman and sophomore years. It was their first time using AWE software in their writing classes. As English majors, most of them were highly motivated to develop their English writing skills.

AWE Software

MY Access! is a web-based AWE program using the IntelliMetric automated essay scoring system developed by Vantage Learning. The scoring system has been calibrated with a large set of pre-scored essays with known scores assigned by human raters. These essays are then used as a basis for the system to extract the scoring scale and the pooled judgment of the human raters (Elliot, 2003). It can provide holistic and analytic scores on a 4-point or 6-point scale along with diagnostic feedback on five dimensions of writing: focus and meaning, content and development, organization, language use and style, and mechanics and conventions. The program offers a wide range of writing prompts from informative, narrative, persuasive, literary, and expository genres for instructors to select for writing assignments. It can be used as a formative or summative assessment tool. When used for formative learning, the program allows for multiple revisions and editing. Students can revise their essays multiple times based on the analytic assessment results and diagnostic feedback given to each essay draft submitted to the program. When run for summative assessment, the system is configured to provide a single submission with an overall assessment result.

In addition, the program provides a variety of writing assistance features, including *My Editor*, *Thesaurus*, *Word Bank*, *My Portfolio*, *Writer's Checklist*, *Writer's Guide*, *Graphic Organizers*, and *Scoring Rubrics*. The first four features were most commonly used in the three writing classes: 1) *My Editor* is a proofreading system that automatically detects errors in spelling, grammar, mechanics and style, and then provides suggestions on how such problems can be corrected or improved; 2) *Thesaurus* is an online dictionary that offers a list of synonyms for the word being consulted; 3) *Word Bank* offers words and phrases for a number of writing genres, including comparison, persuasive, narrative, and cause-effect types of essays; 4) *My Portfolio* contains multiple versions of essays from a student along with the automated scores and feedback. It allows students to access their previous works and view their progress.

Data Collection and Analysis

The data included the students' responses to the end-of-the-course questionnaire made by the researchers, focus group interviews with the students, individual interviews with writing instructors, and the students' writing samples along with the scores and feedback generated by *MY Access!*. The questionnaire surveyed the students' perceived effectiveness of using *MY Access!* for writing improvement, with a primary focus on the adequacy and helpfulness of its automated scores, feedback, and writing assistance features. Four major writing assistance features – *My Editor*, *Word Bank*, *Thesaurus*, and *My Portfolio* – were chosen to be evaluated, as they were the most commonly used modules in the three classes. The questionnaire contained both multiple-choice questions using a Likert scale and open-ended questions. In total, 53 out of 68 students (21 from Class A, 18 from Class B, and 14 from Class C) responded to the questionnaire.

To triangulate the questionnaire results and recapitulate the learning process with the AWE program, three focus group interviews with the students from each class and two interviews with the instructors of two classes were conducted after the courses ended. The students participating in interviews were all volunteers and given a small gift certificate to thank them: five from Class A, five from Class B, and six

from Class C. The instructors of Class A and Class B agreed to be interviewed, but the instructor of Class C declined; therefore, the information about Instructor C was obtained only through the survey results and the interview with the students from Class C. Each interview lasted approximately one hour and was conducted in Mandarin Chinese. The interviewees were asked to talk about how *MY Access!* was used in their writing classes, how they felt about the value of the program, and what factors affected their perceived effectiveness of using the program. All the interviews were audio-taped and then transcribed in Chinese and translated into English. In addition, the students' writing samples along with their automated scores and feedback documented in their online portfolios were used to validate their self-reports.

FINDINGS

Pedagogical Practices with AWE

The three instructors implemented *MY Access!* as an integrated part of their writing instruction, but they did not have the same pedagogical practices using the program. According to the interview data, the differences in the AWE implementation among the three classes were particularly noteworthy in three respects: 1) ways of integrating the AWE program into each instructor's process-writing curriculum, 2) uses of the automated scores and feedback for student writing improvement, and 3) decisions on when and when not to use the AWE program.

Ways of Integrating AWE into the Process-Writing Curriculum

Instructors A and B both designed two stages for students' drafting and revising process. At the first stage, students worked with the program independently, pre-submitting their essay drafts and revising their writing according to the automated scores and feedback they received for each draft. Instructor A, however, required her students to achieve a minimum satisfactory score of 4 out of 6 before submitting their essays for teacher assessment and peer review. At the second stage, she gave written feedback to the students' essays that had achieved the required level and conducted peer review activities through in-class guided discussions. The students then had to revise their essays based on the instructor's feedback along with their peers' comments. Finally, they resubmitted their revisions to the instructor for a final check and published their work on a web-based course platform for everyone in the class to read. Instructor A designed the two-stage process for two major reasons. First, she had less confidence in the usefulness of the AWE program: she believed that the program could only help students improve some basic linguistic forms and organization structures. Second, she believed human input was imperative for students' revising process especially in the areas of content development and coherence of ideas.

