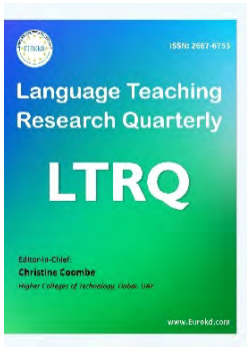


Language Teaching Research Quarterly

2024, Vol. 45, 86–105

A Systematic Review of Automated Writing Evaluation Feedback: Validity, Effects and Students' Engagement

Wen Liu

Faculty of Humanities and Social Sciences, City University of Macau, China School of Foreign Languages, Guangdong University of Science and Technology, China

Received *10 August 2024* ***Accepted*** *01 November 2024*

Abstract

Automated writing evaluation feedback (AWE) has become popular in writing classrooms. However, few studies have conducted a comprehensive review of the employment of AWE in learning areas. This study aimed to provide a systematic review of the current research on AWE feedback, including its validity, effects, and students' engagement with AWE feedback. A total of 68 articles were collected from SSCI and A & HCL-indexed journals, which were published from 2008 to 2023. Concerning the validity of automated feedback, some research reported that the accuracy rate of AWE error flagging was inconclusive and students showed more preference for instructor feedback than AWE feedback. Regarding the effects of AWE feedback, most studies supported the positive evidence of AWE feedback on students' writing outcomes. Additionally, some research found less positive and mixed results, highlighting AWE feedback's shortcomings. Most students held positive feelings toward AWE feedback and perceived it as practical under the intervention of students' engagement. This review has made some implications and suggestions for the stakeholders in this field.

Keywords: *Automated Writing Evaluation Feedback, Writing, Systematic Review, Student's Engagement*

How to cite this article (APA 7th Edition):

Liu, W. (2024). A Systematic review of automated writing evaluation feedback: Validity, effects and students' engagement. *Language Teaching Research Quarterly*, *45*, 86-105. https://doi.org/10.32038/ltrq.2024.45.05

Introduction

Formative assessment focuses on assessing the quality of students' responses, and feedback is treated as a critical factor in formative writing assessment as it narrows the gap between what they write and what is anticipated of them to write, along with suggestions and information for enhancing learners' writing (Biber et al., 2011). Feedback refers to an agent's information

* Corresponding author.
E-mail address: h24092110021@cityu.edu.mo
https://doi.org/10.32038/ltrq.2024.45.05

on evaluating students' performance (Hattie & Timperley, 2007). The agent is often served by the person such as teachers and peers to deliver the feedback. Given the importance of instant feedback in writing pedagogy and teachers' heavy workload of evaluating students' writing, researchers have attempted to design technologies to achieve immediate feedback. With the advent of artificial intelligence, automated writing evaluation (AWE) comes to birth and is widely used in formative assessment in pedagogical activities (Stevenson & Phakiti, 2014; Weigle, 2013) with sophisticated language processing algorithms.

AWE can give students holistic scores and offer instant feedback based on input essays. The widely used AWE programs, such as *Criterion*, *Grammarly* and *Writing Pal*, have expanded their features, providing model essays and analytical scores. Meanwhile, online automated writing assistants have become the research focus. Empirical research has been done to explore AWE feedback from several aspects, including the accuracy, effectiveness, and learners' engagement with AWE feedback. The accuracy of AWE feedback was usually investigated from two aspects: the accuracy of AWE error flagging capability measured by precision and recall and comparison of human feedback and AWE feedback with regard to forms, functions, and types (e.g., Bai & Hu, 2017; Ranalli & Yamashita, 2022). Generally speaking, research exploring the effectiveness of AWE feedback on students' writing outcomes has gained tremendous momentum. Some studies, for example, found positive evidence for automated feedback on improving students' writing accuracy and scores (e.g., Barrot, 2023; Cheng, 2017; Z. Li et al., 2014). On the contrary, some studies revealed no significant difference in improving students' final essays between the teacher feedback group and the automated feedback group (Wilson & Czik, 2016; Wilson & Roscoe, 2020). Additionally, some studies have begun to dig into the way students engaged with AWE feedback from behavioral, cognitive, and emotional conceptualization when it comes to the revision process (e.g., Zhang, 2017; Zhang & Hyland, 2018). Despite some researchers having conducted reviews on AWE (e.g., Nunes et al., 2022), to the best of my knowledge, it seems that no systematic review has presented a comprehensive description of AWE-relevant studies from all these areas. Therefore, this study tried to conduct a systematic literature review of AWE studies, explore the limitations, and provide suggestions for further research.

Previous Reviews of Automated Writing Evaluation Feedback

Among the few synthesis reviews of AWE research, their review focus was mainly concerned with the efficacy of AWE feedback on learners' writing outcomes (Fu et al., 2024; Graham et al., 2015; Nunes et al., 2022; Stevenson & Phakiti, 2014), the validity of AWE feedback (Ding et al., 2024; Shi & Aryadoust, 2024; Strobl et al., 2019) and learners' acceptance toward AWE (Zhai &Ma, 2023). Specifically, the study conducted by Stevenson and Phakiti (2014) delved into the efficacy of the online automated feedback by reviewing book chapters and unpublished articles, which indicated the effects of AWE were quite different due to different factors such as the number of participants, different methods and designs. Moreover, little evidence was found to prove the positive effects of AWE feedback. Graham et al. (2015) embarked on an evaluation of learners' essays from grades one to eight from the angle of formative assessment. They surprisingly discovered that the impact of instant automated feedback on students' essays was relatively small compared to teacher and peer feedback. Two recent review studies have been made by Fu et al. (2024) and Nunes et al. (2022). Nunes et al. (2022) got down to review

eight articles and utilized a quantitative method to assess the impact of AWE on K12 students, which proved that AWE feedback was effective in improving learners' writing outcomes. This finding was similar to the conclusions made by Nunes et al. (2022). Unlike the previous research, Zhai and Ma (2023) confirmed the strong evidence for the positive overall effects of the AWE system. Furthermore, learners' acceptance might affect the efficacy of automated feedback on learners' composition. Apart from the focus on the effectiveness of AWE program, Strobl et al. (2019) examined the validity of 44 automatic assistant writing tools applied to writing instruction. Their research showed that students got benefits from AWE flagging errors at the level of spelling, mechanics and grammar while the AWE feedback related to structure, context and rhetorical issues was hardly provided. Identically, Ding et al. (2024) scrutinized the characteristics and validity of *Criterion*, *Pigai* and *Grammerly*, finding that these prominent AWE tools demonstrated their positive impact on improving student's writing capability. Shi and Aryadoust (2024) found, however, less favorable results in investigating the validity of AWE.

From the above synthesis reviews, it was evident that these few reviews heavily centered on the effects of automated feedback on learners' composition. The synthesis review of students' engagement with AWE feedback was somewhat ignored. Dealing with AWE feedback was regarded as the process by which students interacted from behavioral, cognitive and emotional aspects (Zhang & Hyland, 2018). In the meantime, the feedback was deemed effective in improving writing only under the intervention of students' engagement (Fatawi et al., 2020). In addition, students' interaction with AWE feedback was rather complicated and rarely explored (Ranalli, 2021).

