The Impact of Online Automated Writing Evaluation: A Case Study from Dalian

Changhua Sun RICH
CTB/McGraw-Hill, USA

Abstract

Automated writing evaluation (AWE) technology is being adopted in classrooms in China and the USA. This paper presents the results of a case study of the application of AWE in Dalian, China. The quasi-experimental study was conducted in 2010 in Chinese middle school English language classrooms. An effect size of 0.30 was found in favor of the experimental group using AWE as an online formative assessment. Student survey responses and teacher observations are presented as convergent evidence to illustrate the impact AWE technology has on teachers, students, and student achievement in English writing.

Key words: Automated Writing Evaluation; Automated Essay Scoring; Online Formative Assessment; EFL classroom research in China

1. Introduction

1.1 Application of Automated Writing Evaluation (AWE)

In both China and the United States (USA), education is undergoing profound changes. A 21st century model of learning, educational reform, and development demands acceleration in utilizing technology to enable, motivate, and inspire students and teachers and to create student-centered learning (Gu, 2010; Office of Education Technology, 2010). English language writing is considered one of the most challenging subject areas among language teaching professionals. Research shows that the majority of high school teachers in the USA assign less than one paper per week across all subject areas (Page, 2003). Even when there is a writing assignment, feedback to the students is often limited. In China, it is not uncommon for high school teachers to provide only a letter grade and/or a few underlined spelling or grammar mistakes as feedback to English writing assignments. The
lack of writing assignments and the lack of sufficient feedback to student writing can be attributed partly to the heavy work load of grading papers. If a teacher has two classes of 60 to 80 students, one paper assignment per week may take the teacher an entire weekend to grade. Automated writing evaluation (AWE) technology offers great potential as a solution to assist English writing instruction.

1.2 History of AWE Applications
In the 1960s, the AWE tool Project Essay Grade (PEG) was created to provide a solution for teachers. Based on a sample of student writing, typical current AWE software such as Criterion, MyAccess, and Writing Roadmap 2.0 provide many desirable features as follows

- Assignment of a numerical score to reflect the overall quality of writing.
- Specific, diagnostic feedback to help the student improve writing along a number of key dimensions:
  - Grammar
  - Word choice
  - Sentence structure
  - Organization and development of ideas
  - Writing style

AWE software such as Writing Roadmap 2.0 also provides instructional features on grammar, vocabulary usage, spelling, and capitalization as well as classroom score reports and student writing improvement reports.

2. Literature Review

2.1 English Essay Grading Feedback in the Classroom
A research study by Zhou and Jia (2010) on the impact of teacher feedback on Chinese English as a foreign language (EFL) student writing shows an example of poor teacher feedback:

The composition would be better if it was developed as required!

Their study also provides exemplary teacher feedback that provides corrections on misspellings and word choice as well as summary comments:

You have written an interesting detailed story. I hope you can improve the vividness of your expressions.

The first type of teacher feedback was not considered helpful by students and was often ignored; the second type was preferred by students, as shown by the survey in the same study. Teachers’ formative feedback to student writing is critical for helping students improve their English writing.

Tang (2010) studied the Beiwai Online Institute tutors’ feedback to students who study English online. Her study points out the importance of formative assessment
to student learning and the potential use of online technology for effective feedback to student assignments. Black and Wiliam (1998), and Wiliam et al. (2004) argue that formative assessment can improve overall student performance. Gu (2010) demonstrated the use of multimedia and multimodal methods in EFL teaching. In the context of online EFL instruction, formative feedback to essay writing can be provided in a blended mode from both teachers and a computer program of AWE. Grimes and Warschauer (2010) described the benefits and challenges of using automated writing evaluation in schools in California. Tang and Rich (2011) describe a research study of the use of Writing Roadmap 2.0 in Chinese EFL classrooms. In the pilot study in Dalian, China, AWE was used as an online formative assessment in the classroom for students to practice writing and receive real-time or near real-time score feedback and reporting.

2.2 Credibility of AWE
Ever since its development in the 1960s, AWE has encountered criticism and skepticism. Page (2003) summarized the objections to computer grading into three types: humanistic, defensive, and construct objections. The humanist objection to AWE can be illustrated by some teachers’ doubts about using technology to replace their own functions. Critics using the defensive objection may say that a “bad-faith” essay can “fool” the computer software into giving a high score. The construct objection is represented by critics concerned that computers may measure the wrong things. All three types of objections are valuable in pointing out the technological limitations of AWE. In addition, some teachers pointed out that AWE may limit student creativity and cultivate mechanical writing styles. These are valid concerns that should be respected in the process of integrating technology in classrooms.

