Technology-enhanced English Language Writing Assessment in the Classroom*

TANG Jinlan
Beijing Foreign Studies University
Changhua Sun RICH & WANG Yihong
CTB/McGraw-Hill

Abstract
Technology-enhanced formative assessment integration in classrooms transforms traditional teaching, learning and assessment. This paper introduces a pilot study as part of a three-year research project funded by China’s Ministry of Education on the use of technology-enhanced formative assessment in the classroom. Initial pilot-study results from nine elementary schools and three junior high schools showed positive writing gains with the experimental group and revealed that students became more motivated and autonomous and tended to write and revise more, and teachers switched their roles from spoon-feeding to that of a supporter and facilitator.

Key words: technology-enhanced assessment; online writing assessment; English language teaching; Writing Roadmap

1. Introduction
This paper introduces a pilot study as part of a three-year research grant funded by the Chinese Ministry of Education 2009 Key National Education Sciences Planning research project. The project “Using Automated Writing Assessment in China’s EFL Classrooms” intends to investigate the impact of the use of automated writing assessment on the

* This study is a part of the Ministry of Education Key National Education Sciences Planning research project (2009-2012) (Grant number GFA097005). It is presented at 2011 American Educational Research Association Annual Meeting, April 8-12, New Orleans, USA.
learning process and outcome in China’s EFL classrooms. As a part of the project, this study examines its application in the primary and secondary English class.

2. Literature Review

The history of Automated Essay Scoring (AES) or Automated Writing Evaluation (AWE) can be traced back to the 1960s in America with the development of Page Essay Grade (PEG), a program that used multiple regression analysis of measurable features of text, such as essay length and average sentence length, to build a scoring model based on a corpus of essays previously graded by hand (Page, 2003). With the widespread use of computers and the Internet, the development of Artificial intelligence and natural language processing technologies, and the global emphasis on writing instruction, the 1990s witnessed the emergence of more AWE systems such as E-Rater, Intellimetric and Intelligent Essay Assessor (for a more in-depth overview of AWE, see Warschauer & Ware, 2006). In the past, the main use of these AWE systems was in the grading of the standardized tests. More recently, development in AWE software made its use as a classroom instructional tool possible. Each of the main scoring engines discussed has been incorporated into one or more programs directed at classroom use. These AWE programs such as Criterion, My Access! and Writing Roadmap combine the scoring engine, an editing tool (offering grammar, spelling, and mechanical feedback) and some support resources (such as dictionaries and thesauruses) and are both an assessment and a learning tool.

Research on AWE mainly falls into three types. The first is psychometric studies on the validity of the software, with the aim of examining whether machine scoring is consistent with human scoring. These studies have been mostly conducted by developers and have yielded very positive findings, with machine and human scoring correlations between .80 and .85 (see Rich, Harrington, Kim & West, 2008; Wang & Brown, 2007). The second type of research concentrates on the learning outcome and explores whether AWE brings writing gains on students’ performance (e.g., Elliot & Mikulas, 2004; Rich, et al., 2008; Shermis, Burstein & Bliss, 2004; “Vantage Learning”, 2007; White, Hixson, D’Brot & Perdue, 2010). However, it has been argued that due to certain weaknesses in the research design of some studies such as the lack of a control group or the pre-experiment test data, it is difficult to claim the impact that AWE exerted on writing in some studies (Tang & Wu, 2011; Warschauer & Ware, 2006).

The third type of research has focused on the process of using AWE in the classroom and has yielded mixed findings. Attali (2004) examined the use of Criterion with the sixth and 12th year American students in 2002-2003. Among 33,171 essays, 71% of them were submitted only once for scores, indicating the limited use of the system. For those essays submitted to the system more than once, scores rose from 3.7 on the first submission to 4.2 on the final submission (on a 6-point scale), however, major revisions fell mainly on spelling and vocabulary, not on essay structure.

