Received: 30 April 2021 | Revised: 2 November 2021 | Accepted: 21 November 2021

DOI: 10.1111/jcal.12635

REVIEW ARTICLE

Journal of Computer Assisted Learning WILEY

Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020

Andreia Nunes | Carolina Cordeiro | Teresa Limpo | São Luís Castro

Faculty of Psychology and Education Sciences, University of Porto, Porto, Portugal

Correspondence
São Luís Castro, Faculdade de Psicologia e de Ciências da Educação, Universidade do Porto, Rua Alfredo Allen, 4200-392 Porto, Portugal.
Email: slcastro@fpce.up.pt

Funding information
This work was supported by the Portuguese Foundation for Science and Technology (grants UIDB/00050/2020 and SFRH/BD/139195/2018).

Abstract

Background: Automated Writing Evaluation (AWE) systems to aid writing learning and instruction in primary and secondary education are growing increasingly popular. However, their effectiveness is hardly known. We conducted a systematic review focusing on the effects of these systems providing writing feedback to students in school settings.

Objectives: Our goal was to identify and characterize AWE systems tested in the last 20 years for Grades 1–12 and examine their impact on text quality and other writing-related outcomes.

Methods: The review followed PRISMA guidelines. We identified eight studies reporting the effects on writing of six AWE systems on 1659 students 11–17 years of age.

Results and conclusions: Our review supported the usefulness of AWE systems for writing learning and instruction. Except for one, all studies showed a positive effect of automated feedback in at least one writing-related measure. The integration of AWE systems into more extensive instructional programs, the amount of writing practice provided to students, the type of the control groups, and the role of teachers are factors influencing their impact on students' writing outcomes.

Relevance: Our review generally supported the value of AWE systems in the teaching/learning process of writing. A closer look into the conditions in which AWE systems are put to practice suggested that they are particularly effective when embedded into comprehensive instructional programs providing ample writing opportunities. Findings from this review expand knowledge on AWE systems as valuable tools to enhance writing in school settings.

KEYWORDS
automated evaluation systems, feedback, systematic review, teachers, writing

1 | INTRODUCTION

Writing is a fundamental skill to live in society, as recognized by UNESCO (2011). Through writing, individuals can express their feelings, heal psychological wounds, acquire new knowledge, record information, entertain themselves and others, and create imaginary worlds (Graham, 2018, 2019). Additionally, writing is critical for students to succeed in school, for workers to succeed in their jobs, and, ultimately, for people to succeed in their everyday lives (Graham, 2019). The development of good writing skills is, therefore, a central aim of education. Learning to write well requires the mastery of several basic and complex cognitive processes (Kellogg, 2008). Early on, students are expected to acquire the basic processes of transcription (i.e., handwriting/typing and spelling). Progressively, they develop more complex processes, such as planning (i.e., idea generation and organization), translating (i.e., the transformation of ideas into language) and revising (i.e., modification and reorganization of writing) (Graham & Perin, 2007; Limpo & Alves, 2013a). Students struggle to acquire and develop these processes throughout schooling, and the use of appropriate instructional practices in the classroom can be a powerful tool to diminish those difficulties (Graham, 2006; Harris & Graham, 2016).

An important instructional practice in the classroom is formative writing assessment. Formative assessment refers to "how judgements about the quality of student responses (performances, pieces, or works) can be used to shape and improve the student's competence by short-circuiting the randomness and inefficiency of trial-and-error learning" (Sadler, 1989, p. 120). The quality of student responses relies on writing quality evaluations, which are based either on professional or personal opinions about what is good writing or on scoring rubrics gauging specific attributes of good writing (Graham et al., 2015; Rowntree, 1987; Sadler, 1989).

A key element of formative writing assessment is feedback, which provides information that students can use to enhance their learning process, as it closes the gap between what they write and what is expected of them to write (Biber et al., 2011; Graham et al., 2015). Feedback is defined as the information provided by an agent regarding aspects of performance or understanding (Hattie & Timperley, 2016). The feedback provided varies in terms of (a) the agent, that is, the person who delivers the feedback (teachers, peers), technology, or self-assessment, (b) the mode of feedback delivery, that is, by pen-and-paper, electronic, or automated, and (c) the types of feedback, such as commentaries, responses, or corrections (Nurmukhamedov, 2009; Sadler, 1989). Meta-analytic findings showed that formative writing assessments that provided daily feedback to students enhanced writing quality in Grades 1–8, regardless of being delivered by persons or technology (Graham et al., 2015).

2 | TECHNOLOGY AND WRITING

Within educational settings, technology refers to electronic tools supporting the learning process (e.g., computers, interactive whiteboards, multimedia, and the internet) (Cheung & Slavin, 2012). Technological support for writing may be particularly beneficial to struggling writers as it can help their understanding of spelling and text organization and thus scaffold writing ability (Peterson-Karlan & Parette, 2007). However, a meta-analysis including studies from 1992 to 2002 showed that word-processing technology had a low to moderate positive effect on students' writing ability (Goldberg et al., 2003). Since then, the advantages of including technology into writing instruction have been observed in several studies (Cheung & Slavin, 2012; Goldberg et al., 2003; Little et al., 2018; Meyer et al., 2010; Stevenson & Phakiti, 2014; Wijekumar et al., 2017; Wijekumar et al., 2018). This kind of technological support promotes writing by scaffolding planning, translating, and revising (Peterson-Karlan & Parette, 2007). For example, the effectiveness of planning instruction can be enhanced when combined with computer-based tools (Cheung & Slavin, 2012; Little et al., 2018; Meyer et al., 2010; Stevenson & Phakiti, 2014). A recent meta-analysis confirmed the positive effect of technology-based writing instruction on writing outcomes in K-12 students (Little et al., 2018). Using technology may also promote positive attitudes towards writing and enhance motivation to write (Camacho et al., 2020; Ekholm et al., 2017). Due to recent technological advancements, technology is now being used to provide feedback in formative writing assessments. This form of technology-based feedback was made to help teachers assess their students' writing quickly and cost-effectively (Cotos, 2014). Widely used electronic systems to provide feedback rely on Automated Writing Evaluation (AWE).

AWE is defined as the capability of a computer technology to evaluate and score written text (Shermis et al., 2013). AWE systems were originally developed with a twofold goal: to overcome the cost in time and effort of having humans evaluate large-scale testing products and to create impartial scoring systems free of human fallibility (Stevenson & Phakiti, 2019; Wang et al., 2020). Technological evolution in the last decades allowed this type of technology to go beyond summative scoring and provide students with detailed feedback in an interactive format (Hockly, 2019; Shermis & Burstein, 2003). To automatically evaluate the content, structure, and/or quality of writing, AWE systems use a scoring engine labelled automated essay scoring (AES) (Shermis & Burstein, 2003). The most well-known types of AES systems are Project Essay Grade (PEG), Intelligent Essay Assessor™ (IEA™), Electronic Essay Rater (e-rater®), IntelliMetric®, AutoScore, Bokette, CRASE™, Lexile® Writing Analyser, and LightSIDE (for more detailed information about these systems see Cotos, 2014; Dikli, 2006; Shermis, 2020). AES systems comprise a set of different computerized methods to assign scores to the written texts. Except for PEG and LightSIDE, which use statistical techniques, and IEA™, which is based on latent semantic analysis (LSA), the AES systems mentioned above rely on natural language processing (NLP) (Shermis, 2020). Depending on the method, AES systems can focus on a variety of writing-related features (Shermis, 2020). Whereas LSA methods focus on content rather than mechanical aspects (e.g., spelling and grammar), NLP methods provide feedback on a large range of aspects, such as grammar, usage, mechanics, style, discourse structure, vocabulary usage, sentence variety, source use and discourse coherence quality (for a review on these methods see Deane, 2013; Landauer et al., 2003; Landauer & Psotka, 2000; Page, 2003). In addition to the scoring engine that evaluates the text and provides a quantitative holistic score, AWE systems also include a feedback engine that provides qualitative feedback on how to improve writing and raise the quantitative score (Allen et al., 2016). The feedback from AWE systems is generally displayed in an engaging graphic interface, using several writing assistance tools, such as graphic organizers, text and chart-based feedback on writing, and online dictionaries (Franzke et al., 2005; Ware, 2014). Recent tools, such as Writing Pal and Writing

Mentor, resort to animated agents (Allen et al., 2016; Cahill & Evanini, 2020).

The use of AWE systems in education gained popularity in school and university settings (Dikli, 2006). AWE systems designed for the classroom provide students with many opportunities to plan, write, and revise with the help of the feedback generated by the system (Cotos, 2014; Grimes & Warschauer, 2010). Several AWE systems have been developed throughout the years. Some are web-based, such as Criterion, MyAccess!, and WriteToLearn (Allen et al., 2016). Others are add-ons to existing platforms, such as Grammarly®, which can assist writing in e-mails and social media, or the Writing Mentor application, which is a writing revision assistant for Google Docs (Burstein et al., 2020; Cahill & Evanini, 2020). The last few years have also seen the development of intelligent tutoring systems (ITS), which provide the most sophisticated form of computer-based writing instruction, which may or may not include automated feedback (Allen et al., 2016). For example, Writing Pal is an ITS that combines individualized formative feedback with writing strategy instruction and game-based practice (Allen et al., 2016; Jacovina & McNamara, 2016; Roscoe & McNamara, 2013).

The evaluation of writing by computer-based systems has several advantages. These systems can analyse students' writing as a human rater would, without being affected by factors that typically influence humans such as fatigue or distraction (Weigle, 2013). Moreover, the feedback can be given as soon as the written product is finished, allowing students to receive instant feedback, every time they want, and not when the teacher is available (Cotos, 2014). As the automated feedback is anonymous and does not require face-to-face interactions with feedback agents, AWE systems may also reduce students' evaluation-related anxiety and allow them to rely on trial and error to improve their writing (Weigle, 2013). Ultimately, these features of automated feedback may foster students' motivation to write and revise, and increase writing practice (Grimes & Warschauer, 2010). These systems can also help teachers. By using AWE to target writing mechanics, teachers can focus on higher-level features of writing. They can adapt the system to provide specific feedback according to students' age and proficiency level by using grade-appropriate prompts and scoring models (Cotos, 2014; Dikli, 2006; Jacovina & McNamara, 2016). Online AWE systems can also help second language writing (L2) (Warschauer & Ware, 2006). Recognizing the differences between L1 and L2 writers—which vary in their control over syntax, morphology, and vocabulary, as well as their experience in writing (Weigle, 2013)—AWE systems have been designed in languages other than English (e.g., IntelliMetric® provides versions in Chinese, Hebrew, and Bahai Malaysian) (Shermis, 2020). This is an welcome progress given that, in the early 2000s, AWE systems for foreign languages were based on English grading engines (Shermis, 2020).