Instructor B, who also implemented *My Access!* in two stages, did not require her students' writing to reach a certain level at the first stage; rather, she allowed them to revise their essays using the program as often as needed and only set a deadline for them to submit their essays to her. In the second stage, she gave written feedback to the students' essays; in addition, she conducted individual teacher-student conferencing sessions, where they discussed what the students could do to improve their writing with reference to both the automated feedback and the teacher feedback, yet the automated feedback was given less emphasis. Then, the students revised their essays again and resubmitted their essays to the instructor for a final check. Instructor B also emphasized the importance of human input and considered teacher assessment and feedback much more important and meaningful than that given by the AWE program.

Unlike Instructors A and B, who played substantial roles in facilitating students' writing process, Instructor C did not provide feedback or much consultation during students' drafting and revising process. Rather, she used *MY Access!* to do all the assessing and responding and also asked the students to work with the program autonomously. In addition, she asked them to post their essay drafts in "Yahoo! Groups" for online peer review. However, she did not give specific guidelines on how peer review should be done; as a result, the students felt that their peers' comments usually lacked substantial suggestions and then they lost interest in doing online peer review after a couple experiences. Except for conducting language

exercises and pre-writing activities, Instructor C's involvement in the students' writing process was minimal. What she did was give summative assessments for all the submitted essays at the end of the semester.

Uses of AWE's Scores and Feedback for Student Writing Improvement

Both instructors A and B reported that they did not have strong trust in the automated scoring and feedback system; they had more confidence in their own assessment. Nevertheless, Instructor A set a score of 4 as a threshold for turning in essays for human assessment. She chose this score on the basis of the "MY Access Six Point Holistic Rubric": "4" indicates that the essay achieves an "adequate" level of communicating the writer's message, while "5" indicates "good" and "6" indicates "very effective." She also asked the students to take advantage of the automated feedback to help them reduce mistakes or problems in grammar, language use, and organization. In so doing, she was able to lessen her grading load and focus more on the meaning and content development in her own feedback.

Instructor B gave less emphasis to automated scores and feedback; instead, she asked her students to pay greater attention to the grades and feedback given by her. This was because she found the automated scores that her students obtained were very similar (i.e., most of them obtained a score of 3 or 4 for their first drafts) and doubted that such scores provided valid evaluations of their writing. Also, she found that the information in the automated feedback was vague and formulaic, and thus regarded such feedback as of little help to them.

Instructor C counted the automated scores as part of the students' actual grades, which suggested that she might have had more confidence that the automated scores were able to reflect students' writing performance to a reliable extent. Moreover, without giving teacher feedback, she asked the students to rely only on the automated feedback during their revision process, which implied that she seemed to trust such feedback to provide sufficient and useful information in guiding students to improve their writing.

Decisions on When and When Not to Use AWE

Both Classes A and C implemented *MY Access!* for sixteen weeks, but Class B used it for only six weeks. Instructor B decided to discontinue the AWE implementation for three reasons. First, she pointed out that the automated scores and feedback were of little help in improving students' writing, as mentioned earlier. Second, due to the vagueness of the automated feedback, she had to spend more time explaining what the feedback really meant by providing more specific details, which she thought did not reduce, but increased, her workload. Third, she and her students felt frustrated in working with the program due to many technical problems they encountered but could not solve. She reported frankly that she did not spend much time figuring out how the program worked and that her competence in working with computer technology was generally not very good. She then decided to ask the students to use MS Word instead, which they all knew how to work with.

Instructors A and C had no difficulties in working with the program. However, Instructor A did not ask her students to use the program for all the writing assignments. She commented that the AWE program imposed constraints on the topical content, the organizational structure, and the discourse style the students could use in their writing. She felt that these constraints restricted students' creativity and idea development. Therefore, she provided them other opportunities to practice writing with more freedom. For instance, for their final assignment she asked them to write an essay about Taiwan using any style they preferred. Her purpose was to allow students to write this essay in a personally more meaningful and stylistically less restrained manner.

In contrast, Instructor C used the program for both formative and summative assessment of all the writing assignments. In addition, she used it as a writing assessment tool for the final exam to measure students' timed essay writing performance. Her purpose, according to the students' account, was to give them a timed writing exam experience, which was similar to the writing exams given in TOEFL or GMAT.

However, after the exam, quite a few of her students complained to her about using this method to determine their final exam grades, since they had not practiced timed essay writing before, nor did they trust the fairness of machine scoring based on their experience of using the program for the whole semester. Instructor C was persuaded by the students' entreaties and decided to allow them to submit the revision of their final-exam essays to her for re-assessment. This decision reveals that her confidence in machine scoring seemed to be lowered because of the students' reactions to the use of AWE.

