IJOLAE Indonesian Journal on Learning and Advanced Education

http://journals.ums.ac.id/index.php/ijolae

Transformative Practices: Integrating Automated Writing Evaluation in Higher Education Writing Classrooms - A Systematic Review

Indri Astutik[1], Utami Widiati[2], Devinta Puspita Ratri[3], Peggy Magdalena Jonathans[4], Nurkamilah[5]✉, Yeni Mardiyana Devanti[6], Zaldi Harfal[7]

[1,2,3,4]Faculty of Letters, Universitas Negeri Malang, Indonesia
[1,5,6]Faculty of Teacher Training and Education, Universitas Muhammadiyah Jember, Indonesia
[3]Faculty of Cultural Studies, Universitas Brawijaya, Indonesia
[4]Teacher Training and Education Faculty, Universitas Kristen Artha Wacana, Indonesia
[5]College of Education, the Pennsylvania State University, USA
[7]Warner School of Education, University of Rochester, USA

DOI: 10.23917/ijolae.v6i3.23675
Received: May 15th, 2024. Revised: August 16th, 2024. Accepted: August 27th, 2024
Available Online: September 20th, 2024. Published Regularly: September, 2024

Abstract

This systematic literature review explores the utilization of Automated Writing Evaluation (AWE) as a writing scoring tool over a five-year period from 2016 to 2020, focusing on its role in the transformation and integration of learning tools for pedagogical purposes. Transformation refers to the significant changes and advancements in teaching methods, particularly in adapting to new educational technologies and approaches, while integration involves the seamless incorporation of AWE systems into these evolving instructional practices to enhance the effectiveness of writing instruction. The study aims to analyze the various types of AWE employed in academic research, track trends in AWE technology strategies, and investigate students' perceptions of AWE in both scoring and instructional contexts. Additionally, it aims to uncover the benefits and limitations associated with AWE implementation in writing instruction. Examining 19 journal articles, this review identifies fourteen types of AWE utilized by researchers and tracks advancements in machine learning within the field. The findings reveal positive student perceptions of AWE, citing its usefulness, efficiency, and linguistic accuracy in scoring and instruction. Benefits of AWE implementation include improved linguistic accuracy, enhanced writing performance, increased student engagement, and the provision of reliable and valid feedback. Moreover, AWE demonstrates effectiveness in scoring and feedback provision, with potential short- and long-term effects on student learning. However, limitations of AWE are also noted, including student distrust of feedback and a preference for human raters over AWE-generated scores. This review provides valuable insights into the multifaceted role of AWE in writing instruction, highlighting its potential benefits and areas for improvement.

Keywords: automated writing evaluation, educational technology, employs holistic, integration of learning, student engagement, transformation learning, transformative pratice

✉***Corresponding Author:***
Nurkamilah, Faculty of Teacher Training and Education, Universitas Muhammadiyah Jember
Email: nurkamilah@unmuhjember.ac.id

1. Introduction

A good learning process can improve the quality of education (Abidin et al., 2024). The effects of Information Communication and Technology on education cannot be over emphasized (Onojah et al., 2021). Technology has an important role in the world of education (Sulistyanto et al., 2022, 2023). The

integration of technology into language learning has significantly impacted classroom instruction and the assessment of learners' language proficiency. This trend has driven advancements in educational technology, leading software designers to develop and expand tools for assessing learners' receptive and productive skills. The inception of language assessment technology in the 1960s aimed to streamline the assessment process (Chapelle & Voss, 2016). This technology addresses several drawbacks of traditional paper-based testing, offering faster, more efficient, and cost-effective alternatives (Laborda, 2007). Moreover, it improves the standardization of essay assessments and the provision of timely and valid feedback (Wang et al., 2020).

Automated Writing Evaluation (AWE) is one popular manifestation of technology integration in writing assessment which uses computer systems generates scores and feedback automatically (Stevenson & Phakiti, 2019). It is widely employed in educational settings and standardized tests. High-stakes tests such as the Test of English as a Foreign Language (TOEFL) and the Graduate Management Admissions Test (GMAT) exemplify the utilization of AWE (Stevenson, 2016; Stevenson & Phakiti, 2019). These tests are proofs that the technology has provided an effective and efficient alternative to time-consuming and resource-intensive paper-based tests in educational settings.

However, controversies persist regarding the use of AWE, particularly in assessing productive skills like speaking and writing. Many researchers question the accuracy of scoring, the technology's feedback capabilities, and the implications of writing for a non-human audience (Stevenson, 2016). Despite lingering doubts, AWE has found its way into writing classrooms to aid teachers and learners in evaluating writing competence and providing writing instruction. Numerous research studies and analyses have been conducted to explore the impact of AWE on learners' writing proficiency (Liao, 2016b, 2016a; Lim & Phua, 2019; Roscoe et al., 2017; Silva, 2017; Stevenson, 2016). Stevenson, for example, emphasized that a key feature of AWE lies in its scoring engine, which utilizes techniques such as artificial intelligence, natural language processing, and semantic analysis to generate automated scores. Liao asserted that employing AWE to scaffold students' writing abilities led to a reduction in grammatical errors in L2 writing. Silva underscored AWE's pedagogical nature, noting its integration with the assessment development process and its role in scaffolding student learning. Liao further reported a significant improvement in learners' grammatical performance, indicating that AWE feedback prompted learners to interpret and internalize English grammatical rules through iterative revision processes. This integration of procedural skills ultimately facilitated learner automatization and long-term improvement. Additionally, Roscoe et al. found that learners perceived AWE as accurately scoring their writing and providing appropriate recommendations, thereby enhancing students' confidence in the scoring process.

Nevertheless, the consistency of Automated Writing Evaluation (AWE) in assessing learners' writing competence remains variable, even with teachers' intervention, particularly in feedback provision. This inconsistency stems from the design of AWE software, which often employs holistic scoring scales intended to provide scores reflecting overall text quality. Programs like My Access! and Write to Learn utilize holistic scoring scales, aiming to offer comprehensive scores. Although these programs are equipped to provide analytical scores for

specific aspects of text quality, such as language use, organization, and mechanics, they are not infallible. AWE scoring engines can be prompt-specific or generic, with prompt-specific engines limited to evaluating texts written in response to trained prompts, thereby contributing to scoring inconsistencies (Stevenson and Phakiti 2019). Additionally, inconsistencies may arise from scoring errors, where screeners fail to maintain consistent interpretation or apply scoring criteria uniformly (Godshalk, et al., 1966; Wolfe 2005).

