Ajabshir and Ebadi
Asian. J. Second. Foreign. Lang. Educ. (2023) 8:26
https://doi.org/10.1186/s40862-023-00201-9

Asian-Pacific Journal of Second and Foreign Language Education

RESEARCH Open Access

The effects of automatic writing evaluation and teacher-focused feedback on CALF measures and overall quality of L2 writing across different genres

Zahra Fakher Ajabshir[1*] and Saman Ebadi[2]

*Correspondence: fakherzahra@yahoo.com

[1] University of Bonab, Bonab, Iran
[2] Razi University, Kermanshah, Iran

Abstract

This study investigates the effects of teacher-focused feedback (TF) and automatic writing evaluation (AWE) on global writing performance as well as syntactic complexity, accuracy, lexical diversity, and fluency (CALF) of English as a foreign language (EFL) learners' narrative and argumentative writings. The participants were randomly assigned to TF and AWE groups. During the treatment, the teacher delivered instruction on the narrative and argumentative genres, followed by the participants' engagement in writing texts and getting feedback either from the teacher or AWE. The results revealed improvements in overall writing performance (formal aspects) as well as CALF measures. While there was no significant difference between the two groups in their overall writing performance, AWE yielded better scores in lexical diversity and syntactic complexity, and the TF group outperformed in fluency. Moreover, an interaction was found between feedback types (TF vs. AWE) and text genres in CALF measures. The narrative writings were characterized by higher lexical diversity, syntactic accuracy, and fluency, and the argumentative genre, yielded higher scores in syntactic complexity. The results suggest that both human and machine assessments were beneficial in improving written products in EFL contexts. Also, engaging students in writing various genres is likely to result in improvement in different CALF aspects.

Keywords: Automatic writing evaluation, Argumentative genre, Narrative genre, Teacher-focused feedback, Writing

Introduction

So far, numerous studies have documented the effectiveness of corrective feedback in the improvement of second language (L2) writing skills (Huisman et al., 2019; Lv et al., 2021; Cheng et al., 2021; Zhang & Zhang, 2022). Written corrective feedback is a vital source of information in English as a foreign language (EFL) contexts. It assists students in incorporating new information into their original texts and revising their texts at form and content levels, bridging the gap between their prior and intended knowledge. Given ample evidence on the efficacy of feedback in improving writing, the concern in

feedback research is no longer whether to provide feedback but how to best provide it (Link et al., 2022).

While traditionally providing feedback to written texts was done by teachers or peers, with increasing technological advancements and the devising of automated writing evaluation (AWE) tools, this responsibility has been delegated to online editing and proofreading platforms. These platforms serve as learning affordances that scaffold teachers by providing immediate feedback on micro-level writing features like grammar and spelling. Thus, teachers and students can allocate more time and attentional resources to macro-level writing skills such as organization and content (Li, 2021). AWE compares student writing with an extensive informational database; extracts linguistic, structural, semantic, and rhetorical features of the writing using statistical modeling and algorithms; and offers both a holistic score on writing quality as well as qualitative feedback on micro and macro aspects of the text (Zhai & Ma, 2021). AWE represents a viable assistant tool in a number of ways. First, it offers individualized feedback (Link et al., 2022) tailored to each student's needs and areas of difficulty. Secondly, it fosters learners' autonomy by increasing their self-learning opportunities where they can individually manipulate the whole writing task (Stevenson, 2016). Thirdly, AWE evaluation is more consistent and objective than human evaluation, which may rely on some construct-irrelevant features like neat handwriting or text (Lewis, 2018).

AWE is still in its infancy and deserves ample research to arrive at conclusive and robust findings on its impact on learning. Moreover, given that there are a variety of AWE tools, their relative efficacy should be examined to choose the appropriate tool. As yet, some studies have explored the impact of AWE on students' writing skills (e.g., Dikli and Bleyle, 2014; Ranalli, 2018; Shang, 2019; Li, 2021; Lv et al., 2021; Ti & Nikolov, 2022). However, no study has documented how the effect of AWE (if any) may vary across different genres. This study thus adds its contribution to writing evaluation research and fills the gap in the literature by investigating the differential impacts of AWE as compared with teacher-focused feedback (TF) on syntactic complexity, accuracy, lexical diversity, and fluency (CALF) measures in students' narrative and argumentative performances. The researchers chose these two genres to examine the students' writing ability in recounting a series of events and experiences and making justifications and reasoning on a given topic. Moreover, some previous studies (e.g., Ahmadi & Parhizgar, 2017; Zabihi et al., 2020) argued that Iranian EFL students faced challenges in the composition of narrative and argumentative drafts.