In the past decades, AWE has evolved due to the increased availability of classroom empirical feedback and the ongoing development of technology. Researchers and teachers are gaining understanding in utilizing AWE technology effectively to benefit both teachers and students (Grimes, 2008; Grimes & Warschauer, 2010; Chen & Cheng, 2008; Tang & Rich, 2011). The increased understanding and new developments may lead to the following responses to critics who have questions about bad-faith essays, technology replacing teachers, and how computers grade essays.

2.3 Questions About Bad-Faith Essays
Shermis (2007) demonstrated at a U.S. National Large Scale Assessment Conference that it took a good writer to produce a bad essay that could fool a computer scoring engine. Given the prompt “Christopher Columbus”, the following essay would have received a high score from the AWE software, but it takes time and talent to create such an essay.

Queen America sailed to Santa Maria with 1492 ships. Her husband, King Columbus, looked to the Indian explorer, Nina Pinta, to find vast wealth on the beaches of Isabella, but would settle for spices from the continent of Ferdinand.

Other types of bad-faith essays now can be detected by AWE software. AWE such as Writing Roadmap 2.0 has a flagging function built in to catch some types of bad-faith
The Impact of Online Automated Writing Evaluation: A Case Study from Dalian essays (Rich, Harrington, Kim & West, 2008). For example, the following essay from a 2010 classroom pilot study in Beijing was flagged for teacher review.

My Lingling teuthioerjepogpfjkgmdflkopghkf, hlhl,nn ; ; , jlk,jh;jg ; ;, f’hg’gh;lk,l,ljlhk,;ljh,k;l,k,r,ul;t;j;l,pl,jl,hg[lj,[o]u,l,h;jlk;mj:lg,l,,lkjoghk hkg 675uy8trpjpofj-9jfdkljsifjklshauyinsfjksbuvhsnvjdbjkygufshvdjksb.vugdksbd;kjyghduhvcvguguguthi ohrddoqfjkhokrhlmkshiuajbcxbxh bchxdbcxhbxh hvdfdyukjgfdlk[ohgvbm hdjkgdfffhjkhjhhkhugyftfyyddtdtdhejwflhwflwhigugj.

The AWE system also flagged essays for human review when the engine could not reliably predict a score. Teachers need to be aware that students can copy text from internet sources, which can be a problem for classroom practice or homework. Note that this type of copying is not just limited to writing using AWE technology.

2.4 Questions About AWE Replacing Teachers
During the August 2010 China National Conference in Foreign Language Education, teachers raised questions such as how the computer (AWE) evaluates Chinglish words or whether the computer could teach the wrong types of writing. AWE is a technology tool intended to assist teachers and students; it cannot and should not replace teachers’ knowledge and their critical role in English language education. Depending on how active or inactive teachers are in guiding students, AWE can be an effective technology or a not-so-effective technology in the classroom. In our study, many teachers found that using AWE in their classrooms reduced the grading burden and provided students with more effective practice of English writing (Tang & Rich, 2011). On the other hand, some teachers felt frustrated with the imperfections of the technology and also felt burdened to have to learn a new method of instruction. Effective use of technology in an online- and offline-blended learning environment in the 21st century classroom will continue to be an important ongoing area of research to understand changes in teachers and students and classroom transformation.

2.5 Questions About How the Computer Grades an Essay
Other common questions about AWE are about how the computer scores an essay and the accuracy of the computer scores. Literature on AWE (Burstein, 2003; Dikli, 2006) provided some detailed answers to these questions by addressing the question of what the conceptual dimensions of writing are and how they are measured. The AWE program identifies and extracts the relevant features from written text that correlate with widely recognized writing qualities using natural language processing (NLP) and statistical feature selection techniques. These features of the written text can then be used to build computer algorithms to grade essays. Two main technical approaches used in building computer scoring models are NLP and latent semantic analysis (LSA). The NLP approach works by identifying and extracting the relevant features of writing quality such as rates of errors in grammar, sentence structure, and semantics. The LSA approach analyzes words and
passages in a document, using a document by term matrix to describe the co-occurrences of the words in the document. In combination with NLP and/or LSA, statistical methods such as nearest neighbor classifier and/or linear regression are often used in AWE software to predict essay scores. In summary, most of the work in computerized grading involves text classification, which focuses on searching and not on interpretation of the meaning of the essays.

