English Language Teaching; Vol. 15, No. 7; 2022
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Effectiveness of the Automated Writing Evaluation Program on Improving Undergraduates' Writing Performance

Jingxin Geng[1] & Abu Bakar Razali[1]

[1] Faculty of Educational Studies, Universiti Putra Malaysia, Selangor, Malaysia

Correspondence: Jingxin Geng, Faculty of Educational Studies, Universiti Putra Malaysia, Selangor, Malaysia.

Received: May 20, 2022 Accepted: June 10, 2022 Online Published: June 13, 2022

doi: 10.5539/elt.v15n7p49 URL: https://doi.org/10.5539/elt.v15n7p49

Abstract

Automated Writing Evaluation program (AWE) has gained increasing ground in ESL/EFL writing instruction because of its instructional features, such as the instant automated writing score system and the diagnostic corrective feedback in real-time for individual written drafts. However, there is little known about how the automated feedback provided by the AWE program can impact students' writing performance in an authentic classroom and how to make the most of it to improve students' writing performance effectively, especially for ESL/EFL undergraduate students. This paper attempts to offer an overview of the investigation of the effectiveness of automated feedback via a literature review. According to the inclusion and exclusion criteria, eleven articles published in the past five years were finally included for the analytical synthesis. The literature review matrix for the synthesis reveals the research gaps of the previous literature in the levels of the effectiveness of the automated feedback, including the lack of the design of delayed post-test, writing performance in terms of writing traits, and students' writing strategies regarding the use of AWE program. The conclusion highlights the need for future research by bridging the gaps of exploring the long-term internalized impact of the embedded use of automated feedback and an advanced teaching method on improving both students' overall writing performance and analytic writing scores.

Keywords: Automated Writing Evaluation, automated feedback, ESL/EFL writing, academic writing

1. Introduction

With the support of sophisticated language processing technology, i.e., Natural Language Processing (NLP) and Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998), as well as the innovation of artificial intelligence technology, the Automated Writing Evaluation (AWE) Program has captured ESL/EFL writing researchers' eyes in recent years. The existing myriad types of AWE programs consists of two main systems: 1) One is the automated essay scoring (AES) / the automated essay evaluation (AEE) system providing a holistic score and even analytic scores in terms of the weightings coming from extracting several linguistic features from the essay (Page, 2003; Liu & Kunnan, 2016), such as the corpus-based assessment (Lang, Li, & Zhang, 2019) or a machine-learning algorithm (Bunch, Vaughn, & Miel, 2016; Shermis, Burstein, Elliot, Miel, & Foltz, 2016) that produce scores which simulate human raters (Shermis & Hamner, 2013; Palermo, & Thomson, 2018); 2) The other is the timely individualized diagnostic artificial intelligent feedback (Chapelle, Cotos, & Lee, 2015; Li, Link, & Hegelheimer, 2015) on writing samples. Also, the AWE program has the feature of being an interactive learning platform. Most of the AWE programs afford both build-in and "customizable prompt" (Palermo & Wilson, 2020) for a teacher to assign and a variety of forms for the teacher to give comments, such as the general comment in the macro perspective and text-embedded comment in the micro perspective. Moreover, student-users can revise their drafts according to the received feedback from the source of the AWE system, teacher, and even peers, i.e., MI Write and the Pigai program.

The writing practice via the AWE program is labeled as one essay with multiple drafts (Dikli & Bleyle, 2014; Palermo & Wilson, 2020). Therefore, in order to address the validity of the functions mentioned above, a large body of prior scholarship follows with interest the issues on the consistency or the agreement between the AES/AEE system and human raters, users' perceptions, as well as users' strategies. However, little research attempts to investigate the pedagogical value of its automated feedback on improving students' overall writing competence (Stevenson & Phakiti, 2014; Wilson & Czik, 2016). While, its instant diagnostic feedback on multiple linguistic features (i.e., grammar, vocabulary, organization, mechanics, and so on) provides the potential

values for pedagogical use in English writing class which should be investigated thoroughly.

To address this matter, this literature review aims to identify main indicators that are related to the effectiveness of automated feedback provided by the AWE program on improving undergraduate students' writing competence, hoping to shed some lights on future research, particularly in the ESL/EFL context.

2. The Implementation of Automated Feedback Provided by AWE in ESL/EFL Writing

2.1 Accuracy of the Automated Feedback in AWE

The majority of the previous scholarship has a strong disposition to explore the accuracy of the automated feedback by investigating the consistency between the AWE system and human-raters (i.e., Zhang, 2020; Ranalli, Link, Chukharev-Hudilainen, 2017; Lang, Li, & Zhang, 2019), due to the insufficient literature that investigates the promotional effect on students' writing ability that observed in long-term.