Hoon (2006) reported a group of 34 students using My Access! for one semester
and the research results indicated that nearly 80% of the students held that the system had helped them improve their writing, and more than 90% indicated that they used the feedback provided to improve their writing. The instructor considered the immediacy of feedback positive with scores and the provision of various writing support tools and online resources. Contrary to Hoon’s (2006) study, AWE was not deemed very positive by the students in Chen and Cheng (2008). Despite this, they discovered that the use of AWE was perceived as more effective in the early drafting and revising processes, followed by teacher and peer feedback during the later processes. Teachers’ attitudes toward AWE use and their own technology skills might also impact the efficacy of AWE.

In the study by Warschauer and Grimes (2008), though AWE seemed to relieve teachers of the essay marking workload and motivate students to write more, the use of AWE seemed not to result in more revision or greater attention to content and organization, and this reinforced earlier findings in Attali (2004). It was also discovered that teacher perceptions of the writing pedagogy and the social-economic status of the school determined the use of the software. In his study of five schools using AWE software, Grimes (2008) discovered five paradoxes, which pinpointed the fact that teacher and student enthusiasm for AES would not necessarily lead to its more frequent use in the classroom and students making more essay revisions, and a gap existed between perception and practice.

It may be argued that one main implication drawn from the previous research is that the successful use of AWE in the classroom calls for a paradigm shift from the perspectives of teachers, learners and administrators. Teacher beliefs, attitudes and technology use, and administrator support are of crucial importance.

The previous studies were mostly located in American schools. Most of the process-based studies only revealed problems, with no experimental studies designed to tackle the problems. Moreover, the research was conducted by researchers from outside the school, and no classroom teachers were involved in the research.

3. Current Research Project

3.1 Research questions

The automated writing assessment tool investigated in this study is Writing Roadmap (WRM) from CTB/McGraw-Hill. It is an online essay-scoring tool providing writing practice and automatic feedback for six writing dimensions and a holistic score for student essays, available for Grades 3-12, college and adult students. Two studies on the use of WRM in West Virginian schools indicated positive gains for students who used WRM in practicing writing versus those who did not (Rich, et al., 2008; White, et al., 2010). However, these studies did not study the process of how WRM was used by teachers and students in the classroom.

WRM was first introduced into China in 2007. After several pilot studies of WRM with Chinese students in two cities, Wang (2008) found that teacher attitude and administrative support are crucial to the successful integration of WRM into teaching.
Another interesting finding from Wang (2008) was that Chinese EFL students, particularly those at the school level, lacked sufficient English input, therefore experienced problems in writing. In the light of this fact, she proposed an integrated reading and writing course to resolve the problem.

Drawing on implications from the previous literature on AWE (e.g., Grimes, 2008; Rich, et al., 2008; Wang, 2008; Warschauer & Grimes, 2008; White, et al., 2010), the current study concentrated on investigating both the process of integrating AWE in the classroom and the impact on teaching and learning (Tang & Wu, 2011), because the process-oriented research would yield reasons regarding how the integration of AWE affected the learning outcome. In the light of this understanding, the study combined action research with experiments in project design. Target classrooms were Chinese EFL grades 5-12. The main objectives were to understand the process of integrating online formative assessment into the classroom. Specifically it aimed to answer the following questions:

1. What are student attitudes toward and experiences with the use of AWE?
2. How do teachers view the efficacy of AWE in Chinese EFL classrooms?
3. How does the use of AWE impact student writing performance in Chinese EFL classrooms?

3.2 Research Context
The study was situated in nine primary and three junior high schools in a coastal city in China. Students in this study started their formal English language learning in Year 1 of the primary school. Usually there were five English classes per week, and the textbooks integrated four skills, with more emphasis given to listening and speaking to develop students’ abilities for communication. There are usually about 40 students in one class in the primary school and teachers have more freedom in the design of classroom activities. As junior high students have to take the graduation exams at the end of the junior high school, the English language teaching is much affected by the exams, the class is usually dominated by teachers explaining texts and new words, and students have to do a lot of multiple choice exercises on grammar and reading comprehension. Similar to the primary school group, the size of the class is usually between 40-50 students in the junior high school.

