the jaltcalljournal

ISSN 1832-4215
Vol. 13, No.2 Pages 117–146
©2017 JALT CALL SIG

Automated writing evaluation in an EFL setting: Lessons from China

Jinlan Tang
Beijing Foreign Studies University, China
tangjinlan@beiwaionline.com

Changhua Sun Rich
CTB/McGrawHill, USA
Changhua.rich@act.org

Regular Paper

This paper reports a series of research studies on the use of automated writing evaluation (AWE) in secondary and university settings in China. The secondary school study featured the use of AWE in six intact classes of 268 senior high school students for one academic year. The university study group comprised 460 students from five universities across the country. The teaching experiment included the introduction of an AWE tool, Writing Roadmap, to the English as a Foreign Language classroom and offering support to the teachers. A mixed-methods approach in the form of quasi-experimental research design, questionnaires, interviews and journals were undertaken to evaluate the efficacy of the teaching experiment. In this paper, we summarize the results of these studies in relationship to implementation, teacher and student attitudes, effects on writing and revision processes, and impact on writing test score outcomes. We also discuss the key factors affecting the successful integration of this technology in the classroom.

Keywords: Automated writing evaluation (AWE); Writing instruction; Technology and language learning; EFL teaching and learning

Introduction

Though automated writing evaluation (AWE), which employs artificial intelligence to evaluate essays and offer feedback, has been in existence since the 1960s, its use in assessment and instruction remains controversial. Some argue that use of such software dehumanizes the writing process, violating the social and communicative nature of writing (CCCC Executive

Committee, 2004; Ericsson, 2006). Others argue on behalf of automated evaluation, referring to research evidence demonstrating its high reliability that is comparable to human scoring (Elliot, 2003; Klobucar, Elliot, Deess, Rudniy, & Joshi, 2013; Page, 1994; Ramineni, 2013; Shermis & Burstein, 2003). Despite the ongoing debate, there has been growing use of AWE softwares. Educational Testing Service (Heilman & Tetreault, 2012) reports that its e-rater program scored some 5.8 million scripts in 2010 and was currently utilized in more than 20 applications, including as part of the GRE and TOEFL tests. AWE products for the classroom have been used in schools (e.g., Grimes & Warschauer, 2010; Rich, Harrington, Kim, & West, 2008; Warschauer & Grimes, 2008; White, Hixson, D'Brot, Perdue, Foster, & Rhudy, 2010), and universities (e.g., Chen & Cheng, 2008; Li, Link, Ma, Yang, & Hegelheimer, 2014; Li, Link, & Hegelheimer, 2015; Link, Dursun, Karakaya, & Hegelheimer, 2014; Tang & Rich, 2011; Tang & Wu, 2012; Wang, Shang, & Briody, 2013; Wu & Tang, 2012).

This paper aims to enrich the current literature via summarizing the main impact of AWE in Chinese EFL classrooms by analyzing a series of studies conducted on the use of AWE by secondary and university students who learn English as a foreign language (EFL) in China. We first review relevant research on the use of AWE in the classroom, and then report and discuss the results of these studies.

Prior research on AWE

Warschauer and Ware (2006) provided a comprehensive review of AWE and its related research studies a decade ago and it may be argued that their broad categorization of different types of AWE research still holds true, with some studies concerning the validity of AWE and comparison of the machine scoring with the human scoring (e.g., Deane, 2013; Klobucar *et al.*, 2013; Liu & Kunnan, 2016; Ramineni, 2013; Ramineni & Williamson, 2013); others on the use of AWE in improving students' standardized writing test scores (Attali, 2004; Rich *et al.*, 2008; Vantage Learning, 2007; White *et al.*, 2010). However, what they consider most important is for more process-product research on the use of AWE to reveal the process of AWE application and how it impacts writing instruction (Warschauer & Ware, 2006).

In fact after their call for more classroom research on AWE (see Warschauer & Ware, 2006), the last ten years witnessed an increasing body of studies published in international peer-reviewed journals investigating the use of AWE in the classroom (e.g., Chen & Cheng, 2008; Grimes & Warschauer, 2010; Li *et al.*, 2014; Li *et al.*, 2015; Link *et al.*, 2014; Wang *et al.*, 2013; Warschauer & Grimes, 2008) and even a special CALICO issue on automated writing evaluation published in 2016 (cf. Hegelheimer, Dursun, & Li, 2016), whose findings seemed to support Grimes and Warschauer's suggestions of AWE's "utility in a fallible tool" if deployed effectively (Grimes & Warschauer, 2010, p. 4).

Chen and Cheng (2008) investigated the use of an AWE program with three parallel classes of three teachers over one semester. It may be argued that the most important contribution of their research to the field is their perceptions of the AWE utility in the early drafting and revising process of writing instruction, followed by teacher and peer feedback in the later process. Moreover, it is them who for the first time suggested the potential usefulness for setting up a minimum score requirement as a prerequisite for submission to AWE. For example, one teacher in their study used the AWE score and feedback as a reference in her grading, and required her students to revise their essay in the system until they had achieved a minimum score of 4 out of 6 before they submitted it to teacher assessment and peer review.

Warschauer and Grimes (2008)'s mixed-methods exploratory case study of four schools in their use of two AWE softwares revealed that although the program encouraged students to revise more, the revision was limited to language forms only, few on content or organization. In addition, teachers' use of AWE varied from school to school and was determined most by teachers' prior beliefs about writing pedagogy, which arguably called for the necessity of teacher training on writing pedagogy if AWE were to be successfully used in the classroom.

Grimes and Warschauer (2010) conducted a 3-year longitudinal study on the use of AWE in eight schools in California and concluded that AWE motivated students to write and revise more and promoted learner autonomy. They attributed the successful use of AWE partly to the maturity of the AWE programs in the study, but more importantly to the local social factors such as technical, administrative and teacher support, which seemed to verify the claim that the key to technology use might be neither hardware nor software, but rather human ware (Warschauer & Meskill, 2000).

In the EFL context, Wang *et al.* (2013) investigated the impact and effect of using AWE on freshmen writing with a group of 57 students from a university. They used a quasi-experimental pre-post test research design and the results showed a significant difference between the experimental group and the control group in terms of writing accuracy, with the experimental group demonstrating obvious writing gains in terms of writing accuracy and learner autonomy awareness. In discussing the pedagogical implications, they suggested that teachers should be more actively involved from teaching students structure and to teaching students models of writing so that students knew how to improve their language accuracy and how to improve their writing content and structure.

In examining the impact of AWE corrective feedback on writing accuracy with 70 non-native ESL students in a US university, Li *et al.* (2015) found that the corrective feedback had helped increase the number of revisions and improve the accuracy. Their study seemed to support the claim of the usefulness of the practice proposed by Chen and Cheng (2008) of requiring a minimum score before submission to AWE. Moreover, similar to the previous studies (e.g., Grimes & Warschauer, 2010; Warschuaer & Grimes, 2008; Wang, *et al.*, 2013; Wu & Tang, 2012), their study reinforced the important role of teachers and suggested that the instructor's ways of implementing AWE might impact how students engage themselves in revising in AWE.

It might be argued that except for one study (i.e., Wang *et al.*, 2013), the rest did not have a control group, therefore the claimed AWE effect on writing performance needed to be interpreted with caution. More importantly, though the importance of teacher pedagogical roles has been implied or suggested in some of the studies (e.g., Li *et al.*, 2015; Wang *et al.*, 2013; Warschauer & Grimes, 2008), no systematic training was provided to teachers regarding the writing pedagogy in those studies reviewed. Furthermore, none of these AWE studies so far seemed to suggest a tentative procedure of using AWE effectively in the classroom, in most cases, the ways of using AWE mainly depended on teachers (e.g., Link *et al.*, 2014).