AWE systems also raise concerns. To effectively score written material, AWE requires libraries with a large number of essay samples (Cotos, 2014)—most systems between 300 and 500 (Dikli, 2006; Foltz et al., 2013)—and because it is trained based on human raters, AWE scoring can replicate the biases present in the original human ratings (Deane, 2013). Regarding scoring proper, some systems use prompt-specific scoring engines, which, despite providing more accurate feedback, are less flexible and do not allow teachers to assign custom prompts (Jacovina & McNamara, 2016). Generic engines are more flexible and allow teachers to create prompts, but enhanced flexibility comes at the cost of accuracy (Schneider & Boyer, 2020; Shermis & Hamner, 2013). Besides prompts, several AWE systems also allow teachers to customize the level and type of feedback or impose time limits (Cotos, 2014; Dikli, 2016); however, what is measured by a particular AWE system is typically not manipulated (Cotos, 2014). A frequent criticism is that these systems tend to overemphasize the writing product and neglect the underlying processes enacted by writers to produce the text (Deane, 2013), but this is also true to writing without technology (Shermis et al., 2013). Another limitation of automated scoring is that it falls short of the one of a trained human rater because the scoring engine cannot evaluate contextual aspects of writing, such as information that depends on shared background knowledge (e.g., reference to well-known literature or famous people) or assumptions between readers and writers (e.g., humour and irony) (Weigle, 2013). Indeed, students can manipulate the AWE by using their knowledge of technology (Allen et al., 2016). For example, they were able to trick the system into obtaining higher scores by repeating the same paragraph throughout the text (Powers et al., 2002). Finally, an important concern is that AWE systems might inhibit the social nature of writing by replacing teachers or other feedback agents (Conference on College Composition and Communication, 2014). However, as noted by Attali (2013), AWE systems should be taken as a complement to other types of feedback (e.g., teacher, peers) and not as a replacement of human scoring.

Previous attempts to gather evidence on the effectiveness of AWE systems in improving students' writing quality have not been particularly successful. Morphy and Graham (2012) conducted a meta-analysis of studies published between 1983 and 2011 examining the effects of word processors in struggling writers/readers in Grades 1–12. They found only one study showing beneficial effects of automated feedback (Franzke et al., 2005). This study was also located by Little et al. (2018) in a meta-analysis on technology-based writing instruction (not focused on automated feedback only). To the best of our knowledge, only one critical review specifically examined the effects of AWE systems on text quality from Grade 1 to university (Stevenson & Phakiti, 2014). Findings on the benefits AWE systems were mixed, probably due to the heterogeneity of the studies examined, many without peer review. Stevenson and Phakiti (2014) noted methodological limitations, such as lack of control groups or no statistical testing in the studies included in their review.

3 | PRESENT STUDY

AWE systems are promising means to deliver effective and valid feedback on students' writing. However, there are no sound evidence syntheses of empirical research testing the effects of automated feedback. Most past meta-analyses addressed the general use of technology in writing instruction in Grades 1–12, and they only located a single study testing the effectiveness of AWE systems (Goldberg et al., 2003; Little et al., 2018; Morphy & Graham, 2012). The single review that addressed the effects of computer-generated feedback (Stevenson & Phakiti, 2014) presents several limitations, namely, broad inclusion criteria resulting in the inclusion of non-published studies, without control groups, no reference to having followed PRISMA guidelines (Moher et al., 2009), and no assessment of the risk of bias in the included studies. These limitations constrain conclusions emerging from that work.

Given the lack of reliable evidence syntheses in the field and the increasing use of automated feedback to support writing instruction, a review of the effects of AWE systems is timely. In the present work, we conducted a systematic review exclusively focused on research testing the effects of technology capable of providing writing feedback to students in Grades 1–12. Our goal was to examine the effectiveness of AWE systems to promote Grade 1–12 students' writing quality as well as other writing-related outcomes targeted by the selected studies. We developed the following research question using the PICO (Population/participant, Intervention/indicator, Comparator/control, Outcome; Miller & Forrest, 2001; Schardt et al., 2007) framework: In students from Grades 1–12 (P), is the use of AWE systems (I), compared to other types of feedback or no feedback (C), effective to improve writing quality or other writing-related outcomes (O)? To better understand the characteristics of these systems, we examined their features, including how they articulate with human feedback agents. Generated findings will expand the current understanding of the use of AWE systems in educational settings. Given the pandemic situation we are currently living in, which urged relying on technological tools for writing instruction, knowing whether they work or not is especially meaningful.

4 | METHOD

4.1 | Search strategies

A comprehensive search was conducted between January and February 2020 using the following databases: EBSCOhost (Academic Search Complete, Education Source, ERIC, APA PsycARTICLES, Psychology and Behavioural Sciences Collection, Psychology and Behavioural Sciences Collection and APA PsycINFO), PubMed, and Web of Science. The search was also conducted in the web search engine Google Scholar. The search focused on the last 20 years (2000–2020). Keywords used in the search were as follows: writing technology OR word processing software OR computer-adapted writing technology OR effectiveness of writing technology OR text composition AND feedback OR technology feedback OR automatic feedback OR automated writing evaluation AND intervention OR training OR program AND education OR children OR school OR classroom.

4.2 | Eligibility criteria

Eligibility criteria were defined with the goal of finding the best available evidence while overcoming the limitations of previous reviews in the field (e.g., large variability of selected studies with poor methodological rigour). To achieve good scientific validity, empirical studies should follow three gold standards: manipulation of an independent variable, comparison between an experimental condition with at least a control condition, and randomization of subjects to groups (Thompson & Panacek, 2006). Because this last standard is hard to follow in studies conducted in school settings, we adopted only the first two as inclusion criteria: (1) implementation of an intervention with a technological component providing feedback on writing; (2) inclusion of an active or a passive control group against which the intervention was compared. With this set of inclusion criteria, we aimed to make sure that the selected studies had internal validity, which is critical to achieving trustworthy findings (Campbell & Stanley, 1963). A key aspect of internal validity when testing the effectiveness of an intervention is the inclusion of control groups. In intervention studies without control groups, it is virtually impossible to determine whether the changes arose from the treatment or from confounding variables (Slack & Draugalis, 2001). In line with our research question, four additionally inclusion criteria were: (3) empirical studies with children aged 6–18 years (Grades 1–12); (4) inclusion of quantifiable measures to assess writing quality as an outcome; (5) publication date between January 2000 and January 2020; and (6) publication written in English. Exclusion criteria were: (1) reviews, meta-analyses, editorials, opinion papers, dissertations, and book chapters; (2) use of technology for non-educational purposes and/or outside of the writing domain and/or without feedback on writing; (3) lack of writing quality as an outcome measure; and (4) papers not published in English.

4.3 | Selection of the studies

The selection of studies to include in this review followed PRISMA guidelines (Moher et al., 2009). We used Rayyan, a free website that assists systematic review authors to expedite the initial screening of titles and abstracts (Ouzzani et al., 2016). The database search found 2845 articles, and five additional ones were identified through manual search. After removing duplicates, we obtained 1685 articles. The first two authors independently read their titles and abstracts in view of the inclusion and exclusion criteria and agreed on all but two studies, which were then discussed by both, resulting in the identification of 113 articles. The full texts of these articles were further inspected to confirm full compliance with the selection criteria. As a result, eight articles were selected for the current review (see Figure 1).

4.4 | Study quality

The Cochrane Collaboration's Risk of Bias tools, RoB 2.0 for randomized trials (Sterne et al., 2019) and ROBINS-I for non-randomized trials (Sterne et al., 2016), were used to assess the quality of the studies included in the review. RoB 2.0 assesses bias due to randomization, deviations from intended intervention, missing data, outcome measurement, and selection of reported results. ROBINS-I considers

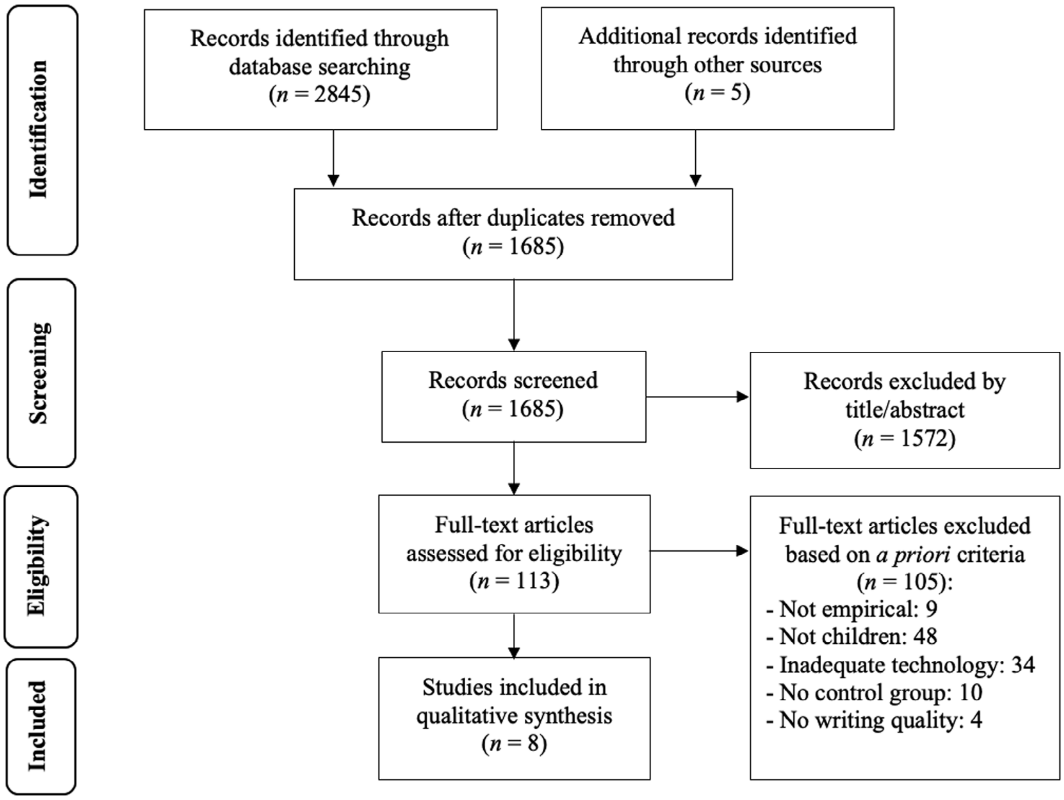

Identification
Screening
Eligibility
Included
Records identified through database searching (*n* = 2845)
Additional records identified through other sources (*n* = 5)
Records after duplicates removed (*n* = 1685)
Records screened (*n* = 1685)
Records excluded by title/abstract (*n* = 1572)
Full-text articles assessed for eligibility (*n* = 113)
Full-text articles excluded based on *a priori* criteria (*n* = 105):
- Not empirical: 9
- Not children: 48
- Inadequate technology: 34
- No control group: 10
- No writing quality: 4
Studies included in qualitative synthesis (*n* = 8)