3.3 Participants
The participants in the pilot study were 1,243 students: 769 grade 5 primary school students (409 in the experimental group; 360 in the control group), and 474 grade 1 junior high students (247 in the experimental group; 227 in the control group); 16 teachers (10 primary school teachers and 6 junior high teachers) and 12 schools (9 primary schools and 3 junior high schools) from a coastal city in China. The control and experimental groups were divided on a by-class basis, with each teacher responsible for two parallel classes (one as the control group and the other as the experimental group). Careful attention was given to the selection of the two classes to make sure that they were equivalent in language proficiency. In most cases, the control and experimental groups were selected from two classes taught by the same teacher.
3.4 Research methods and procedures
A mixed method of qualitative and quantitative analysis was applied, with the main aim of exploring learner and teacher perceptions of and experiences with the use of automated assessment in the Chinese EFL classroom. Research methods in the form of questionnaires, journals, quasi-experimental pre- and post-tests were undertaken to collect the pertinent data.

The experiment consisted of the following phases. Before the experiment, a learner questionnaire was administered to acquire a basic understanding of the students’ computer skills, study attitude, beliefs and their attitude toward assessment. The teachers were also asked about their current English teaching modes, teaching beliefs and computer skills via a questionnaire. A pre-treatment writing test from WRM was administered to all the students, who had to finish the test online within 40 minutes.

The intervention measures varied between the primary and the junior groups, with an integrated skills course of reading and writing for the primary group, while a writing course only for the junior group. Throughout the experimental period, the primary group wrote two essays based on the readers in the WRM in the computer lab at the schools. The junior group wrote three essays in the WRM. The control group wrote the same essays, however not in the WRM. The experiment extended for eight weeks for the primary group and seven weeks for the junior group. Both control and experimental groups participated in writing a pre-test and a post-test essay in the WRM system.

To ensure the smooth implementation of the project, the project included a teacher training workshop on the project overview, on the methodology of teaching integrated reading and writing English courses and the use of WRM prior to the experiment. During the experiment, an online synchronous conference and asynchronous email communications were organized to tackle problems.

The teachers’ experiences with the use of automated assessment were mainly collected through reflective journals. Reflection is a key component in the constructivist learning process; teachers and learners learn by observation, processing, and interpretation, and then personalize the information into personal knowledge (reflection). The teachers were asked to write journals after each reading or writing class. Therefore, each teacher completed four journal entries during the piloting stage. A journal template was designed in advance to guide the teachers in writing down their reflections using the WRM system. Thirty-five journal entries were collected from the primary teacher group, and 11 entries were collected from the junior teacher group.

A post-treatment learner questionnaire was administered to the experimental groups on their experiences of WRM use. Questions on perceptions of and experiences with WRM were based on Scharber, Dexter and Riedel (2008) and Chen and Cheng (2008). The post-treatment questionnaires for the two groups were administered online. Of the primary group, 266 students finished the questionnaire online, accounting for 65% of the total sample; with invalid questionnaires deleted, the final number amounted to 265. Of the junior group, 139 filled in the questionnaire, accounting for 56% of the total sample. The number of valid questionnaires was 131. The data was keyed in into SPSS11.5 for analysis.
3.5 Experimental data analysis
A quasi-experimental non-randomized control/experiment group pretest-posttest design was implemented in the study. This decision was consistent with educational research in examining the average growth of two groups using gain score analysis rather than determining impact at individual level.

To reduce the prompt difficulty as a confounding factor, an A/B-prompts cross-over design was used; if a student took prompt A during the pre-test, he or she would take prompt B in the post-test. In this way, the score differences analysis of the control and experimental groups would not be confounded by prompts difficulty. The students’ pretest-posttest writing prompts were administered in the Writing Roadmap1 online system and scored automatically using the generic scoring algorithm.