In the light of the aforementioned understanding, the current research sought to contribute in these areas by investigating how a group of teacher researchers explored the use of AWE in their EFL classrooms and lessons. Our study differs from the previous studies in that the intervention measures include not just AWE but the AWE-integrated teaching experiment and teacher training on writing pedagogy and ongoing support. For the use of AWE, our study introduced a tentative procedure of integrating AWE in the classroom based on the previous studies (e.g., Chen & Cheng, 2008; Li *et al.*, 2015; Link *et al.*, 2014;

Tang, 2014), therefore teachers are not left alone, but rather provided a reference working framework on how to use the tool in the classroom.

The current situation in China

Use of technology in education is required at both the secondary and tertiary levels in China. Though China college English teaching requirements stipulate high standards for students' English writing competence, students perform lowest in the writing portion of national college English examinations in China (Jin, 2010). It may be argued that the development of students' English writing abilities are hindered due to an array of factors such as the large class size, lack of writing practice, lack of teacher feedback and lack of qualified writing teachers (Tang, 2012). With the development in artificial intelligence and the widespread use of the Internet, computerized feedback by automated writing evaluation (AWE) software is having an increasing influence on writing instruction due to its immediacy and round of the clock availability as noted in the previous AWE studies. The last few years have witnessed increasing efforts on either developing AWE tools for Chinese EFL learners (e.g., Li, 2009; Liang, 2011; Liang, 2016) or applying AWE tools in China EFL classrooms (Liu & Kunnan, 2016; Tang, 2014; Tang & Wu, 2012).

This paper reports on the first large-scale study of AWE use in the Chinese EFL classroom, examining the use of AWE by high school and college students and their teachers. Like most AWE software, Writing Roadmap (abbreviated as WRM) used in this study, is designed primarily for native speakers of English. The suitability of such software for EFL students is an area worthy of exploration and research. Two studies on the use of WRM in the West Virginia schools in the US indicated positive gains in practicing writing for students who used WRM versus those who did not (Rich *et al.*, 2008; White *et al.*, 2010). However, these studies did not research the process on how WRM was used by teachers and students in the classroom, as this study aims to do.

Theoretical approaches

Following Grimes and Warschauer (2010), the current study adopted a social informatics theoretical approach toward the use of AWE, assuming technologies, people, and organizations as a "heterogeneous socio-technical network" (Kling, 1999). Rather than the "tools" view which focuses only on the technology per se, this approach embraces a more complicated and locally situated process of technology integration. Social informatics theory informed the current research design and drew our attention to the more important local factors such as people and organizations in shaping the use of technology. In the light of this understanding, teacher training not only on technology but also on writing pedagogy, and continuous teacher support were provided throughout the experiment in this study. In addition, participatory design (PD), commonly used in human-computer educational research engaging the users of computer systems in designing and revising the computer systems (Steen, 2013), and exploratory practice (EP), a practitioner-based research combining research and classroom teaching in the natural setting with the aim to resolve teacher and students' "puzzles" or "problems" in the classroom (Allwright, 2003), were also pertinent to our study design in that our teachers, rather than subjects to be investigated, but were supported to be active researchers and encouraged to explore the best way of using AWE in their own teaching context.

The current study concentrates on the classroom use of an AWE software tool, Writing Roadmap, in the Chinese EFL context. The mixed methods quasi-experimental study draws on questionnaires, journals, interviews and pre- and post-tests. It aims to investigate the following questions:

1. How does use of AWE affect students' writing performance in English in China?
2. What is the impact of AWE use on teaching and learning processes?

Method

The participants

The participants reported in this study consisted of 268 senior high school students from six intact classes, 460 tertiary students from five universities, and ten teachers (three teachers from the senior high school and seven from across the five universities).

The high school group comprised three cohorts with the first one from Senior High 1 (the first grade of the senior high school), with one class as the experimental and the other as the control (see Table 1). The second and third cohorts are both from Senior High 2 (the second grade of the senior high school). The students' age ranged from 15 to 17 years old and they have received at least nine years of English language education with six years in the primary school and three years in the junior high school. The students upon junior high graduation should be able to use English for basic communication and have a vocabulary size of 1500 words. The initial language proficiency of the experimental and the control groups were of parallel levels based on either student performance on the high school entrance exam in the case of cohort 1 or end of the term exams from the previous year for cohorts 2 and 3.

Table 1. An overview of high school participants under study[1]

Cohort	Grade	Duration (unit:term)	Experimental	Control	Teachers
1	Senior 1	2	46	42	1
2	Senior 2	2	46	47	1
3	Senior 2	2	46	41	1
Total			138	130	3

For the university group of 460 students, 224 were in the experimental group and 236 in the control group (see Table 2). The type of universities varied from teacher education, polytechnic to comprehensive. The majority of the students are from non-English majors (390 students, 85%), of arts major (including English) (327 students, 71%) and of second year in the university (306 students, 67%). For the first-year students, they have at least completed 12 years of English language learning prior to the university and they could use English for communication, and they are required to master at least 3500 words. The second-year students have studied English for one more year in the college and their vocabulary size is expected to reach 4500 words.

Table 2. An overview of university participants under study

No. of universities	University type	Duration (unit:term)	Major	Year	Experimental	Control	Total	Teachers
A	Teacher Education	2	English	Second	37	33	70	1
B	Comprehensive	2	Arts	Second	45	41	86	1
C	Polytechnic	2	Sciences	First	59	74	133	2
D	Comprehensive	2	Arts	First	9	12	21	1
E	Comprehensive	2	Arts	Second	74	76	150	2
					224	**236**	**460**	7

Three teachers participated in the high school study with each teacher assigned one experimental group and one control group. The three teachers ranged in their teaching experiences from two to five years and they all had Master of Arts' degrees in the area of applied linguistics or English language teaching methodology. They were all under 30 years old at the time of this study and were interested in exploring the use of technology to enhance instruction.

The seven university teachers varied in their teaching experiences from five to fifteen years, with one teacher having a PhD degree, and the remaining six teachers having Master of Arts degrees in the areas of applied linguistics or language education. They were interested in using technology to improve teaching and volunteered to participate in this research project.

Research interventions

The intervention measures included introducing a teaching experiment using the Writing Roadmap software, and offering support to the participating teachers. The teaching experiment extended from September 2010 to July 2011. The students were divided into two groups: the experimental and the control group. Each group is a natural intact class, and the experimental and the control groups were made up of two classes of parallel language proficiency levels (based on their end of term tests or high-stake exams such as the senior high school entrance test and the college entrance test). In addition, the same teacher taught both groups to reduce the number of variables affecting the efficacy of the teaching experiment. The two groups participated in a pre-test before the experiment and a post-test after the experiment. The tests were in essay format and administered in the AWE system (see Appendix A).

The software. The automated writing assessment tool investigated in this study is Writing Roadmap (WRM) from CTB/McGraw-Hill. It provides a set of six-trait writing rubrics or assessment criteria (AC), each with a set of indicating components. The six traits are "Ideas and Content", "Organization", "Voice", "Word Choice", "Fluency", and "Conventions." WRM offers immediate online feedback through highlighting problematic sections, narrative comments, discrete (trait-specific) and holistic scores, and remarking and rescoring on revised versions. It also provides a set of writing assistance tools such as "hint," "tutor," "thesaurus," and "grammar tree" to offer tips on improving writing, on grammar and syntax and choice of words (sentences with grammar errors are highlighted in blue, and words with spelling errors are in red).

The AWE-integrated teaching experiment. The teaching experiment extended for two semesters for both the university and the high school group. For the university group, except for the English majors (who wrote 11 essays), the remaining four non-English majors wrote seven essays each, three essays in semester one and four essays in semester two. The high school group wrote seven essays on average throughout the experiment. Writing was a standing-alone individual course for the English majors, which explained why students could write more essays, however, writing was only a component of the general English course for the non-English majors and also the high school group.

Both the experimental and control group students knew that they were using a software to help them with their writing, but they were not informed whether they were in the control or experimental group, nor did they know about the details of the experiment.