With the wide application of AWE in writing activities, studies of AWE writing assistants have gradually increased. It is necessary to make a systematic review of current research on AWE to offer tendencies and implications for future research (Zhang & Zou, 2020). Stakeholders and researchers, thus, can get insights from the comprehensive review. We hope to offer a panoramic view of AWE studies from the accuracy, effects and students' engagement. The following questions of this study are proposed:

RQ_1: What were the main findings regarding the validity of AWE feedback?

RQ_2: What were the effects of AWE feedback on students' writing outcomes?

RQ_3: What results were found in studies on students' engagement with AWE feedback in the revision process?

Method

Data Collection and Inclusion Criteria

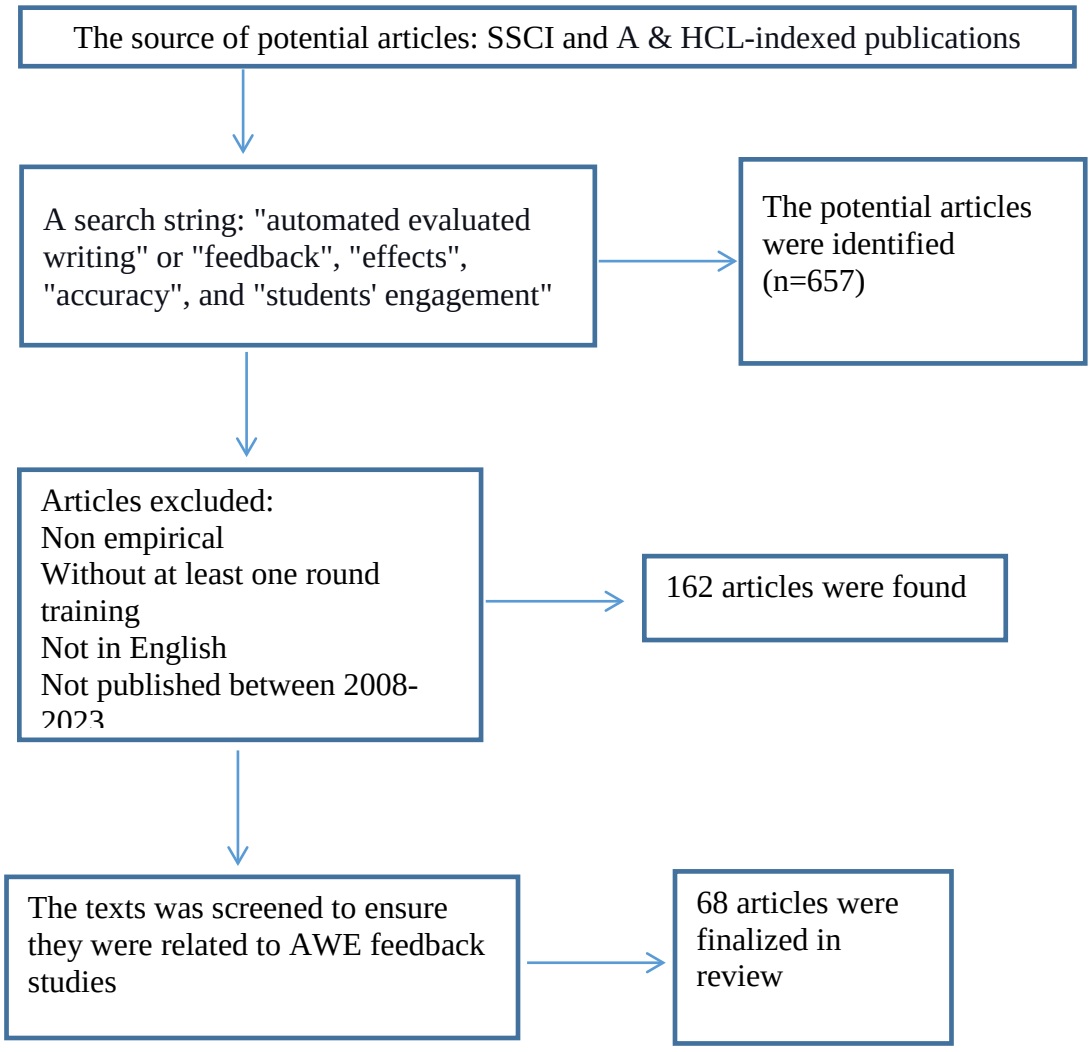
A literature review was limited in retrieving SSCI-indexed publications because they have a strict peer-reviewed process to ensure the quality of articles before publication and greatly influenced on the academic community due to their excellent reputation (Duman et al., 2015). The benefit of retrieving articles from SSCI-indexed journals could reduce the variation of selected articles to strengthen the validation of review findings. Furthermore, A & HCL-indexed publications were also included in the search for related target articles because they demonstrated significant impact in academic areas. This study selected target articles in SSCI and A & HCL-indexed journals by using a search string like "automated evaluated writing" or

"feedback", "effects", "accuracy", and "students' engagement" in the search engine Web of Science and Google Scholar.

Inclusion criteria were framed to select target articles related to the purpose of this research. All the potential articles should follow the strict criteria: 1) be empirical articles published in peer-reviewed journals with the exclusion of book chapters, dissertations, and conference articles; 2) ensure participants in the experiment have trained with at least a complete round of receiving AWE feedback and revised according to the feedback; 3) have students' writing products evaluated after the intervention of AWE; 4) be published in the period from 2008 to 2023; 5) be written in the English language. The preliminary research was 657 papers. After the removal of the articles that failed to meet inclusion criteria, 162 articles were found. Next, two researchers undertook a screening process by reading titles, abstracts, and questions to ensure that selected articles centered on AWE feedback studies. Finally, 68 eligible articles were collected for this review, which was presented in the reference marked with an asterisk. The selection process of eligible articles was presented in Figure 1.

Figure 1

Flow Chart Presenting the Selection Process of Eligible Articles



Coding Scheme

Conducting a review study to answer the above questions, we designed a coding scheme that involves three categories demonstrated in Table 1. The author developed the initial coding scheme. After that, two other teachers with more than a decade of writing teaching experience were involved in coding several articles based on the initial scheme. The issues of accuracy, inconsistency and comprehensiveness have been discussed, and the final coding scheme has been revised to reach an agreement between the coders after two rounds of revision. This coding scheme aimed to help us classify the articles and encode the research contents according to the following categories:1) the validity of AWE feedback, 2) the effects of AWE feedback, and 3) students' engagement with AWE feedback. Regarding assessing the validity of AWE feedback, the coding items were designed with reference to the studies conducted by Bai and Hu (2017) and Ranalli (2018); that is, the accuracy rate of AWE error flagging and the forms, functions and types of automated feedback versus teachers' feedback. In light of the effectiveness of AWE feedback, this study delineated two coding items based on Stevenson and Phakiti (2014): within- and between-group designs. As for students' engagement with AWE feedback, it was coded according to the three conceptualizations put forward by Zhang and Hyland (2018): behavioral, cognitive, and emotional dimensions.