FIGURE 1 Preferred reporting items in systematic reviews and meta-analyses (PRISMA) flow diagram of study selection

seven types of bias: bias due to confounding, selection of participants, classification of interventions, deviation from intended intervention, missing data, measurement of outcomes, and selection of reported results. For RCT and non-RCT papers, there is a set of questions (e.g., "Were interventions groups clearly defined?") and pre-defined responses ("yes", "no", "probably no", "probably yes", or "no information") for each domain of risk bias. Based on the responses, risk bias is classified with a minus (−) indicating a low risk of bias; a plus (+) indicating a high risk of bias; and a question mark (?) indicating the risk of bias is unclear. The ratings by domain are used to determine the overall risk of bias of the study: for RCT, high risk, some concerns, or low risk; for non-RCT, critical, serious, moderate, low risk, or no information. Low overall risk of bias means the study is considered well-performed; high or critical/serious risk of bias indicates that findings from the study should be interpreted with caution. The two first authors independently graded the risk of bias, and disagreements were solved in discussion with the third author.

4.5 | Coding of included studies

The included studies (Franzke et al., 2005; Mørch et al., 2017; Palermo & Thomson, 2018; Tang & Rich, 2017[1]; Wade-Stein & Kintsch, 2004; Ware, 2014; Wilson & Czik, 2016; Wilson & Roscoe, 2020) were coded in three categories: characteristics of the studies, AWE systems, and effectiveness of AWE systems.

4.5.1 | Characteristics of the studies

We gathered information about publication settings, such as authors and year, the scientific journal where the study was published, and the country where it was conducted. Additionally, we identified the study's research goals and gathered information about participants, such as sample size, grade range, and language (English as a native language—L1, English as a Foreign Language—EFL, and English as a Learning Language—ELL). We also extracted information on the study design (viz., inter- or intra-subject design, experimental or quasi-experimental, conditions, and testing sessions), including the use of mixed-methods or exclusively quantitative or qualitative designs. Finally, we coded the intervention duration and agents (i.e., whether students worked alone with the system or teachers also delivered the intervention).

4.5.2 | AWE systems

We registered the following information: name and developer of the system, writing tasks used to provide feedback (e.g., argumentative essays in response to a prompt), description of the system, its engine, the method used for analysing the meaning of the text (e.g., LSA), targeted features of writing quality (e.g., organization, style), how feedback is computed and delivered, and teachers' role—if the teacher can interact with the AWE system or not.

4.5.3 | Effectiveness of the AWE systems

We collected information about the measures used to test the effectiveness of the AWE systems along with the results achieved. We noted how writing quality was measured (e.g., holistic vs. analytic score) and identified other writing measures (e.g., essay elements). Finally, we summarized the effects of AWE systems on all assessed outcomes.

5 | RESULTS

5.1 | Quality of the studies

All studies except one (Ware, 2014) were non-randomized trials. The plots obtained from the risk of bias analyses designed with the *robvis* web app (McGuiness & Higgins, 2020) are presented in Figures 2 and 3.

Despite the lack of information on the randomization process and the intervention not being blind for participants or instructors, the randomized trial study presented low concerns. Concerning the non-randomized trial studies, some uncontrolled events interfered with the instructional schedule in Palermo and Thomson's (2018) study; the instruction procedures varied between experimental groups, and there was no information about human raters being blinded or not in Tang and Rich's (2017) study; missing data were not described in Mørch et al.'s (2017) and Wade-Stein and Kintsch's (2004) studies. Overall, four studies presented a low risk of bias (Franzke et al., 2005; Wade-Stein & Kintsch, 2004;Wilson & Czik, 2016; Wilson & Roscoe, 2020), two presented a moderate risk (Mørch et al., 2017; Palermo & Thomson, 2018), and one a serious risk (Tang & Rich, 2017).

5.2 | Characteristics of the studies

Below we present the main characteristics of the selected studies (see Table 1 for details).

Selected studies were published between 2004 and 2020. Only two were conducted outside the US with L2 writers (in China and Norway). Six studies focused on middle school (Grades 6–9; $n = 1343$), and two focused on high school (Grades 10–12; $n = 316$). Only one study was experimental (Ware, 2014); the remaining used quasi-experimental designs. All studies included evaluations before and after the intervention (pre- and post-test), but none had a follow-up. The intervention time ranged between 11 days and 10 months, even though not all studies provided that information. Six AWE systems were used (see description in Table A1): PEG Writing (Wilson & Czik, 2016; Wilson & Roscoe, 2020), NC Write (Palermo & Thomson, 2018), Writing Roadmap (Tang & Rich, 2017), EssayCritic (Mørch et al., 2017), Criterion (Ware, 2014) and Summary Street (Franzke et al., 2005; Wade-Stein & Kintsch, 2004). The systems targeted different types of texts: narrative, informative, and argumentative essays, memoir writing, and summary writing.

In the intervention groups, the automated feedback was tested alone (Franzke et al., 2005; Mørch et al., 2017; Wade-Stein & Kintsch, 2004; Ware, 2014) or complemented by teacher feedback (Tang &

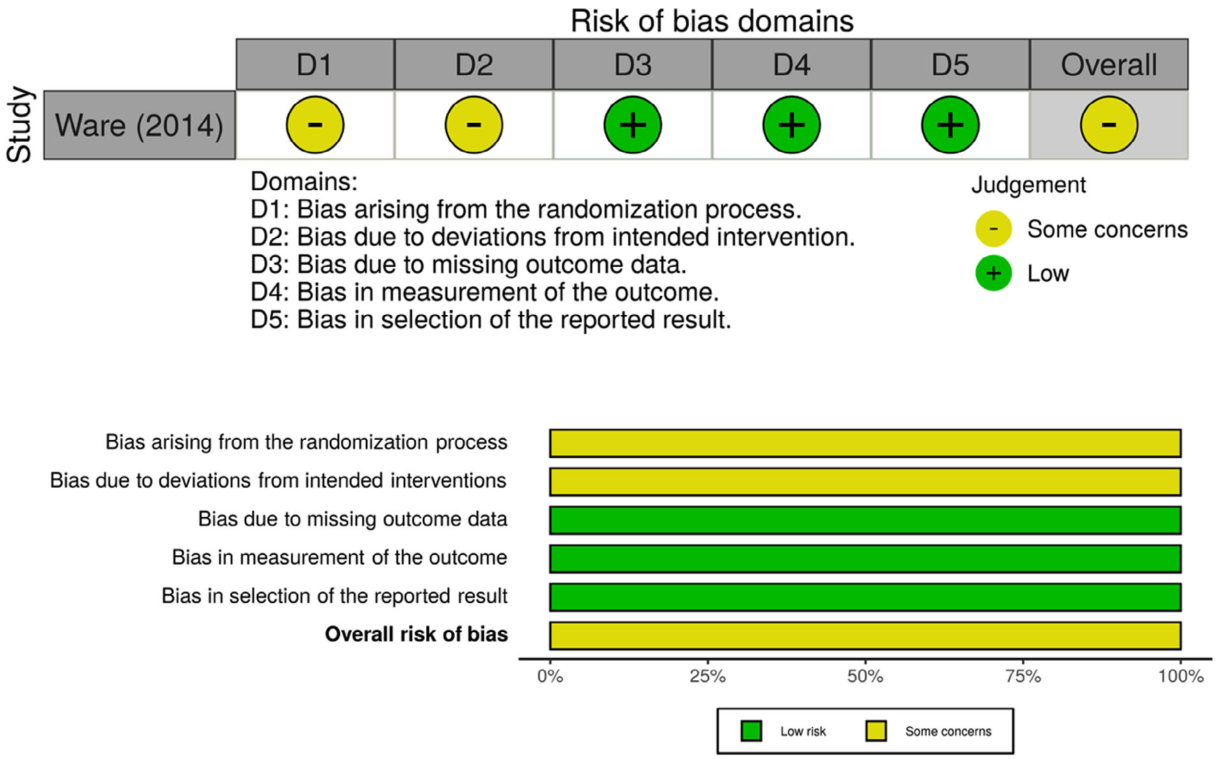
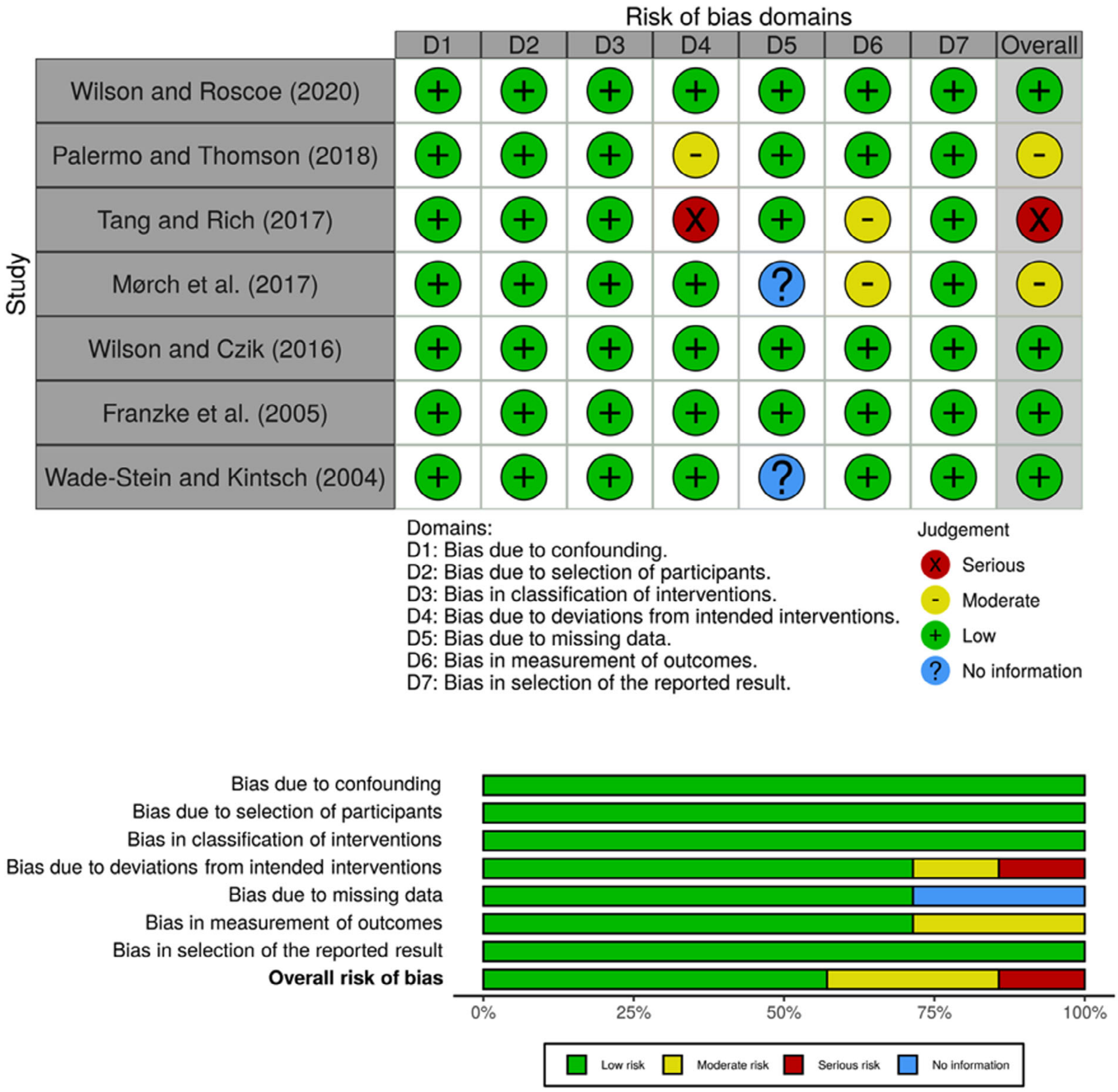


