ReCALL (2024), 36: 2, 187–209
doi:10.1017/S0958344023000265

RESEARCH ARTICLE

A systematic review of AI-based automated written feedback research

Huawei Shi
Yantai University, China (shihuawei01@126.com)

Chinese University of Hong Kong, China

Vahid Aryadoust
National Institute of Education, Nanyang Technological University, Singapore (vahid.aryadoust@nie.edu.sg)

Abstract
In recent years, automated written feedback (AWF) has gained popularity in language learning and teaching as a form of artificial intelligence (AI). The present study aimed at providing a comprehensive state-of-the-art review of AWF. Using Scopus as the main database, we identified 83 SSCI-indexed published articles on AWF (1993–2022). We investigated several main domains consisting of research contexts, AWF systems, feedback focus, ways of utilizing AWF, research design, foci of investigation, and results. Our results showed that although AWF was primarily studied in language and writing classes at the tertiary level, with a focus on English as the target language, the scope of AWF research has been steadily broadening to include diverse language environments and ecological settings. This heterogeneity was also demonstrated by the wide range of AWF systems employed ($n = 31$), ways of integrating AWF ($n = 14$), different types of AWF examined ($n = 3$), as well as varied research designs. In addition, three main foci of investigation were delineated: (1) the performance of AWF; (2) perceptions, uses, engagement with AWF, and influencing factors; and (3) the impact of AWF. We identified positive, negative, neutral, and mixed results in all three main foci of investigation. Overall, less positive results were found in validating AWF compared to results favoring the other two areas. Lastly, we grounded our findings within the argument-based validity framework and also examined the potential implications.

Keywords: artificial intelligence (AI); automated written feedback (AWF); systematic review; argument-based validity

1. Introduction

Automated written feedback (AWF), often provided by automated writing evaluation (AWE) systems, is gaining increasing popularity in language learning and teaching (e.g. Burstein, Riordan & McCaffrey, 2020; M. Li, 2021; Ranalli & Hegelheimer, 2022). Unlike automated scoring, which is aimed at marking or scoring writing for summative purposes, AWF is applied in formative assessment in pedagogical contexts (Stevenson & Phakiti, 2019; Weigle, 2013). This feature renders them specifically appropriate and useful for learning and teaching contexts.

AWEs refer to a type of artificial intelligence (AI) that uses natural language processing algorithms to automatically assess and provide feedback on written language productions. Typical AWE systems include Criterion and Pigai, to name but a few. However, the advancement of technology in different areas such as natural language processing has greatly expanded the scope

Cite this article: Shi, H. & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL* 36(2): 187–209. https://doi.org/10.1017/S0958344023000265

of systems that can provide AWF. For example, word processors such as Microsoft Word and Google Docs, intelligent tutoring systems like Writing Pal, and the recently developed online writing assistants such as Grammarly have all become research objects in studies of AWF (e.g. Ranalli & Yamashita, 2022; Strobl *et al*., 2019). In addition, with the emergence of generative AI such as ChatGPT and Microsoft Copilot, the potential for AWF to utilize these advanced large language models (LLM) for more accurate and comprehensive feedback is highly anticipated (e.g. Barrot, 2023).

Meanwhile, a line of empirical research has been done to investigate AWF, centering on several strategic areas including validity, user perception, impact, and factors influencing students' or instructors' use of AWF (M. Li, 2021). The validity of AWF has often been examined using the accuracy rate and coverage of feedback types in previous research (e.g. Bai & Hu, 2017; Ranalli, 2018; Ranalli & Yamashita, 2022). In addition, a stream of research has explored users' perception and use of AWF, as well as factors that influence their perception and/or use of AWF (e.g. Jiang, Yu & Wang, 2020). The main area of investigation of AWF, perhaps, has been the effect of AWF. For example, a line of studies has found AWF helped increase writing score and performance (e.g. Cheng, 2017; Link, Mehrzad & Rahimi, 2022); further, Zhai and Ma (2023) did a meta-analysis of 26 studies and found that AWF had a large positive overall effect on the quality of students' writing. Nevertheless, some studies also found no significant improvements in writing after language learners use AWF (e.g. Saricaoglu, 2019). Despite some AWF research synthesis, such as Fu, Zou, Xie and Cheng (2022) and Nunes, Cordeiro, Limpo and Castro (2022), to our knowledge, no systematic review has been conducted to provide a panoramic description of AWF research integrating all these areas. Thus, the present study sought to fill this gap in knowledge by conducting a systematic review of the current state of AWF research, identifying gaps and limitations, and proposing recommendations for future research directions and pedagogy.

2. Past reviews of AWF

Few synthesis reviews conducted on AWF research have mainly examined the effects of AWF (Fu *et al*., 2022; Nunes *et al*., 2022; Stevenson & Phakiti, 2014; Strobl *et al*., 2019; Zhai & Ma, 2023). Stevenson and Phakiti (2014) examined the effects of AWF in the classroom by reviewing 33 articles, including unpublished papers and book chapters. Their review revealed modest evidence to support the positive effect of AWF on students' writing. Similarly, Nunes *et al.* (2022) reviewed eight articles that adopted quantitative measures to investigate the effectiveness of AWF on K-12 students and found positive evidence. The study done by Zhai and Ma (2023) meta-analyzed 26 studies and, unlike former studies, found large positive overall effects. A more recent review was conducted by Fu *et al.* (2022) on the outcomes of AWF. In their paper, Fu *et al.* examined 48 studies, finding that past researchers focused most of their attention on second language (L2) writing, whereas the first language (L1) writing context was rarely researched. They further reported that AWF can positively impact learners' writing, but it is not as effective as feedback provided by humans (e.g. teachers and peers). In another study, Strobl *et al.* (2019) examined digital tools used for providing AWF. They found 44 tools that supported writing instruction, but most focused on micro-level features of writing, such as grammar, spelling, and word frequencies. These tools would be suitable for providing feedback on argumentative writing in English, but other writing genres, other languages, and micro-level features, such as structures and rhetorical moves, were underrepresented.

It can be seen that the past reviews predominantly centered on the effects or outcomes of AWF, whereas other areas of AWF studies seem neglected in these reviews. Another point to note is that in the past decade, a multifaceted concept of validity consisting of six validity inferences covering multiple aspects related to AWF studies has been proposed and applied (for details, see Chapelle, Cotos & Lee, 2015; Kane, 1992; Shi & Aryadoust, 2023; Weigle, 2013; Williamson, Xi & Breyer, 2012; Xi, 2010). Similarly, in this study, we aim to provide a comprehensive view of existing AWF research by encompassing all relevant areas. By doing so, we hope to offer a more complete understanding of AWF and its various aspects. The research questions of the study are as follows:

1. What are the research contexts in AWF studies, in terms of ecological setting, language proficiency level, language environment, educational level, and target language(s)?
2. What types of AWF provided by AWF systems were investigated, and in what ways was AWF utilized in these studies?
3. How can the published AWF research be characterized in terms of the research design, including genres, participants, time duration, sample size, data source, and research methods?
4. What are the foci of investigation in the studies, and what results are found? (While effects of AWF might be one focus of investigation, we anticipate accuracy issues and many other areas will also be among the foci of investigation.)

3. Method

3.1 Data set

The relevant research on AWF was identified in Scopus. Scopus is the "largest abstract and citation database of peer-reviewed literature" (Schotten *et al.*, 2018; p31). We identified the following search items based on the relevant literature (e.g. Stevenson, 2016; Stevenson & Phakiti, 2014, 2019): *automat* writ* evaluation*, *automat*writ* feedback*, *automat*essay evaluation*, *automat*essay feedback*, *computer*writ* feedback*, *computer*essay evaluation*, *computer *essay feedback*. We also used some common AWF systems such as MY Access! listed in Strobl *et al.* (2019) as search items. The complete search terms and the screenshot (retrieved on January 2, 2023) of our search and the results are presented in Appendix A in the supplementary material. We set the article type as peer-reviewed journal articles and set the end year at 2022. This preliminary search yielded 375 papers. To meet our research purpose, a further screening process was conducted by the two researchers together.

The first criterion was the relevance of the topic in the identified research: by reading the title and the abstract, we excluded those articles irrelevant to AWF ($n = 165$). Further screening by rereading the abstracts and the research questions excluded studies that solely focused on automated writing scoring ($n = 20$). The second criterion was that studies should be empirical in nature (e.g. Zou, Huang & Xie, 2021), and non-empirical studies such as review articles were excluded ($n = 34$).

