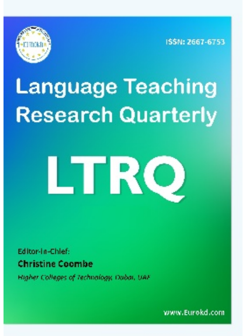


ISSN: 2667-6753

Language Teaching Research Quarterly

LTRQ

Editor-in-Chief:
Christine Coombe
Higher Colleges of Technology, Dubai, UAE

www.Eurokd.com

Language Teaching
Research Quarterly

2022, Vol. 28, 53–73

The Effects of Automated Writing Evaluation (AWE) Feedback on Students' English Writing Quality: A Systematic Literature Review

Ning Fan*, Yingying Ma

The School of Foreign Languages, Zunyi Medical University, Zunyi, China

Received *28 March 2022*　　***Accepted*** *01 June 2022*

Abstract

The purpose of this review is to examine the effects of automated writing evaluation (AWE) feedback on students' English writing performance. We systematically reviewed studies that have empirically focused on this purpose. This review uses several combinations of key words to search in the databases of JSTOR, SSCI, and ERIC for peer-reviewed articles published from 2005 to April 2020. The systematic review produced 22 eligible studies categorized as within-group and between group studies based on Stevenson and Phakiti's (2014) categorization. The results indicated that AWE feedback might be helpful for student writing under certain conditions. Specifically, the feedback was helpful when it was provided for one single group of students. The feedback was also helpful when the writing performance of a group of students receiving the feedback was compared to the writing performance of the other group of students receiving no such feedback. Moreover, AWE feedback should be continuously offered to help students benefit most from it. This review is an update about the effects of AWE feedback on student writing and may serve as a guide for researchers and instructional practitioners through informing them of the latest research on AWE feedback.

Keywords: *Automated Writing Evaluation Feedback, Writing Quality, Systematic Review*

Introduction

Writing is a difficult skill to acquire in language learning for students at all proficiency levels (Kurt & Atay, 2007). English learners often find it hard to articulate their ideas with correct written language (Evans & Green, 2007), resulting in the fact that they tend to make various types of errors in their writing. To help students enhance their writing performance, the provision of corrective

1157263956@qq.com

doi:10.32038/ltrq.2022.28.03

feedback (CF) on student writing becomes necessary. However, despite the potentially helpful role of CF in student writing, the use of CF is not without debate. According to Mohebbi (2021), a group of studies claimed that CF was harmful for student writing (e.g., Kepner, 1991; Polio et al., 1998; Truscott, 1996). Advocating the positive effects of CF on student writing, researchers also conducted experimental studies to demonstrate that CF was helpful for the improvement of student writing (e.g., Chandler, 2003; Ferris, 1999, 2004, 2006; Maleki & Eslami, 2013). More recently, it seemed that additional evidence was reported as to the positive correlation between the adoption of CF and the improvement of student writing. For example, Kang and Han (2015) conducted a meta-analysis to investigate the effects CF on students' writing accuracy. The results yielded an effect size of .54, indicating that CF could exert a substantive impact on students' writing accuracy. Therefore, the focus of research on CF might be shifted from whether it is generally effective to how to provide CF to help students benefit most from it (Ferris & Kurzer, 2019).

A number of studies have been conducted to investigate how to provide students with effective CF, including direct and indirect CF, selective and comprehensive CF, and CF provided in different explicitness levels, etc. As the advancement of natural language processing technology and corpus linguistics, automated writing evaluation (AWE) is widely used to offer CF for students' writing. For example, *Pigai*, one type of AWE system designed for English learners in China, has been used to provide CF for 400 million essays produced by 20 million students from 6,000 schools in China since 2010 (Zhang & Zhang, 2018). With such unprecedented application of AWE to the domain of English writing instruction, it is not surprising that a large number of empirical studies have been conducted to examine the effects of different AWE systems on students' writing performance. In this sense, it becomes necessary and worthwhile to systematically review the studies in order to inform educational practitioners, administrators, and AWE system developers of how AWE systems can influence student writing and in which ways they can be improved to better serve as a useful tool for students' writing. As a matter of fact, several review articles have been published with regard to the use of AWE in instructional settings. For example, Hegelheimer and Lee (2013) summarized three categories of AWE research from a pedagogical aspect. The first category reviewed effects and frequencies of AWE programs in the improvement of students' writing performance. The results indicated that AWE programs could help students improve their writing, and the different frequencies of using AWE programs could have different impacts on student writing. The second category reviewed how students and teachers in instructional practices actually employed AWE programs, while the third category investigated the relationship between the effectiveness of AWE and different instructional settings. Stevenson and Phakiti (2014) reviewed studies that explored the effects of AWE on student writing, reporting modest evidence as to the positive role of AWE feedback in student writing. In addition, they further divided the reviewed studies into within-group and between-group studies and respectively examined the effects of AWE feedback on these two groups of studies. The results showed that AWE feedback could help increase students' writing scores and decrease the number of errors for the former group of studies. In contrast, mixed results were reported for the latter group of studies. They attributed such results to the scarcity of research and the diverse characteristics of participants, contexts, and designs.

More recently, Hibert (2019) reviewed studies about the use of AWE in ESL/EFL classrooms, focusing on the analysis of theories and methodologies used in these studies. The results indicated that the introduction to theoretical framework and the consideration of technology afforded by AWE programs were ignored in most prior AWE studies. Stevenson and Phakiti (2019) also discussed studies about the effects of AWE feedback on student writing. However, it seemed they did not conduct a systematic review on this type of studies and did not center on studies that presented actual effects of AWE feedback on student writing. In other words, they approached the effects of AWE feedback from diverse aspects, such as student engagement with AWE, different types of error corrections with the help of AWE, etc.

Taken together, the above discussion demonstrated the importance of CF in student writing, the wide application and potential merit of AWE to English writing, and the paucity of updated reviews on AWE's effects on student writing. Based on the discussion, the present review aims to answer the following research question: What are the effects of AWE feedback on students' English writing quality?

Methodology

Literature Search

We employed three databases (i.e., JSTOR, SSCI, ERIC) to conduct a literature search regarding the effects of AWE feedback on student writing. The time span of publication included in this review ranged from 2005 to April 2020. We applied multiple key word combinations to search articles relevant to our research question. Specifically, we used the combinations of "automated writing feedback" OR "automated writing evaluation" AND "L2 writing" AND "effects" to search articles in JSTOR. We used key words of "automated electronic feedback" OR "automated writing evaluation" OR "automated writing feedback" to conduct the search in SSCI. We also used these three key word chains to search in ERIC. In addition, after the initial search was completed and the eligible articles were identified, we searched reference lists in each of the articles to help our review cover as many relevant articles as possible.