FIGURE 2 Risk of bias summary and graph for the included randomized control trial



Risk of bias domains

Study	D1	D2	D3	D4	D5	D6	D7	Overall
Wilson and Roscoe (2020)	+	+	+	+	+	+	+	+
Palermo and Thomson (2018)	+	+	+	-	+	+	+	-
Tang and Rich (2017)	+	+	+	X	+	-	+	X
Mørch et al. (2017)	+	+	+	+	?	-	+	-
Wilson and Czik (2016)	+	+	+	+	+	+	+	+
Franzke et al. (2005)	+	+	+	+	+	+	+	+
Wade-Stein and Kintsch (2004)	+	+	+	+	?	+	+	+

Domains:
D1: Bias due to confounding.
D2: Bias due to selection of participants.
D3: Bias in classification of interventions.
D4: Bias due to deviations from intended interventions.
D5: Bias due to missing data.
D6: Bias in measurement of outcomes.
D7: Bias in selection of the reported result.

Judgement
X Serious
- Moderate
+ Low
? No information

Bias due to confounding
Bias due to selection of participants
Bias in classification of interventions
Bias due to deviations from intended interventions
Bias due to missing data
Bias in measurement of outcomes
Bias in selection of the reported result
Overall risk of bias
0%
25%
50%
75%
100%
Low risk
Moderate risk
Serious risk
No information

FIGURE 3 Risk of bias summary and graph for the included non-randomized control trials

Rich, 2017; Wilson & Czik, 2016; Wilson & Roscoe, 2020). In the control groups, students received feedback from the teacher (Tang & Rich, 2017; Ware, 2014; Wilson & Czik, 2016; Wilson & Roscoe, 2020), from peers (Mørch et al., 2017; Ware, 2014) or no feedback (Franzke et al., 2005; Wade-Stein & Kintsch, 2004). Palermo and Thomson (2018) compared the use of automated feedback during a Self-Regulated Strategy Development (SRSD) instruction with traditional writing instruction with or without automated feedback.

5.3 | Effectiveness of the AWE systems

5.3.1 | Measures

In half of the studies, trained raters evaluated writing quality using holistic rating scales (Franzke et al., 2005; Mørch et al., 2017; Wade-Stein & Kintsch, 2004; Ware, 2014). Three studies used human and automatic scores (Tang & Rich, 2017; Wilson & Czik, 2016; Wilson & Roscoe, 2020), and one study relied solely on the overall automatic score. As detailed in Table 2, additional writing outcomes were assessed in all studies except in Tang and Rich's (2017).

5.3.2 | Impact on writing

Only one of the eight studies under review did not find any benefit of the AWE systems. Ware (2014) showed no evidence that using the Criterion improved writing quality, length, and mechanical aspects of the written texts compared to online teacher feedback or peer feedback. Surprisingly, students receiving automated feedback had poorer scores than their peers on writing outcomes such as genre elements

TABLE 1 Coding of the included studies: Main characteristics

	Publication settings			Participants			Design-related variables		Intervention characteristics	
Study	Journal	Country	Research goals	*N*	Grade	Language	Study design	Groups	Time	Agent
Franzke et al. (2005)	Journal of Educational Computing Research	USA	Compare two groups: students who practiced writing summaries with and without Summary Street	121	MS Grade 8	L1	- Inter-subject quasi-experimental design with two conditions (one experimental and one control group) and two testing sessions (pre and posttest) - Mixed-methods design	E-group (*n* = 52; 46.2% females) C-group (*n* = 59; 64.4% females) Age range: 13.5–14.5 (average age: N/A)	4 weeks (twice a week)	AWE (experimental)
Mørch et al. (2017)	Educational Technology & Society	Norway	Compare two feedback conditions: automated feedback from EssayCritic and feedback from collaborating peers.	48	HS	EFL	- Inter-subject quasi-experimental design with two conditions (one experimental and one active control group) and two testing sessions (pre and posttest) - Mixed-methods design	E-group (*n* = 24) C-group (*n* = 24) Gender ratio and average age: N/A	N/A	AWE system (E-group) Peers (C-Group)
Palermo and Thomson (2018)	Contemporary Educational Psychology	USA	Examine the effectiveness on students' writing of integrating NC Write into an SRSD intervention or traditional writing instruction	829	MS grades 6–8	L1	- Inter-subject quasi-experimental design with three conditions (experimental 1, experimental 2, and passive control group) and two testing sessions (pre and posttest) - Mixed-methods design - Large-scale study (10 schools in different districts)	- E-group NC + SRSD (n = 287; 54.9% females; M_{age} = 156.4 months; SD_{age} = 12.3) - E-group NC + TRAD (n = 272; 51% females; M_{age} = 155.5 months; SD_{age} = 9.1) - C-group (n = 270; 51.9% females; M_{age} = 156.5 months; SD_{age} = 9.9)	8 weeks (45 min, 2 days/week)	Classroom Teacher + AWE (E-groups) Classroom Teacher (C-group)
Tang and Rich (2017)	The Jalt Call Journal	China	Examine how Writing Roadmap' feedback affects students' writing (along teacher feedback)	268	HS	EFL	- Inter-subject quasi-experimental design with three cohorts of two conditions each (one experimental and one active control group, in each cohort) and two testing sessions (pre and posttest) - Mixed-methods design	E-group (*n* = 138) C-group (*n* = 130) Age range: 15–17 years (no mean) Gender ratio: N/A	Two semesters (10 months)	Classroom Teacher + AWE (E-group) Classroom Teacher (C-group)

TABLE 1 (Continued)

	Publication settings			Participants			Design-related variables		Intervention characteristics	
Study	Journal	Country	Research goals	*N*	Grade	Language	Study design	Groups	Time	Agent
Wade-Stein and Kintsch (2004)	Cognition and Instruction	USA	Examine Summary Street' feedback on writing quality	52	MS Grade 6	L1	- Intra-subjects quasi-experimental design, with two conditions: one experimental group and one control group - Mixed-methods design	Total sample: $N = 52$ (two classes, divided in two conditions each; no further info available)	2 weeks	AWE (experimental condition)
Ware (2014)	Writing & Pedagogy	USA	Examine the impact of Criterion feedback on students' writing and compare it with teacher and peer feedback	82	MS Grade 8	L1 & ELL	- Inter-subject randomized experimental design with three conditions (one experimental, and two active control groups) and two testing sessions (pre and posttest) - Quantitative research design	E-group AWE ($n = 26$) C-group 1 ($n = 27$) C-group 2 ($n = 29$) Gender ratio: 47.6% females Average age: N/A	Six-weeks (full), 45-min class	AWE (E-group) Online teacher (C-group 1) Peers (C-group 2)
Wilson and Czik (2016)	Computers & Education	USA	Examine the efficacy of PEG Writing on students' writing (along teacher feedback)	145	MS Grade 8	L1	- Inter-subject quasi-experimental design with two conditions (one experimental and one active control group) and two testing sessions (pre and posttest) - Quantitative research design	E-group ($n = 72$; 46% females; $M_{age} = 168.6$ months; $SD_{age} = 4.9$) C-group ($n = 73$; 53% females; $M_{age} = 169$ months; $SD_{age} = 6$)	11 days (1 h/day)	Classroom Teacher + AWE (E-group) Classroom Teacher (C-group)
Wilson and Roscoe (2020)	Journal of Educational Computing Research	USA	Examine the efficacy of PEG Writing on students' writing (along teacher feedback)	114	MS Grade 6	L1	- Inter-subject quasi-experimental design with two conditions (one experimental and one active control group) and two testing sessions (pre and posttest) - Quantitative research design	E-group ($n = 56$; 64% females; $M_{age} = 140.4$ months; $SD_{age} = 4.6$) C-group ($n = 58$; 62% females; $M_{age} = 140.1$ months; $SD_{age} = 4.1$)	7 months	Classroom Teacher + AWE (E-group) Classroom Teacher (C-group)

Abbreviations: MS, middle-school; HS, high-school; L1, English as native language; EFL, English as foreign language; ELL, English as learning language; E-group, experimental group; C-group, control group; PEG, project essay grade; SRSD, self-regulated strategy development; N/A, not available.