Below is a summary of the major differences in pedagogical practices with *MY Access!* among the three writing classes (see Table 1).

Table 1. Pedagogical Practices with *MY Access!* in Three Writing Classes

	Class A	**Class B**	**Class C**
Purpose of Use	- Formative assessment	- Formative assessment	- Both formative and summative assessment
Ways of Use	- Students wrote multiple drafts using the program during the early drafting and revising process. - Both the assessment and assistance features were emphasized.	- Students wrote multiple drafts using the program during the early drafting and revising process. - The assistance features were emphasized more than the assessment features.	- Students wrote multiple drafts using the program during the whole writing process. - The program also served as a writing assessment tool for the final exam. - The assessment features were more emphasized than the assistance features.
Grading Policy	- Students were not allowed to turn in their essays to the teacher until they gained a score of 4 or above. - Students' actual grades were determined by the teacher.	- The automated scores were unimportant and used for reference only. - Students' actual grades were determined by the teacher.	- The automated scores accounted for 40% of students' final grades. - Students' actual grades were determined by both the program and the teacher.
Human Feedback	- The teacher gave feedback during students' later revising process. - In-class peer review was conducted.	- The teacher gave feedback and provided teacher-student conferencing during students' later revising process. - No peer review was conducted.	- The teacher gave feedback to all the essays at the end of the semester but provided little consultation during students' revising process. - Online peer review was loosely conducted.
Duration of Use	16 weeks	6 weeks	16 weeks
Number of Essays Written using AWE	4 take-home essays	2 take-home essays	6 take-home essays & 1 timed in-class essay

The three different AWE implementations were likely affected by four important factors: 1) the teachers' attitudes toward AWE scores and feedback, 2) their views on the role of human feedback, 3) their conceptions of the teaching and learning of writing, and 4) their technology-use skills in working with the AWE program. For instance, a combination of AWE facilitation with human feedback, as was found in Classes A and B, was probably due to the instructors' limited confidence in AWE and their greater emphasis on the importance of human feedback, whereas the autonomous use of AWE in Class C might have resulted from the instructor's elevated level of confidence in the AWE program. However, Instructor C's confidence in the AWE's assessment function may have also been reduced by her students' reactions at the end of the semester. On the other hand, a mixture of machine-governed writing and more personally meaningful writing, as found in Class A, probably resulted from the instructor's perception of the limitations of the AWE program in facilitating writing and her view of how writing should be taught and learned. Finally, the quick abandonment of the AWE implementation in Class B was likely attributed to the instructor's disapproval of the AWE program and her incompetent technology-use skills.

Student Perceptions of AWE Effectiveness

Overall Evaluation of AWE

According to the questionnaire survey, none of the students found *MY Access!* greatly helpful, 55% of them considered it either moderately or slightly helpful, and the other 45% found it unhelpful (see Table 2). Of particular note are the different ratings among the three writing classes. In Class A, 86% of the students thought that the program was more or less helpful for their writing improvement, compared to 28% in Class B and 43% in Class C. The high positive response from Class A suggested that the way *MY Access!* was implemented in Class A seemed to be the most effective. On the other hand, the very negative reactions from Class B probably resulted from the short period of implementation, which may not have given students sufficient opportunities to experience any benefits of the program. Additionally, their reactions could have been negatively influenced by the instructor's apparent disapproval of the program.

Table 2. Overall Perceived Effectiveness of Using *MY Access!*

To what degree do you think using *MY Access!* can help you improve your writing?	Greatly helpful	Moderately helpful	Slightly helpful	Not helpful
Class A (N=21)	0	6 (29%)	12 (57%)	3 (14%)
Class B (N=18)	0	1 (6%)	4 (22%)	13 (72%)
Class C (N=14)	0	2 (14%)	4 (29%)	8 (57%)
Total (N=53)	0	9 (17%)	20 (38%)	24 (45%)

(Note: This question was designed purposely not to include "Neutral" as a choice option)

Adequacy and Helpfulness of AWE Scores and Feedback

It is noteworthy that none of the respondents showed agreement regarding the adequacy of automated scores (see Table 3). A very high percentage of disagreement, 83%, was found in Class B, whereas 57% in Class A and 42% in Class C. Class B's widespread distrust of the computer-rated scores was probably due to the instructor's negative attitude toward those scores. As one student from Class B remarked, "How is it possible for us to trust this kind of scores if our teacher does not think they are valid?" As for Class A, since they often found discrepancies between the automated assessment results and their

instructor's assessment and their peers' comments, their confidence level in these scores was not high. Only in Class C did close to 60% of the respondents hold a neutral attitude toward the automated scores, which may be due to the fact that they received only automated scores and thus had nothing to compare them with.