Table 1

Coding Scheme

Category	Coding Items	References
1. the validity of AWE feedback	1) forms, functions and types of AWE feedback versus teacher feedback 2) accuracy rate of AWE error flagging	Bai & Hu, 2017; Ranalli, 2018
2. the effects of AWE feedback	1) within-group design 2) between-group design	Stevenson & Phakiti, 2014
3. students' engagement with AWE feedback	1) behavioral engagement 2) cognitive engagement 3) emotional engagement	Zhang, 2017; Zhang & Hyland, 2018

Coding Results

Regarding the validity of AWE feedback, the accuracy of automated feedback and the agreement between automated feedback and teacher/peer feedback were the main focus. For the former study, if the precision and recall rates of AWE feedback were less than 0.50, the results of these studies were labeled negative; otherwise, they were coded as positive. If some types of feedback were identified with high accuracy while the accuracy of other errors was relatively low, the outcomes of studies were coded mixed. Concerning coding the results of the latter study, the results could be tagged positive when the coverage of errors identified by AWE was more extensive than that of human raters. When AWE identified the correct linguistic items as wrong or its error flagging was limited to some specific errors compared with teacher feedback, the results in this area were considered negative. If both AWE feedback and teacher feedback showed their advantages in identifying different types of errors, the results were coded mixed.

Concerning the effects of AWE feedback on students' quality, if the learners receiving AWE feedback intervention outperformed students who received teacher, peer feedback or no

feedback, the outcomes of studies in this line were considered positive. The results were coded as neutral if no significant difference between groups was reported. If the students with AWE feedback underperformed the students with traditional feedback, these outcomes could be categorized as negative. The findings were regarded as mixed when two or more results were reported in studies.

In the third category, relevant research concerned students' actions to AWE feedback from their emotional, behavioral and cognitive engagements. If the students who used AWE feedback reported more negative feelings than those receiving human feedback, the outcomes were classified as negative. If the students using automated feedback showed more willingness to revise their writing than those with human feedback, the results would be marked positive. When the students tried to understand AWE feedback by themselves and even utilized cognitive strategies, the outcomes were positive.

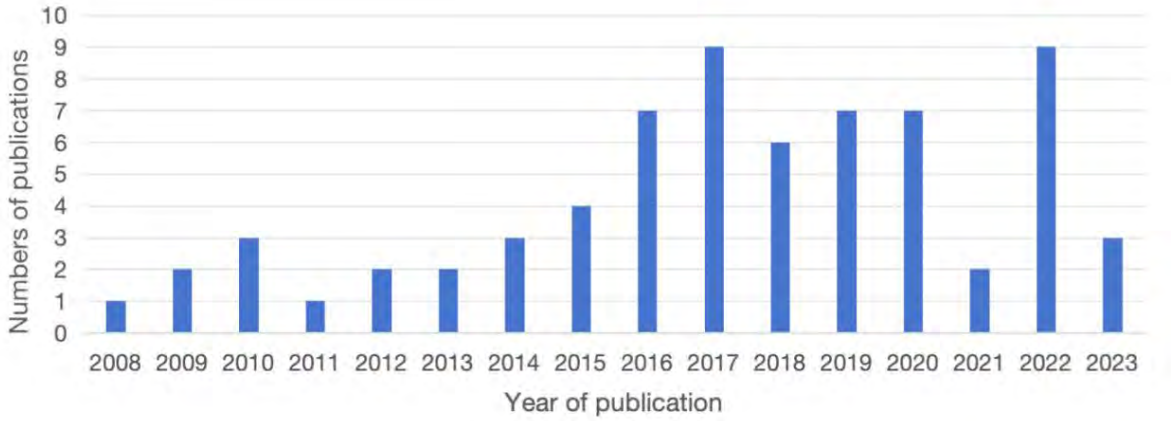
All the articles collected in this review were classified according to their results. Two teachers coded these articles respectively and then all the participants checked the coding results collectively. When the discrepancies occurred, the articles would be rechecked. All the participants finished the discussion until they reached an agreement.

Results

The 68 finalized articles were collected in this review, which was demonstrated in Figure 2.

Figure 2

The Distribution of the Eligible Articles



Validity of AWE Feedback

The results of the validity of automated feedback were shown in Table 2. These articles aimed to evaluate the accuracy of AWE error detection, which could be divided into two lines: the accuracy rate of errors identified by AWE and functions, forms and types of automated feedback versus teacher feedback. Fifteen articles were reported in this respect. In terms of former studies, twelve articles were summarized, five of which showed positive results, four negative results, and three mixed findings. Concerning comparing AWE feedback's functions, forms and types with human feedback, three articles have been targeted with two negative findings and one mixed result.

The accuracy of AWE error flagging was usually assessed by precision and recall rate. Precision measurement was the proportion of corrected detected errors to the total errors

detected by the AWE tool. Oppositely, recall focused on the AWE program's coverage, which indicated the proportion of correctly detected errors to the total errors by human annotators (Burstein et al., 2003). Specifically, four studies showed negative results with relatively low precision and recall rates in quantitative studies. For example, in Liu and Kunnan's (2016) research, they analyzed the accuracy of *WriteToLearn*, the representative AWE assistant system, and concluded that precision and recall were 0.49 and 0.19, respectively. *WriteToLearn* was reported to mislabel specific types of errors as ungrammatical forms, such as prepositions and articles. Feng et al. (2016) explored *CyWrite*'s capability of error detection and concluded similar results that *CyWrite* had a low level of recall at identifying certain types of errors. Meanwhile, five studies showed positive findings in this regard. For example, to deeply understand the knowledge appropriateness of the AWE system, Ranalli and Yamashita (2022) evaluated the capability of *Grammarly*'s error detection. *Grammarly* was found to identify 1412 errors, in which more than 110 error types have precision and recall rates of 0.88 and 0.83, respectively. *Grammarly*'s error detection capability was much higher than that of Microsoft Word NLP in the essays written by the same group of students. Chukharev-Hudilainen and Saricaoglu (2016) investigated the accuracy of the AWE tool in analyzing students' causal discourse in cause-and-effect essays. This research showed that the AWE analyzer had a precision of 0.93 and recall of 0.71, which proved its promising pedagogical application in providing formative feedback in causal discourse. Although some findings showed negative and positive results, three studies reported mixed results. That is to say, the accuracy varied across different error types because some were satisfactory while others had a low accuracy rate. For example, Lavolette et al. (2015) found that the *Criterion* system had a high precision rate of flagging errors such as ungrammatical-formed verbs, capitalization and wrong words, while 47% of errors were identified as wrong with a low recall rate, such as run-on sentence and the improper use of articles. Similarly, Ranalli et al. (2018) examined the accuracy of *Criterion* feedback across ten different error types. They concluded that the accuracy rate was between 71%-77% in general, but the precision rate of identifying extra comma errors was only 51%. Likewise, 75% of errors were detected correctly by *CyWrite*, and it has a low level of recall in identifying specific types of errors (Feng et al., 2016).

Apart from focusing on accuracy rate, another research line compared the types, forms, and functions of AWE feedback with human feedback. Three articles were collected in this area, showing two negatives and one mixed result. Specifically, in Dikli and Bleyle (2014), *Criterion* feedback was compared with teacher feedback and it showed that many error types were misidentified by *Criterion*. Even errors that often occurred in students' writing were misidentified. Compared with *Criterion* feedback, teacher feedback tended to be high quality. Another study also concluded a negative result, which was carried out by Thi & Nikolov (2023). This research centered on an investigation of *Grammarly* feedback versus teacher feedback and the findings showed that teacher feedback identified 410 errors, covering a wide range of writing aspects such as linguistic and content issues. While, *Grammarly* only indicated 281 errors with a heavy focus on superficial linguistic issues of grammar and mechanics. *Grammarly* was a grammar-checking tool. Thus, combining *Grammarly* with teacher feedback was better for improving the efficacy of providing individual feedback. Among AWE feedback versus human feedback, one study (Zhang & Hyland, 2018) presented mixed results, showing that both feedback conditions have their advantages.

