Examining the Effects of Automated Writing Evaluation (AWE) Feedback on EFL Learners' Revision and Writing Quality

Chen Shen[1(✉)], Jiwei Tian[2], and Sheming Qu[3]

[1] Foreign Languages Department, Xi'an Jiaotong University City College, Xi'an, Shaanxi, China
chenshen1109@163.com

[2] Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an, Shaanxi, China

[3] School of English Teacher Education, Xi'an International Studies University, Xi'an, Shaanxi, China

Abstract. Recent years have witnessed an increasing interest in automated writing evaluation (AWE) systems, which provides students with immediate scores and comments based on the use of natural language processing, linguistic corpora, and educational metrics. The paper examines the potential effects of *Pigai*, an AWE system widely used in China, on EFL learners' revision processes and writing products. Participants included 15 first-year students from a college in inland China. A textual analysis of the participants' essays and the feedback they received was conducted to identify students' revision scopes under the influence of AWE feedback. A quasi-experimental design was also employed to measure the lexical and syntactic complexity of first and final drafts to investigate the impact of automated feedback on writing quality. The results indicate that AWE is more helpful in the grammatical and lexical dimensions but does not significantly improve the syntax and higher levels.

Keywords: automated writing evaluation · revision scope · writing quality · lexical and syntactic complexity

1 Introduction

It is common practice to provide instructional feedback to students when teaching writing. Given the excessive workload of teachers and the limited efficiency among peers, various Automatic Writing Evaluation (AWE) systems have emerged in recent decades, making computerized feedback an available pedagogical alternative for writing instructors. AWE, which uses natural language processing techniques, can provide immediate scores and instant feedback to students regardless of their writing genres. However, AWE research, especially its application in EFL classroom contexts, is still in the infancy stage [1]. The majority of research has focused on demonstrating the validity and reliability of

X. Yuan et al. (Eds.): ICEKIM 2023, AHCS 13, pp. 558–566, 2023.
https://doi.org/10.2991/978-94-6463-172-2_60

AWE, as well as the high consistency between machine and human scoring [2]. Scholars have also raised concerns regarding AWE's potential positive or negative impact on English writing performance, but the results remain inconclusive [1]. The most commonly used measures of writing performance have been holistic scores and accuracy [3]. Nevertheless, as equally significant measures of text quality, lexical and syntactic complexity have often been neglected [4]. Few studies have evaluated the role of AWE from the perspective of both EFL learners' revision process and production.

Thus, an investigation was conducted in this study to examine the effects of the AWE program on Chinese EFL learners' revision and writing quality through a process- and product-oriented approach.

2 Literature Review

2.1 Effects of Automated Feedback on Students' Revision

Studying AWE-supported revision can shed light on the instructional and assessment value of AWE systems. Cotos et al. coded revision scopes to analyze the process by which students make decisions on revision using feedback from AWE tools, which were categorized into five types: content, vocabulary, grammar, structure, and mechanics [5]. There was, however, a lack of exploring the relationship between feedback scopes and revision practice. Based on the prior research, this study aims to analyze how students' revision practice related to AWE feedback in an attempt to gather insight into this association.

2.2 Effects of Automated Feedback on Writing Quality

A growing body of research supports the positive potential of AWE for writing quality. The meta-analysis conducted by Morphy and Graham concluded an average weighted effect size of 1.46 for three research on AWE [6]. The studies showed a generally consistent finding that automated feedback appears to support writing improvement while reducing the frequency of errors in successive revisions. To capture changes in writing quality, it is more advisable to use fine-grained linguistic indices rather than holistic scores [7]. Li et al. provided evidence to support the effectiveness of Criterion in promoting accuracy from a rough to a final draft [8]. Li et al. reported positive effects of Criterion on accuracy in the short term, but no significant changes were observed over the long term [9]. However, it is worth noting that lexical and syntactic complexity should be taken into account in addition to accuracy when measuring students' language development and writing performance.

Although the literature reviewed above is thought-provoking, further research is still needed. This study aims to thoroughly examine a Chinese AWE program from the perspective of students' revision process and production in order to shed some light on theory, methodology, teaching, and learning. Following are the research questions.

Q1. What scopes do Chinese EFL students revise in response to AWE feedback?

Q2. What are the effects of AWE feedback on Chinese EFL students' writing quality?

3 Methods

3.1 Participants and Contexts

15 freshmen, who major in English at a college in mainland China, participated in the research, consisting of 12 females and 3 males. All participants have no overseas study experience and have not yet taken the College English Test Band 4 (CET-4). According to their college entrance examination English scores, they had no significant difference in language proficiency.