The results of a generic scoring engine validation study can be found in Rich (2012), where weighted kappa statistics range from 0.77 to 0.82 for the prompts used in this study. Using the automated generic scoring had the benefit of making scores comparable across different prompts in a same genre. The engine scoring also removed the rater effect since the computer scoring algorithm was identical for research pilot essays submitted in the system.

For the experimental quantitative data, this paper only reports the results concerning the junior high schools. Primary schools data were not included due to two factors. First, grade 5 EFL students writing tended to be too short (50 words or less) for obtaining accurate engine scores. Second, some of primary schools conducted the post-test in a non-standard condition during the 2010 pilot. Schools went on summer vacation in July and students took the post-test at home un-supervised in many cases. For the middle school data, we conducted data cleaning to remove the records that did not have both the pre-test and post-test. We also conducted a human review of the engine scored essays. Essays with a large increase or decrease in scores were read by human reviewers, and essays with potentially copied text were excluded from data analysis. In the end, the study included 162 students in the experimental group and 130 students in the control group.

4. Results and Discussion
4.1 Student attitudes toward and experiences with the use of AWE
Table 1 demonstrates student responses on the perceived efficacy of WRM in the classroom. Nearly 65% of the students in the primary group liked to use WRM to practice their writing, and the figure was 85% for the junior group, demonstrating students’ positive attitude toward WRM despite the short length of the experiment. Over 60% in the primary group and 68.7% in the junior group held that WRM encouraged them to write more. Two-thirds of the students in both groups reported revising more in WRM, one possible reason being that students could gain immediate feedback through WRM and could improve scores after revision.

Nearly three quarters students (73.3%) in the junior group held that WRM helped to build up their confidence in writing, while only 61.8% of the students in the primary
group assumed the same view. The discrepancy here might be caused by the different time span these two groups were engaged in using WRM. The junior group wrote and revised three essays within seven weeks, while the primary group wrote two essays within four weeks (The remaining weeks were spent on reading).

Nearly 60% of the students in the primary group and 73% in the junior group considered scores provided by WRM fair, which meant that WRM met students’ expectations about scoring to some extent. This finding contrasted with that in Chen and Cheng (2008). In their study of students’ experiences with an AES tool, My Access!, they discovered that none of the students found My Access! fair in marking. One reason could be traced back to their teachers’ understanding of the technology and writing pedagogy. One student in their study said, “How can we believe its marking if our teacher does not trust it?” Nevertheless, it was worth noting that in the present study, 37% of the students did not think the scores offered by WRM fair, and this calls for further study.

Of the junior group, 87% held that WRM helped them improve their writing skill, while only 58.5% of the students in the primary group held this opinion. The difference might be caused by the time span of the two groups practicing in WRM as mentioned in the preceding part. Moreover, the writing objectives were different in these two stages, with the primary group having the objectives of developing writing interest and basic writing conventions, and the junior group having the objectives of communicating ideas. Therefore, the junior group paid more attention to the improvement of writing skills and made better use of WRM.

More than half of the students from the primary group (56.3%) could understand the scores provided in WRM, while 40% could not, indicating the necessity of enhancing their understanding of WRM rubrics. On the contrary, only 17% of those from the junior group had problems understanding the scores offered in WRM, while 77% had no problems.

Sixty percent of those from the primary group held that the advice given by WRM could help them with their revision, while 85.5% from the junior group held this opinion. Given the short duration of the experiment (which was only 7 to 8 weeks long), this result was promising. As to the same question, the three groups in Chen and Cheng’s study (2008) had much lower percentages of those who viewed the software as being helpful (33%, 29% and 6%).