Inclusion and Exclusion Criteria

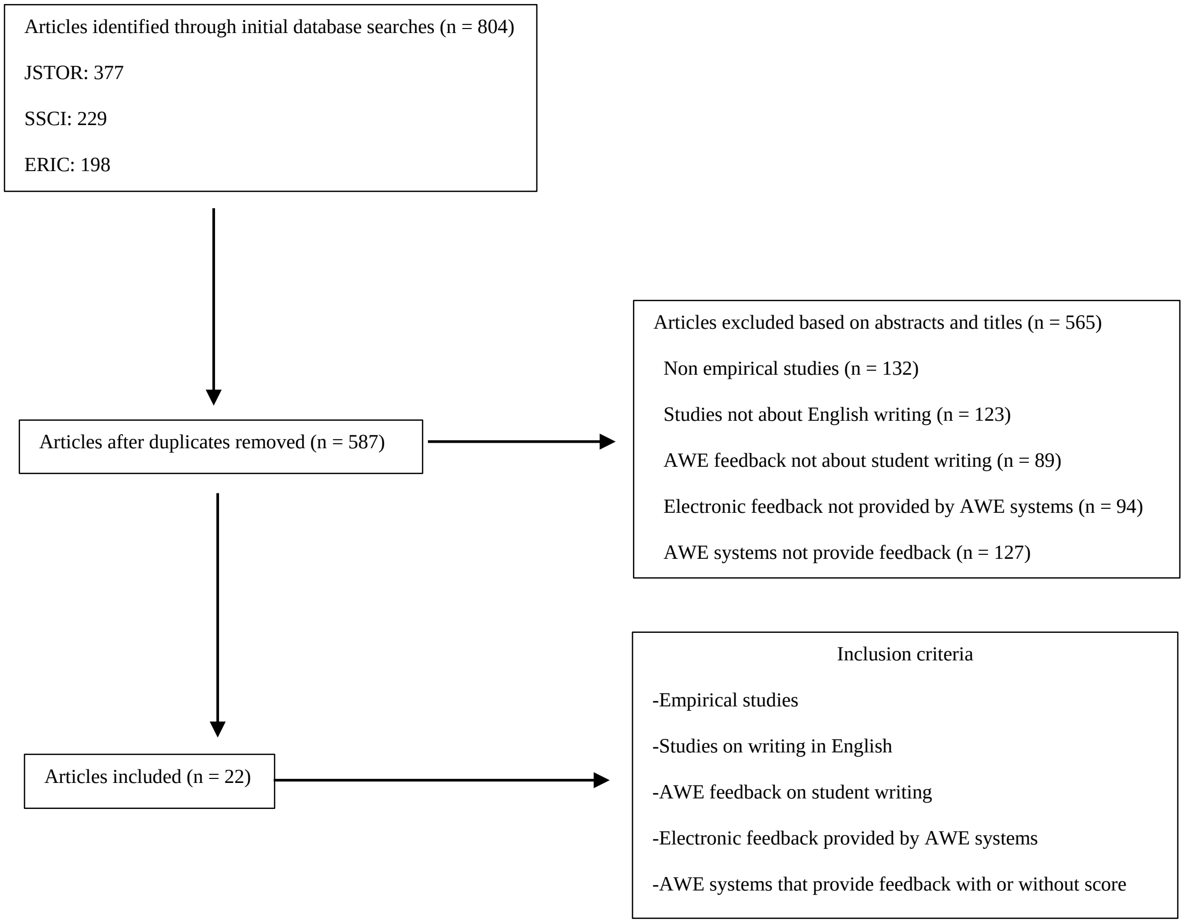
The present review consists of five inclusion and exclusion criteria, which are specified as follows. First, this review included empirical studies published in peer-reviewed journals, so it excluded dissertations, theses, and conference proceedings. The second criterion was that studies were included only if they focused on English writing, which excluded studies of writing on other languages. The third criterion was that AWE feedback should be provided on student writing, which allowed us to exclude studies in which feedback was offered on other learning aspects (e.g., grammar exercises). Fourth, this review centered on electronic feedback provided by AWE systems, excluding electronic feedback given by people. Fifth, this review included studies on AWE systems that provided feedback with or without scores for student writing, so it excluded studies on AWE systems that only provided scores.

Data extraction

A total number of 804 studies were identified after initial searches. Specifically, the searches in JSTOR, SSCI, and ERIC yielded 377, 229, and 198 results, respectively. Eliminating duplicates left us with 587 studies. Subsequently, two independent researchers screened titles and abstracts of these 587 studies based on our inclusion and exclusion criteria, and a third one was involved in a discussion if any disagreement about the eligibility of the studies arose. After this stage, we excluded 565 studies because they did not accord with our inclusion criteria, leaving us with 22 eligible articles for full review (see Figure 1).

Figure 1

Flowchart of Article Selection Process


Articles identified through initial database searches (n = 804)
JSTOR: 377
SSCI: 229
ERIC: 198
Articles after duplicates removed (n = 587)
Articles excluded based on abstracts and titles (n = 565)
Non empirical studies (n = 132)
Studies not about English writing (n = 123)
AWE feedback not about student writing (n = 89)
Electronic feedback not provided by AWE systems (n = 94)
AWE systems not provide feedback (n = 127)
Articles included (n = 22)
Inclusion criteria
-Empirical studies
-Studies on writing in English
-AWE feedback on student writing
-Electronic feedback provided by AWE systems
-AWE systems that provide feedback with or without score

Results

The present review included 22 studies that were summarized into two broad categories: within-group studies and between-group studies. The former category referred to a study in which all subjects were in one group throughout the study, while the latter one referred to a study in which more than one group of subjects were involved. Moreover, these two broad categories could be subcategorized into various study groups based on different types of feedback implemented in certain studies. Specifically, there were eight studies that were included in the category of within-group studies, among which six examined the effectiveness of AWE feedback, one investigated the comparative effectiveness of AWE feedback to online peer feedback, and the other one explored the comparative effectiveness of AWE feedback to hybrid feedback (i.e., AWE + teacher feedback). For the category of between-group studies, a total of 16 studies were contained. The 16 studies could be divided into five subcategories pertaining to the comparisons of AWE feedback to instructor feedback, of AWE feedback to no AWE feedback, and among peer, teacher, and AWE feedback. In addition, the 16 studies also compared the effects of AWE feedback provided in different ways, of AWE feedback provided for students at different proficiency levels, and of feedback provided by different AWE systems. It is necessary to point out that two of the 22 studies included in this review belong to the two broad categories simultaneously, making the total number of reviewed articles 24. Specifically, one study assigned its subjects to one single group and three groups of different proficiency levels as well, and the other study first respectively investigated the effects of two AWE tools on student writing, and then compared the effects between the two tools.