Pertaining to the accuracy of the automated feedback, there is a strong consensus that the AWE system is weak in detecting the deep level errors, i.e., in the aspects of content and organization, but superior in providing surface-level error detection, i.e., grammar, syntax, and mechanics (Wang, 2020; Lang, Li, & Zhang, 2019). To put it more specifically, since the AWE system offers real-time automated corrective feedback to each draft submission, sufficient diagnostic feedback helps students to reduce the frequency of recurring errors in order to improve their writing accuracy. However, in some cases, students find even if they made all the corrections accordingly, they failed to improve the automated score (Li, Link, & Hegelheimer, 2015). That explains the revision of the writing process is far beyond the error correction in the microstructural level (Flower, & Hayes, 1981). Regarding the invalid deep level feedback, Liu and Kunnan (2016) point the AWE program -WriteToLearn hardly detects off-subject essays and its automated feedback on the macrostructural level is kind of general and vague. Moreover, in the Chinese EFL context, the AWE system fails to detect typical Chinese EFL students' errors that root in the negative transfer from Mandarin – Chinglish (Liu & Kunnan, 2016). In addition, Chinese EFL students also feel frustrated because the monolingual automated feedback (in English - the target language) cannot be understood clearly (Ding, 2008).

2.2 Users' Perceptions on the Use of AWE

Users' perceptions toward the use of the AWE system are one of the powerful indicators to reflect its potentials and effectiveness (Wilson & Roscoe, 2020). From the review of literature, it is found that interviews and questionnaires are the frequent methods to reveal this fact either from the perceptions of the students or the teachers. Since students are the main body of the users, most studies tend to focus on investigating students' perceptions. On the whole, students hold a positive perception when they receive automated scores and artificial intelligent feedback is accurate and valid (Roscoe et al., 2017; Ranalli, 2018). That said, their perceptions towards automated feedback mainly depend on how accurate the automated feedback they receive is, which is also positively correlated to users' perception of the automated score. In addition, students confirm that compared to the teacher's feedback, automated feedback outperforms in the aspects of language use, syntax, and mechanics (Wang, 2020). On the other hand, on the surface, the inaccuracy of the automated feedback can lead to negative attitudes causing the low uptake rate (Liu & Kunnan, 2016) and students may feel the interaction with an artificial intelligence rater lacks social human interaction (Wang, Shang, & Briody, 2013). But on a deep level, some scholars argued that these phenomena support the combination instruction of the AWE program and teacher's feedback. Zhang (2020) found students' negative behaviors for the automated feedback do not automatically mean they have not benefitted from using AWE. From a meta-process for screening automated feedback, it reveals a higher level of corrective ability from the students in their writing after receiving automated feedback.

As for the teachers, their perceptions on the use of AWE are rather mixed. Due to the lack of accuracy of AWE scoring, some teachers are concerned that it may impede the improvement of students writing accuracy (Li, Link, & Hegelheimer, 2015). On the other hand, other research has found that teachers do benefit from the AWE systems. First, the automated scoring system reduces teachers' daily burdensome scoring work (Warschauer & Ware, 2006; Li, Link, & Hegelheimer, 2015; Stevenson & Phakiti, 2019; Palermo, & Wilson, 2020), even in the standard language test, i.e., E-rater has been implemented in GMAT (Lang, Li, & Zhang, 2019). Moreover, the adequacy of the automated feedback has the potential to allow the teacher tend to focus more on providing higher-level feedback, such as feedback on content and organization (Jiang, Yu, & Wang, 2020; Link, Dursun, Karakaya, & Hegelheimer, 2014; Li, Link, & Hegelheimer, 2015), which makes up for the weakness of the automated feedback.

3. Objective of the Literature Review

This review focuses on current scholarly works on the impact of the automated feedback provided by the AWE program on students' writing performance in the ESL/EFL context to identify the most salient potentials to be highlighted and problems to be addressed in future studies. To be specific, this review aims to answer the following research objective:

To determine the levels of effectiveness of the AWE's automated feedback on improving students' writing performance.

4. Reviewing the Literature

This review was carried out sticking to the guideline of the PRISMA-P (Preferred Reporting Items for Systematic reviews and Meta-Analyses-Protocols) statement (Moher et al., 2009), which represents the PRISMA-P 2015 checklist that contains 17 numbered items, including 26 sub-items, in the aspects of administrative information, introduction, and methods. Also, in order to prevent the final literature was identified from selection bias, the inclusion and exclusion criteria that were proposed by Gough, Oliver, & Thomas (2012) were adopted to help the researchers to rule out the irrelevant studies, where the authors selected the scholarly works by "study design and the population, intervention/issue, comparison, outcome and context/time" (p.125) (see Table 1). Then, a separate discussion was made based on the results from the literature review matrix, which shows the understanding of what lessons can be learned from past studies to create possible and purposeful directions for future research.

