International Conference "ICT for Language Learning" 6th edition

Automated Scoring of EFL Learners' Written Performance: a Torture or a Blessing?

Roya Khoii[1], Amir Doroudian[2]
[1]Islamic Azad University, North Tehran Branch; [2]Islamic Azad University, South Tehran Branch (Iran)
roya_kh@yahoo.com, amirdoroudian@aol.com

Abstract
Automated Writing Evaluation (AWE) systems, which score essays and generate feedback, have been developed to meet the challenge of evaluating learners' written performance. The present research compares the effects of automated and human scoring of L2 written performance with each other. Twenty two EFL learners equally divided in two essay writing classes participated in this study. The experimental class used the AWE program My Access!® as the scorer of essays and students' main source of feedback, while the control class used a teacher for similar purposes. Both classes had the same teacher and used the same materials.The 10-week treatment period commenced with an essay writing pre-test to measure the students' writing ability. At the end of the treatment, they received an essay writing post-test to evaluate their level of progress in terms of writing. A questionnaire was also given to the experimental students to learn about their attitudes toward AWE. Lastly, a delayed, timed paper-pencil essay writing post-test was administered to them to check the effect of the treatment in a test situation. Based on the results, the experimental group had significantly outperformed the control group on the post-test; however, no significant difference was found between the two groups' timed post-test essay lengths and scores. The students had positive attitudes towards AWE's practicality, and its effect on writing improvement; however, they were barely satisfied with AWE feedback.

1. Introduction
With the widespread availability of the Internet, the educated urban population in much of the world needs to learn to write in English for different purposes [1]. Compared to other language skills, writing is considered the most complex and difficult not only to master but also to teach. A primary challenge in most writing programs is the provision of effective feedback, which is immediate, fair, detailed, and frequent [2]. The recent advances in Intelligent Computer-Assisted Language Learning (ICALL) have resulted in a the development of a range of Automated Writing Evaluation (AWE) programs to evaluate learner's writing, assign a score to it, and provide them with immediate feedback. AWE, according to its developers, can ease teachers' workload and, thus, allow more writing practice and faster improvement [3]. However, despite the growing interest in integrating AWE in instruction, relatively little research has been conducted on its effectiveness, and, even less, on the students' perceptions of it.

2. Feedback to Writing
Feedback to writing is defined as "input from a reader to a writer with the effect of providing information to the writer for revision" [4]. The functions of feedback include the evaluation of students' achievements, development of students' competences, and elevation of students' motivation and confidence. As effective feedback needs to be timely, constructive, motivational, personal, and given comprehensively throughout the writing process, its provision is extremely challenging for the teacher and is impossible with large numbers of students. Moreover, although learners value teachers' form-focused corrections, they are mostly dissatisfied with the quality of teachers' feedback on other aspects of their writing [5]. Developments in Natural Language Processing (NLP), the changing role and importance of L2 writing, the relatively low reliability and the high cost of human rating have all led to a growing interest in the use of machine scoring as a supplement, and even a replacement, to human scoring of essays [6].

3. Automated Writing Evaluation (AWE)
AWE is the use of computer programs to evaluate the quality of an essay by providing a score, a detailed evaluation of essay features, or both [7]. The earlier AWE systems, like the pioneering PEG™, used statistical procedures to analyse surface features of a text and predict the score given by

human raters to a set of similar essays [8]. One of the most important shortcomings of these systems is a disregard for content which undermines the construct validity of its assessment. Newer AWE systems, on the other hand, employ an NLP technique called Latent Semantic Analysis (LSA), which, unlike the statistical models, is based on comparing the semantic content of words used in essays. Despite the dramatic overhaul LSA brought to AWE systems in terms of content evaluation, at the current stage, the AWE tools are still weaker than human raters when it comes to scoring the content of essays and in evaluating works written in non-testing situations [1].
A key advantage of AWE could be its immediate feedback on various aspects of writing, which may lead to more revisions and writing practice [2]. Instructional AWE programs also offer a range of writing tools that can assist the writer during the writing process. However, AWE's critiques maintain that, aside from its limited capability to evaluate content, it eliminates the human element from the process of writing assessment [1], which may send the wrong message to students that their writing is not important since their audience was replaced by a machine. It may also result in formulaic writing. AWE has also been criticized for limiting teacher's range of topic choice and stifling teacher's creativity, because it can only accurately evaluate essays written to specific prompts that come with the program [2]. Finally, AWE discriminates against students who are less familiar with using technology to write or complete tests [9].

3.1. My Access!®

My Access!® is a web-based instructional writing program by Vantage learning that relies on IntelliMetric™, an NLP-based scoring engine also developed by Vantage Learning. It provides immediate feedback, along with holistic and analytical scoring on submitted essays. My Access!® also offers instructional writing tools for students to use during the writing process, including Venn diagrams, progress-tracking tools, pre-writing activities, an spell-checker, and a word bank [10]. It offers over 700 unique prompts in narrative, persuasive, and informative genres for different levels. It has also made it possible for teachers to create their own prompts and write their comments on the students' essays [10].