In this study, two types of feedback were employed: high-level (HL) writing skills, encompassing aspects such as ideas and elaboration, organization, style, and self-feedback directed at the author(s)' writing process or experience, and low-level (LL) writing skills, including spelling, capitalization, punctuation, sentence structure, grammar, formatting, and word choice. The study revealed that the utilization of AWE alongside teacher feedback did not significantly affect the provision of HL feedback, whereas teacher-only feedback resulted in a greater quantity of LL feedback compared to AWE + teacher feedback. Additionally, learners exhibited a tendency to revise LL feedback provided by teacher-only feedback more than that offered by AWE + teacher feedback. Interestingly, learners taught using AWE + teacher feedback demonstrated long-term retention of their accuracy improvement, while those taught using teacher-only feedback showed short-term retention of accuracy improvement (Link, et al., 2022). However, Wilson and Czik (2016) reported slightly different findings regarding HL feedback, indicating that students introduced to AWE received more HL feedback than LL feedback in the teacher-only-feedback condition. Hence, the efficacy of AWE intervention in writing assessment and classroom instruction remains open to question regarding its beneficial impact on learners' writing competence.

Given the varying findings across research studies and diverse applications in language classrooms, AWE is understandably a compelling and pertinent topic for research due to its technological innovations, efficiency, pedagogical benefits, ongoing research, and the controversies it faces. The current review, therefore, aims to clarify the controversies exist in the current studies by exploring a five-year-AWE practice as a tool for scoring English language learners' writing and writing classroom instruction dated from 2016 through 2020. Considering previous researchers' findings, the present research questions were formulated as "1) What types of AWE have researchers used from 2016 to 2020 as tools to score students' writing and provide writing classroom instruction? 2) What are students' perceptions of AWE used for scoring their writing and writing classroom instruction? 3) What are the benefits and limitations of using AWE for scoring students' writing and facilitating writing classroom instruction?

The present review is expected to contribute valuable insights for further research on AWE's role in technology-driven writing assessment and classroom instruction. Additionally, it may serve as a resource for advancing AWE devices to aid teachers in providing valuable feedback beyond the capabilities of technology alone, thus facilitating learners' independent improvement in writing ability.

2. Method

A comprehensive and systematic literature search was conducted to identify relevant primary sources for this review. The focus was on published journal articles from 2016 to 2020, capturing recent advancements

and trends in Automated Writing Evaluation (AWE) systems. The selection of articles was not limited to a specific regional context, as technology is a global issue with widespread usage across countries worldwide. However, the search predominantly targeted studies within the higher education context. Both qualitative and quantitative research articles were included to provide a comprehensive understanding of the effectiveness, implementation, and perceptions of AWE in writing assessment and instruction. Opinion pieces, non-peer-reviewed articles, and studies outside the specified timeframe or educational context were excluded.

The search was conducted using the "ScienceDirect" database, which served as a systematic search engine across relevant journals from 2016 to 2020. The key terms employed in this study included "Automated Writing Evaluation and Students' Writing" OR "Automated Writing Evaluation and Writing Skills" OR "Automated Writing Evaluation and Higher Education Writing Skills." The search parameters were restricted to journal articles published between 2016 and 2020. Key journals in the fields of language learning technology, information writing, and education were selected for inclusion in the review.

The initial search conducted using "ScienceDirect" and the keywords "Automated Writing Evaluation and Students' Writing" yielded a total of 10,223 articles. To narrow down the scope and focus on more recent and relevant research, the search was restricted to articles published between 2016 and 2020, resulting in 3,948 articles. Given the considerable volume of findings and to further hone in on specific aspects of writing skills, the search was refined using the keywords "Automated Writing Evaluation and English Writing Skill" within the same timeframe, which produced 1,001 journal articles. This refinement aimed to target studies that specifically addressed the evaluation of English writing skills, a critical area in higher education. Subsequently, the search terms were adjusted to "Automated Writing Evaluation and Higher Education Writing Skill," resulting in 582 articles from various journals. This adjustment was made to ensure the focus remained on higher education contexts, which is the primary scope of this review. Finally, after thorough screening based on each year of publication and relevance to the research objectives, 19 articles were found to meet the criteria for inclusion in the study. The overall procedure can be seen in Figure 1.

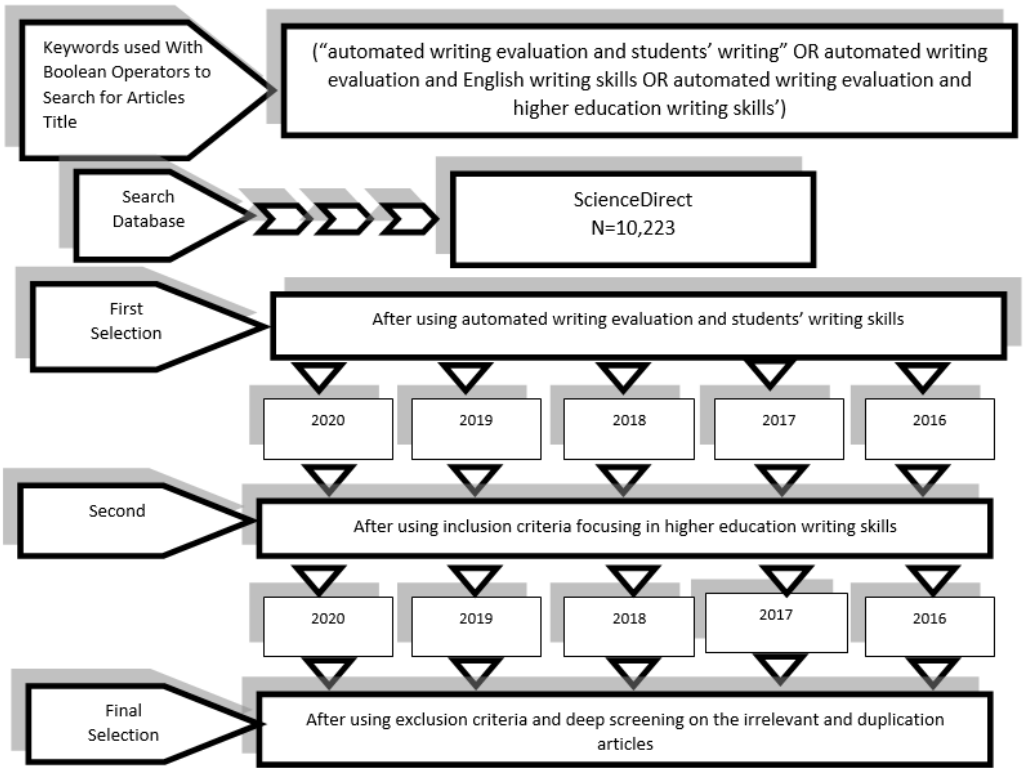

Figure 1. Criteria and Procedure of Searching Literature

Key information was extracted from each of the studies to contextualize the data regarding Automated Writing Evaluation (AWE) usage as both an assessment tool for writing and classroom instruction. This extraction process involved identifying and presenting essential details such as the authors' names, publication year, product utilized, primary data source, participants involved, and the affiliation(s) of the author(s). The results of this analysis were compiled and presented in tabular form for clarity and ease of reference. For a better view of the extracted key information, please see Table 1.