4.1 Search Strategy

A structured search strategy was divided into two stages. The initial search stage of this literature review was undertaken in February 2021, included peer-reviewed articles published in the past five years in English investigating the AWE software in any aspect. Consequently, 664 articles were identified initially, applying the following steps:

a) Electronic databases (i.e., Web of Science, Taylor & Francis, JSTOR, and ProQuest)

b) By following up on the reference sections of the identified articles.

The primary result of the searching was identified by utilizing the Boolean operator "OR" and the keywords automated writing evaluation, automated essay evaluation, and automated essay scoring. The articles were searched up to the saturation point so that no new literature was reported no matter using complete or widely accepted abbreviations.

Table 1. Inclusion and exclusion criteria by Gough, Oliver, & Thomas (2017)

Inclusion criteria	Exclusion criteria
1. Published in 2016 – 2021 February	Published before 2016
2. English language	Not in English
3. ESL/EFL undergraduate students' writing performance	Participants are not undergraduates in ESL/EFL context
4. Controlled trial study	Experimental study is not involved
5. Indexed in Web of Science, Taylor & Francis, JSTOR and ProQuest	Not a journal article
	Cannot be searched via Internet
6. English academic writing	Not on English academic writing
7. The effectiveness of Automated feedback of the AWE program in students' internalization	Not the design of the AWE software
	Not automated scoring ability of the AWE
	Not the accuracy of the automated feedback
	Not the users' perception of the automated feedback
8. Primary research	Not primary study (i.e., book and review)

In the second stage, the data were screened step by step. (Figure 1. illustrates the logic and data screening process of this review). During the data screening process, two coders were invited to screen the titles and abstracts with high inter-rater reliability (k = .88) on identifying the reasons for inclusion and exclusion criteria

(see Table 1.) that were examined in the pilot study by testing Cohen's kappa coefficient k (Cohen, 1960). Consequently, 655 articles were excluded, and there are only remaining 11 articles were included for synthesis.

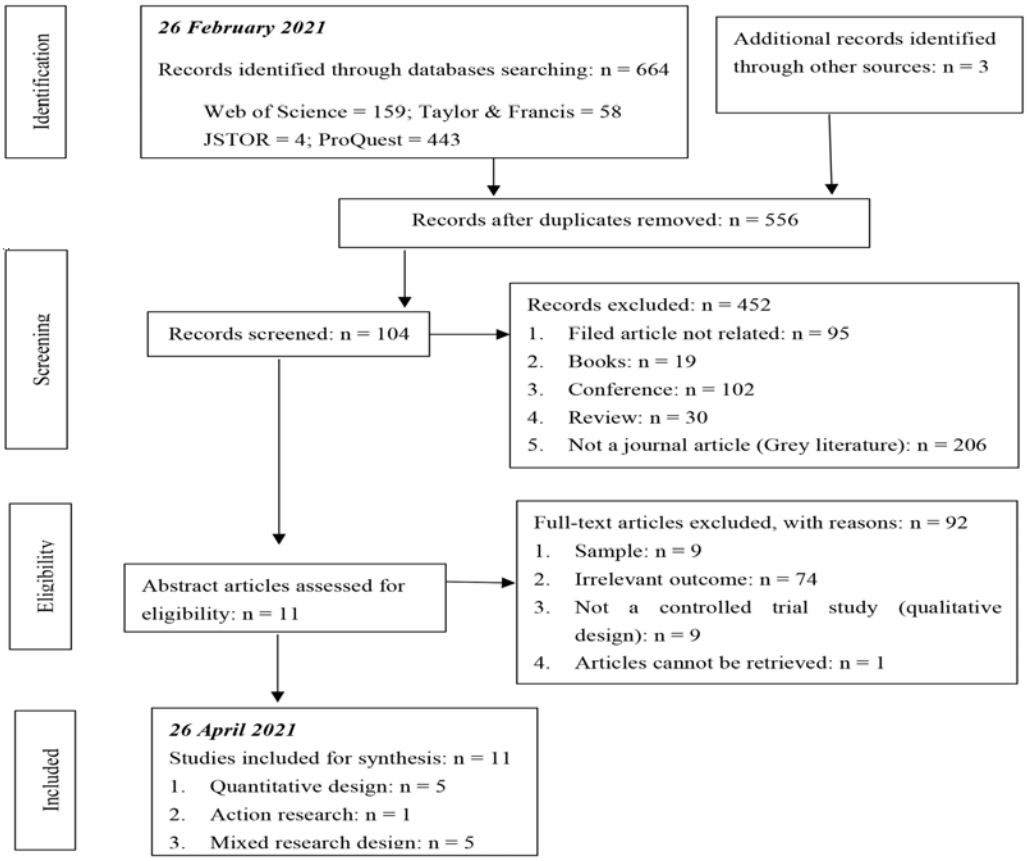

Figure 1. PRISMA Flowchart (Slightly modified after Moher, Liberati, Tetzlaff, & Altman, 2009, p. 8)