3.2. Classroom Research on AWE

In the recent years, AWE programs have increasingly influenced writing instruction [1]. The body of research on AWE, however, has mainly focused on its psychometric issues in summative assessment, usually on its construct validity and rate of agreement with humans to validate it for high-stake purposes [11]. Funded large-scale research on AWE as a classroom tool for native speakers generally reports high rate of student/teacher satisfaction and positive effects on writing quality [7]. However, Warschauer and Grimes [2] found AWE to be "a modest addition to the arsenal of teaching tools and techniques at the teacher's disposal". The scant number of AWE studies on EFL learners have resulted in conflicting findings; some report that students find AWE a satisfactory and effective tool in increasing the students' scores [12], while others report that it creates no statistically significant effect on L2 learners' writing or attitudes towards this skill [13].

4. The Study

4.1. Research Questions

The present research was conducted in order to answer the following questions:

- Does the employment of AWE and human scoring of essays produce significantly different effects on the improvement of EFL learners' writing skill?
- What are the student-writers' attitudes toward using an AWE program as the scorer and the main source of feedback to their essays?

4.2. Participants

The participants of this study consisted of 22 Iranian advanced EFL learners with low to average computer skills in two intact essay writing classes randomly assigned to a control group and an experimental group.

4.3. Instrumentation

The following instruments were used to collect the required data:

a. An essay writing pre-test

b. An essay writing post-test
c. A delayed, timed writing post-test
d. A 20-item attitude-to-AWE questionnaire
e. *Vantage 6-point Independent Writing Rubric*

4.4. Materials

Both groups had the same teacher and used the same textbook, *College Writing Skills* by John Langan, 2004, and were given seven essay writing prompts. In the experimental group, My Access!® was used as the essay scorer and the main source of feedback. A My Access!® user guide, along with individual training sessions, was also provided to the experimental students before the treatment.

4.5. Procedure

Initially, an essay writing pre-test was given to both groups to measure their writing ability at the outset of the study. During the 10-week study, a process-oriented approach to writing and oral generic feedback was used to teach writing to both groups 90 minutes a week. They wrote 7 essays on topics chosen from My Access!® prompts and other sources. The control group received scores and feedback by the teacher who used the same rubric as My Access!®. The experimental students could use My Access!® for 90 minutes a week to score, receive feedback and revise their essays; however, they did not use it for more than 40 minutes a week. The experimental students submitted an original and a revised draft per essay. At the end of the treatment, a post-test was administered to both groups to measure the effect of the treatment on their writing skill. The experimental students also completed an AWE-attitude questionnaire. Finally, to measure the effects of the treatment in a test situation, the students sat a paper-pencil timed post-test over a three-week interval.

5. Data Analysis and Results

5.1. Pre-test

Initially, an essay writing pre-test was administered in order to measure the participants' writing skill before the treatment and ensure their homogeneity. As shown in Table 1, the results of an independent-samples *t*-test (t(20)=.179, *p*=0.86 (two-tailed)) revealed that there was no significant difference between the two groups' means scores (4.20 and 4.25).

Table 1. Independent-Samples t-test for the Pre-test

t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
-.179	20	.860	-.043	.244

5.2. Post-test

At the end of the treatment, an essay writing post-test was administered to both groups in order to measure the impact of the treatment on the students' writing skill. The mean scores of the experimental and control groups on this test were 5.16 and 4.43, respectively. As given in Table 2, the results of another t-test (t(20)=4.64, *p*=.000 (two-tailed)), indicated that there was a significant difference between the means of the two groups, with the experimental group having outperformed the control group.

Table 2. Independent-Samples t-test for the Post-test

t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
-4.64	20	*.000	-.73	.157

* The mean difference is significant at the .05 level.

5.3. Delayed, Timed Post-test

To investigate the impact of the treatment on the students' writing skill in a test situation, a delayed, timed paper-pencil essay writing post-test was administered to both groups three weeks after the post-test. The teacher scored the scripts of both groups; however, the experimental group's essays were

also scored by My Access!® and an inter-rater reliability of .79 was estimated. The mean scores of the experimental and control groups on this test were 4.22 and 4.23, respectively. The results of an independent-samples *t*-test (t(22)=.04, *p*=.96 (two-tailed)) indicated that there was no significant difference between the two groups' mean scores on this test (Table 3).