Based on previous studies (e.g., Chen & Cheng, 2008; Li *et al.*, 2015; Tang, 2014; Tang & Rich, 2011) and the local teaching context, our team proposed and implemented the following procedure of using AWE in the classroom:

1. teacher and student understanding of AWE assessment criteria (i.e. the writing rubrics, abbreviated as AC)
2. teacher-led pre-writing discussion on the writing topic;
3. autonomous writing and revision in AWE with the support of AWE writing assistance tools until arriving at the required score;
4. teacher feedback based on the AC and on the AWE-generated report of students' writing performance and peer feedback in the light of the AC;
5. revision based on teacher and peer feedback;
6. submission of essays in AWE.

The process features the AC comprehension and application throughout, and integrates self, teacher and peer feedback, precipitating the students' autonomy, self revision and formative learning. Moreover, a required score for revision was introduced during the autonomous writing stage to motivate students to revise to achieve a satisfactory mark before teacher assessment. The idea of requiring a minimum score before teacher assessment was proposed in Chen and Cheng (2008), and its efficacy was verified in Li *et al.* (2015), however the authors also found flaws with this practice and called for further investigation of the issue.

To summarize, the three striking features of the suggested AWE integration procedure were the requirement for achieving a minimum score during the autonomous writing stage; combination of self, peer and teacher feedback during the process; and the AC comprehension and application throughout the whole process. It may be argued that though the general procedure is followed, variations also exist with different classes. Table 3 demonstrates the writing instruction procedure in different classes.

Table 3. The writing instruction procedure in different classes

Level	Teacher/ researcher	Control class		Experimental Class		
High school (H)	Teacher 1	Using WRM AC to guide the writing instruction; autonomous writing before class; in-class teacher feedback		WRM AC interpretation in writing instruction	Autonomous writing in WRM before class; in-class peer revision, teacher feedback based on WRM, and self-revision	
	Teacher 2	Autonomous writing before class; in-class peer revision and teacher feedback			Autonomous writing in WRM before class; in-class teacher feedback based on WRM feedback, and self-revision	
	Teacher 3	In-class autonomous writing and teacher feedback			In-class autonomous writing in WRM and teacher feedback based on WRM feedback	
University (U)	U-A/Teacher 1	Autonomous writing, teacher marking and feedback			Autonomous writing and revision in WRM, teacher feedback based on WRM feedback	
	U-C/Teachers 3–4					
	U-D/Teacher 5					
	U-E/Teachers 6–7					
	U-B/Teacher 2		in-class peer revision			in-class peer revision

Table 3 shows that though teachers varied in their writing instruction, the WRM AC interpretation was a core component in all the teachers' experimental class. Second, three high school teachers demonstrated more variation in their writing instruction procedure than the university group. Except one teacher from University B, who used in-class peer revision in both her control and experimental class, the remaining teachers seemed to have adopted a similar procedure

According to the teacher reports, peer feedback for the control group was conducted in a paper format, with students reading and commenting on each other's essay according to the assessment criteria given by the teacher. In-class peer feedback for the experimental group was done in the computer lab, with two students' reviewing their essays together on the computer.

Teacher support. In response to previous research studies' call for offering more pedagogical assistance to teachers (e.g., Grimes & Warschauer, 2010; Warschauer & Grimes, 2008), our team made arrangements to provide timely support to teachers participating in the research. First, a 4-member head group (HG) of the project was established to be responsible for research design, implementation, and evaluation; monitoring the experiment/research process; and providing ongoing academic and practical support. Second, technical support was provided from CTB/McGraw Hill and the Institute of Online Education of Beijing Foreign Studies University for WRM system operation and maintenance. The HG held interactive lectures at the different high schools and universities, network and telephone conferences, symposia, and conducted on-going individual interactions with the participating teachers.

The key issues discussed during the support process included orientation about the software tool WRM and the process of teaching writing; perceptions of the role of WRM in teaching and learning (Tang & Wu, 2011); effective ways of integrating WRM in teaching and learning (Beatty *et al.*, 2008), for example, curriculum-based instructional design input at the school's level; and challenges to AWE feedback and the role and implementation of AWE scoring and assessment criteria in teaching writing. Teachers were encouraged to explore different ways of integrating AWE into their teaching practice based on their individual class needs and their local teaching context.

Research methods

It has been noted that multi-method approaches are increasingly used in research because "mixed methods offers strengths that offset the weaknesses of separately applied quantitative and qualitative research methods" (Creswell & Plano-Clark, 2006, p. 18). The mixed-methods approach has been adopted in previous AWE studies such as Wang *et al.* (2013), Li *et al.* (2015), Warschauer and Grimes (2008), and Grimes and Warschauer (2010).

Hence the current study employed a mixed-methods qualitative and quantitative research approach, with the main aim of examining pre- to post-changes in writing proficiency, student and teacher perceptions of and experiences with the use of automated assessment in the China EFL classroom. Questionnaires, teacher journals, interviews, and quasi-experimental pre- and post-tests were used to collect the pertinent data. The use of a mixed-methods approach is justified as the quantitative study of students' pre- and post-test score comparison will answer the first research question, that is, whether the AWE use affects students' writing performance, while the qualitative data drawing from questionnaires, journals and interviews will help to reveal the impact of the AWE-integrated teaching experiment on the teaching and learning process (i.e., the second research question).

We used a quasi-experimental non-randomized control/experiment group pre- post-test design in examining the average growth of the two groups using gain score analysis. Students' pre- post-test writing prompts were administered in the Writing Roadmap online system and scored automatically using the generic scoring algorithm first, then by human scorers to ensure the reliability and fairness of the scores.

The post-experiment student questionnaire (see Appendix B) centred on students' beliefs toward English learning and writing and evaluation of the AWE-integrated teaching experiment. Group interviews were undertaken on students' experience with WRM and how they used it to revise and improve their writing. Each group interview consisted of three students and lasted thirty minutes (see Appendix C for student interview prompts).

The teacher questionnaire concentrated on teacher perceptions toward assessment, teacher and student communication, teaching mode, beliefs toward writing instruction, teaching methodology and teachers' perceptions of student learning autonomy in the application of AWE (see Appendix D). Teacher interviews were conducted in groups and were semi-structured (see Appendix E for teacher interview prompts). Two group teacher interviews involving five teachers were undertaken, one with the sciences major university and one with the high school. Each interview lasted for about 45 minutes and was recorded and transcribed. Teacher journals mainly concerned teacher experiences and reflection during the experiment (see Appendix F for the journal template).

Data analysis

To address the first research question of how the AWE-integrated experiment impacts students' English writing performance in China, we examined gain scores of students' pre-post writing tests with effect sizes for university and high school groups. While the independent sample t-tests and paired-samples t-tests showed clear evidence on the statistically significant effect of writing performance improvement from the AWE-integrated teaching experiment (e.g., Tang, 2014; Tang & Wu, 2012), in this paper we used a General Linear Model (GLM) procedure to further explore the intervention effect by university student study major and by classrooms in senior high school. In particular we investigated whether the English major or non-English major studies contribute to the differences in the observed score gains in the university sample and what is the classroom effect of the three different teachers on the observed differences.

In response to the second research question of how the AWE-integrated experiment impacts the teaching and learning process, we collected and analyzed student and teacher questionnaire, interview and journal data. The student questionnaire was conducted online and the submission rate was 67%. For the high school group, 71 out of 138 experiment group students (51%) submitted their questionnaires. For the university group, 185 out of 224 experimental group (83%) completed their questionnaires. The teacher questionnaire was sent to the teachers via email to fill in and all ten teachers returned their answers. Multiple choice responses were analyzed via SPSS, while responses toward open-ended questions and interviews along with journals were examined through content analysis. Common themes were extracted, discussed and exemplified to illustrate how the teaching experiment affects the teaching and learning process (see the results below).

Results

The research results are reported in the order of the two questions. The GLM analysis of pre and post test scores will answer whether the use of WRM in EFL instruction will result in students' improved writing performance. Data from questionnaires, interviews and teacher journals will indicate how teaching and learning process might change during the AWE-integrated experiment.

Impact on writing performance

The effects of the AWE-integrated experiment on students' writing performance are discussed in relation to the two groups of students under study, the university group and the high school group.