In the final stage, there are 11 articles that were selected based on the PRISM-P flow chart in total. The articles were chosen from journals such as, *Theory and Practice in Language Studies*, *International Journal of Instruction*, *Innovation in Language Learning and Teaching*, *Computer Assisted Language Learning*, *Calico journal*, *the Journal of EUROCALL*, *English Teaching (영어교육)*, *ReCALL*, *Language Learning & Technology*, and *System*. Moreover, based on the reviewed articles, the effectiveness of automated feedback provided by the AWE program was examined at four levels: students' overall writing performance, analytic writing performance, revision ability, and writing knowledge. From these 11 identified studies (N=11) where three research was conducted using quantitative research design (N=3), seven used mixed-method research design (N=7), and 1 study employed action research design (N=1). Since the controlled trial study is one of the inclusion criteria, there is no qualitative study involved. In regards to the quantitative research design, nine out of eleven of the sample chose the quasi-experimental research design.

4.2 Data Abstraction and Analysis

In order to address the RO, the levels of effectiveness of the automated feedback provided by the AWE program on improving students' writing performance that was investigated by the 11 identified studies were coded. Then, the report of the analysis was based on a literature review matrix (see Appendix) which was developed by extracting the following information from each journal article: author(s), year of publication, journal, the title of the article, underpinning theory or theoretical framework, research method, participants or sampling, instruments, and main findings.

4.3 Limitations

Even though this literature review was carried out in a rigorous way to a large extent, the search strategy may be limited in some aspects. First, the time scope for the searching may cause publication bias. Articles published after February 2021 are not included, so those research findings may vary due to the different time scopes. Furthermore, although the exhaustive search was undertaken in four databases only until the data has reached

saturation, conference proceedings and grey literature were excluded. In addition, the inclusion and exclusion criteria of the sample, research methods, and the research settings of the implementation of the AWE program narrow down the review scope. Therefore, further research could expand the search scope by increasing the number of databases, publication categories, and research settings.

5. Findings and Discussion

After all the different themes were coded, four main indicators of the effectiveness of automated feedback provided by the AWE program emerged, i.e., students' overall writing performance, analytic writing performance, revision ability, and writing knowledge.

5.1 Students' Overall Writing Performance

Studies (N=3) place special focus on look into the influence of automated feedback on improving learners' overall writing performance. Wang (2019) investigated the effect of the use of Pigai program through comparing two different treatment conditions (i.e., students in the control group (CG) only received teacher feedback and those in the experimental group (EG) who received both teacher feedback and automated feedback provided by the Pigai program) at a vocational college in China. The results revealed that students' writing performance of the experimental group outperforms those of the control group after an 18-week intervention. Wang (2019) contributed such significant improvement of students' overall writing competence to students' constantly and independently revision in terms of the real-time automated feedback which reverses students' passive learning position. Parra and Calero (2019) conducted a quasi-experimental study to compare the effectiveness of two different brands of the AWE programs (i.e., experimental group 1 used Grammarly and experimental group 2 used Grammark) to facilitate teacher's instruction in Ecuador. Findings show that after the 8-week intervention, both two experimental groups improved their overall writing performance significantly from pre-test to post-test and that the effects of Grammarly and Grammark on improving students' overall writing competence are positive and similar. In addition, they came to the same conclusion as Wang (2019) that the immediacy and privacy of the diagnostic automated feedback are particularly useful in developing students' learner autonomy. However, as far as the methodology is concerned, the lack of the control group cannot make the conclusion be drawn from the current quasi-experimental research without any alternative explanations. In another research, Lee (2020) adopted the test-retest design to test the long-term effect of automated feedback provided by the AWE- Criterion on improving undergraduate students' writing development in South Korea. According to the descriptive statistical results, both two participants improved their holistic scores across the pre- and post-tests ranging from 10 to 40. It is quite interesting to note that Lee (2020) used the test-retest design and found these findings in his study. This goes to show that the test-retest design is more appropriate to test the reliability of the instrument, other than a research design (Fraenkel, Wallen, & Hyun 1993).

5.2 Students' Analytic Writing Performance

There are two studies that paid particular attention to students' analytic writing performance which provides a specific lens to study the effects of automated feedback on students' writing performance. Link, Mehrzad, and Rahimi (2020) investigated the changes in students' writing performance in two conditions (i.e., Control Group: Process-oriented writing approach (POWA), Experimental Group: POWA + Criterion) in the aspects of three analytic writing traits, which are: syntactic and lexical (assessed by dependent clause per T-unit and mean length clause), accuracy (assessed by coordinate phrases per clause), fluency (assessed by complex nominal per clause) across pre-, post, and delayed tests at an Iran university. The results show that students in Experimental Group retained a notable enhancement only in the accuracy in a long run. However, students in the control group made a more comprehensive enhancement, except for word frequency, and the rest 8 writing traits showed significantly different performances. It is an interesting phenomenon that the embedded use of the Criterion with POWA made students narrow down their focus on grammatical performance.