A tool called *Pigai* (http://www.pigai.org/) is examined. *Pigai* is the most widely used AWE tool in China, similar to *Criterion* and *My Access*. The system can give students overall scores, specific feedback, and general comments based on their drafts. Additionally, it can display the ranking of each student's score for instructors. Students can constantly revise their essays in response to automated feedback until they are satisfied with them.

3.2 Data Collection and Data Analysis

Each participant was asked to write an argumentative essay of at least 120 words within 30 min. The instructor had previously trained all participants on how to use *Pigai*. After completing the writing task, students were required to submit their first draft to *Pigai* and self-revise according to the feedback an unlimited number of times. The deadline for submitting the final draft was set seven days later.

Data analysis contains a textual analysis of student drafts and automated feedback. To answer the first research question, the study marked and coded students' revision scopes and the feedback they received (see Table 1). Significantly, the feedback provided by *Pigai* can be classified as corrective and non-corrective feedback. *Pigai* can identify 11 kinds of errors in writing and present them to students through direct or indirect feedback. Meanwhile, it also supplements students with synonyms to expand their knowledge, which mainly includes three levels of word, phrase, and sentence, so as to help students polish their language.

Lexical and syntactic complexity are two key measures of EFL learners' written language output. To answer the second question, both measures were employed to assess the change in students' writing after receiving feedback via automatic language complexity analyzers developed by Lu and his team. The Lexical complexity analyzer (LCA) was adopted to obtain the lexical complexity data, and the syntactic complexity data was processed by the L2 syntactic complexity analyzer (L2SCA). Lu measured lexical complexity across three dimensions, including lexical density, lexical sophistication, and lexical diversity [10]. He also divided syntactic complexity into the length of production, sentence complexity, subordination, coordination, as well as particular structures [11].

In the research, a representative index was selected from each category for measurement. The lexical dimensions were measured with LD, LS1, NDWERZ, MSTTR, and LV, and syntactic complexity was measured with MLT, C/S, C/T, CP/T, and CN/T. To ensure that the analyzers operate correctly, punctuation, capitalization, and spelling errors were corrected.

4 Findings and Discussions

4.1 Q1. What Scopes Do Chinese EFL Students Revise in Response to AWE Feedback?

Figures 1 and 2 convey three key messages: the frequency of various types of feedback provided by *Pigai*, the frequency of successful corrections made by participants, and the frequency of unsuccessful corrections made by participants. Of the ten categories of errors, participants made the most common errors in punctuation (27), followed by grammatical errors related to sentences (21) and verbs (16). During the data collection process, the researcher found that the major reasons for irregular punctuation were that students have used electronic devices less for writing and have not yet developed an awareness of spaces after punctuation marks. The errors in sentences were basically reflected in incomplete sentence components and the lack of conjunctions. Verb errors were mainly due to subject-verb inconsistency and the incorrect use of modal verbs.

Among the non-error feedback students received, word (84) was prioritized, followed by phrases (32) and sentences (8). It can be seen that the program provides a considerable number of additional learning points in addition to errors, primarily focusing on synonyms of words or phrases because of the inherently large corpus of *Pigai*. Under such circumstances, students not only expand their language knowledge but also have the opportunity to use different or low-frequency vocabulary.

102 of the 121 errors were corrected successfully by the students, but some grammatical problems regarding parts of speech (2), prepositions (3), verbs (3), collocations (4), and sentences (9) were not revised. The researchers discovered that this could be due to different types of feedback and the difficult degree of modification. The software generally gives direct feedback, such as the redundancy of the article "the" and the need to use the plural form of the noun "problem". Thus, students only need to complete the revisions based on the answers provided. In some contexts, however, the system only displays indirect alerts, e.g., the sentence is not grammatically correct. It is difficult for students to make a decision in the process of self-revision. Students may thereby give up revising more difficult errors as a result of their narrow knowledge systems.

Compared to error correction, the unsuccessful rate of non-error feedback was significantly higher, exceeding 40 percent of the total. Most students still showed a positive willingness to respond to feedback and revise their drafts, but they may not consider some substitutions necessary and choose no change.

Overall, students determine the scope of their modification based on the prompts given by the system. The idea that students, influenced by the characteristics of *Pigai*, focus more on words, phrases, and simple grammatical levels and neglect the mastery of higher-level writing skills such as organization, discourse, and content development.

4.2 Q2. What Are the Effects of AWE Feedback on Chinese EFL Students' Writing Quality?

To investigate the effects of feedback from *Pigai* on students' performance of lexical and syntactic complexity, paired samples t-tests were employed. Table 2 only shows the results of the indices with significant differences.