Table 2

Results of Validity of AWE Feedback

Focus	Positive	Negative	Mixed
Validity of AWE feedback			
1.Accuracy rate	Calma et al., 2022; Chukharev-Hudilainen & Saricaoglu, 2016; Ranalli & Yamashita, 2022; Thi et al., 2023; Wang, Harrington & White, 2012	Dikli, 2010; Feng et al., 2016; Hoang & Kunnan, 2016; Liu & Kunnan, 2016	Bai & Hu, 2017; Lavolette et al., 2015; Ranalli et al., 2018
2.Functions, forms, types of AWE feedback versus teacher feedback		Dikli & Bleyle, 2014; Thi & Nikolov, 2023	Zhang & Hyland, 2018

Effects of AWE Feedback on Students' Writing Quality

Concerning the findings of the effects of AWE feedback, thirty-two eligible articles were examined based on within-group and between-group design by Stevenson and Phakiti (2014), which has been presented in Table 3. As for within-group designs, nine articles were finalized, with five studies (three were positive results and two were mixed findings) investigating the effectiveness of automated feedback under comparison of AWE feedback condition with no feedback condition in a writing activity. Two of nine studies compared AWE feedback with teacher or peer feedback, which concluded positive and mixed results. Two studies (one was positive and the other was mixed results) focused on comparing AWE+teacher feedback with teacher feedback.

Regarding between-group studies, twenty-three of thirty-two articles meet such criterion, further divided into six subcategories. Specifically, four studies (three positive and one mixed findings) employed AWE feedback versus no feedback. Five studies were concerned with comparing AWE feedback condition with teacher feedback condition, which four studies concluded positive results and one showed mixed findings. AWE+teacher feedback versus teacher feedback only was the most frequently investigated, with a total of seven studies. Among these studies, five studies showed positive findings and two indicated mixed results. Only one study with negative findings compared three different feedback conditions: student receiving AWE feedback, student receiving peer feedback, and student receiving teacher feedback, respectively. Three studies focused on comparing AWE+peer feedback with peer feedback with two positive and one mixed results. Three studies investigated the effectiveness of different types of AWE feedback, which showed one result was negative and two were mixed.

Two main research lines could be found to examine the effects of AWE feedback on students' writing quality. Generally speaking, the results of these findings were inconclusive. Specifically, regarding within-group studies, three studies found benefits of AWE feedback (Cotos, 2011; Liao, 2016; Kim, 2014), while two studies (Kellogg et al., 2010; Saricaoglu, 2019) concluded mixed results concerning comparing AWE feedback versus no feedback. For example, Liao (2015) concluded that *Criterion* feedback helped students lower the rate of grammatical errors, such as subject-verb agreement and ill-formed verbs. Despite the positive

impact of AWE on writing activity, Saricaoglu (2019) explored the impact of AWE feedback on improving L2 students' causal explanations across pre- and post-test essays and indicated mixed results regarding AWE feedback. The findings showed great changes in students' causal explanations in a cause-effect composition but no significant enhancement across pre- and post-drafts. The second subcategory centered on comparing AWE with teacher or peer feedback. Wang et al. (2013) showed positive results of using AWE feedback. However, Shang (2022) found mixed results under AWE feedback versus online peer feedback, which showed that online peer feedback was more helpful than AWE feedback in polishing sentence writing and reducing grammatical errors. On the contrary, AWE feedback could help student diversify their vocabulary. The third subcategory investigated the effects of hybrid mode in writing classrooms. In this regard, Hassanzadeh and Fotoohnejad (2021) proved that hybrid feedback was more beneficial than traditional feedback in improving students' writing quality. However, Mohsen and Alshahrani (2019) concluded mixed findings and stated that both the combination of teacher and automated feedback and teacher feedback have their strength.

Another line of this research was between group studies, covering six subcategories. Firstly, three studies (Barrot, 2023; Cheng, 2017; Lachner et al., 2017) compared AWE feedback with no feedback condition, presenting positive results, while one showed mixed results (Lee et al., 2009). For example, Barrot (2023) explored whether L2 students with access to *Grammarly* AWCF could improve their writing accuracy. Their findings revealed that L2 students getting *Grammarly* AWCF improved accuracy and outperformed students without feedback. Although some studies proved the potential of AWE feedback on advancing learners' writing quality, Lee et al. (2009) indicated no difference in essay length and scores between the group with AWE feedback and the group without feedback. Differences existed in the number of arguments, which indicated that the average number of arguments in the group with AWE feedback was higher than that of the control group.

Comparing automated feedback with traditional teacher feedback could be classified into the second subcategory. Specifically, four studies (R. Li, 2023; Y. J. Wang et al., 2013; M. Liu et al., 2017; S. Wang & Li, 2019) provided positive evidence. In the meantime, one study (Reynolds et al., 2021) showed mixed results in this respect. Liu et al. (2017) concentrated on exploring the impact of AWE system-assistant feedback and corrective feedback given by teachers on students' writing. They found a positive impact on students' writing, particularly in supporting arguments, organization, and conclusion in the short term during the intervention of AWE ICF. Moreover, AWE feedback promoted students outperformed self-correction over feedback provided by the teacher. However, students exposed to AWE feedback could not continue improving their writing performance. The group receiving AWE feedback showed better writing quality than the essays written by the teacher feedback group in terms of the second and third writing. However, the difference in writing performance was gradually demonstrated in the third and fourth essays written by the group with teacher-feedback condition. This might result from students' perceptions toward different feedback conditions.

The third subcategory that has been given some attention was the investigation of the effects of hybrid mode (automated-teacher feedback) versus pure teacher feedback condition on students'composition. Five studies (Link et al., 2022; Lu, 2019; Palermo & Thomson, 2018; Tang & Rich, 2017; Zhai & Ma, 2023) supported the positive potential of hybrid mode, and two studies (Wilson & Czik, 2016; Wilson & Roscoe, 2020) demonstrated mixed results. Lu

(2019) divided students into the experimental group (*JuKu* AWE+teacher feedback) and the group accepting teacher feedback and employed pre- and post-tests to examine the extent to which the hybrid feedback affected students' writing outcomes. The findings suggested that the group with hybrid feedback scored higher than the pure teacher feedback group. Although AWE positively affected writing issues, not all AWE feedback could improve students' writing quality. Compared with the students who got teacher feedback from *Google Docs*, Wilson and Czik (2016) and Wilson and Roscoe (2020) found there was no significant difference between the two groups in final writing, but the group using *PEG* writing+teacher feedback showed increasing writing persistence and motivation.

The fourth subline compared the impact of automated assistant feedback versus teacher feedback versus peer feedback. One study (Ware, 2014) showed negative results, which revealed no evidence that employing AWE improved students' writing quality compared with teacher or peer feedback. Furthermore, students receiving AWE feedback got poorer scores than others on their writing products, such as in the genre aspect.