AWE software is built to mimic human expert judgments. Using machine learning techniques and statistical models, the AWE scoring engine is calibrated to predict scores that are a best representation of the true scores given by expert human markers.

For example, in 2009-2010, West Virginia, USA, students in grades 3-11 took sets of writing prompts online in a testing environment, and then a team of researchers from CTB/McGraw-Hill conducted a systematic AWE scoring training and validation for the computer assessment of the writing (CTB/McGraw-Hill, 2010). The computer score validation results of the Writing Roadmap 2.0 scoring engine show the rater agreement for comparison of the human and computer engine scores. Higher correlation is found consistently for the comparison between engine scores and human scores (0.83 to 0.89) than for human-to-human scores (0.64-0.80). Similar results were reported in other AWE studies (Shermis & Burstein, 2003).

2.6 Impact of a Formative Assessment Application of AWE in West Virginia
This section examines the experience the State of West Virginia in the USA had using automated writing software in both summative and formative online writing assessments. Because the West Virginia experience offers large-scale assessment of AWE user experiences across five years, it is possible to present systematic evidence of the AWE computer engine score accuracy presented in the previous section.

In 2007, the State of West Virginia and CTB/McGraw-Hill jointly investigated the relationship between the year-end West Virginia Online Writing Assessment and the formative assessment use of Writing Roadmap 2.0, an AWE tool developed by CTB/McGraw-Hill. The Rich et al. (2008) study found positive relationships between online classroom practice using Writing Roadmap 2.0 and West Virginia online writing scores for both grades 7 and 10. They used the quasi-experimental analysis approach conditioning Writing Roadmap users versus non-users in the classroom, matching samples by high-stakes English Language Arts performance levels. The study examined the differential performance based on community, rural versus urban schools, gender, and ethnicity to provide insight into fairness issues in the use of AWE in West Virginia classrooms. Positive score gains on the state writing test were found for students who used Writing Roadmap to practice writing in comparison to students who did not use the automated writing tool. The results were consistent and statistically significant for rural, urban, female, male, and white subgroups of students. One inference made from the study is that AWE technology could have an important educational impact if it were successfully integrated into teaching and learning in West Virginia English Language Arts writing classrooms, where students could regularly practice writing and receive real-time feedback.

In 2010, after four years of integrating automated writing technology in classrooms
and summative assessments, the West Virginia Department of Education (Office of Assessment, Accountability and Research) published a Research Brief (White et al., 2010). The study investigated the impact of AWE (Writing Roadmap 2.0) on WESTEST 2 Online Writing Assessment scores. The study randomly selected students who had completed five or more Writing Roadmap essay assignments during the 2008-2009 academic school year. A comparison group who did not use Writing Roadmap as a formative assessment was selected using matching variables of grade level, geographic location, and social economic status. The final sample in the study included 8,430 randomly selected students in the treatment group and 8,430 students in the comparison group. The summative online writing test score means are compared for the treatment and comparison groups. Independent sample t tests were conducted and statistically significant mean differences were found for each sampled grade level. Cohen’s $d$ was calculated to estimate the effect size, which ranged from 0.17 for grade 9 to 0.59 for grade 4 (White, Hixson, D’Brot & Perdue, 2010).

2.7 From the West Virginia Study to the Dalian Research Pilot Study
The caveat of the West Virginia study was that the results should be validated via a longitudinal design and a controlled experiment. The methods used in the West Virginia studies have been limited to the quantitative analysis of test scores. Because no classroom studies were conducted to understand the reason for teachers’ choice whether or not to use AWE in their classrooms, many questions concerning AWE application as a formative assessment tool remain unanswered. The West Virginia case study findings inspired the current research study in Dalian, China. To investigate the process of AWE integration in the classroom, we used a method of a controlled experimental design in combination with student surveys and teacher observations.

3. Pilot Research Study in Dalian, China: Online Automated English Writing Assessment in the Middle School EFL Classroom

Research has shown that formative assessments can raise students’ proficiencies (Black & Wiliam, 1998; Wiliam et al., 2004). The West Virginia experience in the past five years provided evidence of a positive impact from the use of AWE in the classroom as a formative assessment tool. From March to July of 2010, researchers from the Beijing Foreign Studies University Online Institute of Education and CTB/McGraw-Hill’s Research Department jointly conducted a pilot study in Dalian middle schools and elementary schools. The pilot study extended the experiences in English Language Arts writing classrooms in West Virginia, USA, to experiences in Chinese EFL classrooms. This section presents results from the Dalian pilot study. The Dalian study is a pilot of the three-year research project titled “Applications of Online Automated Writing Evaluation as Formative Assessment Tool in China EFL Classrooms” 2009-2012 (Grant Number GFA097005).