Table 3. Reactions toward AWE Scores and Feedback

	SA	**A**	**N**	**D**	**SD**
The scores given by *MY Access!* are adequate.					
Class A (N=21)	0	0	9 (43%)	12 (57%)	0
Class B (N=18)	0	0	3 (17%)	13 (72%)	2 (11%)
Class C (N=14)	0	0	8 (58%)	2 (14%)	4 (28%)
The written feedback given by *MY Access!* is helpful for revision.					
Class A (N=21)	0	7 (33%)	4 (20%)	7 (33%)	3 (14%)
Class B (N=18)	0	1 (6%)	8 (44%)	6 (33%)	3 (17%)
Class C (N=14)	0	4 (29%)	3 (21%)	4 (29%)	3 (21%)

(Note: SA=strongly agree; A=agree; N=neutral; D=disagree; SD=strongly disagree)

The reasons that the students gave for such negative reactions toward the automated assessment were similar to what has been found in the literature. According to their self-reports, there were four major problems with the scoring system.

1) It favors lengthiness. The longer an essay is, the higher score it is awarded by the program.
2) It overemphasizes the use of transition words. The score of an essay can be increased immediately by adding more transition words such as 'as a result', 'to sum up', or 'on the other hand' without changing other things.
3) It ignores coherence and content development. An essay that has four or five paragraphs containing "key words" related to the topic can be awarded a high score, even though it has serious coherence problems and illogical ideas.
4) It discourages unconventional ways of essay writing. One student from Class C recalled that she once used a story as an introduction in her essay but received a low score along with the feedback indicating a logical problem between the introduction and the conclusion. She then deleted the story and wrote a conventional introduction instead, which improved her score greatly. This shows that the scoring system values conventional, formulaic essay writing styles. Hence, the AWE program may not only fail to adequately assess students' writing but also restrict the ways students express themselves.

With regard to the helpfulness of the AWE feedback, around 50% of the respondents in each class considered it to be of no help. Although they commented that the automated feedback helped them pay attention to some previously unattended language use problems (such as formality, use of the passive voice, and sentence variety), they also criticized the feedback for being "vague," "abstract," "unspecific," "formulaic," and "repetitive." Thus, it did not provide them useful guidance to help them revise their essays. This was particularly evident in the areas of coherence and content development. Moreover, students would receive no feedback if their essays were judged as "off-topic." A few students who had this experience remarked that they felt very frustrated because the program gave no explanation of why their essays were off the topic. This problem became even worse when teacher feedback was also lacking,

as in the case of Class C. These findings show that the AWE feedback alone was unable to provide sufficient or substantial suggestions addressing students' writing problems.

Helpfulness of AWE Writing Assistance Features

The three classes had slightly more positive reactions toward the helpfulness of AWE writing assistance functions than that of assessment scores (see Table 4). Class A rated the four writing assistance features more positively than the other two classes. The combined percentage of "agree" and "neutral" for each item was over 70%. This result can be attributed to Instructor A's emphasis on both the assessment and assistance functions of *MY Access!* (see Table 1). Moreover, she clearly demonstrated to her students how to use each function of the program. In contrast, Instructor B had difficulty showing the students how to use some of the functions properly, whereas Instructor C asked her students to explore the assistance features by themselves, but some did not know how to use these features.

Table 4. Reactions Toward AWE Writing Assistance Features

	SA	**A**	**N**	**D**	**SD**
It is helpful to use *My Editor* during my writing process.					
Class A (N=21)	0	9 (42%)	8 (38%)	3 (15%)	1 (5%)
Class B (N=18)	0	6 (33%)	3 (17%)	7 (39%)	2 (11%)
Class C (N=14)	0	2 (14%)	6 (72%)	3 (7%)	3 (7%)
It is helpful to use *Word Bank* during my writing process.					
Class A (N=21)	0	6 (29%)	10 (48%)	3 (14%)	2 (9%)
Class B (N=18)	0	6 (33%)	2 (11%)	6 (33%)	4 (22%)
Class C (N=14)	0	2 (14%)	4 (29%)	6 (43%)	2 (14%)
It is helpful to use *Thesaurus* during my writing process.					
Class A (N=21)	0	8 (38%)	8 (38%)	4 (19%)	1 (5%)
Class B (N=18)	0	4 (22%)	6 (33%)	6 (33%)	2 (11%)
Class C (N=14)	1 (7%)	1 (7%)	5 (36%)	5 (36%)	2 (14%)
It is helpful to use *My Portfolio* during my writing process.					
Class A (N=21)	0	3 (14%)	12 (58%)	3 (14%)	3 (14%)
Class B (N=18)	0	2 (11%)	6 (33%)	7 (39%)	3 (17%)
Class C (N=14)	0	3 (21%)	4 (29%)	3 (21%)	4 (29%)