The university group. Tables 4 and 5 present pre- and post-test statistics for the university group using overall sample and subgroups. The mean difference scores of pre-post tests across control and experimental groups indicated that the experimental whole group and subgroups mean scores were higher than the control group mean scores with effect size as

0.79, 1.46, 0.72. The control overall group and subgroups not only had smaller gains in the post tests than the experimental overall and subgroups, but also with smaller effect size as 0.19, 0.31, 0.18 (see Table 4). Moreover, the overall F test is significant with p-value 0.0001 (see Table 5), indicating strong evidence that the students in the experimental group of using Writing Roadmap had statistically significant greater English writing improvement as measured by the pre- and post-tests than that of the students from the control group.

Table 4. Descriptive statistics: university students

Variable	N	Pre-test mean	Pre-test SD	Post-test mean	Post-test SD	Mean post-pre test score gain	Score gain SD	Effect size
Experimental	224	2.34	0.57	2.86	0.74	0.52	0.85	0.79
Control	236	2.40	0.53	2.51	0.62	0.11	0.73	0.19
Experimental English major	37	2.65	0.61	3.56	0.64	0.91	1.05	1.46
Experimental non-English major	187	2.28	0.54	2.72	0.68	0.44	0.79	0.72
Control English major	33	2.70	0.64	2.90	0.67	0.20	1.09	0.31
Control non-English major	203	2.35	0.49	2.45	0.59	0.10	0.66	0.18

Table 5. Gain score GLM analysis for university students

Source	**DF**	**SS**	**MS**	**F**	**Pr > F**
Experimental or control group	1	18.75	18.75	31.46	0.0001
English major or non-English major	1	4.93	4.93	8.05	0.005
Group*English major	1	2.03	2.03	3.32	0.07
Experimental English major or non-English major	1	6.72	6.72	9.66	0.002

For the university sample analysis, we used the GLM to examine the effect of the automated writing evaluation software by two types of students. English major group consists of students who study English as a major in the universities. The other group of students labeled as non-English major study English as a general requirement of other majors in the university. We noticed that the English major students were assigned a set of writing topics that are different from the non-English major students (see Appendix A). Because the two variables overlap, we focused on English major or non-English major students, and the combined group and type of English major interaction effect. Overall the GLM analyses indicated statistically significant higher mean score gains from English major students vs. non-English major students with p-value of 0.005 (Table 5). The experimental/control group and major interaction effect is present but not statistically significant with p-value of 0.07 (Table 5). For the experimental group, English major students made statistically greater improvements in writing than the non-English major students. We believe that the significant difference we observed could be explained by the fact that the curriculum for the English major students centered on English language development, while the non-English major students take English as only one course of the curriculum.

The high school group . For the senior high school sample, the gain score difference of the pre- and post-test scores between the experimental and control group is smaller but still statistically significant with p-value 0.03 (Table 7).

Senior high school has three classes listed as Teacher 1, Teacher 2 and Teacher 3 (Table 6) in this study. Teacher 1's class gain score from experimental and control class is very similar, 0.51 from experimental and 0.48 from the control class. Teacher 2 and Teacher 3's classes had very different results from the experimental and control classes. The following factors might account for this. Teacher 1's class was in Senior 1, that is, the first year of high school (in China, students need to study three years in high school before taking an entrance exam for college), when both teachers and students have no imminent pressure from the high stakes exam such as the college entrance exam, and have more time and are more motivated to participate in the teaching experiment. The journal data revealed that Teacher 1 had made active use of the assessment criteria in WRM to guide her writing instruction (see the section on the impact on teacher process below for details), therefore both of her classes might have benefitted from this additional application of the WRM software. In contrast, Teacher 2 and 3's classes were both from Senior 2, when teachers and students were faced with increasing pressure from the high stakes exam, i.e. the college entrance exam (which is held at the end of the third year of the senior high school). It could be argued that the control groups in particular might feel less motivated to take part in the WRM post-tests, which might result in only slight changes in their gain scores.

Table 6. Descriptive statistics for senior high school

Variable	N	Pre-test mean	Pre-test SD	Post-test mean	Post-test SD	Mean post-pre test score gain	Score gain SD	Effect size
Experimental	138	2.65	0.77	3.05	0.90	0.40	0.79	0.47
Control	130	2.57	0.77	2.76	0.86	0.19	0.93	0.19
Experimental teacher 1	46	3.02	0.68	3.53	0.76	0.51	0.93	0.71
Experimental teacher 2	46	2.00	0.50	2.33	0.72	0.33	0.57	0.53
Experimental teacher 3	46	2.94	0.65	3.29	0.75	0.35	0.85	0.49
Control teacher 1	42	2.83	0.73	3.31	0.84	0.48	1.08	0.61
Control teacher 2	47	1.97	0.46	2.02	0.44	0.05	0.43	0.11
Control teacher 3	41	2.88	0.73	2.77	0.67	−0.11	1.04	−0.16

The three teachers from the senior high school sample each taught a control and experimental class. The GLM test statistics show that gain scores did not have statistically significant differences within experimental group classes, nor was there a significant group and class interaction effect (Table 7). We actually were pleasantly surprised to see that all classes, despite different levels of students, benefited from the AWE-integrated teaching experiment. The descriptive statistics in Table 6 show that the three experimental classes gain scores are 0.51, 0.33, 0.35 vs. the gain scores from the three control classes: 0.48, 0.05, −0.11. In the meanwhile, the experimental group gain score effect size of 0.71, 0.53, 0.49 is much bigger than the control groups' effect size of 0.61, 0.11, −0.16, indicating greater improvements from the experimental group and teachers. Teacher 1's control group made strong post-test improvement, perhaps due to the fact that Teacher 1 may have provided more motivation in teaching and testing for the control group.

Table 7. Gain score GLM analysis for senior high school

Source	DF	SS	MS	F	Pr > F
Experimental or control group	1	3.82	3.82	5.06	0.03
Teachers both groups	2	7.20	3.60	4.95	0.008
Group*teacher	2	2.04	1.02	1.40	0.25
Experimental teacher	2	0.93	0.47	0.73	0.48

In summary, the GLM analysis showed that the writing improvement for the university experimental group and English major students was statistically significant. Similarly, the GLM analysis found statistically significant writing improvement for the senior high school experimental group, and all three teachers' classrooms with different levels of students.

Impact on the learning process

In this part, students' responses from the post questionnaire and interviews were drawn to demonstrate how students' writing process might change during the experiment.

First, the integration of the teacher, student and WRM assessment and feedback seemed to have enhanced interaction and motivated students to rewrite and revise. According to the student questionnaire, 70% of the students held that they would be likely to write more and revised their essays more after using the system. 62.3% reported revising 1–2 times; 27.9% 3–4 times. This finding also coincided with that of Grimes and Warschauer (2010). One of the main reasons for students' continuous revision might lie in that WRM offers writing assistance tools such as "Tutor" which provides suggestions for students to correct grammar errors themselves, through which they can remember better as one student noted:

> The "tutor" function in WRM helped with my grammar. I can remember more clearly when I correct my mistakes through WRM and I will not make the same mistakes again. (Student 1, University C, Data source: Questionnaire)

Feedback, considered the lifeblood of learning (Rowntree, 1987), is an important component in formative assessment. Research has shown that instant and prompt feedback enhances learning the most (e.g., Black & Wiliam, 1998). Large class size is the norm in the current typical English class, regardless of teaching levels (Tang *et al.*, 2012; Warschauer & Ware, 2006). This might prohibit the amount of writing practice and make timely feedback on students' writing assignments hardly possible, consequently affecting students' improvement (Warschauer & Grimes, 2008). In the current experiment, however, the multiple and dynamic assessment and feedback from the teacher, the peer and the AWE system, interacted and motivated the students to write and revise continuously. The AWE feedback was instantaneous and prompt, while the teacher feedback was usually more targeted and could tackle the more difficult problems.