To ensure the quality of research, the last criterion was that the articles should be published in Science Citation Index (SCI) or Social Science Citation Index (SSCI) journals, as these journals are rigorously peer reviewed to ensure the quality of the research (Duman, Orhon & Gedik, 2015); those non-SSCI or non-SCI articles were excluded ($n = 72$). Therefore, a total of 83 articles (see the supplementary material for the complete list) were finalized for this review. The whole screening process is presented in Figure 1 following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Page *et al.*, 2021).

The journals that published these papers are presented in Appendix B in the supplementary material. Together, these papers were published in 40 journals. Figure 2 also displays the trend of publications over the years: a general rising trend of research on AWF is easily noticed, especially in recent years.

3.2 Coding

To answer our research questions, we developed a coding scheme that contains four major categories consisting of 17 main variables demonstrated in Appendix C of the supplementary material. To objectively report the results and avoid subjective bias, for the majority of variables,

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

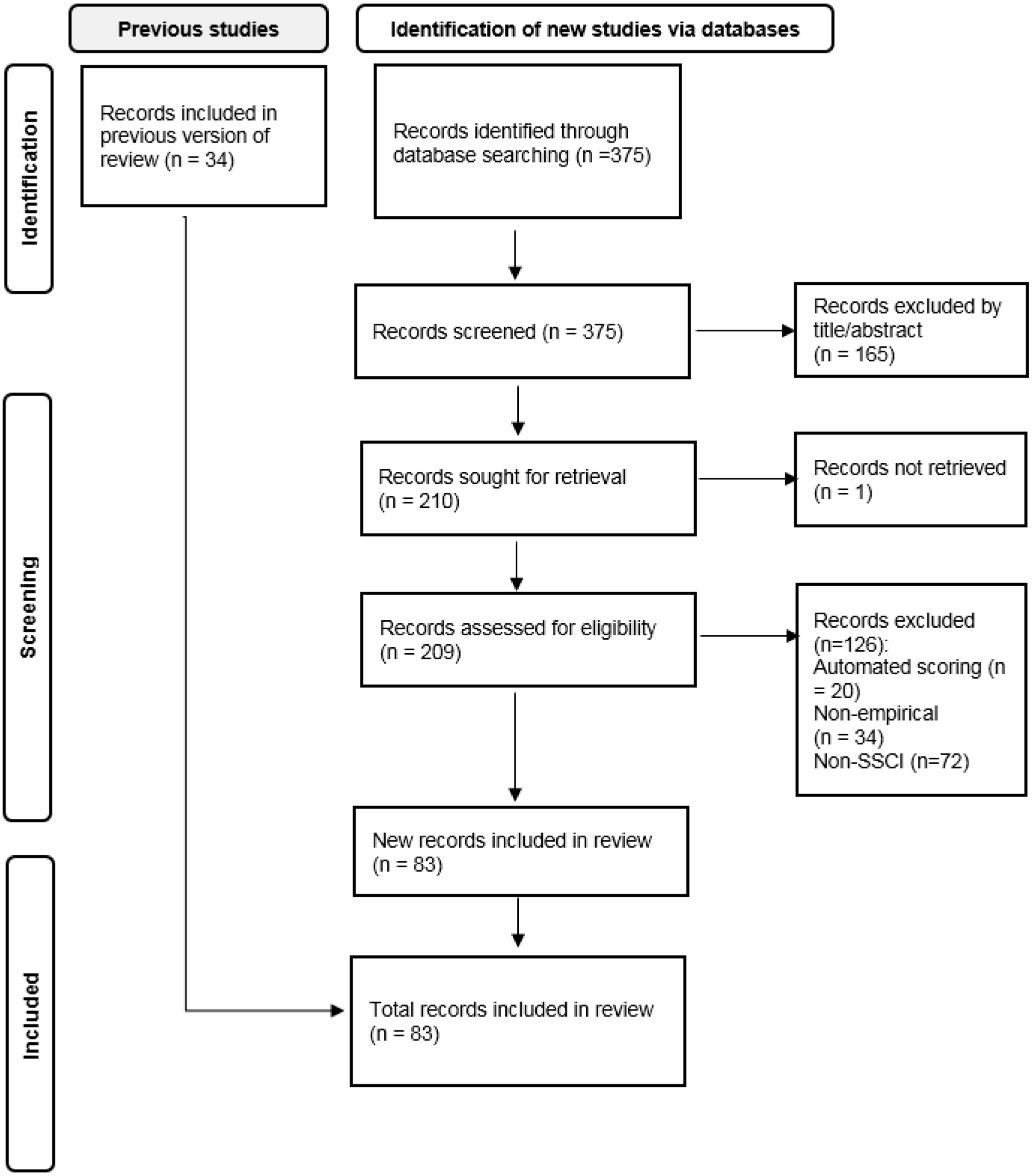
Previous studies
Identification of new studies via databases
Identification
Screening
Included
Records included in previous version of review (n = 34)
Records identified through database searching (n =375)
Records screened (n = 375)
Records excluded by title/abstract (n = 165)
Records sought for retrieval (n = 210)
Records not retrieved (n = 1)
Records assessed for eligibility (n = 209)
Records excluded (n=126): Automated scoring (n = 20) Non-empirical (n = 34) Non-SSCI (n=72)
New records included in review (n = 83)
Total records included in review (n = 83)

Figure 1. PRISMA flow diagram representing the data selection process

we followed the methods of initial coding and focused coding as used in Shen and Chong (2023). These two cycle methods were representative of grounded theory (Charmaz, 2014; Saldaña, 2016). In our review, the first cycle method we used was initial coding; in this cycle, we directly borrowed words or phrases from each paper that corresponded to each variable. For the variables that required a second cycle method, focused coding was used. During this cycle, we searched for the most frequent or significant codes to develop the most salient categories that make the most analytic sense.

The first category is research context, which contains five variables: ecological setting, language proficiency level, language environment, educational level, and target language. The next category is the description of AWF, comprising the name of AWF systems, and the feedback focus, and

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

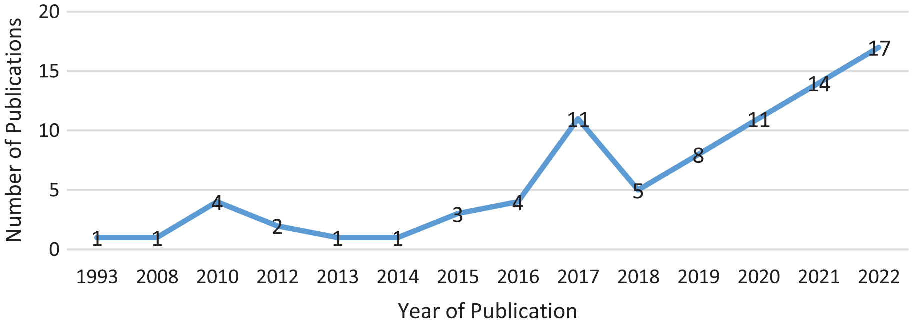

Figure 2. The trend of automated written feedback publications over the years

ways of using AWF. With regard to the feedback focus, we delineated three categories of feedback based on Shintani, Ellis and Suzuki (2014): form-focused feedback, meaning-focused feedback, and comprehensive feedback. Form-focused feedback centers on language, whereas meaning-focused feedback is about content, as well as other aspects related to content. A combination of these two variables was coded as comprehensive feedback. Concerning the ways of using AWF, we developed 13 types. For instance, when AWF was used to compare with human feedback, the study would be coded as "AWF vis-à-vis human feedback," and when AWF and human feedback were combined, we would code it as "AWF + human feedback." Meanwhile, the order of the combination was used to indicate the order of using such feedback.

In the third category, the research design comprised six variables, namely, genres of writing, participants, sample size, time duration, data source, and research methodologies. The last category is made up of three variables: research questions, foci of investigation, and results. For the foci of investigation variable, first we used initial coding to find all research questions and their corresponding results, and then we looked for similarities from the collected data of these two variables and conducted focused coding; through two rounds of focused coding, we identified common categories that surfaced and settled the wording of each subcategory through discussion and with reference to the relevant studies by Chapelle *et al.* (2015), Fu *et al.* (2022), and M. Li (2021). After three rounds of coding, we identified three primary categories that consisted of 20 subcategories.