After extraction, the journal articles have undergone multiple readings and analyses to discern concepts, themes, and ideas pertaining to Automated Writing Evaluation (AWE) as both a scoring tool and as integrated into writing classroom instruction. Upon reading the articles carefully, the key concepts, themes, ideas, and other pertinent information extracted from the reviewed articles were organized into an analysis table, comprising categories such as title, author(s), journal publication, sampling, instrument/theory, findings, and notes. This information guided the results of the analysis, which in turn addressed the research questions and formed the findings of the literature review, which are presented in the Result and Discussion section.

Table 1. Key Information of the Sample Studies

	Author(s)	Year	Product	Primary Data Sources	Respondents	Author Affiliation
1.	Hui-Chuan Liao	2016a	Criterion	Essay composition	66 Taiwanese university students	National Kaohsiung University of Applied Sciences in Taiwan
2.	Svetlana Koltovskaia	2020	AWCF of Grammarly	Pre & Post writing test	2 ESL college students	Oklahoma State University, United States
3.	Zhi Li, Hui-Hsien Feng and Aysel Saricaoglu	2017	Criterion	Essay writing test and interview	63 of intermediate - high-level participants & 72 of advanced-low level participants of academic writing classes	Paragon Testing Enterprises, Inc; Iowa State University; TED University, Turkey
4.	Hui-Chuan Liao	2016b	Criterion	Essay writing and interview	63 participants	National Kaohsiung University of Applied Sciences Taiwan
5.	Stephanie Link, Mohaddeseh Mehrzad & Mohammad Rahimi	2020	Criterion	Pre, Post & Delayed Post Essay writing test	32 participants of undergraduate English majors	Oklahoma State University, USA & Shiraz University, Iran
6.	Sha Liu & Antony John Kunnan	2016	Four Human Raters & WriteToLearn	326 students' essays	163 participants of undergraduate EFL learners	China West Normal University, China & Nanyang Technological University, Singapore
7.	Leyi Qian, Yali Zhao & Yan Cheng	2019	Two expert raters & iWrite	Exposition, Argumentative & narrative essays	332 participants of non-English-major undergraduate students	Hefei University of Technology, China
8.	Rod D. Roscoe, Joshua Wilson, Adam C. Johnson, & Christopher R. Mayra	2017	W-Pal	Essays & Questionnaire	110 undergraduate students of Psychological course	Arizona State University-Polytechnic, USA & University of Delaware, Newark, USA
9.	Aysel Saricaoglu	2018	ACDET	pre- and post-cause & effect	31 students	TED University, Turkey

Author(s)	Year	Product	Primary Data Sources	Respondents	Author Affiliation
			essay tests		
10. Lili Tian &Yu Zhou	2020	Pigai	90 essays	five sophomores of online English writing course	University of Auckland, New Zealand
11. Thomas Daniel Ullmann	2019	Machine Learning	76 essays	76 students of health, business, and engineering students	Institute of Educational Technology, UK
12. Zhijie Wang	2020	automated essay evaluation (AEE: iWrite, Awrite, and Pigai)	Observation, semi-structured interview, and questionnaire	188 students from China Agricultural University	China Agricultural University, Beijing, China
13. Zhe (Victor) Zhang & Ken Hyland	2018	Pigai	Student texts, teacher feedback, AWE feedback, and student interviews.	Two Chinese students of English	The University of Hong Kong & University of East Anglia, Norwich, UK
14. Zhe Victor Zhang	2020	Pigai	Student written texts, AWE feedback, & student interviews	Three Chinese students of English major	The Chinese University of Hong Kong, Hong Kong
15. Brent Bridgeman & Chaitanya Ramineni	2017	e-rater	Students' writings, students' questionnaire, & a faculty member questionnaire	194 graduate students	Educational Testing Service, United States
16. Jim Ranalli, Stephanie Link & Evgeny Chukharev-Hudilainen	2016	Criterion	Argumentative writing tasks	82 students (36 lower-level students and 46 higher level students) of Iowa State University	Iowa State University, United States
17. Andreas Lachner, Christian Burkhart & Matthias Nückles	2017	CohViz	Students' essays	251 students	University of Freiburg, Germany & University of Tübingen, Germany

Author(s)	Year	Product	Primary Data Sources	Respondents	Author Affiliation
18. Mohammed Ali Mohsen & Abdulaziz Alshahrani	2019	MY Access!	Students' essays	6 Arab students of EFL	Najran University, Saudi Arabia
19. Gary Cheng	2017	online automated feedback (OAF)	Students' reflective journals, online questionnaire & focus group interview	138 undergraduate students	The Education University of Hong Kong

3. Result and Discussion

The analysis revealed the utilization of various Automated Writing Evaluation (AWE) software by researchers over the last five years (2016-2020) (Figure 2). Notably, Criterion was employed by multiple researchers during this period (Li, et al., 2017; Liao 2016b, 2016a; Link et al., 2022; Ranalli, et al., 2017), while Pigai was utilized by four researchers (Tian & Zhou, 2020; Wang et al., 2020; Zhang, 2020; Zhang & Hyland, 2018). Additionally, various other software programs were employed by different researchers, including ACDET (Saricaoglu, 2019), WriteToLearn (Liu and Kunnan 2016), iWrite (Qian, et al., 2020; Wang, 2022), Awrite (Wang, et al. 2020), W-Pal (Roscoe et al., 2017), AWCF of Grammarly (Koltovskaia, 2020), Machine Learning (Ullmann, 2019), e-rater (Bridgeman & Ramineni, 2017), CyWrite (Ranalli et al., 2017), CohViz (Lachner, et al., 2017), My Access! (Mohsen & Alshahrani, 2019), and Online Automated Feedback (OAF) (Cheng, 2017). In summary, researchers employed a total of 14 different AWE software programs over the five-year period from 2016 to 2020, highlighting the diverse range of options available for providing corrective feedback to students' writing. These findings corroborate previous research indicating the varied nature of AWE systems (Stevenson & Phakiti, 2019).