Second, with teacher guidance and instruction (e.g., teachers provide feedback by critiquing exemplary essays in the class in the light of WRM six-trait rubrics or assessment criteria) and constant interaction with the system, students seemed to have learned to use AC to guide their own writing, which could be noted in the university group.

> I used to compose a lengthy opening for my English essay. During the experiment, through practicing my essays within the system and understanding the AC, I found the English essays usually state their topics directly in the opening, and with a topic sentence for each paragraph. I think writing practice in WRM helps me think in English when writing essays and ensures smooth cross-cultural communication. (Student 1, University E, Data source: Interview)

It may be argued that compared with what they did in the past, the students in the experiment seemed to have a clearer purpose in writing via using AC to guide and revise essays, during which they gradually internalize the AC and improve their assessment and self-assessment abilities and become a key partner in the assessment process.

Third, it seemed that students had become more autonomous via dynamic interaction with the system and teacher feedback, correcting their mistakes and revising their essays.

> What I found most attractive about the system was that it could force me to practice and revise my own essays, which improved my autonomy and writing. (Student 3, University B, Data source: questionnaire)

Traditional writing instruction follows the linear order of students writing and teacher feedback, during which the students' role might be passive and they might lack the motivation to revise their essays, let alone join the assessment process (Carless, 2006). However, in the WRM-integrated teaching experiment, it seemed that students were motivated to write and revise through continuous interaction with the system, and peer and teacher assessment and feedback. Moreover, the instant feedback from the system along with teacher support with the AC interpretation seemed to have helped students internalize and apply the AC in their own writing and acquire assessment and self-assessment abilities, which can be shown in the following quote:

> The teaching experiment helped me to know better about the ideas and structure of English essays, it also helped to improve my self-assessment ability. Now I can see very clearly the strengths and weaknesses of an essay. (Student 1, University A, Data source: questionnaire)

Consequently, they might change from the traditional role of being assessed to becoming a co-assessor, during which their autonomous learning abilities could be enhanced.

Impact on the teaching process and teachers

Research data from the teacher questionnaire, interviews and journals were used to document how the teaching and teachers might have changed during the experiment.

With language problems largely dealt with by WRM, teachers might not need to spend as much time correcting and commenting on the language mistakes, and the writing instruction seemed to witness a shift of focus from language form to content and discourse, from product to process (e.g., Wang *et al.*, 2013; Warschauer & Grimes, 2008; Wu & Tang, 2012).

> WRM helped to liberate me from the marking workload. I remember I used to mark students' essays every weekend, while students turned a blind eye to my comments. Now with WRM help, I can have time to think how to provide more targeted writing instruction based on their weak points. (Teacher F, Data source: post-questionnaire)

More attention seemed now to be directed on the teaching/learning process. Specifically, a pre-writing phase was incorporated with the main purpose of helping students to brainstorm ideas for writing as specified in the suggested AWE-integrated procedure above.

More importantly, it seems that interpretation of AC has formed a key part of teaching and AC is regarded both as a teaching goal and as a standard to reach. Teacher Q compared what she did in the writing class in the past with the present as follows:

> My writing teaching in the past involves only assigning topics and marking essays. I did not provide specific writing requirements and objectives, nor tell students the assessment criteria. After using this system, I have acquired a better understanding of the importance of writing requirements and assessment criteria. (Teacher Q, Data source: post-questionnaire)

Teacher G (i.e. Teacher 1 who taught the senior 1 group) from the high school group related how AC helped with her writing instruction.

> During the experiment, I used the AC in WRM to guide my writing instruction, and the students became aware of the six traits (i.e. ideas and content, structure, word choice, fluency, voice, conventions and mechanics) of AC and attended to them in their writing. Due to the assistance of WRM, my writing teaching is now more guided and standard. (Teacher G, Data source: journals)

The AWE system feedback seemed to be more effective as it was immediate and could possibly help locate the type of problem, assisting students with language form (cf., Grimes & Warschauer, 2010; Li *et al.*, 2015; Wang *et al.*, 2013; Warschauer & Grimes, 2008). The teacher feedback was more concrete, targeted and contextual. The self and peer feedback might help to empower students in self- and other-assessment, and guide them towards student autonomy. Several teachers commented on how the system and teacher feedback could complement each other in teaching:

> Feedback from the system is relatively general. It can tell me roughly where my students are, with reference to native-speaker performance. My feedback is very concrete, related to the topic concerned and the context, with more concern for content and rhetoric. (Teacher Y, Data source: post-questionnaire)

Concurrent with the teaching method changes observed above, teachers also seemed to change in their roles from the dominator in the class and the only assessor of students' essays, to a facilitator, co-assessor, senior learner, co-manager of learning, and researcher, as noted in the following:

> Teachers now hold new roles: facilitators in learning, assessors to fill in the gaps left by the AWE system in its feedback, senior learners concerning the AC, researchers of their own teaching for the sake of improving teaching and self. (Teacher W, Data source: journal)

It is noted that though the teachers followed the suggested procedure of using AWE in the classroom largely, they were inspired by the theory of Exploratory Practice (EP) (Allwright, 2003), encouraged and supported by the HG research team to do action research to examine the efficacy of the proposed AWE integration process and to explore the proper way of AWE integration that suits their context, and they have become "researchers of their own teaching for the sake of improving teaching and self " as related in the journal (Tang *et al.*, 2012).

Discussion

The research demonstrated that the experimental group seemed to outperform the control group in pre-and post-writing tests along with positive changes in the teaching and learning process, which displayed the usefulness of the technology-enhanced formative assessment on learning, the use of AWE in particular, and again might verify Grimes and Warschauer's suggestion of AWE's "utility in a fallible tool" if deployed effectively (2010, p. 4).

Our research indicated the efficacy of AWE as a formative assessment tool in the early drafting and revising process of writing, which reinforced the findings noted by Chen and Cheng (2008). However, our study seemed to move beyond Chen and Cheng (2008) not only in the subject size (60 university students vs. 460 university students) and the subject range (268 high school students were also included in our study), but also in the experimental process of implementing a procedure of integrating AWE into the writing process and evaluating its efficacy through a mixed-methods research design. In conclusion, we have attempted to display the effectiveness of AWE in the drafting and revising process on a larger scale and with a wider range of student cohorts, and proposed and experimented with a procedure of integrating autonomous writing with AWE support tools and revision goals, AWE feedback, teacher feedback and peer feedback at different stages of writing for future research to follow (see the part "The AWE-integrated teaching experiment" under "Research design").

The key to the success of our project might lie in the introduction of two main interventions: the AWE-integrated teaching experiment and teacher training. The underlying rationale was that the introduction of new technology to teaching is not just an issue of technology, rather it concerns various factors, the core of social informatics theory which informed the current study design (Kling, 1996; Warschauer & Meskill, 2000). Five main factors were identified on the basis of those proposed by Kling (1996) and Warschauer and Meskill (2000), however we developed their three factors of "technology, organization and people" by specifying "people" into "teachers" and "students," and adding the factor of "ways of integrating technology into teaching", which we considered crucial to the introduction of any innovations in teaching.

The study might add to the current literature with the following innovations.

First, it seemed that the study exemplified the role of AWE in the classroom within the social-cultural theoretical framework. The study indicated that as a cultural artifact, the AWE tool regulated the writing process through providing assessment criteria

(AC), instant scaffolding feedback, scores and writing assistance tools within the Zone of Proximal Development (ZPD) (Vygotsky, 1962, 1978). The scaffolding role of AWE might be manifested in the dynamic, formative assessment of the writing process, during which students could interact with AWE through the pre-writing, drafting, revising, rewriting, editing and finalizing stages and through interacting with AC and constant practice, hence improving students' understanding of learning to write and writing skills. Moreover, different from previous research on dynamic assessment (e.g., Lantolf & Thorne, 2006), this study adopted an innovative mediator AWE to provide ongoing continuous assessment and feedback, along with teacher and peer support, which ensured that students received multiple and continuous feedback within the ZPD.