(Note: SA=strongly agree; A=agree; N=neutral; D=disagree; SD=strongly disagree)

Among the four writing assistance features, *My Editor* was viewed slightly more favorably than the others, probably because it helped the students immediately identify their spelling, grammar, and punctuation mistakes as well as some style problems. However, the students reported that the drawback of this feature was similar to the automated feedback: it was not informative enough. For example, it flagged a mistake in verb tense but did not advise which tense to use. Some students even doubted the usefulness of the grammar-check function since the program was unable to take the context of language use into consideration. As for *Word Bank* and *Thesaurus*, the students pointed out two major problems: first, no

explanations or example sentences were provided to illustrate how each word was used; second, the words and phrases collected in these two features were not sufficient to meet their needs. In regard to *My Portfolio*, students enjoyed easy access to their essays along with scores and feedback, but they did not find it of particular help during their writing process.

Appropriateness of AWE Implementation

While many limitations of the program's assessment and assistance functions could have negatively affected student perceptions of the effectiveness of AWE use, the three classes' different reactions were also likely affected by the specific ways the AWE program was implemented. The students' comments on the appropriateness of the implementation of the program centered around four issues: use of automated scores, need for human feedback, students' language proficiency, and purpose for learning writing.

1) Use of Automated Scores. Class B was the only one that disregarded the automated scores, whereas Class A and Class C used those scores for different purposes. The students of Class A generally approved of the policy that they had to pre-submit their essays to the program and achieve a minimum satisfactory score before teacher assessment. The required minimum score (i.e., 4 points) was not too hard for them to obtain in the preliminary stage of their writing process. Therefore, they did not perceive the automated assessment as a threat. Instead, many of them viewed it as a self-confidence builder because it ensured that the grades they received from the instructor would not be too low. Moreover, since their essays still had to be assessed by the instructor, they would not stop revising their essays just because they had obtained a good computer-rated score.

Unlike Class A, the students of Class C were more anxious about the automated scores they received for their essays because those scores accounted for 40% of their final grades. Many of them did not find it an appropriate way of assessing their writing since the AWE scoring system had numerous biases. They suggested that AWE scores should be used only as references to show whether their writing would be improved during the drafting and revising process rather than actual indicators of their writing performance.

2) Need for Human Feedback. All the students in the interviews contended that human feedback was indispensable even when AWE software was used. The students in Class A considered teacher feedback much more valuable than automated feedback because the teacher provided more specific, concrete suggestions and personal comments regarding both form and meaning. In addition to teacher feedback, they also found peer feedback beneficial to them, as it helped them expand or reformulate their ideas from different perspectives. Class A's more favorable attitude toward the AWE implementation (as compared to the other two classes) may have also been related to the availability of both teacher feedback and peer feedback complementing AWE feedback.

In Class B, since the instructor displayed a strong distrust of AWE scores and feedback, the students tended to react similarly; hence, they relied heavily on teacher feedback, which they considered of much greater importance than AWE feedback. As for Class C, many of the students disapproved of the use of the AWE program without teacher feedback. They did not find using the program autonomously to be of much help in improving their writing. Some remarked that the absence of teacher feedback in the AWE learning environment negatively affected their attitudes not only toward the AWE program but also toward the writing class in general. This suggests that the use of the AWE program to replace the teacher's role in assessing and responding to student essays would not satisfy students' needs nor would it motivate them to take such a writing class again.

3) Students' Language Proficiency. A number of students suggested that this program might be more suitable for students whose English proficiency level was lower. They commented that, first, as third-year English majors, they had already learned fundamental essay writing skills and therefore should not have been constrained by so many machine-controlled rules in their writing. Instead, they wished to write with

more flexibility and creativity. Secondly, since the program provided help mainly in the areas of mechanical accuracy and formal organization, students thought that such help would be more useful to beginners or intermediate learners since form tended to be viewed more importantly than meaning in early-stage learning of L2 writing. Thirdly, many of them pointed out that, as more advanced writers, they needed assessment and feedback focusing more on the meaning of their writing. Since the AWE program failed to focus on meaning, students did not find a strong need to use it for the learning of more advanced writing skills. These responses suggest that students who are at a more advanced language proficiency level may not want their writing to be confined by a set of machine-governed criteria, and furthermore they do not find such machine-generated form-focused responses to be valuable.