Table 1. Student experiences with WRM (N=396, Primary=265, Junior =131)

<table>
<thead>
<tr>
<th></th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
<th>Do not Know</th>
</tr>
</thead>
<tbody>
<tr>
<td>I like using WRM to practice writing.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>81 (30.6%)</td>
<td>9 (3.4%)</td>
<td>143 (54.0%)</td>
<td>28 (10.6%)</td>
<td>4 (1.5%)</td>
</tr>
<tr>
<td>Junior</td>
<td>12 (9.2%)</td>
<td>4 (3.1%)</td>
<td>60 (45.8%)</td>
<td>49 (37.4%)</td>
<td>6 (4.6%)</td>
</tr>
<tr>
<td>Total</td>
<td>93 (23.5%)</td>
<td>13 (3.3%)</td>
<td>203 (51.3%)</td>
<td>77 (19.4%)</td>
<td>10 (2.5%)</td>
</tr>
<tr>
<td>I write more after using WRM.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>68 (25.7%)</td>
<td>22 (8.3%)</td>
<td>141 (53.2%)</td>
<td>24 (9.1%)</td>
<td>10 (3.8%)</td>
</tr>
</tbody>
</table>

Table 1. Student experiences with WRM (N=396, Primary=265, Junior =131)
Ninety-five percent of those from the primary group reported revising their essays in WRM at least one time, while the same ratio was 92% for the junior group. Nearly half of those from the primary group and 54% in the junior group revised their essays three times or more in WRM. These results indicate that a majority of the students used WRM to revise their writings (see Table 2). This contrasted with that in Attali’s (2004) study, where 71% of the essays were submitted with no revision. One possible reason for this discrepancy might be due to the fact that in this study, AWE was fully integrated into the English reading and writing course and students were required and guided to use it in writing and revising their essays.
Table 2. Number of revision times per essay in WRM

<table>
<thead>
<tr>
<th></th>
<th>Primary</th>
<th>Junior</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than 5 times</td>
<td>54 (20.4%)</td>
<td>20 (15.3%)</td>
<td>74 (18.7%)</td>
</tr>
<tr>
<td>3-4 times</td>
<td>77 (29.1%)</td>
<td>51 (38.9%)</td>
<td>128 (32.3%)</td>
</tr>
<tr>
<td>1-2 times</td>
<td>123 (46.4%)</td>
<td>50 (38.2%)</td>
<td>173 (43.7%)</td>
</tr>
<tr>
<td>Never</td>
<td>11 (4.2%)</td>
<td>10 (7.6%)</td>
<td>21 (5.3%)</td>
</tr>
<tr>
<td>Total</td>
<td>265 (100.0%)</td>
<td>131 (100.0%)</td>
<td>396 (100.0%)</td>
</tr>
</tbody>
</table>

Four writing assistance tools were provided in WRM: hint (a tool providing suggestions on ideas and paragraph construction), tutor (a tool for correcting spelling, grammar and wording errors), thesaurus and grammar tree (a tool for analyzing sentence structure). Table 3 indicates that half of those in the primary group considered “tutor” most helpful, followed by “thesaurus”. The number of students who chose these two options amounted to three quarters of the total sample. These two functions related to grammar and vocabulary, indicating that the primary group attended more to the language aspect of their essays. The junior group used the four tools at similar frequencies, reflecting the fact that this group valued both content and form. Similar to the choice of the primary group in this study, “tutor” was also deemed the most frequently used function in Chen and Cheng’s study (2008).

Table 3. The most helpful WRM tool perceived

<table>
<thead>
<tr>
<th></th>
<th>Primary</th>
<th>Junior</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hint</td>
<td>48 (18.1%)</td>
<td>29 (22.1%)</td>
<td>77 (19.4%)</td>
</tr>
<tr>
<td>Tutor</td>
<td>146 (55.1%)</td>
<td>35 (26.7%)</td>
<td>181 (45.7%)</td>
</tr>
<tr>
<td>Thesaurus</td>
<td>53 (20.0%)</td>
<td>34 (26.0%)</td>
<td>87 (22.0%)</td>
</tr>
<tr>
<td>Grammar tree</td>
<td>18 (6.8%)</td>
<td>33 (25.2%)</td>
<td>51 (12.9%)</td>
</tr>
<tr>
<td>Total</td>
<td>265 (100%)</td>
<td>131 (100.0%)</td>
<td>396 (100.0%)</td>
</tr>
</tbody>
</table>

Table 4 shows how these WRM tools assisted the students in their writing.