Second, our study undertook a mixed-methods approach in research methodology, among which participatory design and exploratory practice were the most salient methods, which again distinguished our study from previous ones (cf. Chen & Cheng, 2008; Warschauer & Grimes, 2010). Rather than being treated as subjects to be researched (e.g., Link *et al.*, 2014; Li, *et al.*, 2015), teachers (including the HG team in our study) were actively involved in the experimental teaching and research process and become action researchers themselves. Many of them researched their own ways of using AWE pertinent to their individual teaching contexts (see Table 3) and published their research papers reporting the exploratory process (see Tang *et al.*, 2012).

Third, teachers and students have made active use of AC of AWE, which seemed not have been mentioned in any AWE research studies so far. Effective assessment needs to have comprehensible and explicit assessment criteria. Communicating assessment criteria to students is always an important principle of effective assessment (e.g., Brown, Race, & Smith, 1996). During the experiment, many teachers revealed that they did not have a set of clear assessment criteria like WRM AC to mark students' essays prior to our study. The high school teachers tended to use the essay assessment criteria for marking the college entrance exam (Gaokao), and the college teachers for marking the College English Test Band 4 (CET4). Of those who did have AC, the criteria were usually not communicated to the students clearly in the understanding either students had known about Gaokao AC and CET4 AC or students might not bother to know about it. The experimental groups demonstrated through teacher quotes that understanding and interpreting AC constitutes an important part of teaching writing. With AWE serving as both an assessment and teaching tool and AWE AC as both an assessment standard and a teaching goal, teachers and students seemed to become co-assessors, consequently students became more autonomous through interacting with the system and assessing their own works continuously in the system, and teachers changed their roles toward that of a facilitator, co-assessor and co-learner.

Implications and conclusion

It might be argued that only when we have attended to the instructional process, can we understand how technology can assist teaching and learning and how it can effect changes in teaching and learning. AWE-integrated experimental teaching as shown in this study enabled students and teachers to attend to the writing process, during which the AWE tool intervened in the writing procedure and served as a teaching assistant by offering continuous writing assistance and dynamic assessment, thus enhancing writing instruction efficiency. More importantly, via continuous interaction with AWE, it seemed that students have learned to correct their own mistakes and improved their autonomy judging from the student questionnaire and interview data as reported in the "Results" part.

Our research also reiterated the claims that the introduction of technology is not just a technical issue per se (Grimes & Warschauer, 2010; Kling, 1996; Warschauer, 2012), it concerns various interrelated factors relevant to it: organization, technology, teachers, students, and ways of technology integration. It was demonstrated through our research that only when teachers have acquired a good understanding of technology role and can apply it properly, does technology act as a catalyst for positive changes in teachers and teaching.

Notes

1. The number in Experimental and Control columns in Table 1 and 2 refers to the number of students who have completed both the pre- and post-tests. The total number of students enrolled in the class is more than this.

Acknowledgements

This paper is part of a China national educational sciences planning research project titled "Investigating the Application of Online Formative Assessment in the EFL Classroom" (Grant number: GFA097005).

The first author is indebted to China Scholarship Council for sponsoring her research visit to the Open University, UK from September 2016 to July 2017, during which she completed this paper.

The authors would also like to express their sincere gratitude for the great support provided by Professor Yi'an WU, Beijing Foreign Studies University, China; Ms. Yihong WANG, CTB/McGrawHill, the USA; Professor Mark Warschauer, University of California, Irvine, the USA ; and Professor Mike Sharples, the Open University, the UK during the project implementation and manuscript revision.

The authors are also grateful to the two anonymous reviewers for their insightful feedback on the earlier versions of the paper.

References

Allwright, D. (2003). Exploratory practice: Rethinking practitioner research in language teaching. *Language Teaching Research, 7(2),* 113–141.

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

Beatty, I. D., Feldman, A., Leonard, W. J., Gerace, W. J., St. Cyr, K., Lee, H., & Harris, R. (2008). Teacher learning of technology-enhanced formative assessment. Paper presented at *The NARST 2008 Conference.*

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

Brown, S., Race, P., & Smith, B. (1996). An assessment manifesto. 500 Tips on Assessment. Retrieved from http://www.city.londonmet.ac.uk/deliberations/assessment/manifest.html (18 August, 2006)

Carless, D. (2006). Pre-emptive formative assessment. Retrieved from http://www.iaea2006.seab.gov.sg/conference/download/papers/Pre-emptive%20formative%20assessment.pdf (23 July, 2007)

CCCC Executive Committee. (2004). CCCC position statement on teaching, learning, and assessing writing in digital environments. NCTE (National Council of Teachers of English) Position Papers.

Chen, E.E., & Cheng, E. (2008). Beyond the design of automated writing evaluation: pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12(2),* 94–112.

Creswell, John W. & Plano-Clark, Vicki L. (2006). *Designing and conducting mixed methods of research,* Thousand Oaks, CA: SAGE Publications.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18(1)*, 7–24.

Elliot, S.M. (2003). IntelliMetric: from here to validity. In M.D. Shermis & J. Burstein (Eds.), *Automatic essay scoring: A cross-disciplinary approach* (pp. 71–86). Lawrence Erlbaum Associates.

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.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8(6),* 4–43.

Hegelheimer, V., Dursun, A., & Li, Z. (Eds.). (2016). Automated writing evaluation in language teaching: Theory, development, and application. *CALICO Journal, 33(1),* i-141.

Heilman, M., & Tetreault, J. (2012). Using automated scoring to analyze student writing. Paper presented at Georgetown University Round Table on Languages and Linguistics (GURT) 2012, Washington DC.

Jin, Y. (2010). The formative assessment function of standardized language tests. Lecture at Beijing Jiaotong University, Beijing.

Kling, R. (1999). What is social informatics and why does it matter? [J] D-Lib Magazine, 5(1). Retrieved from http://www.dlib.org:80/dlib/january99/kling/01kling.html (28 May , 2012)

Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing, 18(1),* 62–84.

Lantolf, J.P., & Thorne, S.L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford: Oxford University Press.

Li, J.H. (2009). *Using the computer to score Chinese students' English essays via using the latent semantic theory*(Unpublished doctoral dissertation). Guangdong Foreign Studies University.

Li, J.R., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27,* 1–18.

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44,* 44–78.

Link, S., Dursun, A., Karakaya , K., & Hegelheimer, V. (2014). Towards best ESL practices for implementing automated writing evaluation. *CALICO Journal, 31(3),* 322–344.

Liang, M. C. (2011). *Developing the automated system for scoring English essays.* The Higher Education Press, China.

Liang, M. C. (2016). Writing right with iWrite. Paper presented at 2016 International Symposium on Computer-assisted Language Learning. July 22–23, Qingdao, China.

Liu, S., & Kunnan, J. A. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *CALICO Journal, 33(1),* 71–91.

Page, E. (1994). Computer grading of student prose using modern concepts and software. *Journal of Experimental Education, 62(2),* 127–142.

Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing, 18(1),* 40–61.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18(1),* 25–39.

Rich, C. S., Harrington, H., Kim, J., & West, B. (2008). Automated essay scoring in state formative and summative writing assessment, paper presented at Annual Conference of American Educational Research Association, New York.

Rowntree, D. (1987). *Assessing students: How shall we know them?* Kogan Page, London.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, New Jersey: Lawerence Erlbaum Associates, Publisher.

Steen, M. (2013). Virtues in participatory design: Cooperation, curiosity, creativity, empowerment and reflexivity. *Sci Eng Ethics* 19: 945–962.

Tang, J. L. (2014). How to integrate an automated writing assessment tool in the EFL classroom? *Foreign Language Learning Theory and Practice, 1,* 117–125.

Tang, J. L., & Rich, C. S. (2011). Online technology-enhanced English language writing assessment in the Chinese classroom. Paper presented at Annual Conference of American Educational Research Association, New Orleans.

Tang, J. L., & Wu, Y. A. (2012). Using automated writing assessment in the college EFL classroom. *Foreign Languages and Their Teaching, 265(4),* 53–59.