4) Purpose for Learning Writing. This issue is related to the previous one: students at different language proficiency levels may not have the same purpose for learning writing. One key question that is constantly asked in L2 writing classes is whether form or meaning should receive greater emphasis. For the students of the three classes, the consensus was that more weight should be given to meaning generally, and specifically idea development and meaning construction processes. Although many of them agreed that this program helped them to produce grammatically correct and conventionally well-organized essays, they doubted the value of using the program to learn writing if no teacher facilitation was provided. For instance, one student criticized the program for "depriving learners of an opportunity to show human qualities in their writing" and claimed that if students followed all the program's rules to write, their writing would "lose vitality." Moreover, some students, particularly from Class C, questioned why they had to write for a computer program that did not understand what they wrote. They did not see any purpose in learning writing this way nor did they consider it possible to achieve their goal of writing to communicate by interacting with the software.

DISCUSSION

Overall, AWE use was not perceived very positively by students in the three writing classes, and this is likely due to limitations inherent in the program's assessment and assistance functions. The literature on AWE use has also pointed out limitations in the design of AWE technology, such as favoring lengthiness and certain lexico-grammatical features, failing to recognize incoherent or illogical writing, and generating formulaic and unspecific feedback (Cheville, 2004; Herrington & Moran, 2001; Powers et al, 2002; Ware, 2005; Yang, 2004). This study, however, shows that writing teachers' pedagogical practices with AWE software can further affect student perceptions of the effectiveness of AWE in facilitating their learning of writing.

The AWE implementation was viewed comparatively more favorably when the program was used to facilitate students' early drafting and revising process, and when the teacher made a policy of asking students to meet a preliminary required standard and subsequently provided human feedback (Class A). The integration of automated assessment and human assessment for formative learning offers three advantages. First, it can assure students that their writing has achieved a minimum acceptable level before human assessment and can increase their confidence in writing to a certain extent. Second, it can help teachers focus more on meaning negotiation and idea development when giving their feedback, since the program has made suggestions to improve mechanical accuracy and organization. Third, it can reduce the impact of automated assessment whose validity remains questionable because the machine-rated scores serve only as a preliminary assessment, whereas teacher-rated scores are the final evaluation of students' writing performance.

The study has also revealed the necessity of providing human feedback in the AWE learning environment to redress the limitations of AWE feedback. Computer-generated feedback from the *MY Access!* system, though delivered instantly and viewed as helpful in improving some formal aspects of writing, provides only formulaic, generic information that cannot address students' individual writing problems, particularly in the areas of coherence and idea development, whereas human feedback can attend to

meaning, respond to the writer's thoughts, and give specific, personal comments. Moreover, giving human feedback can redress the limitations of the cognitive information-processing model that AWE is based on and take into account the social and communicative dimensions of writing. It is imperative to note that the use of AWE as a surrogate writing coach without human feedback may frustrate students since there is no meaning negotiation between the writer and human readers. As Rothermel (2006) cautions us, although some AWE programs claim to use the process-writing approach to improve students' writing "through a continuous, iterative process of writing and revising," these programs actually "mask a different ideology, one that defines not just writing, but also teaching and learning, as formulaic and asocial endeavors" (p. 200). AWE, if used alone without appropriate human input and feedback, seems unlikely to help students achieve the goal of writing for effective communication in terms of both form and meaning.

Furthermore, any use of CALL software needs to take into consideration learner characteristics and learning goals. The students in this study were all third-year English majors who had already learned fundamental essay writing skills. They questioned the value of AWE since they did not want to be confined by machine-governed rules nor did they want to be assessed by a machine that did not understand what they wrote. These reasons, in fact, are closely related to students' goals for learning writing. As Ericsson (2006) suggests, "during discussions of the machine scoring of writing, participants must carefully consider why we ask students to compose essays and what we expect them to gain from knowing how to compose such texts" (p. 30).

Since many AWE programs currently hold up rather static and formulaic models of "good writing", such as the prototypical five-paragraph essay (Ware & Warschauer, 2006), AWE programs may be of help in training students to write stylistically and structurally well-formed essays that the test graders may prefer to read. This approach could be useful if the learning goal is to strengthen the form of essay writing or to achieve a high score on large-scale institutionalized writing tests. This can also explain why many studies on AWE, particularly those funded by the industry, highlight its effectiveness in improving the scores of students taking standardized writing tests. On the other hand, if the goal is to communicate the writer's thoughts effectively to real audiences and demonstrate the writer's creativity and originality, using AWE is probably not a good choice. This concern can be particularly important for advanced language learners, as they may want to transcend conventional writing styles and develop their own styles to demonstrate the vitality of their writing.