3.1 Research Questions and Methods
Building upon the AWE studies with West Virginia, the current study focused its attention on the classroom process when AWE is used as a technology-enhanced formative assessment.
The study used a multifaceted approach with student surveys and classroom observations in addition to quasi-experimentation of control and experimental groups’ pre- and posttest writing score analysis. The qualitative and quantitative analyses were combined to answer the research questions on AWE as a formative assessment intervention in the Chinese EFL classroom. Preliminary findings are presented from student surveys, teacher journals, and experimental data analysis from the pilot study in a Dalian middle school.

The initial propositions for the study of AWE application in the classrooms were to empower teachers and motivate students in the process of integrating AWE technology in English language classrooms. The goal of the main research project is to systematically understand the classroom teacher and student experience with AWE intervention. During the pilot study in Dalian, we aimed to understand both the research data collection process and the analysis of initial evidence to answer three related research questions:

1. When AWE technology is introduced in the classroom, what happens to the existing teaching and learning strategies?

2. When feedback to student writing comes from a computer rather than a teacher, what is the impact on students and teachers?

3. What is the impact of the integration of AWE software on students’ performance?

The first question explores the classroom experience during AWE integration, the second question focuses on the effect of computer grading, and the last question examines the quantifiable gain in student scores from the use of AWE in an English writing classroom. A major strength of the case study design is that data were collected from multiple sources using multiple methods. Another factor of the research is that teachers are involved as researchers rather than as the object of the study. The study triangulated data collection by using student surveys, teacher journals, and experiments. The multifaceted approach allows for convergence of multiple sources of evidence and non-convergence of evidence (Yin, 1994). The data collection sources in relationship to the research objectives are shown in Table 1.

Table 1. Data Collection Sources in Relationship to the Research Objectives, Dalian Pilot Study from March to July 2010

<table>
<thead>
<tr>
<th>From Students</th>
<th>From Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavior Attitude Perceptions</td>
<td>Classroom Transformation Teaching Transformation</td>
</tr>
<tr>
<td>Teacher Impact</td>
<td>Teacher survey</td>
</tr>
<tr>
<td>Teacher interview</td>
<td></td>
</tr>
<tr>
<td>Teacher journal</td>
<td>Student writing records</td>
</tr>
<tr>
<td>Student diary</td>
<td>Teacher journal</td>
</tr>
<tr>
<td>Student Impact</td>
<td>Student survey</td>
</tr>
<tr>
<td>Student interview</td>
<td></td>
</tr>
<tr>
<td>Student diary</td>
<td></td>
</tr>
<tr>
<td>Pre- and Post-Writing test</td>
<td></td>
</tr>
<tr>
<td>External test</td>
<td></td>
</tr>
</tbody>
</table>
3.2 Data Sources
Nine elementary schools, three middle schools from Dalian, one vocational college, and two universities in Beijing took part in the research pilot study. The data presented in this paper were collected from participants in one of the Dalian middle schools (221 students and three teachers), referred to as “Middle School A”. The data included the observations of the three teachers, who each taught two classes in Middle School A, and student responses to the post-test survey. One class from each teacher was designated as a control group and the other class as an experimental group. The AWE generic engine scores for pre- and posttest writing were used to conduct the statistical analyses.

3.3 Student Surveys
Student surveys were administered online before and after the use of Writing Roadmap in the experimental classrooms.

Table 2 summarizes the student post-test survey responses to six questions concerning experiences in using AWE tools in the classroom. A total of 51 students responded to the survey questions. In general, students had a positive experience using Writing Roadmap in the classroom: 92% agreed or strongly agreed with the statement “I like using Writing Roadmap to practice writing.” Sixty-eight percent of the students agreed they wrote more after using Writing Roadmap. Seventy percent felt they revised essays more after using Writing Roadmap. Three out of four students felt they gained more confidence in their English writing after participating in the experiment. More than 90% of the students responded that they revised at least once before submitting their essays, and 61% reported revising their essays three or more times. As to the four instructional tools available in the Writing Roadmap system, 37% chose Tutor as most helpful, 29% chose Thesaurus, 20% chose Hint, and 14% chose Grammar Tree.