Literature review

Teacher versus AWE feedback

The efficacy of teachers' corrective feedback on improving students' writing performance has been established by ample evidence. There is a wealth of studies suggesting that, as a daily practice, teacher feedback enables improved performance not only in overall quality of writing (De Smedt et al., 2016; Cheng et al., 2021; Lv et al., 2021; Zhang & Zhang, 2022) but also in different aspects and dimensions, including complexity (Barrot & Gabinete, 2019; Lu & Ai, 2015), accuracy (Barrot & Gabinete, 2019), and fluency (Fathi & Rahimi, 2022). As a pedagogical tool, teacher feedback conveys a heavy informational load to learners, offers commentaries on the form and content, and motivates students

to improve their writing (Pourdana & Asghari, 2021). According to Sybing (2021), TF enables students to revise and reformulate their texts effectively by providing an environment for meaningful dialogic teacher-student interaction.TF is crucially important in EFL contexts with limited access to native speakers and few interactional encounters in the target language occur. While acknowledging the advantages of TF, as argued by Jiang and Yu (2021), not all TF necessarily yields improved writing performance; for feedback to mediate quality writing, there should be changes in the intentionality (be intentional and focused), reciprocity (teacher-student interaction), transcendence (transfer learning from one feedback situation to another), and meaning (meaningful learning experience) aspects of TF.

As an alternative to hand-scored writing assessment, AWE has drawn the attention of EFL teachers and scholars in recent years. The advantages of using AWE include its consistency, convenient rating, instant feedback, and opportunities to produce multiple drafts and revisions (Stevenson & Phakiti, 2014). AWE systems assist teachers in providing increased higher-level feedback and expediting the feedback process, reducing the teacher feedback burden and enabling them to be more selective in the type of feedback they deliver (Wilson & Czik, 2016). There is evidence that not only does the AWE feedback affect the multiple revisions of the same text, but also the beneficial effects transfer to subsequent written products, enhancing the writing quality in subsequent submissions of similar texts (Liao, 2016). Nonetheless, its limitations include an emphasis on the micro features of writing, such as mechanics, failing to interpret meaning, make inferences on communicative intentions, or assess the quality of argumentation, hence enjoying a one-size-fits-all nature (Ranalli, 2018). Despite these shortcomings, AWE serves as a viable tool for formative assessment, especially in crowded classes where the constraints of time and heavy workload do not allow teachers to deliver individualized feedback to all students.

So far, AWE research has mainly addressed how teachers and students use and perceive automated feedback (Wang et al., 2013; Ranalli, 2018; Link et al., 2022; Thi Nikolov, 2022). Nonetheless, how varied effects can yield the use of TF versus AWE for error correction in L2 writings remains uncertain, demanding ample empirical evidence to establish robust conclusions on the efficacy of either assessment mode. To date, few studies adopted a comparative stance on the effectiveness of these evaluation tools (Dikli & Bleyle, 2014; Wilson & Czik, 2016). Dikli and Bleyle (2014) investigated the use of an automatic essay scoring (AES) system (Criterion) in a college ESL writing classroom. Fourteen advanced-level students wrote three essays and received feedback from the instructor and the AES system. Both types of feedback were analyzed quantitatively and qualitatively across grammar (e.g., subject-verb agreement, ill-formed verbs), usage (e.g., incorrect articles, prepositions), mechanics (e.g., spelling, capitalization), and perceived quality by an additional ESL instructor. Data revealed an advantage for TF, suggesting that not only was the amount of TF larger, but also it was high-quality feedback compared to the feedback delivered by the automated tool. Wilson and Czik (2016) assigned participants to two conditions: teacher feedback only and teacher feedback + automated essay evaluation (AEE) feedback from Google Docs. Results revealed that while the average amount of feedback provided by the teacher in both conditions was roughly similar, the students subjected to the combined mode received more feedback on higher-level