Whether to focus on form or meaning in L2 writing has been the subject of many debates (Hyland, 2003); however, the choice is never an either-or question, but rather a matter of degree of emphasis. Advanced language learners may expect feedback on their writing to be more meaning-focused and may not be contented with predominantly form-focused responses provided by the machine (Yang, 2004). Therefore, how to strike a balance between form and meaning in L2 writing instruction merits serious attention when AWE is used as a pedagogical tool. In the present study, we see how an instructor attempted to attend to both form and meaning by asking students to tackle their grammatical and organizational problems through the machine-generated feedback while helping them to construct meaning and develop a sense of audience through teacher feedback and peer review. Moreover, this instructor did not require students to use AWE all the time but provided them opportunities to write on personally meaningful topics and with individually preferred styles. This flexibility in integrating facilitation of both machine and human may help students, on the one hand, to enhance their autonomy as learners and to raise their awareness of writing conventions and mechanics by working with AWE independently, and, on the other hand, to learn to write for meaning construction and genuine purposes by interacting with the teacher and peers in the AWE learning environment.

LIMITATIONS OF THE STUDY AND RECOMMENDATIONS

This paper has presented a descriptive study illuminating important insights into the process of teaching and learning writing with AWE and its linkage to student perceptions of learning outcomes. It is, however, an exploratory study and has several limitations. First, the teachers' views of AWE use and instructional rationales were not exhaustively documented. The study would have been more revealing if the viewpoints of the instructor who emphasized the autonomous use of AWE had been obtained. Second, the study would have uncovered more in-depth information showing what was actually happening in the AWE learning environments if classroom observations had been conducted. Third, the findings may be limited to this particular EFL context of advanced English learners. The transferability of the findings depends to a large degree on the similarity of other learning contexts to the present one. Fourth, the learning outcomes were investigated only through students' perceptions; the real gains could be further examined by comparing student essays before and after AWE use. A final limitation is the difficulty of attributing effects to any single factor. This study shows how a combination of contextual and pedagogical factors interact holistically, but how individual factors contribute specifically to the effectiveness of AWE needs further investigation.

Many language scholars and educators have pointed out that it is vital to investigate the pedagogical effectiveness of using technology in various learning contexts, rather than merely to focus on the effectiveness of the technology innovations (e.g., Beatty, 2003; Chapelle, 2003; Levy & Stockwell, 2006; Warschauer & Ware, 2006). Therefore, in addition to the research on how AWE software can be redesigned to address its inherent limitations, more classroom-based research on pedagogical practices with AWE is certainly needed. Further research may employ more rigorously-designed experimental investigations to validate the findings of this exploratory study or conduct longitudinal case studies to probe more deeply into teaching and learning processes with AWE and the long-term effects of using this technology for writing improvement. Moreover, how to measure the "effectiveness" of using AWE in writing classes is a crucial issue. The definition of effectiveness can vary depending on how teachers and students perceive the goals for the learning of writing. Students' language proficiency levels, cultural backgrounds, needs for learning to write in English, and learning contexts may all influence their perceptions of learning goals. AWE research needs to investigate learning outcomes from multifaceted perspectives and take into deliberate consideration learners' characteristics, needs, and goals for learning writing.

CONCLUSIONS

AWE software is still a developing technology that has not yet reached a fully mature stage. With many limitations inherent in the design of AWE technology, the particular pedagogical practices that accompany AWE use are critical. The relationship between technology and pedagogy, as Stockwell (2007) suggests, can be seen "as a symbiotic one, where they are mutually dependent upon each other, potentially to their benefit, but also potentially to their detriment" (p. 118). Therefore, while the design of AWE software needs to be improved to overcome its limitations and to keep up with current writing pedagogical theories, more effective pedagogical practices also need to be developed in order to augment the benefits as well as to minimize the problems that AWE technology may bring to students in their learning of writing.

It is important to emphasize that human facilitation should not be absent in AWE learning environments, because, first, writing is a social-communicative act involving the negotiation of meaning between writers and readers, and, second, writing instruction requires appropriate pedagogical designs capable of responding to contextual changes and learner needs, which cannot be accomplished by machines. AWE, like any existing technology used in the learning of writing, requires well thought-out pedagogical ideas and strategies for its implementation in writing classes. As Hyland (2003) claims, "writing cannot be developed by new tools but only by proper instruction, and this involves providing learners with

appropriate tasks and support" (p. 147). In writing instruction, assessing student writing and providing quality feedback are essential components, yet they are often complex in nature and challenging to writing teachers. When AWE is used, this job does not necessarily become easier for teachers but may become more complicated, requiring more technological competence in working with AWE and more careful pedagogical designs to integrate AWE into writing instruction. This study hopes to increase language teachers' critical awareness of how the design of AWE technology can affect its use and how pedagogical practices can affect the effectiveness of AWE for writing improvement. Writing teachers need to be fully aware of the limitations of AWE technology as well as students' learning needs and contexts in making decisions about how to maximize effective AWE use and to minimize undesirable outcomes.

NOTES

1. The term *automated essay scoring* (AES) is used more broadly and frequently than the term *automated writing evaluation* (AWE). However, we chose AWE instead of AES because the function of current essay grading technology is more than just scoring essays automatically. It also provides other forms of feedback as well as various writing assistance tools.