Table 2. Experimental Group Post-Test Survey Responses to Questions Concerning the Writing Roadmap (WRM) AWE

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>I Don't Know</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I like using WRM to practice writing.</td>
<td>2 (4%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>25 (49%)</td>
<td>22 (43%)</td>
</tr>
<tr>
<td>I write more after using WRM.</td>
<td>0 (0%)</td>
<td>13 (25%)</td>
<td>3 (6%)</td>
<td>21 (41%)</td>
<td>14 (27%)</td>
</tr>
<tr>
<td>I revise my essays more after using WRM.</td>
<td>1 (2%)</td>
<td>10 (20%)</td>
<td>4 (8%)</td>
<td>23 (45%)</td>
<td>13 (25%)</td>
</tr>
<tr>
<td>I feel more confident about my writing.</td>
<td>1 (2%)</td>
<td>9 (18%)</td>
<td>2 (4%)</td>
<td>22 (43%)</td>
<td>17 (33%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How many times did you revise your essays in WRM before submission?</th>
<th>Never</th>
<th>1-2 Times</th>
<th>3-4 Times</th>
<th>5 or More Times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 (8%)</td>
<td>16 (31%)</td>
<td>23 (45%)</td>
<td>8 (16%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instructional Tools</th>
<th>Grammar Tree</th>
<th>Hint</th>
<th>Thesaurus</th>
<th>Tutor</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think the most helpful tool in WRM is</td>
<td>7 (14%)</td>
<td>10 (20%)</td>
<td>15 (29%)</td>
<td>19 (37%)</td>
</tr>
</tbody>
</table>
3.4 Teacher Experiences
In addition to the student survey, teacher journals were collected to document teacher classroom observations and experiences. The biweekly journals were written based on the following template (Tang et al., 2009).

1. Students’ classroom performance (whether students preview the readers before the class; whether they are active during the reading class; whether they like to practice writing in Writing Roadmap; their main writing problems)
2. Student experiences with the class (students’ experiences through observations)
3. Teacher experiences and evaluation
4. Students’ use of Writing Roadmap (students’ attitudes toward Writing Roadmap, what types of students become excited or frustrated by computer-generated feedback)

From the ten weeks of the study in spring 2010, the three teachers in Middle School A submitted nine journal entries. Teachers’ writings were the most valuable and fascinating record of students’ and teachers’ experiences, feelings, problems, and successes in the classroom. Teachers’ journal entries allowed researchers to have a look inside the “black box”. For example, Teacher A wrote (translated from Chinese):

1. Almost all students were very positive towards Writing Roadmap. The first time we went to the computer lab to practice writing, the children were very excited. Their interests in writing English and spirits were soaring; I had never seen children love writing so much.
2. When they saw the immediate score feedback from the computer, they were very happy. They kept asking me, “Teacher, can I revise first before submitting?” “Teacher, I felt that I wrote a good essay, how come I got so few points?” Looking at students’ smiling rosy faces, I felt that this system was wonderful. It stimulated student’s interests in writing. In the meantime, it freed teachers to a very large extent. No need to do mechanical and repeated work.
3. After the first class, 45 of the 48 students finished their writing. But after a few more classes, we found several problems. First, the user interface (UI) was in English only; this was especially difficult for some students who could not understand the UI. The second problem is that there may be too much pressure for students to write in class. But if I assign the writing as homework, only the best students would be able to write the essays on their own. Most students could not write by themselves. The third problem was that the system gave good feedback to students who wrote good essays, while other times some students did not write well and still received high scores from the system.
4. In general, this system can benefit students in improving their writing ability, and also motivate students to learn. It is worthy of adoption. I am very happy to be able to participate in this research. I believe that if we continue, student’s proficiency in writing will be greatly improved.

Integration of AWE in the classroom challenged teachers and students to apply their computer skills. When the first time Writing Roadmap was used in the classrooms, it was quite an unusual and challenging experience for both teachers and students. After the second and third classes, the students and teachers gained confidence in using the online writing and automatic feedback system.
The students were enthusiastic and willing to try writing English online. For many of them, this class was their first experience writing English online. Some students from low-income or migrant workers’ families did not have computers at home. These students had trouble typing English and it took them longer to write than others. The teacher helped these students by pairing them up with other students to learn the functionalities of the online system.