Table 4. Students’ use of WRM  (N=396, Primary=265, Junior =131)

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Seldom</th>
<th>Half the time</th>
<th>Often</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary</td>
<td>26 (9.8%)</td>
<td>98 (37.0%)</td>
<td>65 (24.5%)</td>
<td></td>
</tr>
<tr>
<td>Junior</td>
<td>2 (1.5%)</td>
<td>41 (31.3%)</td>
<td>51 (38.9%)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>28 (7.1%)</td>
<td>139 (35.1%)</td>
<td>116 (29.3%)</td>
<td></td>
</tr>
</tbody>
</table>

I use WRM to correct punctuation and format errors.

I use WRM to correct spelling errors.

<table>
<thead>
<tr>
<th></th>
<th>primary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>primary</td>
<td>21 (7.9%)</td>
<td>83 (31.3%)</td>
</tr>
</tbody>
</table>
More than 60% of those in the primary group and nearly 70% in the junior group used WRM to help with the punctuation and writing conventions “half the time” or “often”, indicating that both groups paid much attention to writing conventions.

Nearly 70% of the students in both groups (67.9% for the primary and 69.4% for the junior) used WRM to correct spelling errors “half the time” or “often”, indicating the frequent use of this function. Three quarters of the students in both groups used WRM to correct grammar errors in their essays “half the time” or “often”. It was also worth noting that nearly 40% in the junior group used WRM to improve their wording “often”.

Over 90% of those from the junior group considered that their writing improved “somewhat” or “a lot”, while more than 54.7% in the primary group considered that their writing improved somewhat and 54.7% reported that their writing improved a lot. However, Chen and Cheng’s study (2008) found that only 55% of the students reported writing gains. The difference could be traced to the different ways AWE was used in the classroom. In this study, AWE along with teacher guidance were built into the experimental teaching, while in Chen and Cheng’s study (2008), AWE use varied among the three teachers from using it in drafting to relying on it solely for marking with no teacher intervention, and this might have resulted in students’ uncertainty and anxiety.

4.2 The teachers’ view of the efficacy of automated assessment

As mentioned before, 46 journal entries were collected from the teachers. The following observations were made in the content analysis.

First, students demonstrated a strong interest in using WRM as noted by the following teacher entry.

Students (regardless of their language proficiency levels) show strong interest in using WRM to write essays. Even those weak students are keen on using the system. They actively recall the new words that they have learnt and seek the teacher’s and student’s help during the writing process. They are eager to have their essays shown in WRM and to learn about the score and
feedback provided by the system. (Teacher A, U primary school)

Most students were very attentive during the writing process, focusing on writing in WRM, which was very rare. As it was machine-scoring, no cheating happened. Many students brought their word notebooks to the class and tried to use them (the words) in their writing. WRM stimulated their interest in writing greatly. (Teacher B, V middle school)

Teacher observations corresponded with students’ responses toward the survey. An overwhelming majority from both groups was satisfied with the use of WRM in the classroom.

Second, WRM has raised students’ awareness of the importance of writing conventions.

WRM can spot errors in spelling. In addition, it can also identify spacing in typing and also punctuation errors, which helps students a lot. (Teacher C, W primary school)

Similar responses can be found in students’ surveys. Nearly two thirds of the students, as shown in Table 4, often used WRM to correct punctuation and format errors.

Third, WRM helped students correct their own errors in writing.

We had a WRM writing class this week and one student wrote his reflection after the class: during our writing process, WRM can help spot our grammar errors and make us correct our errors ourselves and this is what we like in the system best. (Teacher D, X primary school)

Some teachers under study were burdened with a heavy workload with five classes (200 students) to teach and 200 assignments to mark each week. The introduction of the automated writing assessment helped relieve their marking workload. In the meanwhile, through providing individual support in vocabulary, grammar and format, and immediate scoring and feedback, WRM provided scaffolding for students’ learning and helped to develop students’ abilities in self-assessment and in correcting their own errors. With WRM taking care of students’ basic language errors, teachers were able to devote more attention to exploring writing pedagogy.