In line with Shi and Aryadoust (2023), the unit of analysis for coding is one study. In our data set, 81 papers each contained one study, and two papers each contained two studies, so we have 85 studies for coding. To ensure reliability, we first selected 20 varied papers and analyzed them together using initial coding and focused coding; after three rounds of discussion, we developed the initial coding scheme. Based on the coding scheme, the first author coded the rest of the articles twice with an interval of six months. The intra-coder reliability was calculated using Cohen's kappa, the value of which for each variable was above 0.90. All coding was checked by the second author, and the differences were resolved through discussion.

4. Results

4.1 Research contexts

Table 1 presents the research contexts in AWF studies. The first variable concerns ecological settings. Thirty-six studies (42.4%) were conducted in writing classes, and language classes were the second most frequently investigated context. It is noticeable that content classes that focused on subjects other than language or writing were the third most examined context. Regarding language proficiency level, the majority of studies ($n = 57$) did not state the information about the learners' language proficiency level. Among the rest of the studies that reported the language

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

Table 1. Research contexts of automated written feedback studies

Ecological setting	No. of studies	Language proficiency level	No. of studies	Language environment	No. of studies	Educational level	No. of studies	Target language	No. of studies
Writing class	36 (42.4%)	Intermediate	14 (16.5%)	FL	45 (52.9%)	University	70 (82.4%)	English	82 (96.5%)
Language class	22 (25.9%)	Low; high	6 (7.1%)	L1	15 (17.6%)	K-12	15 (17.6%)	German	3 (3.5%)
Content class	9 (10.6%)	Intermediate; high	4 (4.7%)	L2	15 (17.6%)				
Academic writing class	8 (9.4%)	Low; intermediate; high	2 (2.4%)	L1+L2	9 (10.6%)				
Laboratory	2 (2.4%)	High	1 (1.2%)	FL+L2	1 (1.2%)				
Office	1 (1.2%)	Low; intermediate	1 (1.2%)						

Note. Seven studies did not specify their ecological settings and 57 studies did not specify language proficiency level. FL = foreign language; L1 = first language; L2 = second language.

proficiency levels, 14 studies examined users at intermediate level, while the remaining studies examined users of two or more different proficiency levels. Concerning the language environment, over half of the studies ($n = 45$) investigated the foreign language context, followed by the L1 ($n = 15$) and L2 ($n = 15$) context. In addition, nine studies looked into a mix of L1 and L2. In terms of the education level, 82.4% of all studies were conducted at the tertiary level, whereas only 15 studies were carried out in K-12 education. For the last variable, the target language of 82 studies was English, and the rest of the studies ($n = 3$) examined German.

4.2 AWF system, feedback focus, and ways of using AWF

Table 2 presents the systems that provide AWF on writing. In total, 31 different AWF systems were identified. Most of the systems are conventional AI-based AWE systems such as Pigai ($n = 18$, 21.2%), which is the most frequently studied system, followed by Criterion ($n = 16$, 18.8%), and Grammarly ($n = 9$, 10.6%). In addition, 21 systems were studied only once. It is also noted that intelligent tutoring systems such as Writing Pal ($n = 3$) and word processors like Microsoft Word Office ($n = 1$) were also investigated as AWF systems.

The feedback focus on the reviewed studies is presented in Appendix D in the supplementary material. As can be observed, over half of the published research (57.6%) was mainly concerned with form-focused feedback, utilizing 12 distinct AWF systems to provide this feedback. Nineteen studies (22.4%) did not distinguish between different types of AWF, and used it the way it is provided by 10 different AWF systems. The remaining 15 studies explored meaning-focused feedback by 11 AWF systems. It is also noted that some AWF systems, such as Criterion, offered the capability to control feedback focus and were represented in multiple subcategories.

Table 3 shows the application methods of AWF in previous research. Overall, 14 patterns in using AWF were identified. Most research focused solely on AWF ($n = 55$), followed by the comparison of AWF with teacher feedback ($n = 10$) and with peer feedback ($n = 4$). In addition, nine studies also investigated the combination of AWF and human feedback, such as teachers and peers. Integrating AWF with certain pedagogies also appeared in several studies. It is also noted that when integrating AWF with other types of feedback, most studies followed the order of AWF first, except for three studies where teacher feedback was provided to students first.

4.3 Research design

The results of the research design analysis are displayed in Appendix E in the supplementary material. The first variable of interest is writing genres: overall, six different types of genres were investigated, with the genre of essays being the most frequent type ($n = 59$, 69.4%). Nevertheless, 18 studies did not specify the genres of writing, and the rest of the studies examined varied genres, such as research articles ($n = 3$), reports ($n = 1$), etc. With respect to participants, learners were the main subject of the studies ($n = 73$, 85.9%), followed by the studies recruiting both learners and teachers as participants ($n = 7$). Other studies had a more diversified composition of participants ($n = 2$). The next variable is the sample size; we noted that 31 of all studies had a large sample size, over 101 participants, while the number of participants in 12 studies was very small, with less than 10 participants. In terms of duration, nearly 50% of the studies were completed over 10 weeks, while 42.9% of the studies did not indicate their time span. On the other hand, 12.9% of the studies were conducted within two to five weeks, while 11.4% were completed within a time span of six to nine weeks. A mere 5.7% (four studies) were conducted in less than two weeks.

The data source used in the studies is further presented in Figure 3, which represents a variety of data used in the AWF studies. The largest number of studies included participants' writing as

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

Table 2. Automated written feedback systems

Automated writing evaluation systems	# of studies	%
Pigai	18	21.2
Criterion	16	18.8
Grammarly	9	10.6
MI Write	4	4.7
MY Access!	4	4.7
CohViz	3	3.5
Writing Pal	3	3.5
EssayCritic	2	2.4
Glosser	2	2.4
PEG	2	2.4
Ctutor	1	1.2
1Checker	1	1.2
A grammar and spelling-checker program	1	1.2
ACDET	1	1.2
An online automatic classification system	1	1.2
An online science curriculum module	1	1.2
Causal Discourse Analyzer	1	1.2
CorrectEnglish	1	1.2
eRevise	1	1.2
Essay Critiquing System 2.0	1	1.2
Intelligent Academic Discourse Evaluator (IADE)	1	1.2
iWrite	1	1.2
LIWC	1	1.2
Microsoft Word Office	1	1.2
Mosoteach	1	1.2
NC Write	1	1.2
PaperRater	1	1.2
Research Writing Tutor (RWT)	1	1.2
Write & Improve	1	1.2
WRITER	1	1.2
Writing Planet™	1	1.2

Note. One study (Chapelle *et al*., 2015) examined two systems; two studies did not specify the name of the systems.

the only or one of the data sources ($n = 67$), while surveys ($n = 42$) and interviews ($n = 29$) were the next most frequently used data source. Additionally, though not common, 13 other types of data were also used in these studies, such as think-aloud ($n = 5$), observation ($n = 4$), and stimulated recall ($n = 4$).

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

Table 3. Methods of using automated written feedback (AWF)

Methods of using AWF	# of studies	%
AWF	55	64.70
AWF vis-à-vis teacher feedback	10	11.80
AWF vis-à-vis peer feedback	4	4.70
Teacher feedback + AWF	3	3.50
AWF + peer feedback	2	2.40
AWF + teacher and peer feedback	2	2.40
AWF + teacher feedback	2	2.40
AWF + goal setting	1	1.20
AWF + peer feedback vis-à-vis AWF vis-à-vis peer feedback	1	1.20
AWF + reflective thinking promotion mechanism	1	1.20
AWF + Self-Regulated Strategy Development (SRSD) instruction	1	1.20
AWF vis-à-vis online peer feedback	1	1.20
AWF vis-à-vis teacher and peer feedback	1	1.20
AWF + social nudging	1	1.20