The teachers noted that the students were motivated by instant score feedback. They had been concerned that students might only focus on scores but were pleased to see them motivated in the class.

The students liked the online writing where they could work by themselves, although some of them were unfamiliar with using computers and felt uneasy about the online writing class.

Some students asked teachers how to revise their writing. This had rarely happened before the online writing class. Students tended to muddle through the homework assignments, but now in the online writing class, they expected to see their scores improve and were eager to learn how to make it happen.

Several issues were identified by Middle School A teachers in the journals. These issues showed the gradual deepening of the technology integration even in a short period of time. Some teachers began finding solutions to the issues they identified, although others had difficulty adjusting themselves to the challenge. Their experience-sharing meetings, e-mail exchanges, and workshops played an important role in finding solutions.

A list follows some of the challenges students and teachers faced.

- For some students’ essays, scores did not seem to be fair.
- Some students had difficulty with the English user interface and reports.
- Some students figured out the log-in and password for the class and got into other students’ accounts to submit nonsense essays.
- It was difficult to sustain students’ interests if their scores improved slowly or showed no improvement.
- Some students felt that the score feedback was stimulating but teachers were concerned about students only paying attention to their scores.
- There was not enough time in one class period to write, revise, and assess a piece of writing.
- Students needed to read more and master wider vocabulary in order to write better.

3.5 Quasi-Experimentation

The quasi-experimental design (Cook & Campbell, 1979) was implemented in the study. The control and experimental groups were not randomly assigned due to practical limitations. To reduce interaction and confounding factors in the experimental design, the study chose natural classes rather than randomized sampling and requested participating schools to select “comparable” teachers and classes as experiment and control groups. This decision was consistent with educational research in examining the average growth of the two groups.

Two writing prompts were randomly assigned to students within a classroom. The A-B prompts crossover design was used to control for the prompt difficulty factor (Friedman & Sunder, 1994): Students who took prompt A during the pretest took prompt...
B in the posttest. In this way, the mean differences between the control and experimental groups would not be confounded by the level of difficulty of the prompts. The students’ pretest and posttest writing prompts were administered in the Writing Roadmap 2.0 online system and were scored automatically using the generic engine-scoring algorithm. Using automated scoring in the research had the benefit of removing rater effects since the computer scoring algorithm was identical for the essays submitted to the system.

3.5.1 Pretest and posttest data analysis results

Tables 3 through 7 present the results of data analysis for Middle School A. Table 3 shows descriptive statistics and the effect size for comparison between the experimental and control groups. The experimental group’s mean scores were higher than the control group’s mean scores in both pre- and posttests. The gain score, computed by the posttest score minus the pretest score for each student, shows that the control group had a small negative gain. In contrast, the experimental group had a positive posttest score gain. The control-experimental group comparison shows a standardized effect size of 0.30.

Table 4 presents the results of inferential statistics of the mean score differences reported in Table 3. The null hypothesis for the independent samples t test was zero effect of the AWE treatment on student writing outcomes. Independent samples t test shows that the mean score difference was statistically significant at the 0.01 and 0.05 level for the posttest score and gain score comparison across the two groups, indicating the writing score gain for the experimental group was unlikely to be found by chance.

In comparison, the mean difference in the pretest found between the groups was not statistically significant at 0.05. Thus the control and the experimental group pretest scores at the beginning of the intervention were relatively comparable.

The statistically significant gains for the experimental group offered converging evidence for the positive responses from the student surveys and teachers’ observations. The writing performance improvement of the experimental group was encouraging considering the short duration of the 7-10 week experiment in the Middle School A classrooms.

Table 3. Pretest and Posttest Online Writing Descriptive Statistics from Middle School A

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>Control</td>
<td>96</td>
<td>2.45</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>Experimental</td>
<td>125</td>
<td>2.58</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Mean Difference</td>
<td></td>
<td></td>
<td>-0.13</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Posttest</td>
<td>Control</td>
<td>96</td>
<td>2.39</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Posttest</td>
<td>Experimental</td>
<td>125</td>
<td>2.70</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Mean Difference</td>
<td></td>
<td></td>
<td>-0.32</td>
<td>0.59</td>
<td>0.54</td>
</tr>
<tr>
<td>Gain score</td>
<td>Control</td>
<td>96</td>
<td>-0.07</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>Gain score</td>
<td>Experimental</td>
<td>125</td>
<td>0.12</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>Mean Difference</td>
<td></td>
<td></td>
<td>-0.19</td>
<td>0.64</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Mean difference is computed as control group mean score minus experimental group mean score.
Table 4. Middle School A Across Groups Test Statistics (Control – Experimental)