Within-group Studies

Eight studies can be categorized into this group (see Appendix I). Wang's (2013) study investigated whether students' essays improved when scored by *Criterion,* an AWE program, and human raters. A total of 53 English majors were required to write five essays scored by *Criterion* and a pre-test and a post-test essay scored by human raters during one semester. The results revealed that, for each of the five essays scored by *Criterion,* students' scores significantly improved from first to final submissions. Also, students achieved significantly better scores from human raters in their post-test essays than their pre-test essays. Kim (2014) investigated the effects of the *Criterion* feedback on the writings of both high and low proficiency levels of university students. The students at both proficiency levels completed three writing tasks, each including two drafts evaluated by human raters. The scores of the students' first drafts were compared to their second drafts. The results indicated that the students at both proficiency levels significantly improved their scores for each of the three writing tasks. In 2016, Liao conducted two studies to respectively examine at what point in the AWE-assisted process-writing program learners' grammatical performance changed, and whether AWE feedback could help to improve learner linguistic accuracy in revisions and new texts. A nine-week time-series research design was adopted in the former study. Sixty-three sophomores from three intact writing classes participated in this study. The students were assigned four comparison essays that were completed through the process writing approach and were assessed by *Criterion*. The results revealed that students significantly improved their original texts between essay two and three, and between essay three and four. For the latter study, Liao investigated how the *Criterion* feedback affected students' writing accuracy in four error categories (i.e., fragments, subject-verb disagreement, run-on sentences, ill-formed verbs) that were identified as their primary pre-treatment grammatical error types. Different from the former study, this study took into account students' writing accuracy in both revisions and new texts. The results showed that the *Criterion* feedback could significantly help students reduce the number of grammatical errors for revisions and new texts. Parra and Calero (2019) assigned 28 undergraduates to two groups, receiving *Grammark* feedback and *Grammarly* feedback, respectively. The study adopted a pre-test and post-test design. For each writing topic, student participants were asked to submit their first drafts to the designated AWE tool. Then, they submitted their second drafts after revising their first drafts based on AWE feedback. The results revealed that students in both groups significantly improved their writing performance from the pre-test to the post-test. In brief, it seemed that AWE systems might be a useful tool for students to improve their writing performance, particularly for the comparison made before and after the provision of AWE feedback for the same group of students.

Despite the positive role of AWE feedback in student writing demonstrated in the studies reviewed above, several recent studies have suggested that AWE feedback may not be always effective. Specifically, Saricaoglu (2019) explored whether the automated formative feedback provided by *ACDET* could lead to the improvement of 31 ESL learners' written causal explanations within essays and across pre- and post-tests. This pre-experimental, pre-test/post-test study lasted eight weeks. The results showed that learners' causal explanations significantly improved within one cause-and-effect essay, while no significant changes were observed across pre- and post-tests. In Shang's (2019) study, a group of 47 freshmen was asked to complete four writing tasks, among which tasks one and three were provided with online peer feedback (OPF) and tasks two and four were provided with *Cool Sentence* feedback (CSF). The results demonstrated that OPF was more effective than CSF in enabling students to write more sentences, in helping them reduce grammatical errors, and in allowing them to produce more lexical items and a greater variety of words. Similarly, Mohsen and Alshahrani's (2019) study examined the effects of *My Access* feedback and hybrid-mode feedback (i.e., *My Access* + teacher feedback) on students' writing development, and whether there was any difference between the *My Access* feedback and the hybrid-mode feedback in the improvement of student writing. For that purpose, six EFL university students were recruited as a single group to participate in a two-phase experimental study. Each phase consisted of two sessions. In the first phase, the students were required to write their first drafts of an essay in *My Access* in the first session. Then, they revised their drafts based on the feedback provided by *My Access* in the second session. In the second phase, the students were asked to write their first drafts of another essay in *My Access* and revised the drafts in the first session. In the second session, the students revised the drafts again according to the feedback provided by the teacher. The results showed that there was a significant difference in the students' writing performance between the first and second sessions when they received *My Access* feedback. Similar results were also found when the students received hybrid-mode feedback. However, the results demonstrated that the students who received the hybrid-mode feedback significantly outperformed the students who received the *My Access* feedback alone.

Between Group Studies

A total of 16 studies belong to this group (see Appendix II). We divided the 16 studies into five subcategories. The first subcategory consisted of seven studies with regard to feedback provided by teachers, peers, and AWE systems. Among the seven studies, five were about teacher and AWE feedback, one was about peer, teacher, and AWE feedback, and one was about peer and AWE feedback. This part will first present the five studies about teacher and AWE feedback (Lu, 2019; Liu et al., 2017; Wang & Li, 2019; Wang et al., 2013; Wilson & Czik, 2016). Specifically, Wang et al. (2013) assigned 57 EFL university students to an experimental and a control groups, with the former receiving feedback from *CorrectEnglish* and the latter receiving feedback from an instructor. This study adopted a pre-test/posttest design. During the phase of treatment, both groups were required to practice their writings under the guideline of introductory, writing, and revising sessions. While the first two sessions were similar for both groups, the revising session was different in that the experimental group revised their writings with the *CorrectEnglish* feedback and the control group revised with the instructor feedback. The results indicated that the experimental group outperformed the control group in terms of writing accuracy. In Wilson and Czik's (2016) study, 151 eighth grade students were divided into two groups, with the experimental group receiving teacher feedback + automated essay evaluation and the control group receiving teacher feedback only. The students' writing quality was assessed through three scores: PEG Overall and Trait scores, and Holistic Quality. The first two scores were given by *PEG Writing*®, which is a formative writing assessment software. The last score was provided by human raters. The results indicated that there were no statistically significant differences between the two groups for PEG Overall score, PEG Trait score, and Holistic Quality. Liu et al. (2017) examined the impact of indirect corrective feedback (ICF) provided by a web-based automatic feedback generation system and direct corrective feedback (DCF) provided by human teachers on EFL students' writing quality. A sample of 110 EFL students were assigned to two groups, with one group receiving ICF from the system and the other group receiving DCF from teachers. The students in both groups wrote a persuasive essay. Either ICF or DCF was provided for each of the two groups. Then, the students had one week to revise their essays before they submitted them to the system. Based on the seven features of scoring (spelling, grammar, coherence, conclusion, supporting ideas, sentence diversity, organization), two teachers scored the students' essays. The results revealed that the group who received DCF achieved higher linguistic accuracy in the essays than the group who received ICF because the former scored significantly higher than the latter in the features of grammar and spelling.

Lu (2019) divided 114 Chinese EFL university students into an experimental group and a control group, receiving *Juku* AWE feedback with teacher feedback, and only teacher feedback, respectively. This study adopted a pre-test/post-test design. The results indicated that the writing scores of the experimental group were significantly higher than the control group. In Wang and Li's (2019) study, a group of 100 Chinese EFL university students were assigned to two groups. An experimental group was provided with AWE feedback from *Writing Roadmap 2.0* (WRM 2.0), while a control group was offered with only teacher feedback. The study reported results from two parts, depending on how the students' writings were assessed. When the writings were assessed by *WRM* 2.0, the experimental group outperformed the control group in the three aspects of language form, contextual structure, and writing quality. Similarly, when the writings were assessed by teachers, the experimental group also outperformed the control group in the aspect of writing quality. It is necessary to note that *WRM* 2.0 automatically gave scores on each of the three aspects, whereas teachers only gave a holistic score on writing quality.