Table 3. Independent-Samples t-test for the Delayed Timed Post-test

t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
-0.041	20	.968	-.009	.222

5.4. AWE-Attitude Questionnaire

To answer the second research question, a 20-item questionnaire was given to the experimental group. There were 18 Likert-scale items designed to elicit the experimental group's attitude toward AWE regarding its practicality, numerical grading, effect on writing improvement, feedback satisfaction, feedback value, and revision behaviour. There were also two open-ended items asking the students about the best and the worst things about My Access!®. 91% of them considered AWE quite practical; 82% wished it to be integrated in future courses, and 64% of them were satisfied with the grading system, they could not decide if the scores were fair. Besides, the low- and mid-performing students believed AWE increased their writing confidence and improved their writing (64%). 55% of them considered the feedback to be helpful; 73% thought it had good suggestions for improvement, yet most (64%). The students might have had difficulty understanding the feedback, as they received lower scored on revised drafts in several occasions. Finally, 73% of them reported more revisions by adopting the editorial feedback, although almost all the revisions were superficial and partial. The most dissatisfied students with AWE were the highest- and lowest-performing students with computer skills. Interestingly, the most satisfied students were the low- and high-performing students who used the program more extensively.

As for the open-ended items, the most frequent answer to "the best thing" was *"immediate scoring and feedback"* (82%), and the most frequent answers to "the worst things" were *"misdetection of errors," "lack of a personal account"* and *"requiring high-speed connection"* (27%). It should be noted that error misdetection occurred in mechanics due to the program's inability to recognize proper nouns and, sometimes, bound morphemes.

6. Conclusion

Consistent with Ware's study [12], this study found that L2 learners who use My Access!® to score and revise their essays outperformed those who receive scores and feedback from a teacher. However, this finding was not repeated in a delayed paper-pencil test situation, implying that AWE's environment is central to the students' improved performance and casting doubt on AWE's long-term effects. Moreover, the students who used AWE revised their drafts more frequently in response to My Access!® editorial feedback; however, the revisions were almost always partial and limited to the sentence-level superficial features. Thus, it never led to more writing or revising the content and organization. This might be due to the generic nature of AWE feedback on those aspects of writing, which also partly caused the students' dissatisfaction with it, or the students' insufficient experience with the program.

The most positive effects of AWE were boosting the students' writing confidence, involving them in the writing process, motivating them to submit more drafts, and helping struggling writers to reduce their mechanical errors. However, it was observed that AWE users still expected the teacher to read their essays, supporting the idea that "The need for sensitive human readers will not disappear, no matter how closely automated scores approximate scores by expert human graders" [1].

Unfortunately, AWE did not save teacher time, as the teacher still had to read every essay and give oral feedback to the whole class, as well as provide them with accounts at agreed times every week, which required plenty of time and energy. Besides, low-speed connection, frequent error misdetections, and lack of personal accounts, a limitation that also rendered My Access!® writing and progress-tracking tools useless, were among the factors discouraging the students from using My Access!®.

All in all, it was concluded that using AWE did not save teacher time and energy and did not lead to longer essays, profound revisions, or improved content and organization. Yet, it was helpful in boosting learners' writing confidence and motivation and reducing sentence-level errors. Obviously,

AWE's effects on writing development in test situations and outside its environment are questionable, and the program might not safely replace the teacher.

References

[1] Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 1-24. Retrieved from www.gse.uci.edu/person/warschauer_m/docs/AWE.pdf
[2] Leki, I. (1991). The preferences of ESL students for error correction in college level writing classes. *Foreign Language Annals*, 24, 203-218.
[3] Warschauer, M. and Grimes, D. (2008). Automated essay scoring in the classroom. *Pedagogies: An International Journal,* 3, 22–36.
[4] Keh, C. L. (1990). Feedback in the Writing Process: A Model and Methods for Implementation. *ELT Journal*, 44(4), 294-304.
[5] Huxham, M. (2007). Fast and effective feedback: are model answers the answer?. *Assessment & Evaluation in Higher Education*, 32(6), 601-611.
[6] Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, 30(1), 125-141.
[7] Page, E. B. (2003). Project Essay Grade: PEG. In Mark D. Shermis & Jill C. Burstein (Eds.). Automated essay scoring: a cross disciplinary approach. Mahwah, NJ: Lawrence Erlbaum Associates.
[8] Weigle, S. C. (2002). Assessing writing. Cambridge, UK: Cambridge University Press.
[9] Anson, C., Filkins, S., Hicks, T., O'Neill, P., Pierce, K. M., & Winn, M. (2013). NCTE Position Statement on Machine Scoring. NCTE Position Statement Special Issue. 1-12.
[10] Vantage Learning. (2007). My Access!: Efficacy Report. Newtown, PA: Vantage Learning.
[11] Keith, T. (2003). Validity and automated essay scoring systems. In Shermis, M.D. & Burstein, J.C., editors, Automated essay scoring: a Cross-Disciplinary Perspective. Lawrence Erlbaum, 147-167.
[12] Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly*, 45(4), 769-774.
[13] Lee, C., Wong, K., Cheung, W., & Lee, F. (2009). Web-based essay critiquing system and EFL students' writing: A quantitative and qualitative investigation. *Computer Assisted Language Learning*, 22, 57-72.