Fourth, teachers facilitated the effective use of automated writing assessment by providing ongoing support. It was observed that teachers set specific goals for students revising essays in WRM and offered timely support when students experienced difficulties in correcting the errors. Previous research (cf. Chen & Cheng, 2008) has indicated that without teacher intervention and support, students would feel lost and this would limit their writing development. As a result of WRM use, students changed their roles from passively listening to the teachers and reading teacher-marked assignments to actively correcting their essays; at the same time, teachers changed from spoon-feeding to promoting student-centered learning with technology-enhanced assessment, and changed their roles from lecturer and sole assessor to facilitator, supporter and co-assessor.
4.3 The impact on student writing performance

The students’ writing performance in the pre- and post-test is summarized in Tables 7 to 9. Data were aggregated from students pre-/post-test scores from three junior high schools and six teachers. Table 5 shows the descriptive statistics and the effect size for comparison between the experimental and control groups. The mean difference scores of post-test across control and experimental groups indicated that the experimental group mean scores were higher than the control group mean scores with an effect size of 0.40. The gain score showed that on average the control group had a small negative gain comparing the post-test score to the pre-test score. The experimental group had a positive gain score. The control-experimental group gain score comparison showed a standardized effect size of 0.31 in favor of the experimental group.

Table 5. Pretest-Posttest online writing descriptive statistics from the junior high schools

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Min</th>
<th>Max</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>control</td>
<td>130</td>
<td>2.42</td>
<td>0.55</td>
<td>1.13</td>
<td>3.58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>experiment</td>
<td>162</td>
<td>2.46</td>
<td>0.53</td>
<td>1.30</td>
<td>3.87</td>
<td></td>
</tr>
<tr>
<td>Mean Difference</td>
<td></td>
<td></td>
<td>0.04</td>
<td>(0.54)</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>Post-test</td>
<td>control</td>
<td>130</td>
<td>2.30</td>
<td>0.61</td>
<td>1.18</td>
<td>3.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>experiment</td>
<td>162</td>
<td>2.55</td>
<td>0.64</td>
<td>0.67</td>
<td>3.87</td>
<td></td>
</tr>
<tr>
<td>Mean Difference</td>
<td></td>
<td></td>
<td>0.25</td>
<td>(0.63)</td>
<td></td>
<td></td>
<td>0.40</td>
</tr>
<tr>
<td>Gain score</td>
<td>control</td>
<td>130</td>
<td>-0.12</td>
<td>0.68</td>
<td>-1.52</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>experiment</td>
<td>162</td>
<td>0.08</td>
<td>0.62</td>
<td>-1.93</td>
<td>1.45</td>
<td></td>
</tr>
<tr>
<td>Mean Difference</td>
<td></td>
<td></td>
<td>0.20</td>
<td>(0.65)</td>
<td>1.13</td>
<td>3.58</td>
<td>0.31</td>
</tr>
</tbody>
</table>

The SD in parenthesis is pooled standard deviation.

An effect size of 0.31 for the experimental group can be interpreted as being “small” statistically but positive in direction and educationally significant considering the short duration of the intervention.

Table 6 presents the test statistics of the mean score differences found in Table 5. An independent sample t-test showed the mean score difference was statistically significant at 0.01 level for the post-test score and gain score comparisons across the two groups. The test of equality of variances in Table 7 from the control and experimental groups showed the differences in sample variances were not significant. In comparison, the mean difference found from the pre-test between the control and experiment groups was not statistically significant. The test statistic indicated that the control and the experimental group pre-test scores were relatively comparable at the beginning of the intervention. The statistically significant gains for the experimental group over the control group were encouraging considering the short duration of a seven-week experiment of automated writing evaluation in the three junior high schools. The statistically significant score differences also echoed the positive feedback reflected in the student surveys and teacher journals.
**Table 6.** Junior high school across groups test statistics (Control – Experiment)

|                | DF  | t Value | Pr > |t| |
|----------------|-----|---------|------|---|
| Pre-test       | 290 | 0.60    | 0.55 |
| Post-test      | 290 | 3.28    | 0.001** |
| Gain score     | 290 | 2.65    | 0.009** |