In addition to the five studies about teacher and AWE feedback, two studies addressed the effects of peer, teacher, and AWE feedback on student's writing quality (Huang & Renandya, 2018; Ware, 2014). In Huang and Renandya's (2018) study, a sample of 67 Chinese EFL university students was assigned to an experimental and a control groups, with the former receiving peer feedback with *Pigai* feedback, with the latter receiving peer feedback only. The students in both groups were required to write an essay and revise it based on the feedback received. Then, the revision was scored by teachers to examine the impact of adding *Pigai* feedback to peer feedback on the students' revision quality. The results showed that the addition of *Pigai* feedback in the experimental group did not lead to higher revision quality. Ware's (2014) study examined whether the feedback from peer, teacher, and AWE had differential effect on student writing development, which was measured by scores of holistic writings, text length, genre elements, and error rate. Assigned to three groups, 82 eighth-grade students received the three types of feedback after they had completed each of 12 open-ended responses. Then, they were asked to revise the 12 responses based on the feedback received. The results showed that the groups that received different types of feedback did not differ significantly in the scores of holistic writing and text length. However, for the score of genre elements, both the peer and teacher feedback groups achieved significantly higher scores than the AWE feedback group.

The second subcategory included four studies comparing the effect of AWE feedback to no such feedback on student writing (Cheng, 2017; Franzke et al., 2005; Lachner et al., 2017; Lee et al., 2009). In Franzke et al.'s (2005) study, a sample of 111 eighth-grade students was divided into an experimental group who received *Summary Street*® feedback, and a control group who received no such feedback. The students in both groups were asked to summarize 19 short-to-medium length texts during four weeks. The students in the experimental group received the *Summary Street*® feedback after they submitted each of the texts, while the students in the control group wrote and submitted the texts through a word processor. All students were told to work on their own pace across the 19 texts. At the end of the intervention, all students were asked to submit their six best texts for teachers to grade. The study examined both the main effect of condition and interaction effect of text and the condition. For the main effect, the results showed that the students who received *Summary Street*® feedback significantly outperformed the students who did not receive such feedback in measures of overall quality and coverage of the text content. For the interaction effect, the average score of the students' texts one and three were compared with the average score of texts four and six to explore whether there was any interaction effect of text and condition. The results revealed that summaries produced across time with the help of *Summary Street*® feedback were significantly better than summaries produced without such help in measures of overall quality, content, organization, the low amount of detail, and stylistic quality. Lachner et al. (2017) investigated the effect of concept map feedback on university students' explanation writings based on the measures of local and global cohesion, and overall comprehensibility. Forty-two university students were placed in an experimental and a control group, receiving the concept map feedback and no such feedback, respectively. All students were asked to complete a piece of explanation writing. Then, the students in both groups were provided with a writing prompt to revise their explanation writings. The only difference was that the students in the experimental group were offered the concept map feedback in addition to the writing prompt, while the students in the control group were not offered such feedback. The results demonstrated that the students who received the concept map feedback could produce more locally and globally cohesive writing, and more comprehensible explanations than the students who received no such feedback in their revisions. Furthermore, the study also found that the positive effect of the concept map feedback observed in the students' revisions could be maintained in their new pieces of explanation writings. In Cheng's (2017) study, a sample of 138 university students were classified into an experimental and control group, with the former receiving Online Automated Feedback (OAF), while the latter received no such feedback. The purpose of the study was to explore the effect of the OAF on the students' reflective journals. All students were asked to submit three reflective journals in total. For the first and second journals, students in the experimental group received OAF, while students in the control group did not receive any feedback. All the student journals were manually evaluated by the principal investigator of the study. The results indicated that the experimental group significantly outperformed the control group in terms of the overall score for the final reflective journal, and the experimental group also demonstrate a significant improvement in scores across the three reflective journals. Lee et al. (2009) assigned 27 university students to an experimental and a control groups to examine whether Essay Critiquing System (ECS) feedback could help students improve their writing scores. The students in both groups were asked to write an argumentative essay. Then two modes of feedback (i.e., content and organization) from ECS were provided for students in the experimental group, and no feedback for students in the control group. After revising their essays, the students' final submissions were co-marked by two raters. The results revealed that there was no significant difference in students' final scores of their essays between the two groups.

The third subcategory contained two studies with regard to differential amounts of AWE feedback provided (Kellogg et al., 2010; Koh, 2017). Kellogg et al. (2010) investigated the effects of differential amounts of *Criterion* feedback on college students' writing performance. A sample of 59 students was assigned to three groups: intermittent, continuous, and no feedback. All student participants were asked to produce four essays during seven weeks, among which the first three were practice essays respectively written at weeks three, four, and five, and the fourth one was a test essay written at week seven. The participants in the intermittent feedback group received the *Criterion* feedback at week four, while the continuous feedback group received the feedback at weeks three, four, and five. The results indicated that there was not a significant difference in the holistic scores of the practice essays and the test essay between the two feedback groups and the no feedback group. However, the participants in the continuous feedback group significantly reduced their errors of mechanics, usage, and grammar in both the practice essays and the test essay when compared with their counterparts in the other two groups. Koh (2017) examined the potentially different effects of continuous feedback (CF) and non-continuous feedback (NCF) given by *Criterion* on university students' writing performance. All students were asked to produce two argumentative essays, in addition to a pre-test and a post-test. Assigned to two groups, 20 students either received CF or NCF on the two essays, each of which consisted of three drafts. The students in the NCF group had access to *Criterion* feedback only once for each draft submitted, whereas the students in the CF group had unlimited access to the feedback. The students' post-test essays were scored on the aspects of content, organization, grammar, vocabulary, and mechanics. Also, a holistic score of overall writing quality was provided. The results showed that the students in the CAF group significantly outperformed the students in the NCAF group in terms of overall writing quality, content, and grammar.

The fourth subcategory has only one study investigating the effect of AWE feedback on writing performance of students at different proficiency levels (Shang, 2019). In the study, Shang divided 47 freshmen into high, intermediate, and low proficiency levels based a pre-writing task scored by *Cool Sentence Corrective Network*. All students were asked to finish four writing assignments, with 1 and 3 assignments being given online peer feedback (OPF) and two and four assignments being given automated corrective feedback (AOF). The four assignments were scored in four aspects: number of sentences, grammatical errors, lexical items, and types of words. All scores were provided by multiple computer programs. The results revealed that the students of the three proficiency levels did not significantly differ in the scores of the four aspects for both OPF and AOF.

The fifth subcategory compared the effects of different types of AWE feedback on student writing, and two articles were found (Parra & Calero, 2019; Zhu et al., 2020). In Zhu et al.'s (2020) study, 374 seventh to 12th grade students were placed into two AWE feedback conditions, receiving contextualized feedback and generic feedback, respectively. All students were asked to write eight scientific argumentation blocks about climate change. Domain experts scored the students' blocks prior to the study. Then, the human scored blocks were used to train and validate the automated scoring model using the c-rater-machine learning (ML) engine. Both the contextualized feedback and the generic feedback were provided together with automated scores in the scoring model. Students' revisions after receiving the feedback were voluntary, and they were able to make as many revisions as they wanted. The results indicated that there was no significant difference in the mean score changes for the students receiving generic feedback and the students receiving contextualized feedback. Parra and Calero (2019) explored whether two different AWE systems, namely *Grammark* and *Grammarly*, had different effects on undergraduates' writing performance. Assigned to two groups, 28 undergraduates respectively received feedback from *Grammark* and *Grammarly* after submitting their writings to the two systems. The results revealed the two AWE systems did not significantly differ in the effects on the students' writing performance.