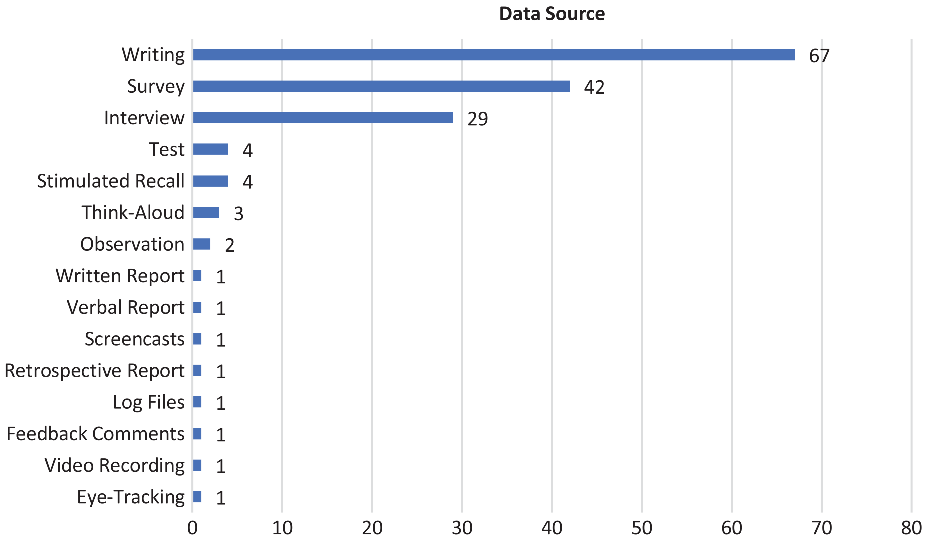

Figure 3. Data source of automated written feedback studies.
Note. Thirty-three studies employed only one type of data, whereas the remaining 52 studies used two or more types of data.

The methodologies used in each study, as presented in Figure 4, were also vast and varied. In total, we identified six categories of research methods adopted in the studies. The largest number of studies ($n = 34$, 40%) adopted mixed methods, followed by quantitative methods ($n = 22$, 26%). The qualitative method was also used in 13 studies, and one study adopted action research. The rest of the studies explicitly stated the experimental design, with 11 studies using quasi-experiment, and four studies using experimental design.

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

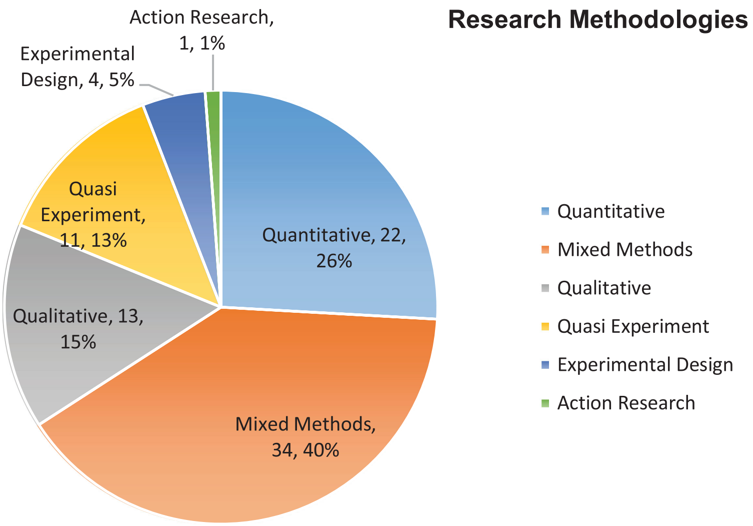

Figure 4. Research methodologies of automated written feedback studies

4.4 Foci of investigation and study results

The foci of investigation and the results of each study are presented in Table 4. As mentioned, three main categories were delineated: the performance of AWF; perceptions, uses, engagement with AWF and their influencing factors; as well as the effects of AWF. Overall, significantly more studies were conducted on perceptions, uses, engagement with AWF and their influencing factors as well as on the effects of AWF than the performance of AWF. Because of space constraints, relevant studies are cited only in the tables.

Regarding the first category, four sub-areas were identified. The first area was domain analysis, which refers to whether the domain of AWF matches the targeted writing construct as indicated in authentic writing tasks (Chapelle *et al.*, 2015). One study (M. Liu, Li, Xu & Liu, 2017) showed positive results, and the other study (Conijn *et al.*, 2022) investigated five groups of stakeholders and found that learners were interested in lower-level behavioral constructs, while teachers, writing researchers, and professional development staff focused on higher-level cognitive and pedagogical constructs such as critical thinking. Accuracy and the issue of agreement of AWF with human raters' feedback were the main foci of investigation in this category. Specifically, five studies provided positive evidence concerning the accuracy of AWF, with the accuracy rate ranging from 0.41 to 0.985. In the meantime, five studies showed negative results in this respect, with the accuracy rate being as low as or less than 50%. Two studies provided mixed results, with the accuracy of some types of errors being satisfactory and the accuracy of other types being low. Three studies that measured agreement between AWF and human raters' feedback discussed the construct represented by AWF; specifically, two studies showed that compared to teacher feedback that covered both language and content issues, AWF covered fewer features of writing, while one study (Zhang & Hyland, 2018) found that both types of feedback had their own strengths.

Perceptions, uses, engagement with AWF and their influencing factors constituted the second focus of the investigation. The concept of engagement was examined in nine studies. Although no studies returned negative results, the study done by Koltovskaia (2020) found that learners have different levels of cognitive engagement when engaging with AWF. Learners' perception of AWF was the main interest of a number of studies. Eleven studies reported positive perception by learners, while only two had negative findings. Interestingly, these two studies found that learners preferred feedback from human raters, whether they were teachers or peers. Two studies produced

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

Table 4. Foci of investigation and results

Foci of investigation	Positive	Negative	Neutral	Mixed
Performance of automated written feedback (AWF)				
Domain analysis	M. Liu *et al*., 2017			Conijn *et al*., 2022
Accuracy	Calma *et al*., 2022; Chukharev-Hudilainen & Saricaoglu, 2016; M. Liu *et al*., 2017; Thi *et al*., 2022; Wang, Harrington & White, 2012	Bai & Hu, 2017; Dikli & Bleyle, 2014; Hoang & Kunnan, 2016		Lavolette *et al*., 2015; Ranalli *et al*., 2017
The agreement with human raters' feedback		Thi & Nikolov, 2022	Zhang & Hyland, 2018	
AWF vis-à-vis the construct of writing	Chapelle *et al*., 2015			Ranalli, 2021
Perceptions, uses, engagement with AWF, and their influencing factors				
Learners' engagement	Lee, 2020; Nazari, Shabbir & Setiawan, 2021; Thi & Nikolov, 2022; Zhang & Hyland, 2022		Ranalli, 2021; Zhang, 2017; Zhang, 2020; Zhang & Hyland, 2018; Zhang & Xu, 2022	Koltovskaia, 2020
Learners' perception	Alnasser, 2022; Dikli & Bleyle, 2014; Fang, 2010; Lu, 2019; O'Neill & Russell, 2019; Palermo & Thomson, 2018; Tan, Cho & Xu, 2022; Wang *et al*., 2020; Wang, Shang & Briody, 2013	Lai, 2010; Sherafati, Largani & Amini, 2020	Bai & Hu, 2017; Calvo & Ellis, 2010	Chen & Cheng, 2008; Cheng, 2017; Dwyer & Sullivan, 1993; J. Li *et al*., 2015; Xu & Zhang, 2022
Learners' use	Chapelle *et al*., 2015; Potter & Wilson, 2021; Ranalli *et al*., 2017; Thi & Nikolov, 2022; Wang, Harrington & White, 2012		El Ebyary & Windeatt, 2019; Jiang & Yu, 2020	Bai & Hu, 2017; Lavolette *et al*., 2015; Link *et al*., 2022; Saricaoglu & Bilki, 2021; Wang *et al*., 2020; Xu & Zhang, 2022
Teachers' perception	J. Li *et al*., 2015; Z. Li, 2021; Lu, 2019; Palermo & Thomson, 2018; Wilson & Roscoe, 2020		Wilson *et al*., 2021	

(*Continued*)

Table 4. (*Continued*)