TABLE 2 Coding of the included studies: Effectiveness of the AWE systems

Study	Writing quality outcomes		Other writing outcomes	
	Measures	Results	Measures	Results
Franzke et al. (2005)	Rater Score: Holistic rating scale	Summaries of the E-group had significantly better quality than C-group.	Organization Mechanics (sentence structure, punctuation and spelling) Minimal use of detail Writing style Summary completion rates: number of texts summarized	Organization, content coverage, minimal use of detail, writing style: advantage of the E-group compared to C-group Mechanics and completion rates: no differences
Mørch et al. (2017)	Rater score: Holistic rating scale (text grades)	Both groups improved from pre- to posttest. No significant between-group differences in the final grades	Number of subthemes included	Both groups improved from pre- to post-test, but the E-group included significantly more subthemes in the final essays than the C-group
Palermo and Thomson (2018)	PEG Overall score	Both E-groups produced higher quality essays than the C-group, with an advantage of the NC Write + SRSD E-group over the NC Write only E-group	Essay length: number of words Essay elements: claim, supporting reasons, elaborations, counterclaims, and conclusion	Students in the NC Write + SRSD E-group produced longer essays and included more basic elements of argumentative essays than students in the other two conditions.
Tang and Rich (2017)	Writing roadmap score Rater score (to ensure reliability)	E-group outperformed the C-group in pre- and post-writing tests	N/A	N/A
Wade-Stein and Kintsch (2004)	Rater Score: Holistic rating scale	Students wrote better summaries in the E-group than C-group.	Time on task: time spent working on summaries Content adequacy	Time on task: students in the E-group spent more than twice as long engaged in the writing task and were more likely to keep revising until the content of all sections had been adequately covered. Content adequacy: summaries on the E-group did a more balanced coverage of the content
Ware (2014)	Rater Score: Holistic rating scale	All groups improved from pre- to posttest. No significant differences between the E-group and the C-groups.	Genre elements score: the raw number of up to 6 total elements deemed central to the genre of open-ended response Length score: the total number of words written during the timed writing period of 45 min Composite error rate score: composite percentage of mechanical errors	Genre elements score: at post-test the two C-groups obtained significantly higher scores than the E-group. Length score: all groups wrote longer essays at post-test, but no significant differences were found between groups. Error rate: no differences found.
Wilson and Czik (2016)	PEG Overall and Trait score Rater Score: Holistic rating scale	No significant differences between groups.	Writing motivation: Writing Disposition Scale (WDS; Piazza & Siebert, 2008)	Students in the E-group reported more writing motivation than those the C-group.
Wilson and Roscoe (2020)	PEG Overall score Rater Score: Holistic rating scale	No significant differences between groups. None of the groups improved significantly from pre- to posttest.	Writing self-efficacy: 22-item Self-Efficacy for Writing Scale (SEWS; Bruning et al., 2013)	Writing self-efficacy: students who used PEG experienced greater gains in posttest self-efficacy

Abbreviations: AWE, automated writing evaluation; E-group, experimental group; C-group, control group; PEG, project essay grade; SRSD, self-regulated strategy development; N/A, not applicable.

(see Table 2). Three studies found positive effects of AWE systems on several writing measures but not on writing quality proper. Mørch et al. (2017) showed that the feedback provided by the EssayCritic resulted in a higher number of relevant topics in the essays than the feedback provided by peers. And Wilson and Czik (2016) and Wilson and Roscoe (2020) showed that, compared to students who used Google Docs to receive teacher feedback, students using PEG Writing had greater writing motivation and writing self-efficacy, respectively.

NC Write (Palermo & Thomson, 2018) produced better, longer, and more complete essays than students receiving traditional instruction either with or without NC Write; importantly, among students receiving traditional instruction, those using NC Write achieved better results than the ones without it. Tang and Rich (2017) showed that the Chinese EFL learners using Writing Roadmap improved more from pretest to posttest than the control group receiving teacher feedback alone. Franzke et al. (2005) showed that, compared to no feedback, using Summary Street led to better results in writing quality, content, organization, degree of detail, and writing style; although it helped all students, the system was especially beneficial for medium- and low-performing students. Confirming the effectiveness of Summary Street, Wade-Stein and Kintsch (2004) found that when students used this system, they produced better summaries, spent more time on task, and provided better coverage of summary contents than when they did not use it. Because this study had an intra-subject design, the authors also showed that students who used the system first kept their gains later in the no-feedback condition.

5.3.3 | Users' perceptions

Students who used AWE systems mentioned that the automated feedback helped them improve writing, made them feel more confident and motivated to rewrite and revise, kept them focused for longer, and increased their enjoyment in writing (Palermo & Thomson, 2018; Tang & Rich, 2017; Ware, 2014). Furthermore, they perceived the systems as valuable, inciting them to reflect upon and be more aware of their writing process (Franzke et al., 2005; Wade-Stein & Kintsch, 2004). Teachers mentioned that AWE systems were appropriate for students and valuable for their practice by helping them focus on teaching rather than on correcting errors (Palermo & Thomson, 2018; Tang & Rich, 2017). They also mentioned that struggling writers might need additional attention and support in interpreting the automated feedback (Palermo & Thomson, 2018; Wilson & Roscoe, 2020). A complete summary of users' perceptions is given in Table A2.

6 | DISCUSSION

Writing is recognized as an essential component of students' academic development, one that requires a considerable amount of time and effort from both teachers and students (Dikli, 2010). More and more research shows the potential of technology to help in the teaching and learning process of writing. Due to technological advancements and the need to adjust this process to new challenges, such as those imposed by the current pandemic, technology capable of providing writing feedback—AWE systems—has gained popularity. However, little is still known about its added value in educational settings. Here, we conducted a systematic review focused on empirical research testing the effectiveness of AWE systems in Grades 1–12 published in the two last decades. In what follows, we discuss several aspects of AWE systems' effectiveness, namely the results obtained by each system and users' perceptions about them. Additionally, we present the studies' weaknesses and recommendations for future research.

6.1 | Effectiveness of AWE systems

Our search revealed that studies testing the effectiveness of AWE systems in Grades 1–12 are scarce. Only eight studies met the inclusion/exclusion criteria and were included in this review. Generally, our review supported the usefulness of AWE systems in the teaching and learning process of writing: all but one study showed a positive effect of automated feedback in at least one writing-related measure. The role of teachers in supporting the use of AWE systems, the time to practice writing, and the type of control groups seem to be critical determinants underlying these findings, as will be discussed next.

6.1.1 | Effectiveness of AWE systems with teachers as main agents

Teachers played a leading role in complementing the automated feedback of three systems, NC Write and Writing Roadmap (that improved text quality) and PEG Writing (no effect). Palermo and Thomson (2018) as well as Tang and Rich (2017) integrated the AWE systems in longer and comprehensive writing interventions, whereas Wilson and Czik (2016) and Wilson and Roscoe (2020) integrated them into a short-length intervention and regular classroom instruction, respectively. The integration of automated feedback in instructional contexts seems a key factor underlying AWE effectiveness. Positive effects of writing intervention programs, such as the SRSD used by Palermo and Thomson (2018), have already been observed (Harris & Graham, 2016; Limpo & Alves, 2013b). This review extends these findings by showing that these programs provide suitable contexts for AWE systems to improve writing and motivation to write. Research with older writers (e.g., university students) has also indicated that providing comprehensive instructional settings, either led by teachers (Liao, 2016) or run through technology (e.g., ITS; Roscoe & McNamara, 2013), is a means to support the use of the AWE systems and boost their impact. These instructional contexts provide students with ample opportunities to practice writing in response to the automated feedback, in supportive and tutor-guided learning environments.

Practicing writing during instructional writing programs is an essential factor to achieve proficiency in writing and boost motivation.

Writing is a complex problem-solving activity (Bereiter & Scardamalia, 1987; Hayes, 2012) that relies on a set of high-level cognitive processes fine-tuned throughout schooling. A considerable amount of research showed that this fine-tuning could be achieved through explicit instruction such as teaching writing strategies (Graham & Perin, 2007). However, explicit teaching is not enough: it must be combined with scaffolded practice connecting instruction to feedback in multiple cycles of drafting, revising, and editing (Graham & Perin, 2007). For automated feedback to be effective, students should have enough opportunity to engage in those writing cycles under the guidance of the AWE system (Kellogg & Whiteford, 2009). This was not the case in the studies with PEG Writing reviewed here (Wilson & Czik, 2016; Wilson & Roscoe, 2020), which may explain why using this system did not improve writing quality (though it improved writing motivation and self-efficacy). The importance of practice when using AWE systems was recently made clear in a case study with a mixed-methods design, in which two undergraduates improved in L2 writing using the Criterion system for one-year, including monthly moments of writing practice (Lee, 2020).

Even guided by AWE systems, students also need the support of teachers during writing practice (Graham, 2019; Graham & Harris, 2016). According to the notion of zone of proximal development (ZPD; Vygotsky, 1978), sufficient exposure to instructional feedback results in gradual internalization and independent performance. But for that to happen, teachers must scaffold the use of AWE systems in the classroom and keep them aligned with instructional goals (Hibert, 2019). Studies with college students showed the added value of teachers when using automated feedback (Li et al., 2015; Parra & Calero, 2019). In our review, this added value was apparent in Palermo and Thomson .'s (2018) and Tang and Rich's (2017) studies. Also, the integration of AWE systems in a supportive environment may explain the improvements in motivation and self-efficacy observed by Wilson and Czik (2016) and Wilson and Roscoe (2020). Other studies have shown that using AWE systems makes students more motivated to write (Warschauer & Grimes, 2008), and increases their independence and writing through self-revisions (Sandolo, 2010); they feel involved and in charge of the writing process and more confident in their skills (Pajares, 2003). Indeed, previous systematic reviews showed that different technological solutions for writing increased motivation and attitudes towards writing, including students' beliefs about their capabilities (Camacho et al., 2020; Ekholm et al., 2017).

6.1.2 | Effectiveness of AWE systems used in isolation

Four studies selected for this review tested AWE systems as the sole source of writing and revision feedback, not supported by teachers nor embedded in comprehensive instructional programs (Summary Street, EssayCritic, Criterion; Franzke et al., 2005; Mørch et al., 2017; Wade-Stein & Kintsch, 2004; Ware, 2014).

Compared to controls receiving no feedback, students using Summary Street showed higher writing quality and better performance in several other writing outcomes (viz., organization, content coverage and adequacy, minimal use of detail, writing style, and time on task) in Grades 6 and 8 (Franzke et al., 2005). This means that Summary Street worked better than no feedback, but the question is if students improved because of the system or simply because they received some form of feedback. To answer this question, AWE systems should be compared with active controls receiving feedback from teachers or peers. The studies examining EssayCritic (Mørch et al., 2017) and Criterion (Ware, 2014) included active control groups, and their results showed that both systems led to better writing quality in high-school and Grade 8 students; however, performance at post-test was not significantly better than that of the control students who had received feedback from other agents. It seems that regardless of being given by computers, teachers, or peers, it is the power of feedback that matters most (Graham et al., 2015). Studies comparing AWE systems with passive control groups with no-feedback provide little, if any, evidence of their specific effectiveness. Passive control group designs are less informative than active control designs because it is difficult to untangle the contribution of the factors under comparison, such as AWE vs. teacher feedback, or common factors such as those kept constant to ensure that conditions are comparable (Lindquist et al., 2007).