Discussion

Main Findings

The purpose of this review was to examine the effects of AWE feedback on students' writing performance. A systematic literature review methodology was adopted for that purpose. Based on the predefined inclusion and exclusion criteria, a total of 22 studies were identified and included in the present review. The studies were categorized into two broad categories: within-group and between group studies, in accordance with Stevenson and Phakiti's (2014) categorization. The between group studies were further classified into five subcategories. For the within-group studies, one potential drawback in terms of design could be that no control group receiving no feedback was included, which made it hard to attribute any observed improvement in student writing to the provision of AWE feedback (see Ferris, 2004). For the between group studies, it may also be difficult to draw solid conclusions with regard to the effects of AWE feedback because of heterogeneity of the studies, such as sources of AWE feedback, students' language proficiency, instructional contexts, etc.

Despite the potential drawbacks noted above, the findings of the present review are in line with Stevenson and Phakiti (2014) study in that AWE feedback given to one single group of students may effectively help decrease their written errors for both revisions and new writings. Moreover, this positive role of AWE feedback in student writing was also observed in one subcategory of between group studies that included both a group of students receiving AWE feedback and a group of students receiving no feedback. The feature of immediate feedback offered by AWE (Fang, 2010) might explain the positive role of AWE feedback. In fact, this type of immediate feedback, along with automatic scores provided by AWE, can serve as incentives for students to revise their writings (Wang, 2013) because they may want to revise and resubmit writings multiple times until they receive a satisfying score. During the revising and resubmitting process, it is obvious that students are given many opportunities to practice their writing capability.

According to skill acquisition theory, this kind of conscious practice may function as a key for students to enhance their language proficiency level (Dekeyser, 2015). Specifically, skill acquisition theory categorizes development as declarative, procedural, and automatic stages (Taatgen et al., 2008) that occur as a sequence. Dekeyser (2015) defined the declarative stage as knowledge THAT and the procedural stage as knowledge HOW. In terms of AWE feedback, the declarative stage may refer to the stage in which students receive the feedback after they submit their first drafts. The procedural stage may refer to the stage in which students revise their drafts based on the offered feedback. The automatic stage may refer to the stage in which students can automatically apply their acquired procedural knowledge to future writing tasks to improve their writing performance. In this sense, the characteristics of immediate feedback and scores given by AWE systems could provide as many practice opportunities as needed by individual student (Grimes & Warschauer, 2010), making it more likely for her to benefit from the feedback in the long run through integrating the three stages. As pointed out by McGregor, Merchant, and Butler (2008), AWE feedback is beneficial to student writing in that it enables students to act on the feedback when it is fresh in their minds. In fact, this advantage of AWE feedback is even more obvious when compared to other sources of feedback, such as teacher or peer feedback in which immediate feedback is almost unlikely due to large class size. Also, the benefit of repeated practice can be used to explain the result of the present review: continuous AWE feedback was more useful than non-continuous AWE feedback in the improvement of students' writing performance, as the former made the repeated practice more like to occur than the latter.

In addition, several advantages of AWE feedback from affective and emotional perspectives may also contribute to the positive effect of the feedback on student writing. For example, AWE feedback can improve students' writing motivation (Grimes & Warschauer, 2010). This advantage is important in that the level of students' motivation in writing can affect how they attend to the feedback they have received and how they use that feedback in their revisions and new writing tasks (Kormos, 2012). Also, AWE feedback can allow students to be more responsible for the completion of writing tasks (Wang et al., 2013), and enhance students' self-confidence in writing (Khoshnevisan, 2019). Students' self-confidence should be emphasized during the provision of AWE feedback, as their writing proficiency can be affected by both "affective and confidence matters" and "cognitive and linguistic factors" (Bruning & Kauffman, 2016, p. 161). Other potential benefits of AWE feedback are that students who used AWE feedback expressed more enjoyment than students who used teacher feedback (Ware, 2014), and AWE feedback are helpful for reducing students' writing apprehension (Waer, 2021). Fang (2010) noted that students often write through AWE systems after school, which may motivate them to become aware of "the value of independent learning outside the classroom" (p. 254). This type of motivation could be beneficial to the development of student autonomy that is defined as a capability of depending less and less on AWE systems for feedback, and focusing more and more on their own writings. After all, AWE feedback is only a tool for students to practice and improve their writing performance (Milton, 1997).

Limitations

With the purpose of exploring the effects of AWE feedback on student writing, this systematic review focused on studies published in peer-reviewed journals between 2005 and 2020, excluding studies published before 2005 and after 2020, and relevant books, chapters, or conference papers. Also, the present review did not identify and categorize distinctive scoring features of different types of AWE systems, making it hard to draw any solid conclusions about the effects of the systems on student writing, as scores of the different systems might focus on different aspects of writing. For example, *Pigai* feedback provides holistic scores based on grammar, vocabulary, and mechanics, while *Criterion* feedback provides scores based on grammar, usage, mechanics, style, and organization and development. Other limitations are that the present review did not distinguish different writing genres and instructional settings, and some studies in the review examined revision effects of student writing, while other studies examined new text effects of student writing. In addition, the present review focused on quantitative studies, which excluded qualitative studies that could provide more insights into the relationship between AWE feedback and students' writing performance.

Future Research Recommendations

This review summarized five recommendations for future studies. First, for the within-group studies, one recommendation is that future studies may want to include a control group in its design to make the possible effects of AWE feedback more valid. The necessity for the inclusion is to minimize the intervening effects of multiple factors, such as the effects of sufficient time students spent on practicing their writing (Gravetter & Wallnau, 2004). Second, future studies may investigate the effects of different AWE systems on student writing because different characteristics are involved in different systems, such as technological affordance, grading methods, accessibility, etc. Such investigation is meaningful in that it may notify teachers with regard to the choice of the most suitable AWE system to their students. Third, some qualitative methods, such as think-aloud protocols or stimulated interviews, may be used to generate more insights into how students interact with AWE feedback to make it useful to their writings. Fourth, according to Ferris and Kurzer (2019), the research on feedback should take into account learners' individual differences (i.e., for whom feedback may work), which include age, gender, language aptitude, working memory, motivation, etc. For example, future studies may explore how students' working memory is associated with the way they use AWE feedback in their writing, or how the use of AWE feedback affects their writing motivation, which, in turn, affects their writing performance. Fifth, teachers' perceptions of AWE feedback may be considered when examining effects of the feedback because the perceptions could affect the way they apply the feedback to teaching practice and the extent to which students benefit from the feedback.