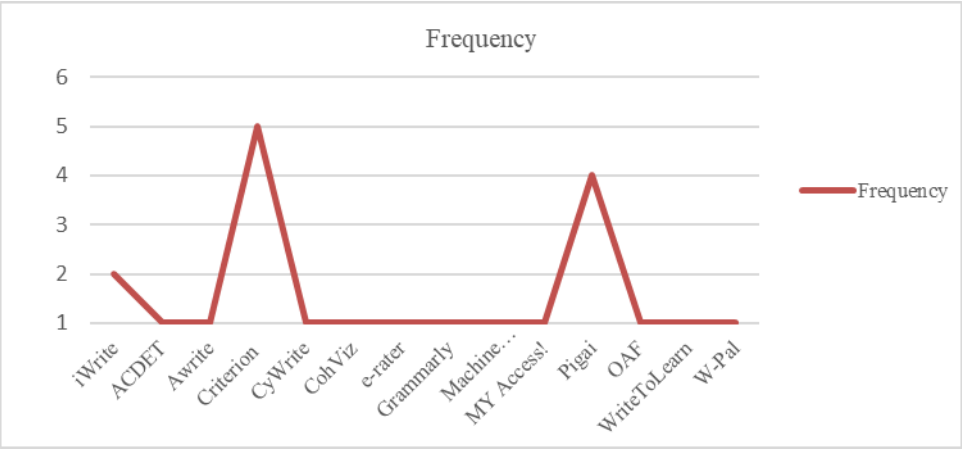

Figure 2. Trend of AWE Software Used by Researchers in 2016 – 2020

Moreover, the findings suggest that Automated Writing Evaluation (AWE) software not only provided corrective feedback on students' essays but also on their reflective journal writing. This observation is supported by research conducted by Cheng (2017) and Ranalli et al. (2017), who investigated the impact of Online Automated Feedback (OAF) and Criterion on students' reflective journal writing. Their studies revealed that OAF and Criterion significantly improved students' scores in writing reflective journals, indicating the effectiveness of these tools in providing feedback on such writing tasks. Consequently, students' understanding and willingness to revise their reflective journal writing were positively influenced by the feedback from the software, resulting in improved scores. Cheng's (Cheng 2017) and Ranalli et al.'s (Ranalli et al., 2017) studies were the only ones to utilize students' reflective journal writing in the higher education context from 2016 to 2020. In contrast, the remaining seventeen studies focused on students' essays, which included argumentative, exposition, and narrative essays. In summary, the writing instruments used in the research studies were primarily categorized into two forms: students' essays and students' reflective journals.

a. Students' Perceptions of the Integration of AWE

The analysis revealed that eight research studies (Cheng, 2017; Koltovskaia, 2020; Li et al., 2017; Liao, 2016b; Mohsen & Alshahrani, 2019; Roscoe et al., 2017; Wang, 2022; Zhang, 2020) focused on examining students' perceptions of Automated Writing Evaluation (AWE) as both a scoring tool and a component of classroom instruction. The results of these studies indicated varied perceptions among students regarding AWE's utility. Students' perceptions were categorized into two main groups in this review: positive and negative perceptions. The Summary is presented on Table 2.

Table 2. The Summary of Positive Perceptions of the Use of AWE

Aspect of Positive Perceptions	Percentage	Kind of AWE	Research Method	Subject	Researcher(s)
Effectiveness in identifying strengths and weaknesses of reflective journal writing	70	OAF	Mixed-method design	138	Cheng (2017)
Satisfaction in giving feedback	80	Criterion	Mixed-method design	135	Lie et al (2017)
Reduce Grammatical errors	-	Criterion	Experimental design	63	Liao (2016b)
Accuracy, relevance, and usefulness in providing scoring and feedback	68.2	W-Pal	Mixed-method design	110	Roscoe et al (2017)
Satisfaction in providing grammar, usage, mechanics, and syntactic complexity feedback	-	iWrite	Experimental design	188	Wang (2022)
Satisfaction in identifying collocation errors	-	AEE: iWrite, Awrite, and Pigai	Mixed-method design	3	Zhang (2020)

The findings indicate that a significant majority (70%) of the 138 students strongly agreed that Online OAF effectively identified strengths and weaknesses in their reflective journal writing. Moreover, more than half (55%) expressed their willingness to use OAF for providing feedback on their reflective writing, while only a small fraction (3%) showed reluctance towards future usage. The most favored feature of Automated Writing Evaluation (AWE) among students was the online automatic classification system, with 60% of respondents citing its usefulness. This system received praise for its ability to identify areas for improvement (40%), provide helpful suggestions and examples (23%), offer user-friendly interface (23%), and provide immediate feedback (13%). Overall, students appreciated the system's content, convenience, and speed. Additionally, a subgroup of students (12) believed that OAF could enhance their understanding of basic aspects of L2 writing. They highlighted two distinct advantages of OAF over teachers: its quick analysis and accessibility, as well as its capability for archiving and restoration. Despite some shortcomings, such as occasional inability to detect certain errors, students remained proactive in seeking solutions to their L2 writing challenges (Dikli & Bleyle, 2014; Graham, et al., 2015).

Further insights into positive perceptions were provided by Li et al. (2017), who surveyed 135 students. The results showed that a small minority (3%) of the 31 students interviewed expressed high satisfaction with Criterion feedback, while the majority (77%) reported satisfaction, and 10% remained neutral. Among these students, 71% expressed satisfaction with Criterion's grammar feedback, while 10% desired more detailed feedback. Grammar feedback was deemed the most helpful type, with 77% of students endorsing it. However, certain categories within grammar, such as run-on sentences (39%), possessives (19%), and prepositions (13%), posed challenges. Corrections were perceived as either easy (45%), difficult (10%), or variable depending on Criterion's clarity (32%). Additionally, 29% of students claimed to have addressed all feedback, while over half (58%) admitted to ignoring some feedback. The majority (71%) of students reported a positive perception of AWE, citing its effectiveness in error identification (See Figure 2). Criterion was particularly instrumental in identifying errors related to articles (58%), wrong verb forms (19%), run-on sentences (19%), subject-verb agreement (13%), fragments (13%), wrong word forms (6%), pronouns (6%), possessives (3%), and faulty comparisons (3%). These findings corroborate previous research highlighting the core components of AWE, such as its scoring engine, which utilizes techniques like artificial intelligence and natural language processing (Stevenson, 2016). Moreover, the use of AWE to scaffold students' writing ability has been shown to reduce grammatical errors in L2 writing (Liao 2016b), indicating that exposure to AWE feedback enhances learners' understanding and application of English grammatical rules over time (Liao, 2016a).