If the control condition uses interventions known to be effective, it will not be surprising if outcomes do not differ between experimental (e.g., AWE) and control groups (e.g., teacher feedback). The automated system can improve writing outcomes, but these improvements may be similar to those produced by receiving feedback from teachers or peers. Experimental and control conditions should be as equivalent as possible, with differences limited to the components under test (Campbell & Stanley, 1963; Higgins et al., 2021; Lindquist et al., 2007), in this case, limited to feedback provider, computer vs. human. In the study testing EssayCritic (Mørch et al., 2017), in addition to varying the feedback agent (computer vs. peers), the control group was provided with more teacher assistance than the experimental group. Differences in the amount of teacher assistance make it difficult to interpret condition effects unequivocally.

The use of research designs comparing automated vs. human feedback is common in the field (Stevenson & Phakiti, 2014). However, this comparison lacks ecological validity (Palermo & Wilson, 2020) as it contrasts two different sources of feedback, system vs. teacher, instead of two equivalent forms, teacher with or without the system. Moreover, these designs emphasize a false dichotomy between teacher vs. system and inadvertently add to the debate pitting the machine against the teacher. As noted in the introduction, the concern that AWE systems replace human feedback agents is real. But as noted by Attali (2013) and confirmed in the current review, AWE systems are more valuable when used as a complement rather than as a replacement of teachers. Machines may not be better judges than teachers, who have knowledge that machines do not, but these are more efficient and productive (Elliot & Klobucar, 1913). Therefore, writing learning and instruction would benefit from the complementary, rather than isolated, strengths of automated and human raters (Cahill & Evanini, 2020).

6.2 | Users' perceptions of AWE systems

Both students and teachers reported that AWE systems could be of great use in the classroom. Perceived usefulness is relevant because users' perceptions impact the adoption, use, and effectiveness of such tools, as proposed in several theoretical models of technology acceptance, such as the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al., 2012). These models propose that people's intended and effective use of technology is influenced by their perceptions (for a review see Venkatesh, 2000). In education settings, a systematic review showed that students are more willing to accept a technological learning tool when they perceive it to be useful and easy to use (Granić & Marangunić, 2019). This finding was replicated in another systematic review targeting AWE systems in college (Na & Ma, 2021). For example, students only utilized the feedback they perceived to be valuable and helpful (Wang, 2015) and perceptions about technical support, usefulness, ease of use, and attitude towards AWE systems, influenced how much writing improved in response to feedback (Tsai, 2014). The students' positive perception observed in the current review is encouraging because it is conducive to the effectiveness of AWE systems.

Teachers are key players to foster the effective use of technology for learning and instruction and stimulate students' acceptance of technology (Davidson et al., 2014; Roscoe et al., 2017; Teo, 2011). Studies in university settings have confirmed the effect of teachers' perceptions on students' acceptance of AWE systems. For example, Li et al. (2015) concluded that students' perceptions of automated feedback depended on their teachers' use and perceptions of AWE. In the studies included in this review, none of the teachers reported avoiding automated feedback in their writing instruction. Instead, teachers relied on these systems to complement their teaching and recognized their many advantages. Contrary to the view that automated systems would dehumanize the writing process and replace teachers, we found no evidence of this concern among teachers. Interestingly, Grimes and Warschauer (2010) advocated that the humanization, or not, of writing instruction by AWE systems depends much more on the teacher than on the machine: if a teacher uses the systems to facilitate the instruction process and overcome students' reluctance to write, using AWE systems is unlikely to dehumanize writing instruction. However, dehumanization could occur if the teacher uses the automated system only to determine grades without integrating it into the teaching process. These claims align with our findings, suggesting that the integration of AWE systems into comprehensive instructional contexts facilitates their effectiveness.

Despite the generally positive perception, teachers identified some caveats in the use of AWE systems, which related to students with more difficulties in writing (Palermo & Thomson, 2018; Wilson, 2017; Wilson & Roscoe, 2020). They noted that to effectively use and understand automated feedback, students with writing needs might require additional attention and proximal assistance from teachers. A similar conclusion emerged from studies with adults. Liao (2016) found that low-performing writers had difficulty understanding and addressing automated feedback. Because struggling writers tend to produce shorter, less-developed, and more error-filled texts than their peers, teacher-oriented feedback should complement AWE systems (Troia, 2006).

6.3 | Studies' weaknesses and future directions

Despite encouraging results, our review also identified weaknesses related to the implementation of the studies. These are presented next, along with suggestions for future studies testing the effects of AWE systems.

6.3.1 | Lack of information

The lack of relevant information was the main problem emerging from the procedure to evaluate study quality. Studies did not inform on how participants were allocated to conditions (Ware, 2014), on coders' blindness to the purpose of the study (Tang & Rich, 2017), and on the extent of missing data (Mørch et al., 2017; Wade-Stein & Kintsch, 2004). This lack of information seems to be a persistent problem in the field, as noted in a previous review of studies from 1990 until 2011 (Stevenson & Phakiti, 2014). Space constraints in many journals can be the reason, but adding supplementary materials with detailed descriptions of the methods is an easy solution, and an important one, because lack of information impedes research replication and limits the generalization of findings.

6.3.2 | Grade-related bias

Although our review targeted Grades 6 to 12, most studies were conducted in middle-school grades. The prevalence of middle-grade studies using AWE systems mimics the writing research field in general (Graham & Perin, 2007). There is thus a clear need for more studies testing the effectiveness of AWE systems across all school grades, namely, high school and primary school. Computer technology to assist early reading and writing in primary school has been successful. For example, Carvalhais et al. (2020) showed the effectiveness of computer-based teaching tools for second graders who struggle to read, indicating that young writers benefit from computer-assisted tools as long as they are adapted to their age. Adapting existing AWE systems to the cognitive and motivational profiles of primary-grade learners looks like a promising next step. Automated feedback may be of great value to scaffold teachers and students in the early phases of writing learning and instruction.

6.3.3 | Focus on English

All AWE systems were in English, either as native language in US contexts (six studies) or as a foreign language (EFL; two studies). This pattern conforms to writing research in the field, which is biased towards English as L1 (Graham & Perin, 2007). Of the two AWE systems tested in EFL, only one (EssayCritic) was developed explicitly for L2 learners. This state of affairs shows that the field needs more studies focused on L2 learners. Note that systems used with L2 school-age writers were developed and tested with native speakers—despite being commercialized for non-native learners as well (Warschauer & Ware, 2006). This situation is problematic because the acquisition of the first and second languages is qualitatively different: L1 is acquired in the early years through naturalistic exposure to language, but L2 learning depends on factors such as age, motivation, L1 and L2 resemblance, among others (Chenu & Jisa, 2009). These differences indicate that the effectiveness of automated feedback on L2 writing should be specifically addressed (Stevenson & Phakiti, 2014) instead of being generalized from L1 studies. Finally, although the use of AWE systems by L2 writers has been studied in university settings (Li et al., 2015; Wang et al., 2013), as shown here, there is barely any research on their effects in earlier educational levels.

6.3.4 | No examination of long-term effects

All studies employed a pre- and post-test design. However, none included a follow-up to ensure the maintenance of the effects. We believe this is a critical limitation of AWE studies. By contrast, in studies addressing reading there is some indication that effects persist over time. For example, the effect of feedback from a web-based instructional tool on reading comprehension persisted after 4 months (Meyer et al., 2010). Similar studies should be conducted to test the long-term effects in the field of writing.

6.4 | Limitations of the current review

This systematic review has several limitations. First, we aggregated findings from independent studies to draw conclusions about the effects of AWE systems on writing quality. However, the value of conclusions depends on several factors, and the quality of the included studies is a major one. Hence, results on the systems' effectiveness should be interpreted considering the weaknesses of the studies. Second, despite using systematic procedures based on PRISMA guidelines (Moher et al., 2009), we may have missed some studies. We defined inclusion and exclusion criteria based on studies published in scientific journals with peer revision. However, some of the criteria could have led to the exclusion of some relevant research, namely, studies published in non-English languages. Third, we selected studies that assessed writing quality with quantitative measures. Therefore, we cannot draw conclusions about studies using non-quantitative measures or studies that targeted non-writing measures only. Fourth, as in any review, we did not directly analyse any of the AWE systems described and based our analysis on second-hand information. Functionalities of the systems not addressed by the authors of the reviewed articles may not have been covered in this review. Fifth, as this review did not cover all grade levels, we cannot assert that using AWE systems is effective at specific grade levels. As noted above, this indicates the need for more research across different school levels. Finally, given the reduced number of studies testing each of the AWE systems targeted here, more research is needed before drawing strong conclusions about their effectiveness.

7 | CONCLUSION

AWE systems are not yet widely adopted by teachers or schools (Graham, 2019). Still, they hold great potential to support writing instruction, mainly with the recent boom of online classes. Resonating with Grimes and Warschauer's (2010) "utility of a fallible tool", we conclude that, despite limitations, AWE systems deserve consideration as a fallible tool that holds promise to enhance students' writing. Even though some caution is needed to interpret the results obtained, our review contributes in three ways to the field of automated feedback in writing. First, we show that there is a lack of studies employing rigorous designs that control for alternative explanations, which are essential to understand the real effectiveness of AWE systems. Second, our findings highlight the importance of the instructional context in which the systems are integrated, which is typically considered a minor feature of the study rather than an influential variable capable of influencing the effectiveness of the AWE systems. Finally, this review put together several limitations in the study of AWE systems, which is helpful to indicate valuable research avenues to move the field forward and stimulate the development of evidence-based AWE systems.

ACKNOWLEDGEMENTS

This work was supported by the Portuguese Foundation for Science and Technology (grant UIDB/00050/2020 awarded to the Center for Psychology at University of Porto; Ph.D. studentship SFRH/BD/139195/2018 awarded to Andreia Nunes).

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare that are relevant to the content of this article.

ENDNOTE

[1] Tang and Rich (2017) report studies in secondary and university settings. For this review, we considered only the study with high-school students.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12635.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Andreia Nunes https://orcid.org/0000-0003-0298-3840
Carolina Cordeiro https://orcid.org/0000-0001-8085-314X
Teresa Limpo https://orcid.org/0000-0002-9903-7289
São Luís Castro https://orcid.org/0000-0002-1487-3596

REFERENCES

*References with an asterisk indicate studies included in this systematic review.