Conclusions

A review about the effects of AWE feedback on student writing is necessary because of its wide application in L2 writing instruction. Specifically, the necessity of conducting this review is two-fold. First, this review is an update of Stevenson and Phakiti's (2014) review with the inclusion of more recent studies; second, this review may also serve as a guide for L2 writing researchers to have a general understanding about studies that have been conducted on the effects of AWE feedback. Under the framework of within-group and between group studies, the essential results of the review were that AWE feedback might be helpful for student writing under certain conditions, such as the feedback was provided for one single group of students or for a group of students that was compared to the other group of students receiving no such feedback. Moreover, AWE feedback should be continuously offered to make students benefit most from it. The present review could shed light on future research that examines the effects of AWE feedback on student writing.

References

Bruning, R. H., & Kauffman, D. F. (2016). Self-efficacy beliefs and motivation in writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 160-173). New York, NY: Guilford Press.

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*, 267-296. https://doi.org/10.1016/S1060-3743(03)00038-9

Cheng, G. (2017). The impact of online automated feedback on 'students' reflective journal writing in an EFL course. *The Internet and Higher Education 34*, 18-27. http://dx.doi.org/10.1016/j.iheduc.2017.04.002

Dekeyser, R. (2015). Skill Acquisition Theory. In B. Vanpatten, & J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction* (pp. 94-112). Routledge, New York, NY.

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes, 6*(1), 3-17. https://doi.org/10.1016/j.jeap.2006.11.005.

Fang, Y. (2010). Perceptions of the Computer-Assisted Writing Program among EFL College Learners. *Educational Technology & Society, 13*(3), 246-256.

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing, 8*, 1-11. https://doi.org/10.1016/S1060-3743(99)80110-6

Ferris, D. R. (2004). The 'grammar correction' debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime …?). *Journal of Second Language Writing, 13*, 49-62. https://doi.org/10.1016/j.jslw.2004.04.005

Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In Hyland, K., & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81-104). Cambridge: Cambridge University Press.

Ferris, D. R., & Kurzer, K. (2019). Does error feedback help L2 writers? Latest evidence on the efficacy of written corrective feedback. In K. Hyland (Ed), *Second language writing* (pp. 106-124). https://doi.org/10.1017/9781108693974

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street®: Computer support for comprehension and writing. *Journal of educational computing research, 33*(1), 53-80.

Gravetter, F. J., & Wallnau, L. B. (2004). *Statistics for the behavioral sciences* (6th ed.). Belmont, CA: Wadsworth.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Language, and Assessment, 8*(6), 1-43.

Hegelheimer, V., & Lee, J. (2013). The role of technology in teaching and researching writing. In M. Thomas, H. Reinders & M. Warschauer (Eds.), *Contemporary computer-assisted language learning* (pp. 287-302). London & New York: Bloomsbury.

Hibert, A. I. (2019). Systematic literature review of automated writing evaluation as a formative learning tool. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming Learning with Meaningful Technologies* (pp. 199-212). https://doi.org/10.1007/978-3-030-29736-7

Huang, S., & Renandya, W. A. (2018). Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in Language Learning and Teaching*, 1-12. https://doi.org/10.1080/17501229.2018.1471083

Kang, E. Y., & Han, Z. H. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *Modern Language Journal*, *99*(1), 1-18. https://doi.org/10.1111/modl.12189

Kellogg, R., Whiteford, A., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research*, *42*, 173-96.

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *Modern Language Journal, 7*, 305-313. https://doi.org/10.2307/328724

Khoshnevisan, B. (2019). The affordances and constraints of automatic writing evaluation (AWE) tools: A case for Grammarly. *ARTESOL EFL Journal*, *2*(2), 12-25

Kim, J. E. (2014). The effectiveness of automated essay scoring in an EFL college classroom. *Multimedia-Assisted Language Learning*, *17*(3), 11-36. Retrieved from http://journal.kamall.or.kr/wp-content/uploads/2014/10/Kim_17_3_01.pdf

Koh, W. Y. (2017). Effective applications of automated writing feedback in process-based writing instruction. *English Teaching, 72*(3), 91-118.

Kormos, J (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing, 21*, 390-403. http://dx.doi.org/10.1016/j.jslw.2012.09.003

Kurt, G., & Atay, D. (2007). The effect of peer feedback on the writing anxiety of prospective Turkish teachers of EFL. *Journal of Theory and Practice in Education 3*(1), 12-23. Retrieved from http://eku.comu.edu.tr/index/3/1/gkurt datay.pdf

Lachner, A., Burkhart, C., & Nückles, M. (2017). Mind the gap! Automated concept map feedback supports students in writing cohesive explanations. *Journal of Experimental Psychology: Applied, 23*(1), 29-46. http://dx.doi.org/10.1037/xap0000111

Lee, C., Wong, K. C. K., Cheung, W. K., & Lee, F. S. L. (2009). Web-based essay critiquing system and EFL students' writing: A quantitative and qualitative investigation. *Computer Assisted Language Learning*, *22*(1), 57-72.

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

Liao, H. C. (2016 b). Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal*, *70*(3), 308-319. https://doi.org/10.1093/elt/ccv058

Liu, M., Li, Y., Xu, W. W., & Liu, L. (2017). Automated essay feedback generation and its impact on revision. *Ieee Transactions on Learning Technologies*, *10*(4), 502-513.

Lu, X. X. (2019). An empirical study on the artificial intelligence writing evaluation system in China CET. *Big Data, 7*(2), 121-129. https://doi.org/10.1089/big.2018.0151

Maleki, A., & Eslami, E. (2013). The effects of written corrective feedback techniques on EFL students' control over grammatical construction of their written English. *Theory and Practice in Language Studies*, *3*(7), 1250-1257. https://doi.org/10.4304/tpls.3.7.1250-1257

Milton, J. (1997). Providing computerized self-access opportunities for the development of writing skills. In P. Benson, & P. Voller (Eds.), *Autonomy and independence in language learning* (pp. 237-263), London: Longman.

Mohebbi, H. (2021). 25 years on, the written error correction debate continues: an interview with John Truscott. *Asian-Pacific Journal Second and Foreign Language Education, 6*(3). https://doi.org/10.1186/s40862-021-00110-9

Mohsen, M. A., & Alshahrani, A. (2019). The effectiveness of using a hybrid mode of automated writing evaluation system on EFL students' writing. *Teaching English with Technology*, *19*(1), 118-131. Retrieved from http://www.tewtjournal.org

Parra, G. L., & 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

Polio, C., Fleck, C., & Leder, N. (1998). 'If only I had more time': ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing, 7*, 43-68.

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

Shang, H. F. (2019). Exploring online peer feedback and automated corrective feedback on EFL writing performance, *Interactive Learning Environments*, 1-13. https://doi.org/10.1080/10494820.2019.1629601

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51-65.