The fifth subcategory compared the effects of the combination of automated and peer feedback versus peer feedback. In this aspect, two studies (Al-Inbari & Al-Wasy, 2022; Mørch et al., 2017) had positive results, and one study (Huang & Renandya, 2020) provided mixed results. Al-Inbari and Al-Wasy (2022) aimed to evaluate the effect of automated evaluation system on cause-effect essays by comparing the automated peer experimental group and peer control group. Al-Inbari and Al-Wasy (2022) conducted a study to assess the impact of an automated evaluation system on cause-and-effect essays written by the automated peer experimental group and peer control group. They found that a significant improvement was shown in essays produced by the L2 students who received automated peer feedback from *WRITER* AWE software. Compared with the positive assumption of AWE in writing instruction, Huang and Renandya (2020) revealed that L2 students at low English proficiency levels held promising attitudes toward Pigai, although no evidence was shown in writing improvement for students receiving AWE feedback.

The sixth subcategory focused on investigating different types of AWE feedback, in which two studies (Parra & Calero, 2019; Zhu et al., 2020) were identified with mixed findings, and one study (Allen et al., 2019) showed negative results. Zhu et al. (2020) revealed no significant score changes obtained by the group with contextualized feedback and generic feedback from the AWE assistant program. Similarly, Parra and Caler (2019) systematically studied the effects of *Grammarly* and *Grammark* automated feedback on students' writing performance. No difference was found in learners' writing quality under the circumstance of receiving the feedback from those two automated evaluation software. One study provided formative and summative feedback and showed no impact was found for the *W-PAL* AWE system.

Table 3

Results of the Effects of AWE Feedback on Students' Writing

Focus	Subcategories	Positive	Negative	Mixed
The effects of AWE				
1.within group	1) AWE feedback vs no feedback	Cotos, E., 2011; Kim, 2014; Liao, 2016;		Kellogg, Whiteford & Quinlan, 2010; Saricaoglu, 2019
	2) AWE feedback vs human (teacher/peer) feedback	Wang, 2013		Shang, 2022
	3) AWE feedback+teacher vs teacher feedback	Hassanzadeh & Fotoohnejad, 2021		Mohsen & Alshahrani, 2019
2.between group	1) AWE feedback vs no AWE	Barrot, 2023; Cheng, 2017; Lachner et al., 2017		Lee et al., 2009
	2) AWE feedback vs teacher Feedback	Liu et al., 2017; R. Li, 2023; Wang & Li, 2019; Wang, et al., 2013;		Reynolds, Kao & Huang, 2021
	3)AWE+teacher feedback vs teacher feedback	Link et al., 2022; Lu, 2019; Palermo & Thomson, 2018;		Wilson & Czik 2016; Wilson & Roscoe, 2020
	4)AWE vs teacher feedback vs peer feedback	Tang & Rich, 2017; Zhai & Ma, 2023		
	4) AWE vs teacher feedback vs peer feedback		Ware, 2014	
	5)AWE+peer/AWE feedback vs peer feedback	Al-Inbari & Al-Wasy, 2022; Mørch, et al., 2017		Huang & Renandya, 2020;
	6) different AWE feedback		Allen et al., 2019	Parra & Calero, 2019; Zhu et al., 2020

Students' Engagement with Feedback in Revision

Students' engagement with AWE feedback was generally related to behavioral, emotional, and cognitive processes (Zhang & Hyland, 2018). Precisely, Affective engagement reflects students' attitudinal changes and emotional reactions to feedback. Behavioral engagement delineates the revision acts, time spent on revision, and the number of submissions. Cognitive engagement refers to understanding the feedback content and cognitive strategies used in revision processes. Forty-four articles were identified in this respect, as shown in Table 4. Regarding emotional engagement, twenty-one was collected in the review, which took the largest percentage of students engagement studies. Ten showed positive feelings by learners, two negative results, two neutral attitudes and seven mixed perceptions. Regarding behavioral engagement, eighteen articles were identified. Eight of eighteen outcomes were positive, two were negative, three were mixed findings, and five had no significant impact on revision. Five studies were summarized concerning the effects of AWE feedback on students' cognitive engagement, four of which showed positive results and only one with negative results.

Specifically, the category of emotional engagement was the main interest of students' engagement studies. Positive attitudes toward AWE were reported. For example, a study by O'Neill and Russell (2019) compared students' reactions to *Grammarly* with students' responses to traditional feedback based on mixed methods. In their study, students responded positively to *Grammarly* and favored of automated feedback compared with traditional feedback. The findings, however, indicated that students did not always take a positive attitude about AWE feedback when compared to human feedback. Take Lai's study (2010) as an example. Students believed feedback from human tended to be more specific and direct with advice, but AWE feedback provided vague and unreadable information. Additionally, students felt stressed when they thought they had entirely accepted suggestions provided by AWE software. Other students held mixed attitudes toward AWE. In Bai and Hu's (2017) research, students could critically select AWE feedback and suggestions based on the accuracy of the different feedback. In other words, students have confidence in some AWE feedback, such as grammar and mechanics. However, they were pessimistic about content-level feedback given by the automated evaluation system. Moreover, students gradually developed an awareness of the shortcomings and potentials of AWE as an assisted writing system. Apart from the aforementioned perception by learners, Calvo and Ellis (2010) found that students' perception toward AWE and traditional feedback was similar because both were either fragmented or cohesive.

Regarding the learners' behavioral engagement, the positive evidence showed that online automated feedback encouraged students to carry out revision actions like rejection, substitution, no change (Koltovskaia, 2020), reorganization, and additions (Roscoe et al., 2017; Sung et al., 2016). Besides, Zhang (2017) found that students using AWE feedback showed a strong willingness to spend more time on revision, and they were more likely to write more drafts as AWE feedback was understandable. Nevertheless, not all AWE feedback came into play when students were revising their writing. For example, Lai (2010) reported that peer feedback involved social interaction and audience awareness as it differed from AWE feedback and promoted more revisions and discussions than automated feedback. In addition to positive and negative results in this respect, some studies concluded neutral results. Sherafati et al. (2020) aimed to explore the effects of computer-mediated teacher feedback and automated feedback on students' writing outcomes according to students' motivation. They reported that students were inclined to utilize computer-mediated teacher feedback in the delayed post-test and emphasized the long-term transfer effects of computer-automated feedback. Mixed results were also identified (Bai & Hu, 2017; Koltovskaia, 2020; Zhu et al., 2017). For instance, the revision pattern was investigated by Zhu et al. (2017) and they found that students' initial scores were not directly correlated with students' behavioral revision actions, such as the number of submissions of revisions and time spent on revision. Compared with students who got lower scores at the first submission, learners with higher scores preferred to notice feedback.

Automated feedback could encourage students to engage with writing activities cognitively. Research into students' cognitive engagement was quite a few. Only five studies concerning students' cognitive engagement were included in this review. Four reported positive results and one showed negative findings. For example, in Zhang's (2017) study, one student tried to get high grades by understanding AWE feedback by herself. The learner paid attention to the suggestions and information from AWE and revised according to the advice. This study found

that this student even attempted to adopt cognitive strategies, which was reflected in selectively adopting AWE feedback rather than receiving all the feedback. Thus, students' writing strategies were gradually developed.