Foci of investigation	Positive	Negative	Neutral	Mixed
Teachers' use	Z. Li, 2021; Wilson, Myers & Potter, 2022		Engeness, 2018	
Factors influencing use/ perception/engagement with AWF	Zhu *et al.*, 2020		Chen & Cheng, 2008; Jiang *et al.*, 2020; R. Li, 2021; R. Li *et al.*, 2019; Ranalli, 2021; Roscoe, Wilson, Johnson & Mayra, 2017; Sun & Fan, 2022; Wilson, 2017; Zhai & Ma, 2021; Zhang & Hyland, 2018	
Effects of AWF				
Effects on writing performance	Al-Inbari & Al-Wasy, 2022; Barrot, 2021; Cheng, 2017; Fang, 2010; Hassanzadeh & Fotoohnejad, 2021; Lachner, Burkhart & Nückles, 2017a; Lachner, Burkhart & Nückles, 2017b; Lachner & Neuburg, 2019; J. Li *et al.*, 2015; Z. Li, 2021; Link *et al.*, 2022; Liu, Hou, Tu, Wang & Hwang, 2021; M. Liu *et al.*, 2017; Lu, 2019; McCarthy, Roscoe, Allen, Likens & McNamara, 2022; Mørch, Engeness, Cheng, Cheung & Wong, 2017; Nazari, Shabbir & Setiawan, 2021; Palermo & Thomson, 2018; Potter & Wilson, 2021; Saricaoglu & Bilki, 2021; Sun & Fan, 2022; Tan, Cho & Xu, 2022; Thi & Nikolov, 2022; Wambsganss, Janson & Leimeister, 2022; Wang, Shang & Briody, 2013; Z. Wang & Han, 2022; Wilson, 2017; Wilson, Potter, Cordero & Myers, 2022; Wilson & Roscoe, 2020; Zhu *et al.*, 2020	Allen *et al.*, 2019	Kellogg, Whiteford & Quinlan, 2010; Wilson, Myers & Potter, 2022	Bond & Pennebaker, 2012; Kellogg, Whiteford & Quinlan, 2010; Reynolds, Kao & Huang, 2021; Reynolds, Kao & Huang, 2021; Shang, 2019; Xu & Zhang, 2022
Effects on genre	Cotos *et al.*, 2017; Lachner & Neuburg, 2019			Saricaoglu, 2019
Effects on grammatical accuracy	Liao, 2016a, 2016b; Waer, 2021			

(*Continued*)

Table 4. (*Continued*)

Foci of investigation	Positive	Negative	Neutral	Mixed
Effects on coherence and cohesion			Chen & Cui, 2022	
Effects on teacher feedback			Jiang *et al*., 2020	Link *et al*., 2022
Effect on metacognitive strategy		Zhang & Zhang, 2022		
Effects on anxiety	Waer, 2021			
Effects on mindset and motivation	Yao, Wang & Yang, 2021			

Effects on self-efficacy	Liu, Hou, Tu, Wang & Hwang, 2021;
	Nazari, Shabbir & Setiawan, 2021
	Wilson, Potter, Cordero & Myers, 2022;
	Wilson & Roscoe, 2020			

Effects on self-regulated learning	Wilson, Potter, Cordero & Myers, 2022			
Effects on identity		Zaini, 2018		

neutral results, which indicated that learners are cognizant of the strengths and weaknesses of both AWF and human rater feedback, while five other studies yielded mixed results; that is, students held positive attitudes towards some aspects or some types of AWF such as feedback on grammar, but expressed negative perceptions towards other aspects or types of AWF like content-related feedback (e.g. Li, Link & Hegelheimer, 2015). A similar pattern also emerged in the use of AWF by learners, where positive ($n = 5$), neutral ($n = 2$), and mixed ($n = 6$) results were found across different studies. In contrast, in terms of teachers' perception and teachers' use, most studies showed that teachers were positive about AWF and were willing to integrate AWF into teaching. No mixed or negative results concerning teachers' perception and teachers' use were reported.

Factors that might influence perception, uses, and engagement with AWF comprised the last subcategory in this area. A myriad of influencing factors was identified. For example, Zhu, Liu and Lee (2020) found that higher initial scores and contextualized feedback positively impacted such interaction, and vice versa. In addition, the instructor's willingness to apply AWF (e.g. Jiang *et al*., 2020), perceived trust in AWF (e.g. Ranalli, 2021; Zhai & Ma, 2021), perceived ease of use (R. Li, 2021), and self-efficacy (e.g. Li *et al*., 2019) are all found to play a role in users' interaction with AWF.

In the category of the effects of AWF, 11 sub-areas were identified. Among these areas, the effect of AWF on writing performance was examined by the highest number of studies ($n = 38$), with most of them ($n = 30$) proving the positive effects except one study by Allen, Likens and McNamara (2019), in which no impact was found for AWF. In addition to the overall performance, subconstructs of writing such as genres, grammatical accuracy, coherence, and cohesion were examined in 12 studies, seven of which yielded positive results. Moreover, the impact of AWF on other aspects related to writing such as strategy, and psychological states including self-efficacy, and motivation was also inspected in nine studies, especially in the last two years, of which, seven showed positive effects.

5. Discussion

This study examined 85 studies in 83 papers on AWF. It aimed to answer four research questions concerning research context, description of AWF, research design, foci of investigation, and results. It also aimed to provide a bigger picture of past AWF research to propose future AWF-related research.

5.1 Research question 1

The first research question dealt with research contexts. Our results showed that AWF was mainly studied in language and writing classes at the tertiary level, with English being the target language, and that the majority of studies did not report language proficiency level. Nevertheless, the heterogeneity of research contexts was demonstrated, which also echoes Stevenson and Phakiti's (2014) review. For instance, it is noticed that AWF applied to a wider ecological setting such as content classes of subjects other than language. This heterogeneity indicates a widespread interest in AWF, on the one hand, and warrants the need for careful consideration of the writing construct, on the other hand (Aryadoust, 2023), as any variation in delineating the writing construct could significantly impact research results. As a result, even minor changes in the research context could lead to vastly different findings, even if other variables remained consistent (Shi & Aryadoust, 2023).

5.2 Research question 2

The second research question dealt with AWF systems, feedback focus, and the way of using AWF. It was found that a wide range of AWF systems have been applied and that AWF providers

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

are not confined to AWE systems, although they are still the dominant provider of AWF. Other systems capable of providing AWF such as word processors (e.g. Microsoft Word) and online writing assistance (e.g. Grammarly) are also gaining traction in research. All this suggests a broader view and definition of AWF systems.

In terms of feedback focus, we found that the majority of AWF studies focused on form-related feedback, and fewer studies provided meaning-related feedback. This may be partly due to the limited affordances of the AWF systems examined. As the writing construct is rather complex (Vojak, Kline, Cope, McCarthey & Kalantzis, 2011; Weigle, 2013), it appears that most AWF provided by AWE systems is unable to represent the entire construct of writing, and therefore relying solely on AWF in language learning would likely result in an incomplete evaluation of students' writing abilities. This could also lead to underrepresenting writing for language learners who are in need of personalized feedback and guidance in areas beyond surface-level errors, such as content organization, coherence, and style. Furthermore, as will be discussed, the integration of generative AI such as ChatGPT and GPT-4 in AWF research offers exciting opportunities for more in-depth and nuanced feedback that transcends surface error correction and addresses higher-order writing skills.

Regarding the ways of using AWF, 14 patterns were identified, with most studies investigating AWF in isolation, highlighting the predominant interest in how AWF performs without other interventions. This attention also signals queries about whether AWF aligns with or closely mirrors human feedback in writing. Nevertheless, 16 studies examined AWF in comparison with human feedback, which might provide more substantial evidence regarding the utility and benefits of using AWF compared with human-generated feedback. Other studies combined AWF and human raters' feedback or learning strategies. The varied ways of utilizing AWF prompt us to reflect on the role of AWF as a complement to teaching or a replacement. Specifically, it is important to investigate the extent of the efficacy of AWF in language learning and whether AWF alone, particularly that provided by generative AI such as ChatGPT, can provide sufficient feedback for learners to improve their writing skills. These are crucial questions to consider in designing future studies and developing effective approaches to AWF implementation in language learning.