Table 1. Code Scheme and Examples of Automated Feedback

	Scopes	**Examples**
Corrective feedback	Punctuation	First of all,parents should put themselves in their children's view. **[Punctuation error]** There should be a space after punctuation marks in English.
	Capitalization	I have following suggestions to deal with problem between parents and Children. **[Capitalization error]** Please check the capitalization of "Children".
	Spelling	In order to balance the relationship, firstly, parents should lower their expections. **[Spelling error]** Please check the spelling of "expections".
	Part of Speech	Then, as long as the parents cannot meet a little desire, will lead to relationship. **[Part of speech error]** The noun "relationship" is misused as a verb.
	Article	Then, parents need to develop with children at the the same time. **[Article error]** Please check if the article "the" is redundant.
	Preposition	I want to we have a best relationship between us and our parents. **[Preposition error]** Please check if the preposition "to" is redundant.
	Noun	I have following suggestions to deal with problem between parents and Children. **[Noun error]** Please check if the noun "problem" should be used in the plural form.
	Verb	Parents and children are two important roles in the family. **[Verb error]** Please check if the verb "are" is correctly used.
	Collocation	When parents communicate with their children, parents should affirm their children's advantages before pointing out their shortcomings. **[Collocation error]** Please check if the collocation "pointing shortcomings" is correctly used.

(*continued*)

Table 1. (*continued*)

	Scopes	Examples
	Sentence	Then, as long as the parents cannot meet a little desire, will lead to relationship. **[Sentence error]** Please check if the sentence components are complete.
Non-corrective feedback	Word	Mutual understanding can ease conflicts. **[Extended learning points]** **"**alleviate/relieve/lighten" is recommended to use, which has a similar meaning to "**ease**".
	Phrase	How to deal with the relationship between children and parents is so difficult. **[Extended learning points]** "tackle/address/figure out" is recommended to use, which has a similar meaning to "**deal with**".
	Sentence	If parents want to maintain a good relationship with their children, only timely communication can solve the problem. **[Extended learning points]** "provided that" is recommended to use, which has a similar meaning to "**if**".

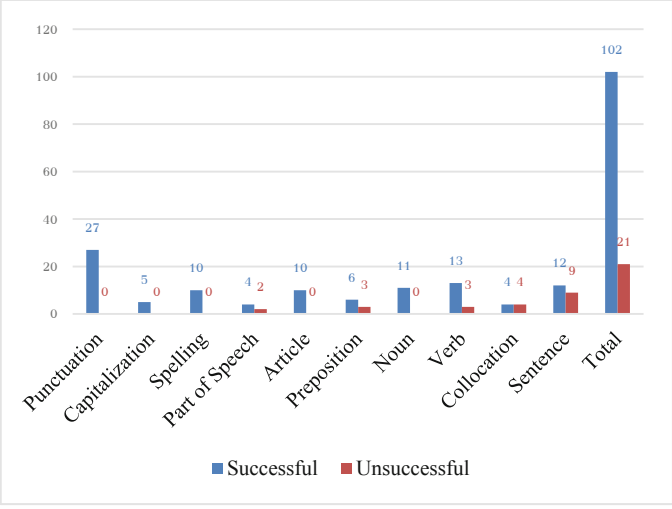

Fig. 1. Frequency of corrective feedback, successful correction, and unsuccessful correction

Within-subjects comparison presented that LS changed differently after writing revisions ($t = -2.71$, $p = .018$). Among the three measures of lexical diversity, there were significant differences in MSTTR and LV between the first and final drafts ($t = -2.84$,

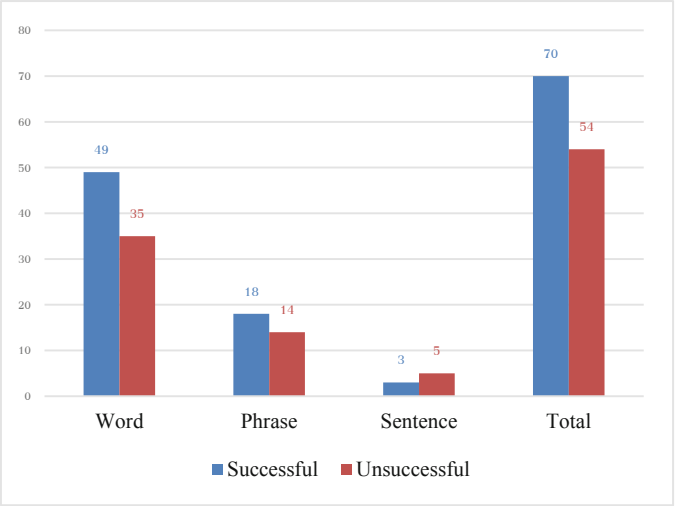

Fig. 2. Frequency of non-corrective feedback, successful correction, and unsuccessful correction

p = .014; t = -3.80, p = .002, respectively). As for the improvement of syntactic complexity, the experimental group only made great progress in CP/T (t = −2.45, p = .031).