Tang, J. L. *et al.* (2012). *English teaching reform in the digital age – the application of educational assessment technology in English writing instruction.* Foreign Language Teaching and Research Press.

Tang, J. L., & Wu, Y. A. (2011). Using automated writing evaluation in classroom assessment: a critical review. *Foreign Language Teaching and Research, 43(2),* 273–282.

Tang, J. L. (2011). A survey on teacher perceptions of ICT and writing instruction. Paper presented at 2011 Symposium on Using AWE in College Writing Instruction, July 21–22, Beijing, Beijing Foreign Studies University.

Vantage Learning. (2007). My access! efficacy report. Retrieved from http: //www.vantagelearning.com/school/research/myaccess.html (8 August, 2010)

Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26(3),* 234–257.

Warschauer, M., & Meskill, C. (2000). Technology and second language learning. In J. Rosenthal (Ed.), *Handbook of undergraduate second language education* (pp. 303–318). Mahwah, New Jersey.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3,* 22–36.

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

Warschauer, M. (2012). Writing to learn and learning to write. Plenary speech presented at 2012 GloCALL Conference and 2012 International Symposium on CALL, October 18–19, Beijing, China.

White, L., Hixson, N., D'Brot, J., Perdue, J., Foster, S., & Rhudy, V. (2010). Research brief, impact of Writing Roadmap 2.0 on WESTEST 2 online writing assessment scores. Retrieved from http://wvde.state.wv.us/oaa/pdf/research/Research%20Brief%20-%20WRM2.0%20Impact%20FINAL%2001.27.10.pdf (20 September, 2010)

Wu, Y. A. & Tang, J. L. (2012). Impact of integrating an automated assessment tool into English writing on university teachers. *Computer-Assisted Foreign Language Education, 146 (4):* 3–10.

Author biodata

Jinlan Tang, PhD, is the Deputy Dean and Professor of English in the School of Online and Continuing Education, Beijing Foreign Studies University, China. Her research has covered the areas of language assessment, tutor feedback, EFL teaching and learning in the e-learning environment. She also serves as the Secretary-General of the China Computer-Assisted Language Learning Association (ChinaCALL, www.chinacall.org.cn) (2016–2020).

Changhua Sun Rich, PhD, is a principal research scientist at ACT, USA. She works collaboratively with ACT research, test development, and international programs to conduct international assessment research. Prior to joining ACT, she was a research director at CTB/McGraw-Hill, USA and worked on automated essay and speech scoring applications in China.

Appendix A

Essay prompts for the pre and the post test

For Senior 1/Non-English Major First Year University Students
Prompt A: Hobby or Sport (Informative)
If you could be good at any hobby or any sport, what would it be? Explain why in detail.

Prompt B: Enjoyable Work (Informative)
Think about one specific career that might interest you. Explain why you would enjoy doing this type of work in the future.

For Senior 2/Non-English Major Second Year University Students
Prompt A: Is Progress Always Good (Informative)
Think about all the new technology we have in our lives. Is technological progress always good? In an essay, write about one kind of technology – maybe MP3 players, or the latest computers. Then give at least one reason why this new technology leads to progress – or why it doesn't lead to progress. Give examples and details to support your reasons.

Prompt B: Act of Kindness (Informative)
In an essay, discuss the effects of a recent act of kindness that you either saw or took part in. Give details about the act itself and its effects.

For the English Major Second Year University Group
Prompt A: Art or Computer Science (Persuasive)
A local school must make a choice between offering art or computer science classes. Which do you think it should be? Write a letter to the principal to persuade him or her to offer the class you choose. Give sound reasons to support your views.

Prompt B: Littering Problem (Persuasive)
Littering can be a problem. Some places look bad because there is a lot of trash lying around. If people did not litter, there would be less trash. Write an essay to convince people not to litter. Your writing should be convincing.

Appendix B

The post-experiment student questionnaire

Student questionnaire on the Use of Writing Roadmap (WRM) in the Class

Dear student,

This questionnaire intends to acquire your experiences with the use of WRM in your class. Please answer the questions truthfully. Your answers will be kept strictly confidential and used for research purposes only. Thank you!

From the project team

I. Personal Information. Please tick ✓ in the appropriate place or fill in the blank space.

1. Name: ________________
2. Gender: A. male B. female
3. Grade: A. Grade one B. Grade two
4. English score at the national entrance test for college __________
5. Major ________________ *
6. Name of your English teacher______________________
7. Name of the university ___________________

* Questions 4–5 do not apply to the high school students. Except for these two, the rest of the questions are the same for both groups.

II. Perceived efficacy of English language teaching. Please tick (✓) in the appropriate box.

		Strongly disagree	Disagree	Agree	Strongly agree
8.	The English class helps me know English history and culture.				
9.	I did not learn much from the English class.				
10.	I have improved my English communication ability through the English class.				
11.	The English class only helps me learn some grammar rules and expressions.				
12.	I always look forward to having the English class.				
13.	Teacher feedback on my assignment is very timely.				
14.	Teacher written feedback on my essay is very helpful to me.				
15.	I like to write in English more.				

III. Perceptions on English language learning. Please tick (✓) in the appropriate box.

		Strongly disagree	Disagree	Agree	Strongly agree
16.	I learn English in order to get a good grade.				
17.	I like learning English.				
18.	I do not like to listen to teachers lecturing too much, but prefer to join the English class activities.				
19.	I am not gifted in learning English.				
20.	English ability can only be improved through use, rather than by listening to lectures.				
21.	My English should be assessed by my teacher, rather than by myself or my classmates.				
22.	Teachers should play the main role in English teaching, and the students should assist the teacher.				
23.	The success or failure of my English learning lies in myself.				
24.	My English will improve if I study hard.				
25.	I do not like quizzes during the term.				
26.	I do not think that self-correction can help improve my English writing.				
27.	I do not think that English writing course is necessary to me.				
28.	The computer and the Internet help with English writing with their functions of storage, searching and revision.				
29.	My English will improve if I got to know good English learning methods.				
30.	Teacher feedback will help with my English writing.				
31.	The computer and the Internet can help improve English learning.				
32.	I get very worried before and during the exam.				
33.	I do not like correcting my own essays.				
34.	Commenting on other students' work help improve my English writing a lot.				
35.	The computer and the Internet cannot provide reliable and effective assessment of English essays.				
36.	Good essays result from continuous revisions.				

IV. Experiences with the English writing course. Please tick (✓) in the appropriate box.

37. How do you feel about the use of WRM in your writing?

A. Very satisfactory B. Satisfactory
C. Moderate D. Not Satisfactory

38. I pay most attention to ________________ after submitting the essay.
   A. Score B. Report
   C. Both score and report D. Neither score nor report
39. Do you like revising essays?
   A. Yes, I like to. B. No, I don't like to. C. Hard to say.
40. Do you try to avoid the problems your teacher has pointed out in your writing?
   A. Yes, I try hard. B. I sometimes try.
   C. Basically I do not try. D. No, I have no idea about it.
41. Do you like to revise essays with teachers and peers in the class?
   A. Yes, I like to. B. No, I don't like to C. Hard to say.
42. Are you willing to continue practicing English writing online?
   A. Yes, I am willing to. B. No, I am not willing to. C. Hard to say.

V. Experiences with WRM. Please tick (✓) in the appropriate box.

		Strongly disagree	Disagree	Agree	Strongly agree	Do not know
43.	I like using WRM to practice writing.					
44.	I write more after using WRM.					
45.	I revise my essays more after using WRM.					
46.	I feel more confident about my writing competence.					
47.	I think the scores given by WRM are fair.					
48.	WRM helps to improve my writing skills.					
49.	I can understand the scores given by WRM.					
50.	The advice given by WRM helps me revise my essays.					