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 316–329). Guilford.

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/Taylor & Francis Group.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.

Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-english and L2-writing development: A meta-analysis. *ETS Research Report RR-11-05 Princeton, NJ- ETS, 2011*. https://www.ets.org/Media/Research/pdf/RR-11-05.pdf i, 99.

Bruning, R., Dempsey, M., Kauffman, D. F., McKim, C., & Zumbrunn, S. (2013). Examining dimensions of self-efficacy for writing. *Journal of Educational Psychology*, *105*(1), 25–38. https://doi.org/10.1037/a0029692

Burstein, J., Riordan, B., & McCaffrey, D. (2020). Expanding automated writing evaluation. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 329–346). CRC Press/Taylor & Francis Group.

Cahill, A., & Evanini, K. (2020). Natural language processing for writing and speaking. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 69–92). CRC Press/Taylor & Francis Group.

Camacho, A., Alves, R. A., & Boscolo, P. (2020). Writing motivation in school: A systematic review of empirical research in the early twenty-first century. *Educational Psychology Review*, *33*, 213–247. https://doi.org/10.1007/s10648-020-09530-4

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally.

Carvalhais, L., Limpo, T., Richardson, U., & Castro, S. L. (2020). Effects of the Portuguese GraphoGame on reading, spelling, and phonological awareness in second graders struggling to read. *Journal of Writing Research*, *12*(1), 9–34. https://doi.org/10.17239/jowr-2020.12.01.02

Chenu, F., & Jisa, H. (2009). Reviewing some similarities and differences in L1 and L2 lexical development. *Acquisition et interaction en langue étrangère*, *17-38*, 17–38. https://doi.org/10.4000/aile.4506

Cheung, A. C. K., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review*, *7*(3), 198–215. https://doi.org/10.1016/j.edurev.2012.05.002

Conference on College Composition and Communication. (2014). Writing assessment: A position statement. http://www.ncte.org/cccc/resources/positions/writingassessment.

Cotos, E. (2014). Automated writing evaluation. In E. Cotos (Ed.), *Genre-based automated writing evaluation for L2 research writing* (pp. 1–64). Palgrave Macmillan.

Davidson, L. Y. J., Richardson, M., & Jones, D. (2014). Teachers' perspective on using technology as an instructional tool. *Research in Higher Education*, *24*, 1–25. https://eric.ed.gov/?id=EJ1064110

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, *5*(1), 1–36. https://ejournals.bc.edu/index.php/jtla/article/view/1640

Dikli, S. (2010). The nature of automated essay scoring feedback. *CALICO Journal*, *28*(1), 99–134. https://doi.org/10.11139/cj.28.1.99-134

Ekholm, E., Zumbrunn, S., & DeBusk-Lane, M. (2017). Clarifying an elusive construct: A systematic review of writing attitudes. *Educational Psychology Review*, *30*(3), 827–856. https://doi.org/10.1007/s10648-017-9423-5

Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 16–35). Routledge/Taylor & Francis Group.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of intelligent essay assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68–88). Routledge/Taylor & Francis Group.

*Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary street®: Computer support for comprehension and writing. *Journal of Educational Computing Research*, *33*(1), 53–80. https://doi.org/10.2190/DH8F-QJWM-J457-FQVB

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *The Journal of Technology, Learning and Assessment*, *2*(1), 1–51. https://ejournals.bc.edu/index.php/jtla/article/view/1661

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). Guilford.

Graham, S. (2018). A revised writer(s)-within-community model of writing. *Educational Psychologist*, *53*(4), 258–279. https://doi.org/10.1080/00461520.2018.1481406

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, *43*(1), 277–303. https://doi.org/10.3102/0091732x18821125

Graham, S., & Harris, K. R. (2016). A path to better writing. *The Reading Teacher*, *69*(4), 359–365. https://doi.org/10.1002/trtr.1432

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, *115*(5), 523–547. https://doi.org/10.1086/681947

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology*, *50*(5), 2572–2593. https://doi.org/10.1111/bjet.12864

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, *8*(6), 1–44. https://ejournals.bc.edu/index.php/jtla/article/view/1625

Harris, K. R., & Graham, S. (2016). Self-regulated strategy development in writing. *Policy Insights From the Behavioral and Brain Sciences*, *3*(1), 77–84. https://doi.org/10.1177/2372732215624216

Hattie, J., & Timperley, H. (2016). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hibert, A. I. (2019). Systematic literature review of automated writing evaluation as a formative learning tool. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming learning with meaningful technologies* (pp. 655–658). Springer.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2021). Cochrane Handbook for Systematic Reviews of Interventions version 6.2. www.training.cochrane.org/handbook

Hockly, N. (2019). Automated writing evaluation. *ELT Journal*, *73*(1), 82–88. https://doi.org/10.1093/elt/ccy044

Jacovina, M. E., & McNamara, D. S. (2016). Intelligent tutoring systems for literacy: Existing technologies and continuing challenges. In R. K. Atkinson (Ed.), *Intelligent tutoring systems: Structure, applications and challenges*. Nova Science Publishers.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1–26. https://doi.org/10.17239/jowr-2008.01.01.1

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist*, *44*(4), 250–266. https://doi.org/10.1080/00461520903213600

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated essay scoring: A cross disciplinary perspective. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring and annotation of essays with the intelligent essay assessor* (pp. 87–112). Lawrence Erlbaum Associates.

Landauer, T. K., & Psotka, J. (2000). Simulating text understanding for educational applications with latent semantic analysis: Introduction to LSA. *Interactive Learning Environments*, *8*(2), 73–86. https://doi.org/10.1076/1049-4820(200008)8:2;1-b;ft073

Lee, Y.-J. (2020). The long-term effect of automated writing evaluation feedback on writing development. *English Teaching*, *75*(1), 67–92. https://doi.org/10.15858/engtea.75.1.202003.67

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

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

Limpo, T., & Alves, R. A. (2013a). Modeling writing development: Contribution of transcription and self-regulation to Portuguese students' text generation quality. *Journal of Educational Psychology*, *105*(2), 401–413. https://doi.org/10.1037/a0031391

Limpo, T., & Alves, R. A. (2013b). Teaching planning or sentence-combining strategies: Effective SRSD interventions at different levels of written composition. *Contemporary Educational Psychology*, *38*(4), 328–341. https://doi.org/10.1016/j.cedpsych.2013.07.004

Lindquist, R., Wyman, J. F., Talley, K. M. C., Findorff, M. J., & Gross, C. R. (2007). Design of control-group conditions in clinical trials of behavioral interventions. *Journal of Nursing Scholarship*, *39*(3), 214–221. https://experts.umn.edu/en/publications/design-of-control-group-conditions-in-clinical-trials-of-behavior

Little, C. W., Clark, J. C., Tani, N. E., & Connor, C. M. (2018). Improving writing skills through technology-based instruction: A meta-analysis. *Review of Education*, *6*(2), 183–201. https://doi.org/10.1002/rev3.3114

McGuiness, L. A., & Higgins, J. P. T. (2020). Risk-of-bias visualization (robvis): An R package and shiny web app for visualizing risk-of-bias assessments. *Research Synthesis Methods*, *12*(1), 55–61. https://doi.org/10.1002/jrsm.1411

Meyer, B. J. F., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P., Meier, C., & Spielvogel, J. (2010). Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth- and seventh-grade readers. *Reading Research Quarterly*, *45*, 62–92. https://doi.org/10.1598/RRQ.45.1.4

Miller, S. A., & Forrest, J. L. (2001). Enhancing your practice through evidence-based decision making: PICO, learning how to ask good questions. *Journal of Evidence-Based Dental Practice*, *1*, 136–141. https://doi.org/10.1067/med.2001.118720

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*, e1000097. https://doi.org/10.1371/journal.pmed.1000097

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

Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/readers: A meta-analysis of research findings. *Reading and Writing*, *25*(3), 641–678. https://doi.org/10.1007/s11145-010-9292-5

Na, Z., & Ma, X. (2021). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*, 1–26. https://doi.org/10.1080/09588221.2021.1897019

Nurmukhamedov, U. (2009). Teacher feedback on writing: Considering the options. *Writing & Pedagogy*, *1*(1), 113–124. https://doi.org/10.1558/wap.v1i1.113

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(210), 210. https://doi.org/10.1186/s13643-016-0384-4

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54) Lawrence Erlbaum Associates, Inc.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly*, *19*, 139–158. https://doi.org/10.1080/10573560308222

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

Palermo, C., & Wilson, J. (2020). Implementing automated writing evaluation in different instructional contexts: A mixed-methods study. *Journal of Writing Research*, *12*(1), 63–108. https://doi.org/10.17239/jowr-2020.12.01.04

Parra, L., & Calero, 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

Peterson-Karlan, G. R., & Parette, H. P. (2007). Supporting struggling writers using technology: Evidence-based instruction and decision-making. In *Department of Special*. Education Illinois State University. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.605&rep=rep1&type=pdf

Piazza, C. L., & Siebert, C. F. (2008). Development and validation of a writing dispositions scale for elementary and middle school students. *The Journal of Educational Research*, *101*(5), 275–286. https://doi.org/10.3200/JOER.101.5.275-286

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*, 103–134. https://doi.org/10.1016/S0747-5632(01)00052-8

Roscoe, R. D., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, *105*(4), 1010–1025. https://doi.org/10.1037/a0032340

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

Rowntree, D. (1987). *Assessing students: How shall we know them?*. Kogan Page.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–140. https://doi.org/10.1007/BF00117714

Sandolo, L. (2010). How can the use of technology enhance writing in the classroom? [Master dissertation, St. John Fisher College]. Fisher Digital Publications. https://fisherpub.sjfc.edu/education_ETD_masters/194/.

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics Decision Making*, *7*. https://doi.org/10.1186/1472-6947-7-16

Schneider, C., & Boyer, M. (2020). Design and implementation for automated scoring systems. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 217–239). CRC Press/Taylor & Francis Group.

Shermis, M. D. (2020). International applications of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 113–131). CRC Press/Taylor & Francis Group.

Shermis, M. D., & Burstein, J. (2003). Introduction. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. xiii–xvi). Lawrence Erlbaum Associates.

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1–15). Routledge/Taylor & Francis Group.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313–346). Routledge/Taylor & Francis Group.