In a similar vein, Saricaoglu (2019) conducted a quasi-experimental study to ascertain the effects of the automated feedback provided by the ACDET-AWE program (i.e., a newly developed AWE tool, ACDET which specializes in analyzing causal discourse and providing formative feedback on causal explanations) by looking into the changes of students' writing performance on written causal explanations in pre- and post-tests and in particular analyzing students' performance of causal language features (i.e., conjunctions, adverbs, prepositions, adjectives, verbs, and nouns). By counting the frequency of each indicator of the written causal explanations, the descriptive and compared statistical result revealed the development of students' improvement within assignments, and the changes in the total number of students' each causal language features for pre- and post-tests uncover the effects of the ACDET on students' writing ability in causal explanation in long-term.

5.3 Students' Revision Ability

A few studies (N=5) focused on the impact of the AWE program on enhancing students' revision ability. To be more specific, since grammar problems are effectively 'treatable' by providing selective error feedback (Bitchener & Knoch, 2009), researchers think there is a potential effect of the AWE program in improving students' grammatical performance via revising.

The normal way of examining the effectiveness of the AWE program on students' grammatical performance is to do the error counts for students' first and last drafts which were submitted through the AWE system (see Li, Feng, & Saricaoglu, 2017; Saricaoglu & Bilki, 2021; Liao, 2016). For example, in 2017, Li, Feng, and Saricaoglu investigated the short-term (within assignment) and long-term (from pre-test to post-test) effects of the AWE-Criterion on two different levels (i.e., intermediate-high level and advanced-low level) of the ESL students' development of the grammatical accuracy in the US. Researchers coded students' grammatical errors into 9 types, which are: word choice, verb form, word form, articles, pronoun, run-on sentence, fragment, sentence structure, and subject-verb agreement. By calculating the error count according to the formula suggested by Chandler (2003): (error count/essay length) ×100 for each error type within and across assignments, they attempted to set up a multilevel growth model to obtain the trajectory of students' writing accuracy across assignments. The results indicate the automated feedback provided by the Criterion has a positive impact on improving students' grammatical accuracy. In other words, Criterion has an advantage in developing students' self-revising skills. Moreover, they found three error categories, namely, fragment, run-on sentence, and subject-verb Agreement were significantly decreased across the three assignments, indicating the Criterion has a positive impact on students' revision ability in sentence level. Since the AWE program is potentially used by students on their own outside of the classroom, Saricaoglu and Bilki (2021) investigated the impact of students' voluntary use of the AWE-Criterion on their revision practice out of the classroom based on two different courses (i.e., Introduction to Sociology (IS) and Introduction to Education (IE)) in a private Turkish university. Without any teacher's monitoring, the utilization of the Criterion was influenced by teachers' attitude which is in line with previous studies (see Roscoe et al., 2017; Li, Link, & Hegelheimer, 2015). Through comparing students' error reduction rate for the first and last drafts within each assignment and across the two assignments, in addition to the significant decreases of four error types which were observed in the low usage group-IE (i.e., Subjective-Verb agreement, Possessive, Missing Article, and Missing Comma), the high usage group (IS) also made significant improvement in a number of aspects, which indicates the high usage rate of the Criterion has a potential to improve students' grammatical revision ability in a thorough way. It is also worth noting that the error types were categorized from the Criterion and that alternative explanations for the results are usually associated with the accuracy of automated feedback (Chapelle et al., 2015) provided by the Criterion.

Another way to investigate students' revision ability is to count students' revision behavior. Link, Mehrzad, and Rahimi (2020) coded students' revision behavior into 6 types: no change, remove, add, delete, change, and transpose. According to the descriptive statistical results, Link, Mehrzad, and Rahimi (2020) found 24% of automated feedback provided by Criterion led to no change in their writing performance which is much higher than the percentage of teacher feedback that resulted in no change (12%). Whether this type of students' negative behavior is linked to the accuracy of the automated feedback or other factors needs researchers to combine qualitative research methods to reveal the fact beyond the numbers.

It is worth noting that there are two other studies (see Huang & Renandya, 2020; Hou, 2020) that investigated the effects of the AWE program on the changes in students writing performance from pre-test to post-test by comparing students' holistic scores of their first drafts and the scores of the last revision drafts. In this situation, it is not appropriate to draw the conclusion from students' revision ability to generalized over students' overall writing ability because in most authentic writing tests students' writing competence mainly refers to students' expression of the knowledge acquired about a topic (Woolfolk, 2013). In addition, usually, there is not too much time left for students to revise their writing pieces. In other words, students' writing performance is more than revising their drafts.