Stevenson, M., & Phakiti, A. (2019). Automated Feedback and Second Language Writing. In K. Hyland & F. Hyland (Eds.), *Feedback in Second Language Writing: Contexts and Issues* (Cambridge Applied Linguistics, pp. 125-142). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108635547.009

Taatgen, N. A., Huss, D., Dickison, D., & Anderson, J. R. (2008). The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General, 137*, 548-565.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, *46*, 327-369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Waer, H. (2021). The effect of integrating automated writing evaluation on EFL writing apprehension and grammatical knowledge. *Innovation in Language Learning and Teaching*. https://doi.org/10.1080/17501229.2021.1914062

Wang, P. L. (2013). Can automated writing evaluation programs help students improve their English writing? *International Journal of Applied Linguistics & English Literature*, *2*(1), 6-12. https://doi.org/10.7575/ijalel.v.2n.1p.6

Wang, S. W., & Li, R. (2019). An empirical study on the impact of an automated writing assessment on Chinese college students' English writing proficiency. *International Journal of Language and Linguistics*, *7*(5), 226-237. https://doi.org/10.11648/j.ijll.20190705.16

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), 234-57.

Ware, P. (2014). Feedback for Adolescent Writers in the English Classroom Exploring Pen-and-Paper, Electronic, and Automated Options. *Writing & Pedagogy*, *6*(2), 223-249. https://doi.org/10.1558/wap.v6i2.223

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

Zhang, Z., & Zhang, Y. (2018). Automated writing evaluation system: Tapping its potential for learner engagement. *Ieee Engineering Management Review, 46*(3), 29-33. https://doi.org/10.1109/EMR.2018.2866150

Zhu, M. X., Liu, O., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*. https://doi.org/10.1016/j.compedu.2019.103668

Appendix 1

Summary of Within-group Studies

Study	Sample (N)	Predictor variables	Outcome variables	Statistical analyses	Principal results	Statistics
Kim (2014)	45 university freshmen	Automated feedback provided by *Criterion*	-Essay scores of all students -Essay scores of students in high level and low level groups	Paired-samples *t*-tests	-All students' essay scores significantly improved after receiving *Criterion* feedback. -Both high level and low level groups significantly improved their essay scores from first to second drafts.	-All students: t (44) = -12.67, *p* = .00 -High level group: t (23) = -8.90, *p* = .00; Low level group: t(20) = -9.16, *p* = .00
Liao (2016 a)	63 university sophomores	Automated feedback provided by *Criterion*	Learners' grammatical performance	Protected *t* tests	-No significant difference in grammatical performance between essay 1 and 2 -Significant differences between essay 2 and 3, and between essay 3 and 4	-t (62) = 1.59, *p* = .059 -t(62) = 1.83, *p* = .036; t(62) = 5.39, *p* = .000
Liao (2016 b)	66 university sophomores	Automated feedback provided by *Criterion*	Learners' grammatical performan	Paired-samples *t*-tests	-Reduced frequencies of grammatical errors for both revisions and new texts	Refer to Liao, 2016 b, p. 316-317

			ce of revisions and new texts in four error categories: fragments, subject-verb disagreement, run-on sentences, ill-formed verbs			
Mohsen & Alshahrani (2019)	12 university EFL learners	-*MY Access* feedback - *MY Access* feedback and teacher feedback	Holistic scores given by *MY access*	Paired-samples *t*-tests	- *MY Access* feedback helped students improve their writing performance. - *MY Access* feedback and teacher feedback helped students improve their writing performance. - *MY Access* feedback and teacher feedback could better help students improve their writing performance than *MY Access* feedback alone.	-*t*(5) = -10.38, *p* = .000 -*t*(5) = -11.6, *p* = .000 -*t*(5) = -9.64, *p* = .000
Parra & Calero (2019)	28 undergraduates	*Grammark* and *Grammarly* feedback	Writing performance of students' post-test essays	Paired-samples *t*-tests	-Both *Grammark* and *Grammarly* feedback improved students' essays from pre-test to posttest.	-t(13) = -3.38, *p* = .0048; t(13) = -3.42, *p* = .0044
Saricaoglu (2019)	31 ESL students	Automated formative feedback provided by ACDET	-Causal conjunctions -Adverbs -Prepositions -Adjectives -Verbs -Nouns	Wilcoxon signed-rank test	For the first essay: -Reduced causal conjunctions -Increased adverbs -Increased adjectives For the pre- and post-tests: -Significant change for causal adverbs	- Z = -2.58, *p* = .01 - Z = -3.11, *p* = .00 - Z = -2.43, *p* = .02; - Z = -2.70, *p* = .01
Shang (2019)	47 university freshmen	-Automated feedback provided by *Cool Sentence* (essays 2 and 4) -online peer feedback (essays 1 and 3)	-Number of sentences -Grammatical errors -Lexical items -Types of words	Paired-samples *t*-tests	-More sentences in online peer feedback (OPF) essays than *Cool Sentence* feedback (CSF) essays. -Fewer grammatical errors in OPF essays than CSF essays -More lexical items and types of words in OPF essays than CSF essays	-t(46) = 2.61, *p* = .014 -t(46) = -2.30, *p* = .028 -t(46) = 2.79, *p* = .009; t(46) = 3.80, *p* = .001
Wang (2013)	53 English majors	Automated feedback provided by *Criterion*	-Five essays scored by Criterion -A pre-test and a post-test essays	Paired sample *t*-tests	-Improved essays scores from first to final submissions for each of the five essays -Higher scores in students' post-test essays than their pre-test essays	-For the five essays: t(52) = -4.36, *p* = 0.000; t(52) = -6.35, *p* = 0.000; t(52) = -4.41, *p* = 0.000; t(52) = -

			scored by human raters			4.53, $p = 0.000$; t(52) = -4.95, $p = 0.000$ -For the pre-test and the posttest essays: t(52) = 3.081, $p = 0.003$

Appendix 2

Summary of Between-group Studies

Study	Sample (N)	Predictor variables	Outcome variables	Statistical analyses	Principal results	Statistics
Cheng (2017)	138 university students	Online automated feedback (experimental group)/no online automated feedback (control group)	Three reflective journals scored in terms of: -Analysis (A) -Strategy application (S) -External influences (E) -Report of events or experience (R)	Independent samples *t*-test	-For the first journal, the control group outperformed the experimental group in A; -For the second journal, the experimental group outperformed the control group in E; -For the third journal, the experimental group outperformed the control group in both A and E.	-$M_e = 1.27$, $M_c = 1.70$, $p < 0.05$; -$M_e = 0.93$, $M_c = 0.35$, $p < 0.01$; -$M_e = 2.32$, $M_c = 1.60$, $p < 0.001$; $M_e = 1.38$, $M_c = 0.43$, $p < 0.001$
Franzke et al. (2005)	111 8th-grade students	*Summary Street* feedback/no *Summary Street* feedback	-Holistic quality -Content -Organization -Mechanics -Detail -Style -Plagiarism	Analyses of variance with orthogonal contrast codes	-Main effect of condition: *Summary Street* feedback significantly improved students' summaries in measures of holistic quality and content. -Interaction effect of text and condition: *Summary Street* feedback significantly improved students' summaries across time in terms of overall quality, content, organization, detail, and stylistic quality.	-Main effect of condition: holistic quality: M = 3.32, SD = 1.06; M = 2.95, SD = 1.30; p < .05; content: M = 1.51, SD = 0.55; M = 1.28, SD = 0.60; p < .01 -Interaction effect of text and condition: refer to Franzke et al. 2005, p. 69.
Huang & Renandya	67 non-English	Peer feedback with *Pigai*	Six elements in student	Independent samples *t*-test	There was no significant	-t (67) = 0.33, *p* = 0.148