Slack, M. K., & Draugalis, J. R. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy*, *58*(22), 2173–2181. https://doi.org/10.1093/ajhp/58.22.2173

Sterne, J. A., Hernan, M. A., Reeves, B. C., Savovic, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hrobjartsson, A., Kirkham, J., Juni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savovic, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H. Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernan, M. A., Hopewell, S., Hrobjartsson, A., Junqueira, D. R., Juni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

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

Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing* (pp. 125–142). Cambridge University Press. https://doi.org/10.1017/9781108635547.009

*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

Teo, T. (2011). Factors influencing teachers' intention to use technology: Model development and test. *Computers & Education*, *57*(4), 2432–2440. https://doi.org/10.1016/j.compedu.2011.06.008

Thompson, C. B., & Panacek, E. A. (2006). Research study designs: Experimental and quasi-experimental. *Air Medical Journal*, *25*(6), 242–246. https://doi.org/10.1016/j.amj.2006.09.001

Troia, G. A. (2006). Writing instruction for students with learning disabilities. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 324–336). Guilford.

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

United Nations Educational, Scientific and Cultural Organization, UNESCO. (2011). *UNESCO and education: Everyone has the right to education*. UNESCO.

Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, *11*, 342–365. https://doi.org/10.1287/isre.11.4.342.11872

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

*Wade-Stein, D., & Kintsch, E. (2004). Summary street: Interactive computer support for writing. *Cognition and Instruction*, *22*(3), 333–362. https://doi.org/10.1207/s1532690xci2203_3

Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., Magooda, A., & Quintana, R. (2020). eRevis(ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, *44*, 100449. https://doi.org/10.1016/j.asw.2020.100449

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

*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

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*(1), 22–36. https://doi.org/10.1080/15544800701771580

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

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.

Wijekumar, K., Graham, S., Harris, K. R., Lei, P., Barkel, A., Aitken, A., Ray, A., & Houston, J. (2018). The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material. *Reading and Writing*, *32*, 1431–1457. https://doi.org/10.1007/s11145-018-9836-7

Wijekumar, K., Meyer, B. J. F., & Lei, P. (2017). Web-based text structure strategy instruction improves seventh graders' content area reading comprehension. *Journal of Educational Psychology*, *109*(6), 741–760. https://doi.org/10.1037/edu0000168

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing: An Interdisciplinary Journal*, *30*(4), 691–718. https://doi.org/10.1007/s11145-016-9695-z

*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

How to cite this article: 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

APPENDIX 1

See Tables A1 and A2.

TABLE A1 Characteristics of AWE systems

Study	AWE system	Writing task	Description	AES engine	Method	Features	AWE feedback	Teachers' role in the AWE system
Franzke et al. (2005)[a] Wade-Stein and Kintsch (2004)[a]	Summary Street Developer: Pearson Knowledge Analysis Technologies/University of Colorado http://lsa.colorado.edu/summarystreet/ http://kt.pearsonassessments.com/	Summary writing	Web-based educational software destined for summary writing. Appropriate for students in grades 5–12.	IEA™	LSA	Coverage of main topics, degree of conciseness, relevance of information, and source-text copying	The system provides immediate feedback in a coloured-based graphical interface. It indicates the appropriate summary length, if the main ideas are covered, and what needs more work, until a pre-set threshold is achieved.	N/A
Mørch et al. (2017)	EssayCritic Developer: Intermedia/University of Oslohttps://www.uv.uio.no/iped/english/research/projects/essaycritic/	Argumentative essays on a specific topic	Web-based program for semantic analysis of short texts, based on previously selected topics. This program was developed to provide a simple feedback to essay writing in English as a foreign language (EFL)	N/A	LSA	Content and organization of short texts; identification of presence or absence of subthemes	The system computes a score for the similarity between the essay and previously identified subthemes. If the subtheme was covered, the writer receives the feedback "Praise", if not the writer receives the feedback "Critique".	Teachers help researchers in the creation of the concept tree representing the essay topic, which will be included in the AWE system and used to define the thresholds.
Palermo and Thomson (2018) Wilson and Czik (2016) Wilson and Roscoe (2020)	NC write (Palermo & Thomson, 2018)[b] PEG writing (Wilson & Czik, 2016; Wilson & Roscoe, 2020)[b] Developer: measurement incorporated https://www.measurementinc.com/miwrite	Argumentative essays in response to a prompt (Palermo & Thomson, 2018) Memoir writing in response to a prompt (Wilson & Czik, 2016) Narrative, informative and argumentative essays in response to a prompt (Wilson & Roscoe, 2020)	Web-based instructional writing tool to help students improve their writing through practice and immediate feedback. Appropriate for students in grades 3–12.	PEG™	Statistical	Development of ideas, organization, style, sentence structure, word choice, and conventions (six traits)	The system generates a trait score (1–5 scale) and an overall score (6–30 scale) report. A descriptive evaluation and feedback for each trait is also given.	Teachers can create customizable prompts, and write in-text or summary comments. Additionally, they are also able to generate customizable reports to monitor students' progress.

TABLE A1 (Continued)

Study	AWE system	Writing task	Description	AES engine	Method	Features	AWE feedback	Teachers' role in the AWE system
Tang and Rich (2017)	Writing Roadmap Developer: CTB/McGraw-Hill https://www.mheducation.com/home.html	Argumentative essays in response to a prompt	Web-based instructional writing tool to help students improve their writing through practice and immediate feedback. Appropriate for students in grades 3–12.	Bookette	NLP	Word choice/ grammar usage, sentence structure, mechanics, organization, development (five traits)	The system provides immediate feedback on specific traits. It highlights problematic sections, provides narrative comments, generates discrete (trait-specific) and holistic scores, and remarks and rescores revised versions.	Teachers can create their own original topics.
Ware (2014)	Criterion Developer: Educational Testing Service (ETS) https://www.ets.org	Open-ended response to an informative text (i.e., students reads a text and draws information from the text to answer a focus question)	Web-based instructional writing tool to helps students to plan, write and revise their essays, through practice and immediate feedback.	e-rater®	NLP	Grammar, usage, mechanics, style, organization, development, lexical complexity, prompt-specific vocabulary usage	The system provides analytic and holistic trait scores on several aspects of writing. The overall score is usually on a 4- or 6-point scale.	Teachers can create their own original topics (Scored Instructor Topic feature).

Abbreviations: IEA, intelligent essay assessor; PEG, project essay grade; LSA, latent semantic analysis; NLP, natural language processing; N/A, not available.

[a]Both studies use the Summary Street system.

[b]NC Write and PEG Writing use the same underlying architecture and automated scoring engine (PEG), and differ on the graphical interface.

TABLE A2 Summary of users' perceptions

		Results	
Study	Assessment	Students	Teachers
Franzke et al. (2005)	Students only. - Interviews: students' perceptions and observations regarding summary writing (both conditions), and the use of Summary Street (E-group)	E-group: students reflected more about their writing process and were more aware of what kind of processing are needed to write better summaries, and mentioned at least two important summarization strategies. C-group: students reflected less, mentioning only one summarization strategy.	N/A
Mørch et al. (2017)	Students only - Field notes and video recording of the activity: analysis of the interactions between students during the writing process - No questionnaires or interviews	E-group: students wrote more idea-rich essays, discussed with more confidence about the ideas to include in their essays and used personal stories to anchor them. However, they focused less on organizing their essays for readability and the variety of ideas surfacing in these students' discussions appeared not to be entirely beneficial. C-group: students worried about teachers' expectations regarding their essays. They were uncertain about their knowledge on the assigned topic and if they were progressing. However, they were more confident in organizing the essays and suggested several strategies (e.g., cutting down text).	N/A
Palermo and Thomson (2018)	Students and teachers - Quantitative questionnaire: Usability, effectiveness - Interview: social validity of NC Write and SRSD procedures (if they liked it, would recommend, experienced difficulties, etc.)	Students perceived NC Write as useful to help them improve writing. They liked the STOP and DARE strategies and using self-statements.	Teachers perceived NC Write as appropriate for the students and said that they would recommend it to other teachers. One teacher mentioned that struggling writers needed significant support to interpret the automated feedback.

TABLE A2 (Continued)

		Results	
Study	**Assessment**	**Students**	**Teachers**
Tang and Rich (2017)	Students and teachers - Quantitative questionnaire: Usability, effectiveness, questions about particular aspects of the system - Interviews: explored the above aspects	Students' motivation to rewrite and revise seems to be positively affected by the combination of teacher and automated feedback. This combination allowed students to be more independent in correcting their mistakes and revising their essays.	Teachers could direct their attention to the teaching/learning process, as the AWE system aided them on correcting and commenting on the language mistakes. They recognized the system as a complement to teaching, as the AWE feedback was immediate and could help the students locate the type of problem, and teacher feedback was more concrete, targeted and contextual.
Wade-Stein and Kintsch (2004)	Students and teachers - Interviews: perceptions regarding summary writing and the use of Summary Street	Students complained that summary writing was more difficult with the system than without it, as it was it was often frustrating to get the right content within the prescribed length. However, they recognized that they wrote better summaries when using Summary Street. The simple graphic display of the content feedback was perceived as easy to grasp. As Summary Street points out problems such as missing content and problematic sentences, this gave students a better sense of what they needed to fix and kept them focused much longer. The students felt more comfortable in receiving the evaluation of their essays by the system than by their teachers.	Teachers liked the coloured graphical feedback display, the redundancy and relevance flags, and the immediate and fast feedback.
Ware (2014)	Students only - Quantitative questionnaire: Usability, effectiveness	Students using technology (i.e., E-group and C-group 1) reported higher enjoyment levels, greater desire to stay in their assigned group, the belief to be better writers, and a higher perception of the usefulness of the feedback than students in the face-to-face group (C-group 2). Students in C-group 1 perceived themselves as better writers after the intervention as compared to students in C-group 2. Students' enjoyment of reading or writing and their perceptions about the clarity of feedback provided by each of the delivery modes did not differ between groups.	N/A

(Continues)

TABLE A2 (Continued)

		Results	
Study	Assessment	Students	Teachers
Wilson and Czik (2016)	Teachers only - Quantitative questionnaire: feasibility, utility, and desirability of providing feedback via the two software programs.	N/A	Teachers indicated that PEG Writing was more efficient than Google Docs, as it allowed them to devote more energy to commenting on content, and was easier, more motivating for students to use, and promoted greater student independence.
Wilson and Roscoe (2020)	Teachers only - Quantitative questionnaire: Usability, effectiveness, and desirability	N/A	Teachers perceived PEG Writing very positively in terms of its usability, effectiveness, and desirability.

Abbreviations: AWE, automated writing evaluation; E-group, experimental group; C-group, control group; PEG, project essay grade; SRSD, self-regulated strategy development; N/A, not applicable.