Additional positive perceptions were documented by Roscoe et al. (Roscoe et al., 2017), where students regarded Automated Writing Evaluation (AWE) as accurate, relevant, and useful in providing scoring and feedback. They expressed satisfaction with the quality of feedback provided by W-Pal, with 68.2% of 110 students expressing a willingness to use it again in the future. Similarly, Wang (Wang, 2022) reported that students exhibited a positive attitude towards Automated Essay Evaluation (AEE)

systems, including iWrite, Awrite, and Pigai, rating them as greatly or moderately helpful. Most respondents appreciated the features of these systems, particularly regarding grammar, usage, mechanics, and syntactic complexity. While they were satisfied with the content analysis provided by AEE systems, they desired more feedback on discourse elements. Zhang's (Zhang, 2020) findings also highlighted students' positive perceptions of AWE feedback, noting its helpfulness in L2 writing, particularly in identifying collocation errors rarely addressed by teachers. The feedback was perceived as immediate and accurate, aiding students in revising their work and fostering an understanding of the importance of revision in the writing process. Moreover, it encouraged students to adopt the practice of multiple drafting when completing writing assignments outside of writing subjects. These results underscore the effectiveness of various AWE software types in enhancing the accuracy of scoring students' writing (Bridgeman, et al., 2012; Shermis, 2014).

Table 3. The Summary of Negative Perceptions of the Use of AWE

Aspect of Dissatisfaction	Percentage	Kind of AWE	Research Method	Subject	Researcher(s)
Less authoritative and inaccurate of feedback	33	Grammarly	Case study	2	Koltovskaia, (2020)
A frustrating tool for a reluctant student to read and comprehend	-	Criterion	Experimental design	66	Liao (2016a)
Quality of feedback, the system's comprehension of human language, scoring methods, and the lack of explanatory reasons	37.68	OAF	Mixed-method design	138	Cheng (2017)
Difficulty comprehending AWE feedback on content and organization	-	My Access	Case study	6	Mohsen and Alshahrani (2019)

To provide a balanced view, alongside the benefits of Automated Writing Evaluation (AWE), students' perceptions reveal significant concerns. Table 3 outlines their dissatisfaction, including feedback quality, system comprehension of human language, scoring methods, and lack of explanatory reasons. Conversely, negative perceptions towards Automated Writing Evaluation (AWE) were highlighted by Koltovskaia's (Koltovskaia, 2020) findings, which revealed that one out of two students believed AWE's feedback to be less authoritative than that of teachers and possibly inaccurate. This skepticism arises from the belief that automated AWE systems lack the depth of understanding and contextual awareness that human teachers offer. Students might think that AWE tools are unable to appreciate the finer details of their writing or the specific context in which it was created, which can lead to doubts about the accuracy and reliability of the feedback. Furthermore, because AWE systems rely on algorithms and fixed criteria, there is concern that they may misinterpret intricate aspects of writing or not address the unique needs of individual students as effectively as feedback from a teacher. Consequently, AWE feedback was described as the most frustrating tool to read and comprehend by a reluctant student (Liao, 2016a). The issues highlighted include the quality of feedback, the system's

comprehension of human language, scoring methods, and the lack of explanatory reasons. Among the surveyed students (52), dissatisfaction stemmed from various aspects, including the quality of feedback (35%), the system's comprehension of human language (27%), scoring methods (19%), and lack of explanatory reasons (19%) (Cheng, 2017). While students expressed a keen interest in using AWE, particularly My Access, they felt the program, particularly its word bank functionality, did not adequately benefit them. They encountered difficulties in comprehending AWE feedback, particularly regarding content and organization, perceiving it as overly general and not tailored to their needs. In contrast, teacher feedback, especially regarding content and organization, was deemed clearer and more diagnostic than AWE feedback (Mohsen & Alshahrani, 2019). These findings echo previous research indicating AWE's limitations in providing feedback that fully meets students' needs (Stevenson & Phakiti, 2014).

b. Benefits of AWE

The final objective of this study is to delineate the advantages and limitations of Automated Writing Evaluation (AWE) as both a scoring tool for students' writing and a facilitator of writing instruction in the classroom. Numerous benefits have been identified, including enhanced linguistic accuracy, improved student performance, increased reliability and validity, effective scoring and feedback mechanisms, and both short-term and long-term impacts on student learning. These advantages have been validated by various studies, underscoring the promising potential of AWE for practical implementation in classroom settings (Table 4).

Table 4. Benefits of Automated Writing Evaluation (AWE)

Aspect of Benefit	Kind of AWE	Researcher(s)
Offering Linguistic Accuracy	Criterion; Grammarly; WriteToLearn.	Liao (2016a,b) and Ranalli et al (2017) ; Koltovskaia, (2020); Liu and Kunnan (2016)
Improving Students' Performance	W-Pal; Criterion; AEE: iWrite, Awrite, and Pigai; CohViz; My Access	Roscoe et al (2017); Link et al (2022); (Zhang 2020); Lachner et al (2017); Mohsen and Alshahrani (2019)
Providing Reliability and Validity	Machine Learning; AEE: iWrite, Awrite, and Pigai; My Access	Ullmann (2019); Wang (2022); Mohsen and Alshahrani (2019).
Offering Effectiveness in Scoring and Giving Writing Feedback	AEE: iWrite, Awrite, and Pigai; Pigai	Wang (2022); (Zhang and Hyland (2018).
Yielding Short-Term and Long-Term Effect on Students Learning		Link et al. (2022); Li et al. (2017)

1) Offering Linguistic Accuracy

The results indicated that Automated Writing Evaluation (AWE) had an early effect on reducing the number of fragments and subject-verb disagreements in new texts, while the reduction of run-on sentences and ill-formed verbs became noticeable towards the end of the study phase. Despite variations among categories in both revisions and new texts, a consistent trend of linguistic growth facilitated by AWE was observed, leading to improved linguistic accuracy (Liao, 2016b). AWE also demonstrated effectiveness in enhancing students' linguistic accuracy at a moderate level (57%) by addressing errors highlighted and suggested by the system, such as word form, articles, punctuation, spelling,

prepositions, and spacing. Errors were visually highlighted in red, while suggestions were presented in green by the Automated Corrective Writing Feedback (ACWF) system (Koltovskaia, 2020). Moreover, AWE notably improved students' grammatical accuracy in both original and revised essays of the final task, underscoring its precision in identifying grammatical errors (Liao, 2016a). WriteToLearn, a form of AWE, demonstrated greater consistency in rating papers compared to human raters, and was more accurate in identifying errors related to capitalization, spelling, punctuation, and connecting words, achieving precision rates of 100% for connecting word errors, 92.3% for capitalization errors, and 83.5% for recall. However, precision rates varied between 70% and 79% for errors related to subject-verb agreement, comma splices, and singular-plural nouns (Liu & Kunnan, 2016). Ranalli et al. (2017) also reported a high accuracy rate of 70% in linguistic feature identification using Criterion. This is achievable because AWE systems employ algorithms designed to identify and rectify grammatical errors, spelling mistakes, and contextual issues, offering uniform feedback and personalized suggestions through sophisticated machine learning and natural language processing methods. These findings corroborate previous studies conducted by Bridgeman et al., (2012), and Shermis (2014), which emphasized the high accuracy of many AWE software types in scoring students' writing (Chapelle, 1999).