**Table 7.** Test of equality of variance between the control and the experimental groups

<table>
<thead>
<tr>
<th></th>
<th>Method</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>Folded F</td>
<td>129</td>
<td>1.06</td>
<td>0.74</td>
</tr>
<tr>
<td>Posttest</td>
<td>Folded F</td>
<td>161</td>
<td>1.11</td>
<td>0.54</td>
</tr>
<tr>
<td>Gain score</td>
<td>Folded F</td>
<td>129</td>
<td>1.20</td>
<td>0.27</td>
</tr>
</tbody>
</table>

5. Conclusions

Being the first study on the use of AWE in Chinese schools, the study yielded findings on how an AWE tool, WRM, impacts students writing performance, and the teaching and learning process.

One purpose of this study was to understand whether online technology-enhanced English language writing assessment is effective in raising students’ achievements. Preliminary quasi-experimental data analysis for the junior high schools showed some encouraging results. Across-group comparison showed an effect size of 0.31 in favor of the experimental group. This effect size for a short period of seven weeks was educationally significant in terms of student English language writing. Data were analyzed using repeated measures across the groups. Positive score gain was observed in the experimental group. Within-group repeated measures test for the experimental group also showed statistically significant results for the post-test score mean differences. However, negative score gain was observed in the control group post-test scores, and this should be explained in further studies.

Most students adopted a positive attitude toward the use of WRM to writing. The use of WRM also helped to build up students’ confidence in writing. They became more motivated and autonomous, and tended to write more. Students made the most frequent use of tutor (tool for correcting grammar and vocabulary errors), and this corresponded with the finding in the previous research (see Chen & Cheng, 2008). The underlying reason might be attributed to the fact that AWE system could underline the grammar and vocabulary errors and offer relevant revision suggestions, while AWE suggestions on content and structure were more general and difficult for students to follow. The participating teachers also acknowledged the positive role that WRM exerted in motivating students to write and revise. During the experimental teaching, they also changed their role from one of spoon-feeding students to one of a supporter and facilitator.

The study also revealed that it was crucial to explain WRM rubrics to primary
students and teachers. Nearly 40% of the primary group could not understand the marking of WRM. A similar problem was also revealed in the teacher journals. The follow-up research should focus on helping teachers and students understand assessment rubrics. Moreover, it has been argued that communicating rubrics to students is an important principle of effective assessment (e.g., Brown & Knight, 1994), which in the meantime can help students internalize the assessment rubrics and promote learner autonomy.

It might be argued that the successful implementation of the findings of this study might lie in the orientation of the teachers towards the project significance and technology use in education prior to the research and towards teachers’ acknowledgement of the positive role WRM in motivating students to write, improving writing abilities and reducing teaching workload, which corresponded with previous research findings that teachers’ attitude and their perceptions were the determining factors in the success of a teaching reform (e.g., Chen & Cheng, 2008; Grimes, 2008). Second, compared with similar research in the US (e.g., Warschauer & Grimes, 2008), the students and teachers in this study did not have imminent high-stake exams to prepare for and, therefore, had more freedom and could make full use of the system. Finally, AWE in this study, rather than being a tool left at teachers’ and students’ disposal, was carefully designed and fully integrated into the experimental teaching, in which teachers guided and exploited its use with their instruction to enhance teaching and learning efficacy.

With the increasing need for the use of information technology in education, it is hoped that the current research on applying automated writing assessment in the Chinese EFL classroom could provide empirical evidence and shed light on people’s understanding of technology-enhanced formative assessment integration in classrooms for improving teaching and learning.

Note


References


(Copy editing: Duncan Sidwell)