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>219</td>
<td>-1.84</td>
<td>0.0673</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posttest</td>
<td>219</td>
<td>-3.93</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain score</td>
<td>219</td>
<td>-2.17</td>
<td>0.0311</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5 presents descriptive statistics, and Tables 6 and 7 show the results of the paired-sample t tests for mean score comparison within each group between the pretest and posttest. The null hypothesis was zero gain between the pre- and posttest scores. The t statistic was large enough to reject the null hypothesis for the experimental group, but it failed to reject the null hypothesis for the control group. The statistically significant result shows that the score gain of the experimental group was unlikely to be found by chance. The within-group t test analysis results supported the findings from the student survey responses and teacher observations that the students were motivated to write, write more, and revise more.

Table 5. Within-Group Gain Score Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-Pretest</td>
<td>Control</td>
<td>96</td>
<td>-0.07</td>
<td>0.69</td>
</tr>
<tr>
<td>Post-Pretest</td>
<td>Experimental</td>
<td>125</td>
<td>0.12</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 6. Control Group Gain Score Test Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>t value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-Pre</td>
<td>95</td>
<td>-0.95</td>
<td>0.346</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7. Experimental Group Gain Score Test Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>t value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-Pre</td>
<td>124</td>
<td>2.28</td>
<td>0.0243</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Discussion

The State of West Virginia has integrated AWE technology in the classroom since 2006. It has also used AWE technology in large-scale assessment in the summative year-end testing. The 2010 West Virginia study shows the empirical evidence of raising student English writing proficiency in every grade from elementary school to high school as the result of practicing writing online with instant formative feedback. The 2010 West Virginia study not only replicated the positive impact found in the 2008 study but also extended it from grades 7 and 10 to every grade from grade 4 through grade 11. Since 2007 AWE was
introduced in Chinese classrooms. The current research attempted to find what works and what does not work when AWE is introduced as a formative assessment tool in the Chinese EFL classroom.

4.1 Lessons Learned from the Dalian Case Study

During 2010, as a pilot study of the Ministry of Education-funded R&D project, we conducted a study of the integration of AWE as a formative assessment tool in middle schools in Dalian. Student survey responses and teacher journals from the spring 2010 Dalian study revealed the great potential for AWE technology’s integration into Chinese classrooms for teaching English. The quasi-experiment results in the single case of a middle school showed statistically significant writing score gain for the experimental group. This finding is encouraging and has educational significance. Students were motivated to write independently and were able to improve their English writing ability within a short period of time.

The AWE software, like any other tool, can be used appropriately or inappropriately in the classroom or in a high-stakes test setting. Some issues that arose in the pilot study will require our attention. The instant feedback from AWE was not sufficient for some students with beginning computer skills or in a very beginning level of English writing ability. These students require more attention from teachers for correcting mechanical mistakes as well as cultivating overall writing skills. Other students’ motivation may wear off after the initial excitement of receiving real-time computer score feedback. English writing is hard work that still requires both teachers’ and students’ sustained efforts. Students’ and teachers’ interaction with AWE formative feedback is a process of learning from the computer system and adjusting expectations.

In the Dalian study, the teachers were action researchers. They reflected on the challenges during the study and solved many problems creatively in their teaching. For example, some teachers used the computer scoring rubric to teach and asked their students to co-teach and learn how the computer arrived at scores (Tian, 2010). Other teachers and some students were frustrated by instances where they felt the computer scores for essays were not fair. Similar experiences were found in Grime and Warschauer (2010) study of southern California schools. It was observed that technology-enhanced formative assessment integration in the classroom means a change in teacher behavior, which is never a simple matter. Beatty et al. (2008), Beatty and Gerace (2009), and Beatty (2011) described the process of teachers learning technology-enhanced formative assessment, and Beatty pointed out that “hidden perspectives underlie most or all of the common difficulties teachers report when learning to use clickers”.

In this case, providing ongoing support to teachers by periodic experience-sharing sessions and solid technical support is crucial for deepening the technology integration in the classroom. In addition, future studies should add a design point for teachers to motivate the control group students in the study to prevent a negative impact on student learning.

From teacher and student feedback, the technology provider also gained valuable understanding of user needs for technology improvement. For example, based on feedback provided by the participants in the study, the user interface of Writing Roadmap software
now has dual-language options for Chinese and English.