5.3 Research question 3

Regarding the research design, our findings further validate the heterogeneous nature of AWF, as showcased in the first research question. The first variable investigated under this heading is the genres of writing, as one of the important features of writing constructs. Our result showed that a range of genres were examined in the previous AWF research, with essays being the most frequently examined type. Although some AWF systems were designed for specific genres (e.g. Research Writing Tutor [RWT]; see Cotos, Link & Huffman, 2017), most AWF systems did not specify whether they are suitable for all genres or just certain types of genres, which, if not taken into consideration, could impact the results of teaching and research. It is plausible that prior to the advent of LLMs, which are foundational machine learning models utilized to process, understand, and generate natural language, existing AWF systems were genre specific and/or limited to specific writing prompts. However, this postulation requires verification, as there is a dearth of research that examines the historical development of AWF systems. Concerning the participants, it is clear that learners were the main users of the AWF systems, followed by teachers. It is interesting that some studies expanded the scope of the study to participants such as workplace representatives (e.g. Conijn *et al.*, 2022; Zaini, 2018), indicating that AWF is not confined to educational settings. The data source and the methodology together imply that researchers have approached AWF from many possible angles. However, although mixed methods and quantitative methods were the common ways, the number of studies adopting an

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

experimental design or quasi-experimental design was relatively small. According to Marsden and Torgerson (2012), experimental and quasi-experimental designs enable researchers to draw more unambiguous conclusions concerning the potentially causal relationship between the variables of interest, so it is advisable to promote the utilization of these two designs in future AWF research.

5.4 Research question 4

Our fourth research question examined the foci of investigation and results of previous AWE research. Unlike the previous reviews that focused on learning outcomes or the effects of AWF (e.g. Fu *et al.*, 2022; Nunes *et al.*, 2022), our review provides a comprehensive view of research related to AWF. Three main areas surfaced: the performance of AWF; perceptions, uses, and engagement with AWF and their influencing factors; as well as the effects of AWF. Within each area, several subcategories were identified. Though seemingly varied and loosely related, these areas were interwoven from the perspective of argument-based validity (Chapelle *et al.*, 2015; Kane, 1992). Argument-based validity (ABV) regards issues of validity not as separate, but as a unitary concept that consists of validation of interpretations and uses of assessment instruments and scores (Kane, 1992). Domain analysis is the first inference in this argument-based validation framework (Chapelle *et al.*, 2015) – that is, whether the writing assessment task represents writing tasks in the target language domain. Our results showed that only two studies explicitly investigated the target language domain (Conijn *et al.*, 2022; M. Liu *et al.*, 2017). This might partly be due to our exclusion of non-empirical studies in the review. Nevertheless, the importance of domain analysis lies in the identification of language knowledge, skills, and abilities that are needed for test design (Im, Shin & Cheng, 2019), and in our context, the design of AWF. In light of the importance of domain analysis, it is suggested that further research be conducted to expand domain definition and specification in AWF studies. The results of a robust domain analysis can also be utilized to examine the efficacy of generative AI across all the identified domains and determine where AI can best complement human feedback in the AWF process.

The issue of accuracy pertains to the evaluation and generalization inferences – that is, whether the AWF is an accurate representation of writers' performance on parallel writing tasks. What we found was a mix of results in this respect; that is, for some AWF providers, positive results were found (e.g. Calma, Cotronei-Baird & Chia, 2022; Chukharev-Hudilainen & Saricaoglu, 2016), while for others, negative results attenuated this inference (e.g. Bai & Hu, 2017; Dikli & Bleyle, 2014). Another compromising piece of evidence was that for some AWF providers, such as Criterion, mixed results were produced in different studies (Lavolette, Polio & Kahng, 2015; Ranalli, Link & Chukharev-Hudilainen, 2017). All this suggests that the accuracy of AWF deserves continuous research attention, particularly because in the ABV framework, low-level inferences function as a premise to high-level inferences. To move up to the next inference without validating a former one would be unjustified and can undermine interpretations and uses of the feedback generated by AWF systems.

In addition, there are some data pertaining to the higher-level duo inferences of explanation and extrapolation. The data draw mainly on the comparison of AWF and writing constructs (explanation) and the agreement between human raters' feedback and AWF (extrapolation). That is, based on the tenets of ABV, AWF should resonate with the construct of interest and human raters' feedback (Chapelle *et al.*, 2015). No conclusive evidence was found to support the effectiveness of AWF in comparison to teacher feedback, indicating that AWF and teacher feedback might not be addressing the same aspects of the writing construct. This suggests that AWF may hardly be the equivalent of or replacement for human raters' feedback. Therefore, for users to choose the appropriate AWF providers, developers of AWF systems should endeavor to disclose the nature and content of the training data. This will result in better compatibility between the educational goals of writing programs and AWF systems.

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

The next two categories – perceptions, uses, engagement with AWF and their influencing factors, as well as the effects of AWF – belonged to the utilization inference, which concerns whether AWF systems provide meaningful information and positive consequences. As mentioned, if a lower-level inference is not robust, the strength of upper-level inferences would inevitably suffer. So it is not surprising to see a mix of results both in perception, uses, and engagement with AWF and its effects. The inconclusive findings of the study would encourage researchers to continuously explore how learners and teachers perceive, use, or engage with AWF, as well as investigate the factors that impact the effectiveness of AWF.

Regarding the effect of AWF, most studies reported positive results in both general writing performance and different subcategories such as grammatical accuracy. Our findings resonate with the results of previous meta-analysis research that reported positive effects of AWF on writing (e.g. R. Li, 2023; Mohsen, 2022). Our review also surfaced increasing research interest in the effects of AWF on constructs related to writing such as anxiety, self-efficacy, and motivation in accordance with Hayes's (1996) model of writing (e.g. Wilson & Roscoe, 2020), indicating the effects of AWF are multidimensional. Accordingly, future research should continue to explore these areas and provide a more comprehensive view of the effects of AWF.

6. Conclusion

With the aim of providing a comprehensive synthesis of AWF research to inform future research, we reviewed 83 papers comprising 85 studies and presented our results in four main areas: research contexts; AWF systems, feedback focus, and ways of using AWF; the research design; and foci of investigation and results.

It was found that although AWF was mainly studied in language and writing classes at the tertiary level, with English being the target language, the heterogeneous nature of AWF research was demonstrated by diverse ecological settings, and language environments in which it was studied. This heterogeneity was also demonstrated in the results of the second and third research questions. That is, it was found that a wide range of AWE systems and ways of using AWF have been applied; three types of AWF were investigated, with form-related feedback being the most frequently examined; and varied research designs were also found in terms of genres of writing, participants, sample size, time duration, data source, and research methods. Looking at it from a different perspective, we argue that this indicates the difference and complexity of the construct of writing for different subgroups from heterogeneous contexts. Meanwhile, three main foci of investigation were observable in the data: the performance of AWF; perceptions, uses, and engagement with AWF and their influencing factors; as well as the effects of AWF. By comparison, limited research was done in the area of validating AWF. For the results, we identified positive, negative, neutral, and mixed results in all three main foci of investigation. Specifically, less conclusive results were found in validating AWF compared to those favoring the other two areas.

We grounded our findings within the argument-based validation framework (Chapelle *et al.*, 2015; Kane, 1992), based on which several suggestions can be offered. First, AWF systems may delineate the application scope with reference to the types of users, the language levels these systems are targeted at, and the genres of writing they are designed for. Second, both AWF providers and researchers may continue to examine the accuracy of AWF, as well as the agreement between AWF and human raters' feedback, especially teacher feedback. Third, it is suggested that AWF systems clearly state the interpretations and uses of AWF; for instance, under what circumstance will the provided AWF be viewed as the replacement of or the supplement to human feedback? Finally, although the effect of AWF on writing continues to be a thriving research area, effects on other writing-related factors such as self-efficacy, self-regulated strategies, and factors that influence perception, use, or engagement with AWF also merit future research (Aryadoust, 2016; Hayes, 1996).

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

This review is not without limitations. A limitation is that the articles reviewed were empirical studies indexed in SSCI peer-reviewed journals, which may have restricted the scope of the review. Another limitation is that our review only included papers published till 2022. The advent of transformer-based generative AI systems like ChatGPT promises more personalized feedback. This innovation has garnered great interest, with a fresh body of research emerging this year. Such advancements should be considered in subsequent research endeavors. Nevertheless, unlike past reviews that mainly examined the effects of AWF, the study presented a broader view of AWEs. It is hoped that the study will serve as a useful guide for future AWF research and practice, and the indictments of AWF as identified in the study will be addressed in rigorously designed studies.

Supplementary material. To view supplementary material referred to in this article, please visit https://doi.org/10.1017/S0958344023000265

Acknowledgements. We would like to acknowledge the support from the Teaching Reform Research Project of Yantai University (No. jyxm2022065). We would also like to thank the anonymous reviewers of the journal for their insightful reviews, which helped us strengthen our arguments and improve the manuscript.

Ethical statement and competing interests. All the studies collected for the research can be accessed online. The authors declare no competing interests.