(2018)	majors	feedback/Peer feedback only	revisions were examined: -Content -Organization -Vocabulary -Language use -Mechanics -Overall score		difference between students who received peer feedback with *Pigai* feedback and students who received peer feedback only in the six elements of student revisions.	-t (67) = 0.33, *p* = 0.74 -t (67) = 0.87, *p* = 0.39 -t (67) = -1.40, *p* = 0.17 -t (67) = 0.33, *p* = 0.10 -t (67) = 0.65, *p* = 0.52
Kellogg et al. (2010)	59 university students majoring in English	Intermittent feedback from *Criterion*/continuous feedback from *Criterion*/no feedback	Students' overall error scores	ANACOVA	Students who received continuous feedback significantly produced fewer errors than students who received intermittent feedback and no feedback.	F(1, 36) = 9.70, *p* < .001, *MSE* = .00007
Koh (2017)	20 university EFL student	Non-continuous automated feedback (NCAF)/continuous automated feedback (CAF)	-Total score of student essays -Scores of content, organization, grammar, vocabulary, and mechanics	Independent-samples *t*-tests	Students who received CAF significantly outperformed students who received NCAF in terms of total score, grammar, and content.	-Total score: t (19) = -2.19, *p* = .04; -Grammar: t (19) = -3.13, *p* = .01; -Content: t (19) = -2.76, *p* = .01
Lachner et al. (2017)	42 university students studying advanced Educational Science	*Concept map* feedback/no feedback	Students' revised explanation measured in: -Local cohesion -Global cohesion -Comprehensibility	ANCOVAs	Students who received *concept map* feedback produced more locally, globally, and comprehensible cohesive explanations than students who received no such feedback.	-F (1, 39) = 6.47, *p* = .02, ŋ$^{2}_{p}$ = .14 (medium to large effect) -F (1, 39) = 4.77, *p* = .04, ŋ$^{2}_{p}$ = .11 (medium effect) -F (1, 39) = 10.31, *p* = .00, ŋ$^{2}_{p}$ = .37 (large effect)
Lee et al. (2009)	27 university students	*Essay Critiquing System* (ECS) feedback/no feedback	A holistic mark based on content and organization	Independent-samples *t*-test	No significant difference in the holistic mark between ECS feedback and no feedback groups	*p* > .05, M = 6.86, SD = 1.55 (ECS feedback group); M = 6.88, SD = 1.56 (No feedback group)
Liu et al. (2017)	110 English majors	Indirect corrective feedback (CF) provided by an automatic feedback generation system/direct CF provided by teachers	The features of evaluating a persuasive essay	Independent samples *t*-test	Students who received direct CF outperformed students who received indirect CF in the features of grammar and spelling.	-t (106) = 18.300, *p* < 0.001; -t (106) = 12.236, *p* < 0.001

Lu (2019)	114 university EFL learners	*Pigai* feedback with teacher feedback/teacher feedback only	Scores of a post-test essay	Independent samples *t*-test	Students who received *Pigai* feedback and teacher feedback significantly outperformed students who received teacher feedback.	t (114) = 8.019, $p < 0.05$
Parra & Calero (2019)	28 undergraduates	*Grammark*/*Grammarly* feedback	Students' writing performance of a post-test writing task	Independent samples *t*-test	No significant differences were observed between students who received *Grammark* feedback and students who received *Grammarly* feedback.	$p > .05$, M = 57.2857, M = 55.0714
Shang (2019)	47 freshmen	*Pigai* feedback provided for high, intermediate, and low levels of students	-Number of sentences -Grammatical errors -Tokens -Types	One-way ANOVA	There was no significant difference among the three proficiency levels of students in number of sentences, grammatical errors, tokens, and types.	-F = 0.895, p = 0.416 -F = 2.349, p = 0.108 -F = 0.515, p = 0.601 -F = 1.360, p = 0.268
Wang & Li (2019)	100 university EFL students	*WRM 2.0* feedback (experimental group)/teacher feedback (control group)	-Language form -Contextual structure -Writing quality	Independent samples *t*-test	-When *WRM 2.0* was used to assess student writing, the experimental group significantly outperformed the control group in language form and writing quality. -When student writing was assessed by teachers, the experimental group significantly outperformed the control group in writing quality.	-Language form: word choice (p = 0.004 < 0.05); fluency (p = 0.010 < 0.05); conventions (p = 0.010 < 0.05); Writing quality: p = 0.027 < 0.05 -Writing quality: p = 0.044 < 0.05
Wang et al. (2013)	57 freshmen	*Correct English* feedback (experimental group)/instructor feedback (control group)	Students' writing accuracy	Independent-samples t-test	Students who received *Correct English* feedback significantly outperformed students who received instructor feedback.	p = 0.000, M = 4.84, SD = 3.02 (experimental group); M = 15.58, SD = 9.90 (control group)

Ware (2014)	82 8th grade students	Peer feedback/teacher feedback/AWE feedback	-Text length -Holistic quality -Elements of the genre of open-ended response	Repeated-measures ANOVA	-No group difference was observed for text length and holistic quality; -Both the peer feedback and teacher feedback groups significantly outperformed the AWE feedback group in elements of the genre.	F (1,81) = 6.13, *p* < .01; effect size = 0.14
Wilson & Czik (2016)	151 eighth grade students	Teacher feedback with automated essay evaluation feedback/teacher feedback only	-PEG overall scores -PEG trait scores -Holistic writing quality	One-way ANOVA	There were no statistically significant differences between the two conditions for PEG overall score, PEG trait scores, and holistic writing quality.	-Overall scores: F (1,142) = 0.10, *p* = 0.749 -Trait scores: refer to Wilson & Czik, 2016, p.104 -Holistic scores: F (1,142) = 0.13, *p* = 0.715
Zhu et al. (2020)	374 seventh to twelfth grade students	Automated contextualized feedback/automated generic feedback	-Scores of students' scientific argument writing -Number of student revisions	Independent samples *t*-tests	-There was no significant difference in score changes between contextualized and generic feedback. -Contextualized feedback led to less revisions than generic feedback when both of them generated similar score changes.	-Generic feedback (M = 0.90, SD = 0.69) and contextualized feedback (M = 0.81, SD = 0.69); t (277) = 1.28, p = .20 -Generic feedback (M = 2.20, SD = 1.29) and contextualized feedback (M = 1.77, SD = 1.08); t (242) = 3.18, p < 0.01

Acknowledgements

Not applicable.

Disclosure statement

Not applicable.

Funding

This work was supported by the Funding of Chinese Foreign Language Education under Grant ZGWYJYJJ10A055.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access