VI. Use of WRM in writing. Please tick (✓) in the appropriate box.

51. Do you revise essays in WRM before submission?
   A Yes B. No
52. How many times did you revise your essays in WRM before submitting to your teacher?
   A. More than five times B. Three to four times
   C. Once or twice D. Never
53. I think the ____________ is the most helpful tool in WRM.
   A. Hint B. Tutor C. Thesaurus D. Grammar Tree
54. Sort the tools below from the most important to the least important, and write your answers in the blank.
   A. Hint B. Tutor C. Thesaurus D. Grammar Tree
55. I ____________ WRM to correct punctuations and format errors.
   A. never used B. seldom used
   C. half the time used D. frequently used

56. I ____________ WRM to correct spelling errors.
    A. never used B. seldom used
    C. half the time used D. frequently used
57. I ____________ WRM to correct grammar errors.
    A. never used B. seldom used
    C. half the time used D. frequently used
58. I ____________ WRM to improve my wording.
    A. never used B. seldom used
    C. half the time used D. frequently used
59. I ____________ WRM to improve my essay content and structure.
    A. never Used B. seldom used
    C. half the time used D. frequently used
60. I feel I have made ______________ progress in my writing skills this term.
    A. great B. some C. no
61. I have made the greatest progress in _______________ through a term's study.
    A. choice of words B. grammar C. spelling
    D. use of punctuation marks E. structure
    F. ideas and content
62. I think _______________ helps me most in writing through a term's study.
    A. feedback from WRM B. teacher feedback
    C. classroom activities D. peer feedback
63. Through a term's study, what do you like about WRM best?
64. Regarding the use of WRM in writing, have you got any other experiences or suggestions that you would like to share?

Appendix C

Student interview prompts

1. What are your experiences with the WRM-integrated teaching experiment?
2. Do you revise your essays in WRM?
3. How many times do you revise your essays in WRM?
4. In revising your essays, what do you pay more attention to: the revision of grammar and spelling or the revision of structure and ideas?
5. Is WRM helpful for improving your writing?
6. What do you think WRM helps your writing the most?
7. Do you think WRM helps with your essay theme and structure?
8. Does WRM or WRM assessment criteria help you understand what is a good essay?
9. What do you think is the teacher role in the experiment?
10. Can WRM replace your teacher?
11. Do you read carefully teacher feedback on your essays?
12. How do you respond to WRM feedback?
13. If you were given a student essay, are you confident in using the WRM assessment criteria to score it?
14. What are your suggestions for improving WRM?

Appendix D

*The post-experiment teacher questionnaire**

Teacher questionnaire on the WRM-integrated teaching experiment

Dear Teacher,

It will be a whole academic year in a week's time after you have undertaken WRM-integrated writing teaching experiment and we would like very much to learn about your experiences about using WRM, hence we design this questionnaire and would appreciate very much your feedback. Thank you!

From the project team

Instructions: The questionnaire consists of two parts: the overall and the specific aspects. Please give brief answers to the questions in the overall aspect and specific answers to the questions in the specific aspect.

I. The overall aspect

1. How did you apply WRM into your teaching? And why?
2. In which areas does WRM assist your writing teaching?
3. Has your writing teaching changed during the process? If yes, what are the changes? Please tick (✓) in the appropriate place.
   1) teaching mode
   2) teacher-learner interaction: form, content, frequency, etc
   3) teaching objectives
   4) assessment criteria
   5) assessor
   6) instruction effect
   7) commitment in teaching
   8) others, please specify.
4. Have you changed your understanding of writing teaching? If yes, what are the differences?
5. What difficulties did you meet in your writing teaching in the past year?

II. The specific aspect

1. Assessment
   1) Do you have writing assessment criteria before this experiment, how do you like it?
   2) Do you agree to WRM assessment criteria? What are the differences between this one and the one you used before?
   3) When did you introduce the assessment criteria to students, how did you do it?
   4) Can your students understand each dimension of the assessment criteria?
   5) How did you help students understand and internalize the WRM assessment criteria?
   6) How did you like the idea of giving students clear assessment criteria at the beginning of the semester?

* The tutor questionnaire was designed by Professor Yi'an Wu, advisor and key member of our project team. The authors were grateful for her approval of its use in the study

7) To what extent do you think your students have mastered the assessment criteria so far?
8) Do your students have the opportunity to assess their own work? What about peer assessment? What are their advantages and disadvantages?
9) Do students agree to the scores offered by WRM? If some students question the scoring, how do you respond?

2. Teacher and learner interaction
   1) Has the interaction between you and the students increased or decreased after you have used WRM? Are there any changes, please describe.
   2) Are there any changes with the interaction pattern? If yes, please describe.
   3) Are there any changes with the interaction content? If yes, please describe.
   4) Are there any changes with the relation between you and the students? If yes, please describe.
   5) After the use of WRM, are there any changes with your understanding of students? If yes, please describe.

3. Teaching mode
   1) Has the introduction of WRM changed your teaching procedure?
   2) If yes, what changes are they?
   3) Do you need to adjust your teaching from time to time in the class? Please specify.
   4) In your class, have you ever encountered the situation in which students respond differently toward WRM. If so, how did you deal with it?
   5) With WRM assistance, do you know your students' problems in the learning process better?
   6) How do you deal with these problems? Please give an example.

4. Perceptions on writing teaching
   1) What difficulties have you met after you used WRM, how did you tackle these difficulties?
   2) Has your understanding of teaching objectives changed before and after the experiment, please specify.
   3) Have your teaching methods changed after you used WRM, what are the changes and why?
   4) What are teacher and students roles in your teaching after you used WRM, how different are they from before?
   5) What do you think are the effective writing teaching methods? Do you also hold this view before?
   6) How do you feel about using technology in writing instruction?
   7) What other puzzles do you have in writing teaching?

5. Perceptions over writing pedagogy
   1) What are the key factors to improve writing ability?
   2) How do you view the role of student correction of essays in writing instruction?
   3) How do you view the role of WRM in writing instruction, how can WRM feedback and teacher feedback complement each other?
   4) How much do your students use your feedback and WRM feedback?

5) How much do your students use WRM?

6. On learner autonomy
   1) Do you find any changes in learner autonomy?
   2) If there are changes, what are the changes?
   3) Does WRM arouse your students' interest in learning? Please specify.
   4) What types of students are more interested in WRM?
   5) What types of students are not interested in WRM?
   6) Does students' interest in WRM relate to their scores, characters, gender, computer skills?

7. Through one-year teaching experiment, what do you think are the key factors for the successful use of WRM in teaching?

8. Do you have any other thoughts or experiences to share?

9. If possible, are you going to continue the use of WRM in your teaching?

10. What suggestions do you have for further improving WRM?

Appendix E

Teacher interview prompts

1. Before the experiment starts, what do you predict WRM can assist your teaching?
2. Do you think that WRM has met your expectations?
3. You mentioned in the questionnaire that WRM is having a far-reaching influence on teaching, what are the influences?
4. Has your writing teaching changed during the process? If yes, what are the changes?
5. Do you have a clear set of writing assessment criteria before the experiment?
6. What are the differences between WRM assessment criteria and the assessment criteria you used before?
7. Do students question WRM scoring? If yes, how do you respond?
8. How do you help students understand and internalize the assessment criteria?
9. Do you still mark students' essays during the experiment?
10. Which will students remember better, WRM correction or teacher correction?
11. After using the system, has teacher-learner interaction increased or decreased?
12. Has the use of WRM enhanced learner autonomy?
13. Have teacher and student roles changed during the process?
14. Have you met any difficulties during the teaching experiment?
15. What do you think is the most effective way of teaching writing? Do you still hold this view?
16. What are the key factors toward improving writing proficiency?

Appendix F

Teacher journal template

School/University:
Marking Date:
Teacher:
Class:
Essay prompt:

Observations and reflections (which can center on the following three points):

1. Students' overall performance in this essay (i.e. what they have done well, what problems they have, and what they need to improve)

2. Experiences and reflections on marking this essay (i.e. in what areas the teacher has done well and what area needs improvement; whether students like to practice writing and revise in WRM, whether they have problems in writing, what type of problems; whether they ask the teacher questions, what type of questions they might ask)

3. Students use of the online assessment system (i.e. students' attitude toward WRM, what type of students like using WRM, what type of students do not like or are afraid of using WRM)