5.4 Student's Writing Knowledge

Since the AWE program offers multiple opportunities for learners to engage in revising their written pieces, which boosts their internalization of grammar knowledge and writing knowledge, there is one study that investigated the impact of the AWE-Grammarly on students' learning about passive structure in writing. Qassemzadeh and Soleimani (2016) conducted a quasi-experimental study to compare students' scores of the writing knowledge tests (i.e., multiple-choice tests) of two groups of learners under the treatment of teacher feedback and the use of AWE-Grammarly feedback respectively across pre-, post, and delayed post-tests. The

results revealed that students who received the intervention of the AWE-Grammarly underperform those who only received teacher feedback in the post-test, but outperform those in the delayed post-test. It demonstrates that compared to the teacher feedback, the effectiveness of the AWE program has an advantage in promoting students' grammatical performance in a long run. On the other hand, it also indicates that the use of the AWE is not always positive enough to back up self-regulated learning.

6. Conclusion

In this literature review, the authors have reviewed a number of previous studies published within the past five years on the levels of effectiveness of the automated feedback provided by the AWE program on improving ESL/EFL students' writing performance in higher education. In general, there are four potentials for the implementation of the AWE program. First, the use of diagnostic automated feedback is a crucial learning resource that supplemented teacher's instruction which improves students' overall writing performance. Secondly, the sufficiency and immediacy of the automated feedback help students revise their draft independently and in their privacy which is able to foster their learner autonomy. Moreover, the automated feedback has an outstanding advantage in improving students' writing in the aspect of the accuracy. In addition, the automated feedback facilitates students to internalize their writing knowledge, such as the passive structure.

Further research is recommended to improve the research that focuses on the effects of the automated feedback provided by the AWE program in the following ways. Future research is needed first to fill the gap of the delayed test design to testing AWE's effects on improving students' overall performance under an advanced teaching method. Secondly, studies are called for investigating its effects on students' analytic writing performance. Specific attention also needs to be focused on the influence of automated feedback on the changes in students' writing strategies.

Acknowledgments

The authors would like to thank Ms. Meng Zhao for her contribution as a coder and the sponsorship from Shanxi University of Finance and Economics under the project named Effectiveness of AWE system in College English Teaching in China's post epidemic Era, grant number SXSKL2021SX0092.

References

Bitchener, J., & Knoch, U. (2009). The value of a focused approach to written corrective feedback. *ELT Journal, 63*(3), 204-211. https://doi.org/10.1093/elt/ccn043

Bunch, M. B., Vaughn, D., & Miel, S. (2016). Automated scoring in assessment systems. In Y. Rosen, S. Ferrara & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 611-626). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch023

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*(3), 267-296. https://doi.org/10.1016/S1060-3743(03)00038-9

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language testing, 32*(3), 385-405. https://doi.org/10.1177/0265532214565386

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback?. *Assessing writing, 22*, 1-17. https://doi.org/10.1016/j.asw.2014.03.006

Ding, T. (2008). *A study of an E-Comment generating system: Development & application*. Unpublished master's thesis. National Tsinghua University of Education, Beijing, China.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College composition and communication, 32*(4), 365-387. https://doi.org/10.2307/356600

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (1993). *How to design and evaluate research in education* (Vol. 8). New York: McGraw-hill.

Gough, D., Oliver, S., & Thomas, J. (Eds.). (2017). *An introduction to systematic reviews*. Sage.

Hou, Y. (2020). Implications of AES System of Pigai for Self-regulated Learning. *Theory and Practice in Language Studies, 10*(3), 261-268. https://doi.org/10.17507/tpls.1003.01

Huang, S., & Renandya, W. A. (2020). Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in language learning and teaching, 14*(1), 15-26. https://doi.org/10.1080/17501229.2018.1471083

Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System, 93*, 102302. https://doi.org/10.1016/j.system.2020.102302

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284. https://doi.org/10.1080/01638539809545028

Lang, F., Li, S., & Zhang, S. (2019). Research on Reliability and Validity of Mobile Networks-Based Automated Writing Evaluation. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC), 10*(1), 18-31. https://doi.org/10.4018/IJMCMC.2019010102

Lee, Y. J. (2020). The Long-Term Effect of Automated Writing Evaluation Feedback on Writing Development. *ENGLISH TEACHING (영어교육), 75*(1), 67-92. https://doi.org/10.15858/engtea.75.1.202003.67

Li, J., 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. https://doi.org/10.1016/j.jslw.2014.10.004

Li, Z., Feng, H. H., & Saricaoglu, A. (2017). The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *Calico Journal, 34*(3), 355-375. https://doi.org/10.1558/cj.26382

Liao, H. C. (2016). Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach. *System, 62*, 77-92. https://doi.org/10.1016/j.system.2016.02.007

Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards best ESL practices for implementing automated writing evaluation. *CALICO Journal, 31*, 323-344. https://doi.org/10.11139/cj.31.3.323-344

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), 605-634. https://doi.org/10.1080/09588221.2020.1743323

Liu, S., & Kunnan, A. J. (2016). Investigating the Application of Automated Writing Evaluation to Chinese Undergraduate English Majors: A Case Study of "WriteToLearn". *Calico Journal, 33*(1), 71-91. https://doi.org/10.1558/cj.v33i1.26380

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Paul Shekelle, & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews, 4*(1), 1-9. https://doi.org/10.1186/2046-4053-4-1

Page, E. B. (2003). Project Essay Grade: PEG. *Automated essay scoring: A cross-disciplinary perspective*.