2. Examples of commercial AWE programs include *Criterion* developed by Educational Testing Service, *MY Access!* by Vantage Learning, *Intelligent Essay Assessor* by Pearson Education, and *WritePlacer* by College Board (see an overview of automatic essay scoring programs in Dikli, 2006 and Phillips, 2007).

ACKNOWLEDGMENTS

We would like to thank most sincerely the three anonymous reviewers as well as the two editors of this special issue, Joel Bloch and Richard Kern, for their valuable suggestions and comments on our manuscript.

ABOUT THE AUTHORS

Chi-Fen Emily Chen is Associate Professor and currently Chair of the Department of English at National Kaohsiung First University of Science and Technology, Taiwan. Her research interests include computer-assisted language learning, electronic literacy, second language writing, and discourse analysis.

Email: emchen@ccms.nkfust.edu.tw

Wei-Yuan Eugene Cheng earned his master degree at National Kaohsiung First University of Science and Technology in 2006. His research interests include second language writing, computer-assisted language learning, and vocabulary acquisition.

E-mail: weiyuan.cheng@msa.hinet.net

REFERENCES

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

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-30.

Beatty, K. (2003). *Teaching and researching computer-assisted language learning*. New York: Longman.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Chapelle, C. A. (2003). *English language learning and technology*. Amsterdam: John Benjamins.

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

Conference on College Composition and Communication (2006). *Writing assessment: A position statement*. Retrieved July 20, 2007, from http://www.ncte.org/cccc/resources/positions/123784.htm.

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. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23-40). Mahwah, NJ: Lawrence Erlbaum.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 1-35.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.). *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum.

Elliot, S., Darlington, K., & Mikulas, C. (2004). *But does it really work? A national study of MY Access! Effectiveness*. Paper presented at the National Council on Measurement in Education in San Diego, CA. April, 2004.

Ericsson, P. F. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 28-37). Logan, UT: Utah State University Press.

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

Flower, L. (1994). *The construction of negotiated meaning: A social cognitive theory of writing*. Carbondale, IL: Southern Illinois University Press.

Grabe, W., & Kaplan, R. (1996). *Theory and practice of writing*. Harlow: Longman.

Grimes, D., & Warschauer, M. (2006). *Automated essay scoring in the classroom*. Paper presented at the American Educational Research Association in San Francisco, California. April 7-11, 2006.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, *63*(4), 480-499.

Hyland, K. (2003). *Second language writing*. New York: Cambridge University Press.

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

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147-167). Mahwah, NJ: Lawrence Erlbaum.

Levy, M., & Stockwell, G. (2006). *CALL dimensions: Options and issues in computer-assisted language learning*. Mahwah, NJ: Lawrence Erlbaum.

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

Phillips, S. M. (2007). *Automated essay scoring: A literature review*. Kelowna, B. C., Canada: Society for the Advancement of Excellence in Education. Retrieved July 24, 2007, from http://www.saee.ca/pdfs/036.pdf.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stamping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*(2), 103-134.

Rothermel, B. A. (2006). Automated writing instruction: Computer-assisted or computer-driven pedagogies? In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 199-210). Logan, UT: Utah State University Press.

Shermis, M. D., & Burstein, J. (Eds.) (2003a). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum.

Shermis, M. D., & Burstein, J. (2003b). Introduction. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. xiii-xvi). Mahwah, NJ: Lawrence Erlbaum.

Stockwell, G. (2007). A review of technology choice for teaching language skills and areas in the CALL literature. *ReCALL, 19*(2), 105-120.

Vantage Learning. (2005). *MY Access!* (Version 5.0) [computer software]. Newtown, PA: Vantage Learning.

Vantage Learning. (2007). *MY Access! Efficacy Report***.** Newtown, PA: Vantage Learning. Retrieved December 2, 2007, from http://www.vantagelearning.com/school/research/myaccess.html.

Ware, P. (2005). Automated writing evaluation as a pedagogical tool for writing assessment. In A. Pandian, G. Chakravarthy, P. Kell, & S. Kaur (Eds.), *Strategies and practices for improving learning and literacy* (pp. 174-184). Selangor, Malaysia: Universiti Putra Malaysia Press.

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave MacMillan.

Whithaus, C. (2006). Always already: Automated essay scoring and grammar-checkers in college writing courses. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 166-176). Logan, UT: Utah State University Press.

Williamson, M. M. (2004). Validity of automated scoring: Prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment, 1*(2), 85–104.

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

Yeh, Y. L., & Yu, Y. T. (2004). Computerized feedback and bilingual concordancer for EFL college students' writing. *Proceedings of the 21st International Conference on English Teaching and Learning in the Republic of China* (pp. 35-48). Taichung, Taiwan: Chaoyang University of Technology.