References

*References marked with an asterisk were included in the review.

*Al-Inbari, F. A. Y. & Al-Wasy, B. Q. M. (2022) The impact of automated writing evaluation (AWE) on EFL learners' peer and self-editing. *Education and Information Technologies.* https://doi.org/10.1007/s10639-022-11458-x

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

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

Aryadoust, V. (2016) Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology*, 36(10): 1742–1770. https://doi.org/10.1080/01443410.2014.950946

Aryadoust, V. (2023) The vexing problem of validity and the future of second language assessment. *Language Testing*, 40(1): 8–14. https://doi.org/10.1177/02655322221125204

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

*Barrot, J. S. (2021) Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, 1–24. https://doi.org/10.1080/09588221.2021.1936071

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

*Bond, M. & Pennebaker, J. W. (2012) Automated computer-based feedback in expressive writing. *Computers in Human Behavior*, 28(3): 1014–1018. https://doi.org/10.1016/j.chb.2012.01.003

Burstein, J., Riordan, B. & McCaffrey, D. (2020) Expanding automated writing evaluation. In Yan, D., Rupp, A. A. & Foltz, P. W. (eds). *Handbook of automated scoring: Theory into practice.* Boca Raton: CRC Press, 329–346. https://doi.org/10.1201/9781351264808-18

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

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

*Chapelle, C. A., Cotos, E. & Lee, J. (2015) Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3): 385–405. https://doi.org/10.1177/0265532214565386

Charmaz, K. (2014) *Constructing grounded theory* (2nd ed.). SAGE.

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

*Chen, M. & Cui, Y. (2022) The effects of AWE and peer feedback on cohesion and coherence in continuation writing. *Journal of Second Language Writing*, 57: 100915. https://doi.org/10.1016/j.jslw.2022.100915

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

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

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

*Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van Waes, L. & van Zaanen, M. (2022) How to provide automated feedback on the writing process? A participatory approach to design writing analytics tools. *Computer Assisted Language Learning*, 35(8): 1838–1868. https://doi.org/10.1080/09588221.2020.1839503

*Cotos, E., Link, S. & Huffman, S. (2017) Effects of DDL technology on genre learning. *Language Learning & Technology*, 21(3): 104–130.

*Dikli, S. & Bleyle, S. (2014) Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, 22: 1–17. https://doi.org/10.1016/j.asw.2014.03.006

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

*Dwyer, H. J. & Sullivan, H. J. (1993) Student preferences for teacher and computer composition marking. *Journal of Educational Research*, 86(3): 137–141. https://doi.org/10.1080/00220671.1993.9941152

*El Ebyary, K. & Windeatt, S. (2019) Eye tracking analysis of EAP Students' regions of interest in computer-based feedback on grammar, usage, mechanics, style and organization and development. *System*, 83: 36–49. https://doi.org/10.1016/j.system.2019.03.007

*Engeness, I. (2018) What teachers do: Facilitating the writing process with feedback from EssayCritic and collaborating peers. *Technology, Pedagogy and Education*, 27(3): 297–311. https://doi.org/10.1080/1475939X.2017.1421259

*Fang, Y. (2010) Perceptions of the computer-assisted writing program among EFL college learners. *Educational Technology and Society*, 13(3): 246–256.

Fu, Q.-K., Zou, D., Xie, H. & Cheng, G. (2022) A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2022.2033787

*Hassanzadeh, M., & Fotoohnejad, S. (2021) Implementing an automated feedback program for a foreign language writing course: A learner-centric study: Implementing an AWE tool in a L2 class. *Journal of Computer Assisted Learning*, 37(5): 1494–1507. https://doi.org/10.1111/jcal.12587

Hayes, J. R. (1996) A new framework for understanding cognition and affect in writing. In Levy, C. M. & Ransdell, S. (eds.), *The science of writing: Theories, methods, individual differences, and applications*. Mahwah: Lawrence Erlbaum Associates, 1–27.

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

Im, G.-H., Shin, D. & Cheng, L. (2019) Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia*, 9(1): Article 14. https://doi.org/10.1186/s40468-019-0089-4

*Jiang, L. & Yu, S. (2020) Appropriating automated feedback in L2 writing: experiences of Chinese EFL student writers. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2020.1799824

*Jiang, L., Yu, S. & Wang, C. (2020) Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System*, 93: Article 102302. https://doi.org/10.1016/j.system.2020.102302

Kane, M. T. (1992) An argument-based approach to validity. *Psychological Bulletin*, 112(3): 527–535. https://doi.org/10.1037/0033-2909.112.3.527

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

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

*Lachner, A., Burkhart, C. & Nückles, M. (2017a) Formative computer-based feedback in the university classroom: Specific concept maps scaffold students' writing. *Computers in Human Behavior*, 72: 459–469. https://doi.org/10.1016/j.chb.2017.03.008

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

*Lachner, A. & Neuburg, C. (2019) Learning by writing explanations: computer-based feedback about the explanatory cohesion enhances students' transfer. *Instructional Science*, 47(1): 19–37. https://doi.org/10.1007/s11251-018-9470-4

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

*Lavolette, E., Polio, C. & Kahng, J. (2015) The accuracy of computer-assisted feedback and students' responses to it. *Language, Learning & Technology*, 19(2): 50–68. http://hdl.handle.net/10125/44417

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

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

Li, M. (2021) *Researching and teaching second language writing in the digital age*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-87710-1

*Li, R. (2021) Modeling the continuance intention to use automated writing evaluation among Chinese EFL learners. *SAGE Open*, 11(4): 1–13. https://doi.org/10.1177/21582440211060782

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

*Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C. & Xiao, W. (2019) Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Computer Assisted Language Learning*, 32(7): 784–804. https://doi.org/10.1080/09588221.2018.1540433

*Liao, H. C. (2016a) 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

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

*Link, S., Mehrzad, M. & Rahimi, M. (2022) Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 35(4): 605–634. https://doi.org/10.1080/09588221.2020.1743323

*Liu, C., Hou, J., Tu, Y.-F., Wang, Y. & Hwang, G.-J. (2021) Incorporating a reflective thinking promoting mechanism into artificial intelligence-supported English writing environments. *Interactive Learning Environments*, 1–19. https://doi.org/10.1080/10494820.2021.2012812

*Liu, M., Li, Y., Xu, W. & Liu, L. (2017) Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, 10(4): 502–513. https://doi.org/10.1109/TLT.2016.2612659

*Lu, X. (2019) An Empirical Study on the Artificial Intelligence Writing Evaluation System in China CET. *Big Data*, 7(2): 121–129. https://doi.org/10.1089/big.2018.0151

Marsden, E. & Torgerson, C. J. (2012) Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, 38(5): 583–616. https://doi.org/10.1080/03054985.2012.731208

*McCarthy, K. S., Roscoe, R. D., Allen, L. K., Likens, A. D. & McNamara, D. S. (2022) Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision? *Assessing Writing*, 52: 100608. https://doi.org/10.1016/j.asw.2022.100608

Mohsen, M. A. (2022) Computer-mediated corrective feedback to improve L2 writing skills: A meta-analysis. *Journal of Educational Computing Research*, 60(5): 1253–1276. https://doi.org/10.1177/07356331211064066

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

*Nazari, N., Shabbir, M. S., & Setiawan, R. (2021) Application of Artificial Intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, 7(5): e07014. https://doi.org/10.1016/j.heliyon.2021.e07014

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

*O'Neill, R. & Russell, A. M. T. (2019) Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology*, 35(1): 42–56. https://doi.org/10.14742/ajet.3795

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffman, T. C., Mulrow, C. D., . . . Moher, D. (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, 10: Article 89. https://doi.org/10.1186/s13643-021-01626-4

*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

*Potter, A. & Wilson, J. (2021) Statewide implementation of automated writing evaluation: Analyzing usage and associations with state test performance in grades 4-11. *Educational Technology Research and Development*, 69(3): 1557–1578. https://doi.org/10.1007/s11423-021-10004-9

*Ranalli, J. (2018) Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, 31(7): 653–674. https://doi.org/10.1080/09588221.2018.1428994

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

Ranalli, J. & Hegelheimer, V. (2022) Introduction to the special issue on automated writing evaluation. *Language Learning & Technology*, 26(2): 1–4. https://doi.org/10125/73473

*Ranalli, J., Link, S. & Chukharev-Hudilainen, E. (2017) Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1): 8–25. https://doi.org/10.1080/01443410.2015.1136407

*Ranalli, J. & Yamashita, T. (2022) Automated written corrective feedback: Error-correction performance and timing of delivery. *Language Learning & Technology*, 26(1): 1–25. http://hdl.handle.net/10125/73465

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

*Roscoe, R. D., Wilson, J., Johnson, A. C. & Mayra, C. R. (2017) Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior*, 70: 207–221. https://doi.org/10.1016/j.chb.2016.12.076

Saldaña, J. (2016) *The coding manual for qualitative researchers*. London: Sage.