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology, 54*, 255-270. https://doi.org/10.1016/j.cedpsych.2018.07.002

Palermo, C., & Wilson, J. (2020). Implementing Automated Writing Evaluation in Different Instructional Contexts: A Mixed-Methods Study. *Journal of Writing Research, 12*(1), 63-108. https://doi.org/10.17239/jowr-2020.12.01.04

Parra, G., & Calero S., X. (2019). Automated Writing Evaluation Tools in the Improvement of the Writing Skill. *International Journal of Instruction, 12*(2), 209-226. https://doi.org/10.29333/iji.2019.12214a

Qassemzadeh, A., & Soleimani, H. (2016). The impact of feedback provision by Grammarly software and teachers on learning passive structures by Iranian EFL learners. *Theory and Practice in Language Studies, 6*(9), 1884-1894. https://doi.org/10.17507/tpls.0609.23

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it. *Computer Assisted Language Learning, 31*(7), 653-674. https://doi.org/10.1080/09588221.2018.1428994

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of

argument-based validation. *Educational Psychology, 37*(1), 8-25. https://doi.org/10.1080/01443410.2015.1136407

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior, 70*, 207-221. https://doi.org/10.1016/j.chb.2016.12.076

Saricaoglu, A. (2019). The impact of automated feedback on L2 learners' written causal explanations. *ReCALL, 31*(2), 189-203. https://doi.org/10.1017/S095834401800006X

Saricaoglu, A., & Bilki, Z. (2021). Voluntary use of automated writing evaluation by content course students. *ReCALL, 33*(3), 265-277. https://doi.org/10.1017/S0958344021000021

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313-346). New York, NY: Routledge. https://doi.org/10.4324/9780203122761.ch19

Shermis, M. D., Burstein, J. C., Elliot, N., Miel, S., & Foltz, P. W. (2016). Automated writing evaluation: An expanding body of knowledge. In C. A. McArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd., pp. 395-409). New York, NY: Guilford Press.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51-65. https://doi.org/10.1016/j.asw.2013.11.007

Wang, J. (2019). A comparative study on the washback effects of teacher feedback plus intelligent feedback versus teacher feedback on English writing teaching in higher vocational college. *Theory and Practice in Language Studies, 9*(12), 1555-1561. https://doi.org/10.17507/tpls.0912.12

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. https://doi.org/10.1080/09588221.2012.655300

Wang, Z. (2020). Computer-assisted EFL writing and evaluations based on artificial intelligence: a case from a college reading and writing course. *Library Hi Tech, 40*(1), 80-97. https://doi.org/10.1108/LHT-05-2020-0113

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language teaching research, 10*(2), 157-180. https://doi.org/10.1191/1362168806lr190oa

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94-109. https://doi.org/10.1016/j.compedu.2016.05.004

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*(1), 87-125. https://doi.org/10.1177/0735633119830764

Woolfolk, A. (2013). *Educational psychology* (12th ed.). New York: Pearson.

Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing, 43*, 100439. https://doi.org/10.1016/j.asw.2019.100439

Appendix

A Summary of Selected Articles for the Literature Review

Author(s) & year & journal	Title	Participants	Research design	Treatment & duration	Main findings
Wang, J. (2019). / Theory and Practice in Language Studies	A Comparative Study on the Washback Effects of Teacher Feedback plus Intelligent Feedback versus Teacher Feedback on English Writing Teaching in Higher Vocational College	105 EFL sophomores majoring in business English in Jiangxi College of Foreign Studies, Jiangxi province, China	Quasi-experimental design	EG: AWE-Pigai + Teacher feedback CG: Teacher feedback / 18-week	Students in EG were able to significantly outperform those in CG after the treatment.
Parra G., L., & Calero S., X. (2019) / International Journal of Instruction	Automated Writing Evaluation Tools in the Improvement of the Writing Skill	28 EFL undergraduate students of the English Teacher Training Program of the Faculty of Social Science in a public university in Ecuador	Quasi-experimental design & documents & questionnaire	G1: AWE-Grammark + Teacher instruction G2: AWE-Grammarly + Teacher instruction (no control group) / 8-week	1. Compared to students' writing performance in the pre-test, both two groups improved their writing performance in post-test. However, there's no significant difference between G1 and G2 in the post-test. 2. Learners held positive attitudes toward the effect of using AWE on the improvement of some aspects of writing skills.
Huang, S. & Renandya, W. A. (2020) / Innovation in Language Learning and Teaching	Exploring the integration of automated feedback among lower-proficiency EFL learners	67 non-English major EFL sophomores in a university in China	Quasi-experimental design & survey	EG: AWE-Pigai + Process oriented writing approach without teacher feedback CG: Process oriented writing approach without teacher feedback / 2-week	1. There's no significant difference of students' revision quality existing between two groups after the intervention. 2. Users held positive attitude towards the feedback provided by AWE- Pigai.
Link, S., Mehrzad, M., & Rahimi, M., (2020) / Computer Assisted Language	Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement	28 senior undergraduates in English majors in Iran	Documents & Quasi-experimental design	EG: AWE-Criterion + Process oriented writing approach CG: Teacher-only-feedback approach	1. The use of AWE-Criterion affects the amount of teacher feedback, particularly in lower-level. 2. The use of AWE positively impact students' revision practices. 3. The significant improvement of accuracy of