Table 4

Results of Students' Engagement with AWE Feedback

Focus	Positive	Negative	Neutral	Mixed
Students' engagement with AWE				
1.emotional engagement	Alnasser, 2022; Cheng, 2017; Dikli & Bleyle, 2014; Garcia-Gorrostieta et al., 2018; Landauer et al., 2009; O'Neill & Russell, 2019; Roscoe et al., 2017; Wang et al., 2013; Wilson & Roscoe, 2020; Zhang, 2017	Lai, 2010; Zaini, 2018	Calvo & Ellis, 2010; Shang, 2022;	Bai & Hu, 2017; Barrot, 2021; Chen & Cheng, 2008; Chen et al., 2016; Li et al., 2015; Xu & Zhang, 2022; Zhu et al., 2017
2.behavioral engagement	Guo et al., 2022; Lachner et al., 2017; Lee, 2020; Proske et al., 2012; Roscoe et al., 2017; Zaini, 2018; Zhang, 2017; Zhang & Hyland, 2018	Lai, 2010 Sung et al., 2016	Lee et al., 2009; Proske et al., 2012; Sherafati et al., 2020; Wang, 2013; Zhang, 2020;	Bai & Hu, 2017; Koltovskaia, 2020; Zhu et al., 2017
3.cognitive engagement	Cotos et al., 2017; Han & Hyland, 2015; Wang et al., 2013; Zhang, 2017	Zhang & Zhang, 2024		

Discussion

This review study indicated that great attention has recently been paid to automated writing assistants. Automated writing evaluation system adopts natural language processing technologies (Zhu et al., 2020), and its advantages have been proven. Despite the benefits of the AWE system, studies in this respect are still relatively scarce, and further AWE-related studies should be enriched.

Research Question 1

The first question focused on the accuracy of errors identified by AWE. It was found that precision and recall were widely applied to measure the accuracy of automated feedback. The advantage of AWE feedback was verified. AWE could accurately target specific types of errors, such as mechanics, grammar, and spelling, which could improve students' writing quality. Despite the positive evidence proving the advantages of AWE feedback, other studies reported that accuracy varied across different error types. That is, some errors were identified with a high accuracy rate, while others were not satisfactory. In addition to positive and mixed results, negative results were reflected in a low precision and recall rate. This low accuracy rate was partly caused by AWE software's limited capabilities of analyzing texts. At present, most AWE software adopt similar natural language processing mechanics, meaning the capability to analyze input text mainly depends on the size of samples in selected corpora (Roscoe et al., 2017). In other words, the low precision was caused by the small sample size (Cook & Hatala,

2015), and the working mechanism behind AWE led to the ignorance of the context-feature feedback. Therefore, employing a repeated measure design (e.g., crossover design) was recommended to achieve high precision in samples of small sizes. Meanwhile, it is suggested that the future research could integrate *ChatGPT* into AWE-related research to provide more accurate and in-depth feedback.

The synthesis review also reported the superiority of human feedback over AWE feedback. This phenomenon might be because AWE feedback omitted human inferencing skills and centered on indirect and vague information (Liu et al., 2017; Shang, 2022), making it hard for students to understand and revise their essays. As a result, students' revision was often conducted at word or sentence level while content-related problems seemly remain unchanged. In this way, students showed an inclination to teacher feedback with social context. Given the negative washback of AWE software, it was recommended to take the AWE feedback as a supplement to human feedback. That is to say, teacher feedback should be involved in writing pedagogy to provide content-related information and AWE was applied to provide linguistic feature feedback in the teaching activity. This co-construction of writing knowledge could help students promote efficient revision.

Research Question 2

The second question mainly investigated the effects of AWE feedback on students' writing quality. This research line could be categorized into within-group and between-group studies (Stevenson & Phakiti, 2014). The promising effects of AWE feedback were observed. AWE feedback helped students reduce grammatical errors in group studies (Liao, 2016). The benefits of AWE feedback were also demonstrated in between-group studies. Access to AWE feedback helped students understand how to use the target language grammatically and developed their learning autonomy (Wang, 2013). The instant feedback and scores provided by AWE would motivate students to further take revision actions when polishing their essays because they tended to revise essays until they got satisfactory scores.

Despite the benefits above, concerning the within-group designs, one limitation was the lack of a control group for comparison, which made it hard to conclude that the provision of AWE feedback accounted for the improvement of students' writing quality (Ferris, 2004). In the aspect of between-group design, limitations also existed in comparison between AWE and teacher feedback. The findings in this field needed to be more solid due to the lack of ecological validity (Palermo & Wilson, 2020). In other words, the comparison between AWE and teacher feedback was not examined from two equivalent forms. Moreover, this design overemphasized the dichotomy between automated software and teachers, which further aroused controversy over replacing AWE with teacher. As stated by Attali (2013), AWE maximized its potential as it was regarded as a supplement tool instead of the teachers' equivalent. Therefore, writing pedagogy would benefit from AWE as a supplement. Further writing instruction should combine teacher feedback for content aspects and AWE feedback for linguistic issues. Besides, AWE was mainly applied in argumentative essays instead of descriptive and expository essays. Thus, some studies recommended that the effects of AWE should be explored in various contexts, such as various genres (Lachner et al., 2017). Examining the affordance of AWE in different settings in future research enhanced students' confidence in employing AWE in writing instruction.

Research Question 3

The third question placed emphasis on the way students responded to automated feedback information in light of behavioral, emotional and cognitive engagement. This systematic review showed that students' perceptions of AWE feedback were positive, negative, and mixed. Students' different perceptions of the usefulness of AWE influence the adoption of technical support. Meanwhile, the perception of AWE determined how much writing quality improved according to automated feedback (Tsai, 2014). Students' positive perceptions toward AWE could maximize the potential of AWE software. That is, teachers should make use of their guiding role to influence what students felt and how they accepted the automated feedback in writing activity (Roscoe et al., 2017). Therefore, teachers' perceptions should carefully guide students' attitudes toward using AWE in writing instruction.

From the above-reviewed articles, the results showed that AWE could promote students to undertake more revision actions (Sung et al., 2016), control their writing process (Proske et al., 2012), spend more time on revision and revise their writing many times (Zhang, 2017). The features of immediacy and interactivity of automated feedback strengthen AWE feedback on cultivating students' writing skills. The problem that students encountered could be solved by the instant feedback from AWE system. However, little evidence was found about how students engaged with feedback and their revision actions. Future research should focus on such a field to provide insights into how writing knowledge was developed via AWE-involved classrooms. Despite the advantages of AWE, some students still preferred teacher feedback to the AWE feedback in taking revision actions. They believed teacher feedback took humans as real audiences (Lai, 2010). Moreover, AWE's form-related feedback would result in students' revision at the superficial level and an incomplete assessment of students' writing capability, which might lead to failure to meet some students' needs beyond form-related issues. Therefore, AWE should be appropriately integrated into writing instruction. Specifically, teacher commentary could be added after the sentence-level feedback provided by AWE. It was optimal to employ such a hybrid model (Wang, 2015).