The findings support that AWE feedback is significantly helpful to students' lexical diversity and sophistication. When commenting on each sentence, the system provides students with synonyms for commonly used words, especially verbs and adjectives, as alternatives to replace words they repeatedly use. It is possible that, in the long run, this will develop students' awareness of using different vocabulary as well as complex words, thereby enhancing their writing language. The finding is similar to other investigators' research [4]. In general, the system has a significant effect on improving EFL learners' lexical complexity.

Among the results of the five measures for appraising syntactic complexity, the experimental group only showed an improvement in the use of coordinate construction. Coordinate sentences and phrases can indeed increase the length of sentences, but coordinate structures are in the low-level development process of syntax. Low-level learners

Table 2. Paired Samples T-test for Different Measures Between First and Final Drafts

	First Draft		Final Draft		t	p
	M	SD	M	SD		
LS1	.46	.08	.49	.06	−2.71	.018*
MSTTR	.75	.03	.77	.03	−2.84	.014*
LV	.67	.06	.71	.05	−3.80	.002*
CP/T	.43	.24	.70	.33	−2.45	.031*

can expand sentences as long as they use conjunctions or commas. Compared with this kind of structure, subordination structures can better reflect the quality of sentence expansion. Students use dependent clauses (e.g., adjective, adverbial, and nominal) to understand the logical relationship between the clauses. This feedback form cannot facilitate learners' ability to apply subordinate structures while writing. The reason for this is that the software lacks instruction on sentence diversity.

5 Conclusions

This study enriches the line of research on the effects of automated feedback on English language learners' revision and language development. The results indicate that AWE helps students more at the grammatical and lexical levels, while it does not help significantly at the syntactic level and beyond. Therefore, this study suggests that developers of AWE systems focus more on programming and upgrading sentence accuracy, sentence structure, content, or other higher aspects, and instructors can effectively combine automated feedback with other feedback methods. This deserves further exploration.

In the future, the researcher will increase the number of samples and the duration of experiments and add teacher feedback as a control condition to further test the hypothesis.

Acknowledgement. It was supported by the project of Xi'an Jiaotong University City College [Grant 2022Q34].

References

1. S. Link, M. Mehrzad, and M. Rahimi, "Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement," *Computer Assisted Language Learning*, vol. 35, no. 4, pp. 605–634, May 2022, https://doi.org/10.1080/09588221.2020.1743323.
2. Z. Li, S. Link, H. Ma, H. Yang, and V. Hegelheimer, "The role of automated writing evaluation holistic scores in the ESL classroom," *System*, vol. 44, pp. 66–78, Jun. 2014, https://doi.org/10.1016/j.system.2014.02.007.
3. R. T. Kellogg, A. P. Whiteford, and T. Quinlan, "Does Automated Feedback Help Students Learn to Write?," *Journal of Educational Computing Research*, vol. 42, no. 2, pp. 173–196, Mar. 2010, https://doi.org/10.2190/EC.42.2.c.
4. J. Xu and S. Zhang, "Understanding AWE Feedback and English Writing of Learners with Different Proficiency Levels in an EFL Classroom: A Sociocultural Perspective," *Asia-Pacific Edu Res*, vol. 31, no. 4, pp. 357–367, Aug. 2022, https://doi.org/10.1007/s40299-021-00577-7.
5. E. Cotos, S. Link, and S. Huffman, "Effects of DDL Technology on Genre Learning," *Language Learning*, vol. 21, no. 3, pp. 104–130, 2017.
6. P. Morphy and S. Graham, "Word processing programs and weaker writers/readers: a meta-analysis of research findings," *Read Writ*, vol. 25, no. 3, pp. 641–678, Mar. 2012, https://doi.org/10.1007/s11145-010-9292-5.
7. J. Wilson and A. Czik, "Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality," *Computers & Education*, vol. 100, pp. 94–109, Sep. 2016, https://doi.org/10.1016/j.compedu.2016.05.004.

8. J. Li, S. Link, and V. Hegelheimer, "Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction," *Journal of Second Language Writing*, vol. 27, pp. 1–18, Mar. 2015, https://doi.org/10.1016/j.jslw.2014.10.004.
9. Z. Li, H.-H. Feng, and A. Saricaoglu, "The Short-Term and Long-Term Effects of AWE Feedback on ESL Students' Development of Grammatical Accuracy," *CALICO*, vol. 34, no. 3, pp. 355–375, Aug. 2017, https://doi.org/10.1558/cj.26382.
10. X. Lu, "The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives," *The Modern Language Journal*, vol. 96, no. 2, pp. 190–208, Jun. 2012, https://doi.org/10.1111/j.1540-4781.2011.01232_1.x.
11. X. Lu, "A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development," *TESOL Quarterly*, vol. 45, no. 1, pp. 36–62, Mar. 2011, https://doi.org/10.5054/tq.2011.240859.