2) Improving Students' Performance

Some students exhibited improved performance following engagement with Automated Writing Evaluation (AWE) systems. This improvement was evidenced by higher revision scores, which correlated with positive changes in students' perceptions of feedback (Roscoe et al., 2017). Another study revealed that AWE had a substantial impact on students' revision (92.31%) and text modifications (97.44%), indicating both short-term and long-term effects on learning and performance (Link et al., 2022). AWE also fostered increased student engagement with learning tools and enhanced their habits of drafting and revising writing (Zhang, 2020). Additionally, Lachner et al. (Lachner et al., 2017) observed that students reported improvements in the global cohesion of their texts, attributed to the formative feedback provided by machine learning. Mohsen and Alshahrani (2019) suggested that AWE systems were valuable for evaluating students' writing and facilitating improvement. AWE improves students' writing ability by promptly identifying errors, offering instructional feedback, encouraging revisions, and tracking individual progress, all of which contribute to enhanced writing practices. While technology can serve as an assistant to instructors in second language learning, it cannot fully replace the role of instructors (Salaberry, 1999).

3) Providing Reliability and Validity

Ulmann's (2019) research identified a comprehensive reflective writing model that demonstrated reliability and validity in detecting reflection in students' writing. The evaluation model's quality was theoretically deemed reliable and valid for detecting reflection. While reliability and validity tests utilized a rule-based approach across various model categories, empirical validation was achieved in only one category. Evaluation detection performance revealed that the machine learning component reliably differentiated between

reflective and descriptive sentences and effectively distinguished categories of sentences with or without elements such as experience, feelings, personal beliefs, awareness of difficulties, perspective, lessons learned, and intention. Automated Writing Evaluation (AWE) systems have been positively perceived for their reliable scoring methods (Wang, 2022). Mohsen and Alshahrani (2019) revealed that My Access's hybrid model enhanced writing accuracy through features like My Editor. These findings align with research by Liu et al., (2018), who utilized a model emphasizing both technical and personalistic aspects across three phases (analysis, description, and critique). Notably, the model's high reliability in inter-rater agreement underscores its effectiveness. Additionally, the results affirm previous findings indicating that AWE is more reliable and consistent than human raters in identifying writing errors (Hutchison, 2007; Shermis & Hamner, 2013).

4) Offering Effectiveness in Scoring and Giving Writing Feedback

AWE is often considered more effective than human scoring and feedback due to its accessibility at any time and its capacity to provide detailed content feedback. Its integration into writing classrooms has proven effective in fostering students' learning autonomy, critical thinking, and overall writing proficiency. Additionally, AWE serves as an efficient tool for sharing learning resources (Wang, 2022). Students have shown preference for eight key characteristics of AEE, including its accessibility at any time, specificity, personalization, and comprehensibility compared to human rating systems (Wang, 2022). Zhang and Hyland (2018) also noted that 'AWE feedback offers discernible advantages over teacher feedback in terms of timeliness, convenience, multiple drafting opportunities, and even potential learner autonomy', corroborating previous findings by Chen and Cheng (2008) and Dikli (2006). The accessibility of feedback at any time and the ability to revise drafts multiple times align with earlier research (Cotos, 2015; Stevenson & Phakiti, 2014; Warschauer & Ware, 2006).

5) Yielding Short-Term and Long-Term Effect on Students Learning

AWE has demonstrated both short-term and long-term impacts on learning and student performance, as indicated by Link et al. (2022). Specifically, students who received instruction using AWE alongside teacher feedback exhibited long-term retention of accuracy improvement by fostering skill development, greater writing proficiency, and continuous improvement through progress tracking, whereas those taught solely with teacher feedback showed short-term retention (Link et al., 2022) by providing immediate feedback and encouraging prompt revisions. Additionally, AWE has shown a positive long-term effect on reducing instances of run-on sentences across all proficiency levels, as well as improvements in subject-verb agreement, with varying degrees of change observed between intermediate-high and advanced-low levels (Li et al., 2017)., whereas long-term, it contributes to skill growth, ongoing proficiency, and sustained development by monitoring progress over time. These long-term effects are attributed to sustained advancements in accuracy, allowing students to internalize knowledge gained from AWE corrective feedback and retain it in their long-term memory for future use (Bitchener, 2012).

The use of Automated Writing Evaluation (AWE) systems presents both advantages and limitations. Despite offering timely feedback, students often express skepticism due to the perceived disparity between AWE and human feedback. Studies show AWE's weaknesses in comparison to human raters, raising questions about its effectiveness in facilitating meaningful revisions and its impact on students' writing development.

c. Limitations of AWE

AWE also presents certain limitations, such as students' skepticism towards the feedback provided (Table 5). Skepticism often arises from the perceived gap between teacher feedback, which is seen as more influential due to its expert knowledge, personalized advice, and interactive nature, and AWE feedback, which lacks these strengths and is criticized for its inability to provide nuanced, authoritative guidance and to prompt thorough revisions (Koltovskaia, 2020). Particularly, lower-level students have expressed frustration with receiving scores and feedback devoid of human interaction (Liao, 2016a). Additionally, students have shown a tendency to prioritize the quantity of feedback over the quality of revision suggestions. While the use of causal verbs decreased from the initial to the final draft, there was no corresponding increase in the use of causal nouns. Furthermore, the findings of Saricaoglu's (Saricaoglu, 2019) research study did not indicate any long-term effects.

AWE also exhibits several weaknesses, notably in comparison to human raters. In a study by Liu and Kunnan (Liu & Kunnan, 2016), human raters outperformed AWE in rating students' writing, assessing 326 essays compared to AWE's 319 essays. AWE missed identifying 7 essays, and its accuracy was lower than that of human raters, detecting only 15 errors compared to the 22 identified by human raters. Additionally, studies such as iWrite (Qian et al., 2020) reported no correlations between AWE and human scores, indicating poor automated scoring quality. AWE consistently yielded lower scores than human raters, and there was no correlation between the presented scores and feedback quality. These findings were echoed by Roscoe et al. (Roscoe et al., 2017), highlighting that the scoring accuracy did not align with the initial expectations of AWE quality. Furthermore, Prompt-specific AWE engines were found to be limited in their applicability, as they could only evaluate texts written in response to pre-trained prompts (Stevenson & Phakiti, 2019).