Writing is a rather complex process (Weigle, 2013), which can be broken down into 'plan, draft and revise' (Hyland, 2009). Students' cognitive engagement refers to how students respond to the AWE feedback and employ revision operations and metacognitive strategies to polish writing outcomes (Koltovskaia, 2020). AWE study (Zhang, 2017) suggested that students' writing strategies could be cultivated as they independently utilized cognitive strategies such as management of their revision process with the help of the AWE feedback. Obviously, the automated feedback could not directly affect the improvement of students' writing quality. It depended on how AWE feedback was processed by students' mental efforts to evaluate and internalize it. To better help student revise their essays, teachers should monitor students' revision process and reasons behind revision actions. Meanwhile, they could provide instant assistance when students did not figure out the feedback or were low motivation in revision. For example, teachers could guide students to critically evaluate the feedback rather than passively receive all the feedback, which further cultivated revision strategies to monitor students' revision process. Besides, creativity is regarded as an essential component of writing (Onkas, 2015) and developing creative thinking skills is conducive to generating elaborate ideas in essays. Teachers, thus, might utilize a case study to examine the impact of AWE from this aspect.

Conclusion

This review tried to make a comprehensive summary of AWE-related study to indicate further research and presented a synthesis review in three lines: the validity, the effects of automated feedback on students' writing outcomes and students' engagement in revision. The findings indicated that automated feedback plays a vital role in improving learners' composition, especially improving linguistic capability. However, compared with human feedback, it could not meet students' need to provide content-related suggestions. We also found that the accuracy of the AWE system varied across different types of errors, and responses to AWE feedback are closely related to students' engagement from behavioral, cognitive and emotional aspects. Such engagement further impacts learners' revision performance. Therefore, further research may continue to explore the accuracy of AWE feedback, how learners engage with automated feedback, and what factor influences students' revision actions.

This review has some limitations. The first limitation is that all the finalized articles in this review belonged to empirical studies and were mainly indexed in SSCI and A & HCL journals, limiting the research scope. Secondly, the research was limited in three aspects, which means the findings of this study presented a small part of published articles instead of all the articles in this field. This limitation is common in other synthesis-reviewed articles (Hung et al., 2018). Further reviewed studies, thus, expand corpora to cover more research to solid validation of synthesis review. Hopefully, this study might serve as an incentive to guide further research, and the limitation above will be handled in carefully designed studies.

ORCID

https://orcid.org/0009-0001-6564-9607

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

References

*Al-Inbari, F. A. Y., & Al-Wasy, B. Q. M. (2023). The impact of automated writing evaluation (AWE) on EFL learners' peer and self-editing. *Education and Information Technologies*, *28*(6), 6645-6665. https://doi.org/10.1007/s10639-022-11458-x

*Allen, L. K., Likens, A. D. & McNamara, D. S. (2019). Writing flexibility in argumentative essays: A multidimensional analysis. *Reading and Writing*, *32*(6), 1607–1634. https://doi.org/10.1007/s11145-018-9921-y

*Alnasser, S. M. N. (2022). EFL Learners' perceptions of integrating computer-based feedback into writing classrooms: Evidence from Saudi Arabia. *SAGE Open*, *12*(3), 215824402211230. https://doi.org/10.1177/21582440221123021

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). Routledge.

*Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

*Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy. *Computer Assisted Language Learning*, *36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Research Report Series*, *2011*(1), i–99. https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Burstein, J., Chodorow, M., & Leacock, C. (2003, August). CriterionSM online essay evaluation: An application for automated evaluation of student essays. *IAAI*, 3–10.

*Calma, A., Cotronei-Baird, V. & Chia, A. (2022). Grammarly: An instructional intervention for writing enhancement in management education. *The International Journal of Management Education*, *20*(3), 100704. https://doi.org/10.1016/j.ijme.2022.100704

*Calvo, R. A. & Ellis, R. A. (2010). Students' conceptions of tutor and automated feedback in professional writing. *Journal of Engineering Education*, *99*(4), 427–438. https://doi.org/10.1002/j.2168-9830.2010.tb01072.x

*Chen, C. F. E. & Cheng, W. Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning and Technology*, *12*(2), 94–112. http://dx.doi.org/10125/44145

*Chen, W., Chen, M., Chen, M., & Ku, L. (2015). A computer-assistance learning system for emotional wording. *IEEE Transactions on Knowledge and Data Engineering*, *28*(5), 1093–1104. https://doi.org/10.1109/tkde.2015.2507579

*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. https://doi.org/10.1016/j.iheduc.2017.04.002

*Chukharev-Hudilainen, E. & Saricaoglu, A. (2016). Causal discourse analyzer: Improving automated feedback on academic ESL writing. *Computer Assisted Language Learning*, *29*(3), 494–516. https://doi.org/10.1080/09588221.2014.991795

Cook, D. A., & Hatala, R. (2015). Got power? A systematic review of sample size adequacy in health professions education research. *Advances in Health Sciences Education: Theory and Practice*, *20*(1), 73–83. https://doi.org/10.1007/s10459-014-9509-5

*Cotos, E., Link, S. & Huffman, S. (2017). Effects of DDL technology on genre learning. *Language Learning & Technology*, *21*(3), 104–130. https://doi.org/10.4324/9781003002901-28

*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, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*, *29*, 1–53. https://doi.org/10.1007/s10639-023-12402-3

Duman, G., Orhon, G., & Gedik, N. (2015). Research trends in mobile assisted language learning from 2000 to 2012. *ReCALL*, *27*(2), 197–216. https://doi.org/10.1017/S0958344014000287

Fatawi, I., Degeng, I. N. S., Setyosari, P., Ulfa, S., & Hirashima, T. (2020). Effect of online-based concept map on student engagement and learning outcome. *International Journal of Distance Education Technologies* (IJDET), *18*(3), 42–56. https://doi.org/ 10.4018/978-1-6684-7540-9.ch040

*Feng, H. H., Saricaoglu, A., & Chukharev- Hudilainen, E. (2016). Automated error detection for developing grammar proficiency of ESL learners. *CALICO Journal*, *33*(1), 49–70. https://doi.org/10.1558/cj.v33i1.26507

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*(1), 49–62. https://doi.org/10.1016/j.jslw.2004.04.005

Fu, Q. K., Zou, D., Xie, H. & Cheng, G. (2024). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*. *37*(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

*Garcia-Gorrostieta, J. M., & Lopez-Lopez, A. (2018). Argument component classification in academic writings. *Journal of Intelligent & Fuzzy Systems*, *34*(5), 3037–3047. https://doi.org/10.3233/jifs-169488

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, *115*(4), 523–547. https://doi.org/10.1086/681947

*Guo, Q., Feng, R., & Hua, Y. (2022). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*, *35*(9), 2312–2331. https://doi.org/10.1080/09588221.2021.1879161

*Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of Second Language Writing*, *30*, 31–44. https://doi.org/10.1016/j.jslw.2015.08.002

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3389/fpsyg.2019.03087

Hassanzadeh, M., & Fotoohnejad, S. (2021). Implementing an automated feedback program for a Foreign Language writing course: A learner-centric study: Implementing an AWE tool in a L2 class. *Journal of Computer Assisted Learning*, *37*(5), 1494-1507.

*Hoang, G. T. L. & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY access. *Language Assessment Quarterly*, *13*(4), 359–376. https://doi.org/10.1080/15434303.2016.1230121

*Huang, S., & Renandya, W. (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

Hung, H. T., Yang, J. C., Hwang, G. J., Chu, H. C., & Wang, C.-C. (2018). A scoping review of research on digital game-based language learning. *Computers & Education*, *126*, 89–104. https://doi.org/10.1016/j.compedu.2018.07.001

Hyland, K. (2009). *Teaching and research writing*. Pearson Longman.

*Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42*(2), 173–196. https://doi.org/10.2190/ec.42.2.c

*Kim, J. E. (2014). The effectiveness of automated essay scoring in an EFL college classroom. *Multimedia-Assisted Language Learning*, *17*(3), 11–36. https://doi.org/10.15702/mall.2014.17.3.11

*Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, *44*, 100450. https://doi.org/10.1016/j.asw.2020.100450

*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. https://doi.org/10.1037/xap0000111

*Lai, Y. H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology*, *41*(3), 432–454. https://doi.org/10.1111/j.1467-8535.2009.00959.x

*Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory into Practice*, *48*(1), 44–52. https://doi.org/10.1080/00405840802577593

*Lavolette, E., Polio, C. & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language, Learning & Technology*, *19*(2), 50–68. https://doi.org/10.1080/1355800770140107

*Lee, C. (2020). A study of adolescent English learners' cognitive engagement in writing while using an automated content feedback system. *Computer Assisted Language Learning*, *33*(1–2), 26–57. https://doi.org/10.1080/09588221.2018.1544152

*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. https://doi.org/10.1080/09588220802613807

*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, R. (2023). Still a fallible tool? Revisiting effects of automated writing evaluation from activity theory perspective. *British Journal of Educational Technology*, *54*(3), 773–789. https://doi.org/10.1111/bjet.13294

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *44*, 66–78. https://doi.org/10.1016/j.system.2014.02.007

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

*Link, S., Mehrzad, M. & Rahimi, M. (2022). 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. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *The CALICO Journal*, *33*(1), 71–91. https://doi.org/10.1558/cj.v33i1.26380

*Lu, 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

*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.

*Mørch, A. I., Engeness, I., Cheng, V. C., Cheung, W. K., & Wong, K. C. (2017). Essay Critic: Writing to learn with a knowledge-based design critiquing system. *Educational Technology and Society*, *20*(2), 213–223.

Nunes, A., Cordeiro, C., Limpo, T. & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, *38*(2), 599–620. https://doi.org/10.1111/jcal.12635

*O'Neill, R., & Russell, A. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology*, *35*(1), 37-95. https://doi.org/10.14742/ajet.3795

Onkas, N. A. (2015). Interpretation theory and creative writing. *The Anthropologist*, *22*(2), 196–202. https://doi.org/10.1080/09720073.2015.11891869

*Palermo, C., & Thomson, M. M. (2018). Teacher implementation of selfregulated 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. *The Journal of Writing Research*, *12*(1), 63–108. https://doi.org/10.17239/jowr-2020.12.01.04

*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

*Proske, A., Narciss, S., & McNamara, D. S. (2012). Computer-based scaffolding to facilitate students' development of expertise in academic writing. *Journal of Research in Reading*, *35*(2), 136–152. https://doi.org/10.1111/j.1467-9817.2010.01450.x

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. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816

*Ranalli, J. & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of writing performance: A quasi-experimental study. *The Asia-Pacific Education Researcher*, *30*(6), 585–595. https://doi.org/10.36892/ijlts.v5i3.495

*Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2018). 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

*Reynolds, B. L., Kao, C.-W. & Huang, Y. (2021). Investigating the effects of perceived feedback source on second language writing performance: A quasi-experimental study. *The Asia-Pacific Education Researcher*, *30*(6), 585–595. https://doi.org/10.1007/s40299-021-00597-3

*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

*Shang, H. F. (2022). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments*, *30*(1), 4–16. https://doi.org/10.1080/10494820.2019.1629601

*Sherafati, N., Largani, F. M., & Amini, S. (2020). Exploring the effect of computer-mediated teacher feedback on the writing achievement of Iranian EFL learners: Does motivation count? *Education and Information Technologies*, *25*(5), 4591–4613. https://doi.org/10.1007/s10639-020-10177-5

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 1–23. https://doi.org/10.1017/s0958344023000265

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

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A. & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, *131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005

*Sung, Y. T., Liao, C. N., Chang, T. H., Chen, C. L., & Chang, K. E. (2016). The effect of online summary assessment and feedback system on the summary writing on 6th graders: The LSA-based technique. *Computers & Education*, *95*, 1–18. https://doi.org/10.1016/j.compedu.2015.12.003

*Tang, J., & Rich, C. S. (2017). Automated writing evaluation in an EFL setting: Lessons from China. *JALT CALL Journal*, *13*(2), 117–143. https://doi.org/10.29140/jaltcall.v13n2.215

*Thi, N. K., Nikolov, M. & Simon, K. (2023). Higher-Proficiency students' engagement with and uptake of teacher and Grammarly feedback in an EFL writing course. *Innovation in Language Learning and Teaching*, *17*(3), 1–16. https://doi.org/10.1080/17501229.2022.2122476

Tsai, Y. R. (2014). Applying the technology acceptance model (TAM) to explore the effects of a course management system (CMS)-assisted EFL writing instruction. *CALICO Journal*, *32*(1), 153–171. https://doi.org/10.1558/calico.v32i1.25961

*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, P. L. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching*, *12*(1), 79–100.

*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, Y., Harrington, M., & White, P. (2012). Detecting breakdowns in local coherence in the writing of Chinese English learners. *Journal of Computer Assisted Learning*, *28*(4), 396–410. https://doi.org/10.1111/j.1365-2729.2011.00475.x

*Ware, P. (2014). Feedback for adolescent writers in the English classroom. *Writing & Pedagogy*, *6*(2), 223–249. https://doi.org/10.1558/wap.v6i2.223

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp.36–54). Routledge/Taylor & Francis Group.

*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

*Xu, J, & Zhang, S. (2022). Understanding AWE feedback and english writing of learners with different proficiency levels in an EFL classroom: A sociocultural perspective. *The Asia-Pacific Education Researcher*, *31*(4), 357–367. https://doi.org/10.1007/s40299-021-00577-7

*Zaini, A. (2018). Word processors as monarchs: Computer-generated feedback can exercise power over and influence EAL learners' identity representations. *Computers & Education*, *120*, 112–126. https://doi.org/10.1016/j.compedu.2018.01.014

*Zhai, N. & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(4), 875–900. https://doi.org/10.1177/07356331221127300

*Zhang, Z.V. (2017). Student engagement with computer-generated feedback: A case study. *ELT Journal*, *71*(3), 317–328. https://doi.org/10.1093/elt/ccw089

*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

*Zhang, J. & Zhang, L. J. (2024). The effect of feedback on metacognitive strategy use in EFL writing. Computer Assisted Language Learning, *37*(5–6), 1198–1223. https://doi.org/10.1080/09588221.2022.2069822

Zhang, R., & Zou, D. (2020). Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. *Computer Assisted Language Learning*, *35*(4), 696–742. https://doi.org/10.1080/09588221.2020.1744666

*Zhang, Z. V. & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90–102. https://doi.org/10.1016/j.asw.2018.02.004

*Zhang, Z. V. & Hyland, K. (2022). Fostering student engagement with feedback: An integrated approach. *Assessing Writing*, *51*,100586. https://doi.org/10.1016/j.asw.2021.100586

*Zhu, M. X., Lee, H. S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668. https://doi.org/10.1080/09500693.2017.1347303

*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. *Computer & Education*, *1433*, 103668. https://doi.org/10.1016/j.compedu.2019.103668