*Saricaoglu, A. (2019) The impact of automated feedback on L2 learners' written causal explanations. *ReCALL*, 31(2): 189–203. https://doi.org/10.1017/S095834401800006X

*Saricaoglu, A. & Bilki, Z. (2021) Voluntary use of automated writing evaluation by content course students. *ReCALL*, 33(3): 265–277. https://doi.org/10.1017/S0958344021000021

Schotten, M., Aisati, M., Meester, W. J. N., Steigninga, S. & Ross, C. A. (2018) A brief history of Scopus: The world's largest abstract and citation database of scientific literature. In F. J. Cantu-Ortiz (Ed.), *Research analytics: Boosting university productivity and competitiveness through Scientometrics* (pp. 33–57). Taylor & Francis.

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

Shen, R. & Chong, S. W. (2023) Learner engagement with written corrective feedback in ESL and EFL contexts: A qualitative research synthesis using a perception-based framework. *Assessment & Evaluation in Higher Education*, 48(3): 276–290. https://doi.org/10.1080/02602938.2022.2072468

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

Shi, H. & Aryadoust, V. (2023) A systematic review of automated writing evaluation systems. *Education and Information Technologies*, 28(1): 771–795. https://doi.org/10.1007/s10639-022-11200-7

Shintani, N., Ellis, R. & Suzuki, W. (2014) Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning*, 64(1): 103–131. https://doi.org/10.1111/lang.12029

Stevenson, M. (2016) A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, 42: 1–16. https://doi.org/10.1016/j.compcom.2016.05.001

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 Hyland, K. & Hyland, F. (eds.), *Feedback in second language writing: Contexts and issues*. Cambridge: Cambridge University Press, 125–142. https://doi.org/10.1017/9781108635547.009

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

*Sun, B. & Fan, T. (2022) The effects of an AWE-aided assessment approach on business English writing performance and writing anxiety: A contextual consideration. *Studies in Educational Evaluation*, 72: 101123. https://doi.org/10.1016/j.stueduc.2021.101123

*Tan, S., Cho, Y. W. & Xu, W. (2022) Exploring the effects of automated written corrective feedback, computer-mediated peer feedback and their combination mode on EFL learner's writing performance. *Interactive Learning Environments*, 1–11. https://doi.org/10.1080/10494820.2022.2066137

*Thi, N. K. & Nikolov, M. (2022) How teacher and Grammarly feedback complement one another in Myanmar EFL students' writing. *The Asia-Pacific Education Researcher*, 31(6): 767–779. https://doi.org/10.1007/s40299-021-00625-2

*Thi, N. K., Nikolov, M. & Simon, K. (2022) Higher-Proficiency students' engagement with and uptake of teacher and Grammarly feedback in an EFL writing course. *Innovation in Language Learning and Teaching*, 1–16. https://doi.org/10.1080/17501229.2022.2122476

Vojak, C., Kline, S., Cope, B., McCarthey, S. & Kalantzis, M. (2011) New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, 28(2): 97–111. https://doi.org/10.1016/j.compcom.2011.04.004

*Waer, H. (2021) The effect of integrating automated writing evaluation on EFL writing apprehension and grammatical knowledge. *Innovation in Language Learning and Teaching*, 1–25. https://doi.org/10.1080/17501229.2021.1914062

*Wambsganss, T., Janson, A. & Leimeister, J. M. (2022) Enhancing argumentative writing with automated feedback and social comparison nudging. *Computers & Education*, 191: 104644. https://doi.org/10.1016/j.compedu.2022.104644

*Wang, Z. & Han, F. (2022) The effects of teacher feedback and automated feedback on cognitive and psychological aspects of foreign language writing: A mixed-methods research. *Frontiers in Psychology*, 13: 909802. https://doi.org/10.3389/fpsyg.2022.909802

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

*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, Article 100449. https://doi.org/10.1016/j.asw.2020.100449

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

*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

Weigle, S. C. (2013) English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18(1): 85–99. https://doi.org/10.1016/j.asw.2012.10.006

Williamson, D. M., Xi, X. & Breyer, F. J. (2012) A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1): 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

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

*Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G. & MacArthur, C. (2021) Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, 168: 104208. https://doi.org/10.1016/j.compedu.2021.104208

*Wilson, J., Myers, M. C. & Potter, A. (2022) Investigating the promise of automated writing evaluation for supporting formative writing assessment at scale. *Assessment in Education: Principles, Policy & Practice*, 29(2): 183–199. https://doi.org/10.1080/0969594X.2022.2025762

*Wilson, J., Potter, A., Cordero, T. C. & Myers, M. C. (2022) Integrating goal-setting and automated feedback to improve writing outcomes: A pilot study. *Innovation in Language Learning and Teaching*, 1–17. https://doi.org/10.1080/17501229.2022.2077348

*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

Xi, X. (2010) Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3): 291–300. https://doi.org/10.1177/0265532210364643

*Xu, J, & Zhang, S. (2022) Understanding AWE Feedback and English Writing of Learners with Different Proficiency Levels in an EFL Classroom: A Sociocultural Perspective. *The Asia-Pacific Education Researcher*, 31(4): 357–367. https://doi.org/10.1007/s40299-021-00577-7

*Yao, Y., Wang, W. & Yang, X. (2021) Perceptions of the inclusion of Automatic Writing Evaluation in peer assessment on EFL writers' language mindsets and motivation: A short-term longitudinal study. *Assessing Writing*, 50: 100568. https://doi.org/10.1016/j.asw.2021.100568

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

Zhai, N. & Ma, X. (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

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

*Zhang, J. & Zhang, L. J. (2022) The effect of feedback on metacognitive strategy use in EFL writing. *Computer Assisted Language Learning*, 1–26. https://doi.org/10.1080/09588221.2022.2069822

*Zhang, Z. & Xu, L. (2022) Student engagement with automated feedback on academic writing: A study on Uyghur ethnic minority students in China. *Journal of Multilingual and Multicultural Development*, 1–14. https://doi.org/10.1080/01434632.2022.2102175

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

*Zhang, Z. V. (2020) Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, 43: Article 100439. https://doi.org/10.1016/j.asw.2019.100439

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

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

*Zhu, M., Liu, O. L. & Lee, H.-S. (2020) The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, 143: Article 103668. https://doi.org/10.1016/j.compedu.2019.103668

Zou, D., Huang, Y. & Xie, H. (2021) Digital game-based vocabulary learning: Where are we and where are we going? *Computer Assisted Language Learning*, 34(5–6): 751–777. https://doi.org/10.1080/09588221.2019.1640745

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press

About the authors

Huawei Shi is an English language instructor at Yantai University, China, and currently a PhD student at the Chinese University of Hong Kong. His research interests include technology-enhanced language learning, and language assessment. His work has appeared in journals such as *Education and Information Technologies*.

Vahid Aryadoust is an associate professor of language assessment at the National Institute of Education of Nanyang Technological University, Singapore; an honorary associate professor at the Institute of Education (IOE) of UCL, London; and a visiting professor at Xi'an Jiaotong University, China. His areas of interest include the application of generative AI and computational methods in language assessment, meta-analysis, as well as sensor technologies such as eye tracking, brain imaging, and GSR in language assessment. He has published his research in *Applied Linguistics*, *Language Testing*, *System*, *Computer Assisted Language Learning*, *Current Psychology*, *International Journal of Listening, Language Assessment Quarterly*, *Assessing Writing*, *Educational Assessment*, *Educational Psychology*, etc.

Author ORCiD. Huawei Shi, https://orcid.org/0000-0001-7872-3047
Author ORCiD. Vahid Aryadoust, https://orcid.org/0000-0001-6960-2489

https://doi.org/10.1017/S0958344023000265 Published online by Cambridge University Press