Table 5. Limitation of Automated Writing Evaluation (AWE)

Aspect of Limitation	Kind of AWE	Researcher(s)
Lacks authority to prompt thorough revisions	Grammarly; Criterion; ACDET	Koltovskaia (2020); Liao (2016a); Saricaoglu 2019
AWE fell short of human raters in rating students' writing	Four Human Raters & WriteToLearn	Liu and Kunnan (2016)
Poor automated scoring quality	Two expert raters & iWrite; W-Pal	Qian et al. (2020); Roscoe et al. 2017
Provide low-level feedback	Pigai	Tian and Zhou (2020)

The research findings revealed that automated feedback primarily provided low-level feedback, with lexical meaning receiving a 66.2% rating and a mere 5.7% uptake rate, while grammar and mechanics garnered a 68.7% rating with a 46.4%

uptake rate. When compared to peer and teacher feedback, automated feedback exhibited the lowest uptake rate, followed by peer feedback, with teacher feedback being the most authoritative among the three (Tian & Zhou, 2020). These findings corroborate previous research highlighting the detrimental effects of AWE when acting as a non-human audience (Stevenson, 2016).

4. Conclusion

This systematic literature review aimed to examine various AWE software used by researchers within five years in teaching writing. It also unveils student perceptions, as well as advantages and disadvantages of using the tools in writing. 14 kinds of AWE software were used in the reviewed articles. The findings showed that AWE tools are generally perceived positively by students, particularly for scoring, enhancing writing skills, and promoting engagement, which encouraged more thorough revision habits. Given the findings, some benefits are obtained, including enhancing language precision and writing performance, providing effective scoring and writing feedback, developing short-term and long-term impact on student learning, and providing reliable and valid feedback. However, some drawbacks were noted, such as distrust in AWE feedback and a preference for human evaluators.

The review's limited scope suggests that future research should encompass broader contexts and longer durations. Additionally, future researchers should explore the integration of AWE with other educational technologies, assess the long-term impact on students' writing skills beyond higher education, and investigate the effectiveness of AWE in diverse educational settings and with varied learner demographics. Comparative studies between AWE and human feedback, as well as the development of more advanced and context-sensitive AWE systems, could also provide valuable insights. Furthermore, qualitative studies on student and instructor attitudes towards AWE could enrich understanding of its practical applications and limitations.

5. References

Abidin, N. L. F., Dwiningsih, K., Jehwae, P., & Sari, C. K. (2024). Leveraging technology to improve learning independence in chemistry: A study on Moodle integration. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 6(3), 365-386.

Bitchener, John. 2012. "A Reflection on 'the Language Learning Potential' of Written CF." *Journal of Second Language Writing* 21(4):348–63. doi: 10.1016/j.jslw.2012.09.006.

Bridgeman, Brent, and Chaitanya Ramineni. 2017. "Design and Evaluation of Automated Writing Evaluation Models: Relationships with Writing in Naturalistic Settings." *Assessing Writing* 34:62–71. doi: 10.1016/j.asw.2017.10.001.

Bridgeman, Brent, Catherine Trapani, and Yigal Attali. 2012. "Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country." *Applied Measurement in Education* 25(1):27–40. doi: 10.1080/08957347.2012.635502.

Chapelle, C. .. 1999. "Research Questions for a CALL Research Agenda: A Reply to Rafael Salaberry." *Language Learning & Technology* 37(1):108–13.

Chapelle, Carol A., and Erik Voss. 2016. "20 Years of Technology and Language Assessment in Language Learning & Technology." *Language Learning and Technology* 20(2):116–28.

Chen, Chi Fen Emily, and Wei Yuan Eugene Cheng. 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.

Cheng, Gary. 2017. "The Impact of Online Automated Feedback on Students' Reflective Journal Writing in an EFL Course." *The Internet and Higher Education* 34:18–27. doi: 10.1016/j.iheduc.2017.04.002.

Cotos, E. 2015. "AWE for Writing Pedagogy: From Healthy Tension to Tangible Prospects. Special Issue on Assessment for Writing and Pedagogy." *Writing & Pedagogy* 7(2–3):197–231.

Dikli, Semire. 2006. "An Overview of Automated Scoring of Essays." *The Journal of Technology, Learning and Assessment* 5(1 SE-Articles).

Dikli, Semire, and Susan Bleyle. 2014. "Automated Essay Scoring Feedback for Second Language Writers: How Does It Compare to Instructor Feedback?" *Assessing Writing* 22:1–17. doi: 10.1016/j.asw.2014.03.006.

Godshalk, F. .., F. Swineford, and W. .. Coffman. 1966. *The Measurement of Writing Ability*. New York: College Entrance Examination Board.

Graham, Steve, Michael Hebert, and Karen R. Harris. 2015. "Formative Assessment and Writing: A Meta-Analysis." *The Elementary School Journal* 115(4):523–47.

Hutchison, Dougal. 2007. "An Evaluation of Computerised Essay Marking for National Curriculum Assessment in the UK for 11-year-olds." *British Journal of Educational Technology* 38(6):977–89. doi: 10.1111/j.1467-8535.2006.00686.x.

Koltovskaia, Svetlana. 2020. "Student Engagement with Automated Written Corrective Feedback (AWCF) Provided by Grammarly: A Multiple Case Study." *Assessing Writing* 44:100450. doi: 10.1016/j.asw.2020.100450.

Laborda, Jesus Garcia. 2007. "Introducing Standardized EFL/ESL Exams." *Language Learning and Technology* 11(2):3–9.

Lachner, Andreas, Christian Burkhart, and Matthias Nückles. 2017. "Formative Computer-Based Feedback in the University Classroom: Specific Concept Maps Scaffold Students' Writing." *Computers in Human Behavior* 72:459–69. doi: 10.1016/j.chb.2017.03.008.

Li, Zhi, Hui-Hsien Feng, and Aysel Saricaoglu. 2017. "The Short-Term and Long-Term Effects of AWE Feedback on ESL Students' Development of Grammatical Accuracy." *CALICO Journal* 34(3):355–75.

Liao, Hui-Chuan. 2016a. "Enhancing the Grammatical Accuracy of EFL Writing by Using an AWE-Assisted Process Approach." *System* 62:77–92. doi: 10.1016/j.system.2016.02.007.

Liao, Hui-Chuan. 2016b. "Using Automated Writing Evaluation to Reduce Grammar Errors in Writing." *ELT Journal* 70(3):308–19. doi: 10.1093/elt/ccv058.

Lim, Fei Victor, and Jean Phua. 2019. "Teaching Writing with Language Feedback Technology." *Computers and Composition* 54:102518. doi: 10.1016/j.compcom.2019.102518.

Link, Stephanie, Mohaddeseh Mehrzad, and Mohammad Rahimi. 2022. "Impact of Automated Writing Evaluation on Teacher Feedback, Student Revision, and Writing Improvement." *Computer Assisted Language Learning* 35(4):605–34. doi: 10.1080/09588221.2020.1743323.