4.2 Key Findings of the Dalian Case Study

4.2.1 When AWE technology was introduced to the classroom, student independent learning occurred.
A total of 51 students from the Dalian Middle School A answered all six questions on the student survey. Ninety-two percent of the students liked using Writing Roadmap to practice writing, 68% said they wrote more, and 70% said they revised more after using Writing Roadmap. Three out of four students felt they had gained more confidence in their English writing ability.

More than 90% of the students responded that they revised their essays at least once before submitting them, and 61% revised their essays three or more times.

4.2.2 When the feedback on student writing came from a computer rather than a teacher, teachers experienced transformation in the classroom.
The teachers’ journals show that students were motivated by the instant scoring feedback. They observed that students liked the online writing where they could work by themselves. Some students asked the teachers how to revise their writing using the feedback given by the online software, and students enjoyed reading stories and then writing about them.

The teachers’ journals noted that students paid more attention to the computer score feedback than to teachers’ grading on the homework assignments. The students expected to see improvements in their scores. The journals also noted that computer scoring freed the teachers to a large extent from the mechanical and repetitious work of grading essays. Teachers began focusing on collaborative learning and in-depth instruction of writing.

4.2.3 Significant writing score gains were found for the Middle School A experimental group that integrated the AWE software.
Pre- and posttests using Writing Roadmap allowed score comparisons for Dalian Middle School A before and after the study. The score data showed that the experimental group had a higher mean score in comparison with the control group with an effect size of 0.30. The effect size was comparable to the mean effect size of 0.32 found by Wiliam et al. (2004) on student achievement with formative assessment intervention. It was also comparable to the average effect sizes found from students’ test performance in reading and in mathematics (Coe, 2002).

4.3 Limitations of the Study
The present study had several limitations. First, it was conducted during a short time period of 7 to 10 weeks. A longer period of study will be necessary to provide more in-depth understanding of AWE implementation issues. Second, the data analyzed included only schools with comparable pre- and posttest conditions. The data from some participating schools were not included because these schools had non-comparable testing conditions. For example, several elementary school participants did not conduct
posttests until the summer vacation had started, and the posttests were done at home by many students. As a result, the data for these schools could not be compared. Third, the control group students’ motivation was important for a valid comparison between the experimental and control groups. In the present study, the control group had a lower score mean for the posttest. It was speculated that the reason for this might be the lack of motivation for the control group to take a writing test using AWE. Future studies would need to design methods to maintain the interest in writing for both the control and experimental groups.

5. Conclusions and Remarks

The Outline of China’s Medium to Long Term National Educational Reform and Development (2010-2020) calls for rapid information technology application in education, improvement of teachers’ application of technology for effective teaching, and students’ application of technology for independent learning. This paper explored issues in the classroom integration of AWE software as a formative assessment tool. It first examined the existing studies and application of AWE in schools in West Virginia, USA, and then described the study in a middle school in Dalian, China. These case studies in the USA and China have provided useful and practical ways to help teachers and students effectively integrate online technology in English language education in China.

John Dewey, a 20th century philosopher and education reformer, advocated using research and theory to transform uninformed practice into intelligent practice. From March to July 2010, the Beijing Foreign Studies University Online Institute and CTB/McGraw-Hill researchers worked closely together with school teachers to gain first-hand experience in applying AWE technology in English language classrooms. The research findings offer relevant and important insights for future research of the integration of AWE into the classroom. Teachers who participated in the research were action-researchers. The integration of teaching with research could become an ongoing practice for online technology-based formative assessment. The research-based practice might not only empower teachers and students in their own classroom teaching and learning but might also have a broader impact on teacher practice in global English language classrooms.

Acknowledgements

The Dalian study was conducted under a three-year research grant funded by the Ministry of Education, People’s Republic of China 2009 Key National Educational Science Research Project (Project Number GFA097005). The project is titled “Using Automated Writing Assessment in China EFL Classrooms”. The Institute of Online Education of Beijing Foreign Studies University and the CTB/McGraw-Hill 2010 R&D Program provided additional funding for the study.

The author wishes to thank the Dalian Education Bureau, the participating schools, teachers and students for supporting and participating in the pilot research study and wishes to thank the
West Virginia Department of Education Office of Assessment for giving permission to reprint figures from the 2010 Research Brief and the 2009 WESTEST 2 Technical Report.

References


(Copy editing: Steve Samuels)