Learning				/ a 16-week semester	EG was maintained from pretest to delayed post-test. 4. Students in CG retained the significant improvement of the syntactic complexity and fluency from pre-test to delayed test.
Hou, Y. (2020) / Theory and Practice in Language Studies	Implications of AES System of Pigai for Self-regulated Learning	6 EFL undergraduates in a university in China	Quasi-experimental design & follow-up interviews	AWE-Pigai program (no control group) / Within 40 mins	1. No statistically significant difference was found due to the small sample size. 2. According to interviews, all the participants felt more motivated after the use of the AWE-Pigai.
Li, Z., Feng, H., & Saricaoglu, A. (2017) / Calico journal	The Short-Term and Long-Term Effects of AWE Feedback on ESL Students' Development of Grammatical Accuracy	135 first-year undergraduate EFL learners in a University in US.	Quasi-experimental design & interviews	G1(intermediate-high level)/ G2 (advanced-low level): AWE-Criterion program + Process-based writing approach (no control group) / 3 assignments	1. The effects of automated feedback provided by Criterion on both intermediate-high and advanced-low ESL students' error reduction were significant in the short term and long term. 2. Majority students hold positive attitude toward the effects of the Criterion.
Saricaoglu, A. (2019) / the Journal of EUROCALL	The impact of automated feedback on L2 learners' written causal explanations	32 first-year undergraduate ESL learners in a University in US.	Quasi-experimental design & documents	AWE-ACDET (no control group) / 8-week	1. Students' revision caused the means of adjectives and adverbs increased and the means of conjunctions decreased significantly within essay 1. 2. Across the first draft to the final draft, students' revision caused the means of causal verbs significantly decreased from the Essay 2. 3. No improvements across pre- and post-tests were significant but for causal verbs.
Qassemzadeh, A., & Soleimani, H. (2016) / Theory and Practice in Language Studies	The Impact of Feedback Provision by Grammarly Software and Teachers on Learning Passive Structures by Iranian EFL Learners	Seventy first-semester EFL students in a university in Iran.	Quasi-experimental design	CG: Teacher feedback only EG: AWE-Grammarly only / 5-week	1. There's no significant difference between CG and EG either in post-test or in post delayed test. 2. The effect of AWE-Grammarly underperforms teacher feedback in posttest, but outperforms teacher feedback in delayed test. 3. Students in both two

					groups improves significantly in post and delayed tests.
Lee, YJ. (2020) / English Teaching (영어교육)	The Long-Term Effect of Automated Writing Evaluation Feedback on Writing Development	2 undergraduates in Korea	A test-retest research design & Interviews & case study	AWE-Criterion / one academic year	1. Two participants improved their writing scores after the use of Criterion. 2. Participants' writing fluency and accuracy improved after the use of Criterion.
Saricaoglu, A. & Bilki Z. (2021) / ReCALL	Voluntary use of automated writing evaluation by content course students	114 students at a private Turkish university	Action research / 1 semester	AWE-Criterion IE Course/ IS Course / 2 assignments	1. Students' use of Criterion was impacted by the instructor. 2. In IE course, there were some significant improvements between students' first and last draft in terms of Subjective-Verb agreement 3. In IS course, students' grammatical errors were improved significantly by students from the first draft to the last draft.
Liao, H. C. (2016) / System	Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach.	63 sophomores at three different universities in Taiwan.	Questionnaire & interview & quasi-experimental design	AWE-Criterion + teacher-student conference + Process-based writing approach (No control group) / 9-week	1. Students made significant error reduction from 1st draft to the last draft in Essay 1/2/4. 2. Significant improvement of students' grammatical performance was detected until the 3rd assignment 3. Students' grammatical improvements were observed in their revisions firstly and later new writing.