Liu, Q., S. Zhang, Q. Wang, and W. Chen. 2018. "Mining Online Discussion Data for Understanding Teachers Reflective Thinking." *IEEE Transactions on Learning Technologies* 11(2):243–54. doi: 10.1109/TLT.2017.2708115.

Liu, Sha, and Antony John Kunnan. 2016. "Investigating the Application of Automated Writing Evaluation to Chinese Undergraduate English Majors: A Case Study of WriteToLearn</Em>." *CALICO Journal* 33(1):71–91.

Mohsen, Mohammed Ali, and Abdulaziz Alshahrani. 2019. "The Effectiveness of Using a Hybrid Mode of Automated

Writing Evaluation System on Efl Students' Writing." *Teaching English with Technology* 19(1):118–31.

Onojah, A. A., Onojah, A. O., Olumorin, C. O., & Omosewo, E. O. (2021). Secondary School Teachers' Accessibility to Internet Facilities for Advanced Instruction in Nigeria. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 3(2), 86-95.

Qian, Leyi, Yali Zhao, and Yan Cheng. 2020. "Evaluating China's Automated Essay Scoring System IWrite." *Journal of Educational Computing Research* 58(4):771–90. doi: 10.1177/0735633119881472.

Ranalli, Jim, Stephanie Link, and Evgeny Chukharev-Hudilainen. 2017. "Automated Writing Evaluation for Formative Assessment of Second Language Writing: Investigating the Accuracy and Usefulness of Feedback as Part of Argument-Based Validation." *Educational Psychology* 37(1):8–25. doi: 10.1080/01443410.2015.1136407.

Roscoe, Rod D., Joshua Wilson, Adam C. Johnson, and Christopher R. Mayra. 2017. "Presentation, Expectations, and Experience: Sources of Student Perceptions of Automated Writing Evaluation." *Computers in Human Behavior* 70:207–21. doi: 10.1016/j.chb.2016.12.076.

Salaberry, Rafael. 1999. "CALL in the Year 2000: Still Developing the Research Agenda." *Language Learning & Technology* 3(1):104–7.

Saricaoglu, Aysel. 2019. "The Impact of Automated Feedback on L2 Learners' Written Causal Explanations." *ReCALL* 31(2):189–203. doi: 10.1017/S095834401800006X.

Shermis, Mark D. 2014. "State-of-the-Art Automated Essay Scoring: Competition, Results, and Future Directions from a United States Demonstration." *Assessing Writing* 20:53–76. doi: 10.1016/j.asw.2013.04.001.

Shermis, Mark D., and Ben Hamner. 2013. "Contrasting State-of-the-Art Automated Scoring of Essays." Pp. 313–46 in *Handbook of automated essay evaluation*. Routledge.

Silva, Pedro. 2017. "Scaffolding Assignments: Analysis of AssignMentor as a Tool to Support First Year Students' Academic Writing Skills." *E-Learning and Digital Media* 14(1–2):86–97. doi: 10.1177/2042753017695652.

Sulistyanto, H., Anif, S., Sutama, S., Narimo, S., Sutopo, A., Haq, M. I., & Nasir, G. A. (2022). Education application Testing Perspective to Empower Students' Higher Order Thinking Skills Related to the Concept of Adaptive Learning Media. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 4(3), 257-271.

Sulistyanto, H., Sumardjoko, B., Haq, M. I., Zakaria, G. A. N., Narimo, S., Astuti, D., Adhantoro, M. S., Setyabudi, D. P., Sidiq, Y., & Ishartono, N. (2023). Impact of Adaptive Educational Game Applications on Improving Student Learning: Efforts to Introduce Nusantara Culture in Indonesia. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 5(3), 249–261. https://doi.org/10.23917/ijolae.v5i3.23004

Stevenson, Marie. 2016. "A Critical Interpretative Synthesis: The Integration of Automated Writing Evaluation into Classroom Writing Instruction." *Computers and Composition* 42:1–16. doi: https://doi.org/10.1016/j.compcom.2016.05.001.

Stevenson, Marie, and Aek Phakiti. 2014. "The Effects of Computer-Generated Feedback on the Quality of Writing." *Assessing Writing* 19:51–65. doi: https://doi.org/10.1016/j.asw.2013.11.007.

Stevenson, Marie, and Aek Phakiti. 2019. "Automated Feedback and Second

Language Writing." Pp. 125–42 in *Feedback in second language writing: Contexts and issues*.

Tian, Lili, and Yu Zhou. 2020. "Learner Engagement with Automated Feedback, Peer Feedback and Teacher Feedback in an Online EFL Writing Context." *System* 91:102247. doi: https://doi.org/10.1016/j.system.2020.102247.

Ullmann, Thomas Daniel. 2019. "Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches." *International Journal of Artificial Intelligence in Education* 29(2):217–57. doi: 10.1007/s40593-019-00174-2.

Wang, Elaine Lin, Lindsay Clare Matsumura, Richard Correnti, Diane Litman, Haoran Zhang, Emily Howe, Ahmed Magooda, and Rafael Quintana. 2020. "ERevis(Ing): Students' Revision of Text Evidence Use in an Automated Writing Evaluation System." *Assessing Writing* 44:100449. doi: https://doi.org/10.1016/j.asw.2020.100449.

Wang, Zhijie. 2022. "Computer-Assisted EFL Writing and Evaluations Based on Artificial Intelligence: A Case from a College Reading and Writing Course." *Library Hi Tech* 40(1):80–97. doi: 10.1108/LHT-05-2020-0113.

Warschauer, Mark, and Paige Ware. 2006. "Automated Writing Evaluation: Defining the Classroom Research Agenda." *Language Teaching Research* 10(2):157–80. doi: 10.1191/1362168806lr190oa.

Wilson, Joshua, and Amanda Czik. 2016. "Automated Essay Evaluation Software in English Language Arts Classrooms: Effects on Teacher Feedback, Student Motivation, and Writing Quality." *Computers & Education* 100:94–109. doi: https://doi.org/10.1016/j.compedu.2016.05.004.

Wolfe, E. M. 2005. "Uncovering Rater's Cognitive Processing and Focus Using Think-Aloud Protocols." *Journal of Writing Assessment* 2(1):37–56.

Zhang, Zhe (Victor). 2020. "Engaging with Automated Writing Evaluation (AWE) Feedback on L2 Writing: Student Perceptions and Revisions." *Assessing Writing* 43:100439. doi: https://doi.org/10.1016/j.asw.2019.100439.

Zhang, Zhe (Victor), and Ken Hyland. 2018. "Student Engagement with Teacher and Automated Feedback on L2 Writing." *Assessing Writing* 36:90–102. doi: https://doi.org/10.1016/j.asw.2018.02.004.