writing features, supporting the argument that by assisting the lower-level revisions, AWE frees the teacher's time to concentrate on higher-level writing skills. However, The two groups' writing quality in their final draft showed no difference. A similar observation was made by Thi and Nikolov (2022). In their study, the feedback provided by the teacher and an AWE tool (Grammarly) on several texts written by intermediate-level students was analyzed in terms of scope. Grammarly was found to provide feedback on surface-level errors, while teacher feedback addressed both lower- and higher-level writing concerns, implying an integration of the traditional teacher-directed and automated feedback.

Narrative and argumentative writing genres

Text genres are generally characterized by distinct formal properties and communicative functions within social contexts (Swales, 1990). As two distinct writing genres, narrative and argumentative genres are characterized by different discoursal features and communicative functions (Berman, 2008). Narratives involve recounting real or imagined experiences or events in the form of detailed chronological contributions of scenes, objects, events, people, and actions to the audience (Loschky et al., 2020). Narration mainly aims to maintain the reader's interest in the course of actions by depicting a personal experience or an event. Contrary to the narrative genre, which is agent-oriented, the argumentative genre represents a topic-oriented discourse (Qin & Uccelli, 2016). The writer uses a logical structure to collect evidence, make reasoning to support his arguments, and establish a stance on the topic. Argumentative essays center on a statement involving clear boundaries and interrelated ideas in a coherent manner, where the writer aims to convince the reader about the correctness of the statement (Hyland, 2009).

Apart from the differences between narrative and argumentative genres in terms of organizational macro structures, they also vary in micro features. Research in L2 writing has documented that the two genres render different levels of form-related features. In terms of complexity level in cross-genre writing performance, the research shows varied findings on syntactic complexity and lexical variety. Way et al., (2000) examined the complexity, accuracy, and fluency of L2 French learners in narrative, expository, and descriptive genres. As measured by words per T-unit, the complexity outweighed in expository than narrative and descriptive genres. Also, the accuracy (as measured by the percentage of error-free T-units) got higher scores by narratives. Along similar lines, Lu (2011) examined the syntactic complexity of the narrative and argumentative compositions of EF learners across a number of measures (length of production, subordination, coordination, embedding, etc.). Out of 14 measures used, 13 measures were found to exceed in argumentative essays than narratives. The increased syntactic complexity of argumentatives over narratives was also reported by Qin and Uccelli (2016) and Chung and Ahn (2020).

At the lexical level, compared to written argumentatives, narrative texts produced by L2 writers tend to display more variety (Olinghouse & Wilson, 2013; Yoon & Polio, 2017; Chung & Ahn, 2020). Olinghouse and Wilson (2013) evaluated the written compositions of fifth-grade students in narrative, persuasive, and informative texts. The written productions were assessed in terms of written quality, lexical diversity, and some further measures. Narrative texts included the highest lexical diversity, followed by persuasive

and informative texts. Yoon & Polio, (2017) reported genre effects on the length of production units and phrase-level complexity measures (argumentatives possessing higher levels than narratives), but no significant effects on subordination or coordination. They also found that while the argumentative texts demonstrated greater lexical sophistication (as measured by longer words and lower word frequency), the narratives contained higher lexical diversity (as measured by varied word use). In terms of accuracy, they found no significant effect. However, the improved performance in most of the above measures was not found to be long-lasting. A roughly similar observation was made by Chung and Ahn (2020) who studied Korean learners' essays written on two topics (one argumentative and one narrative) once using different resources and a week later using Google Translate. The essays written by machine translation possessed higher lexical diversity and syntactic complexity, but lesser lexical sophistication. Regarding genre-specific features, they found better performance in terms of lexical diversity and sophistication in narrative essays and higher scores for syntactic complexity in argumentative texts. While the studies reported above found evidence for the higher lexical complexity of narratives over argumentatives, this is not corroborated by Qin and Uccelli (2016) who reported the overall better quality, lexical diversity, and lexico-syntactic features in argumentative texts written by 100 EFL Chinese secondary school learners as compared to narratives. They argued that lexico-syntactic complexity and diversity of organizational markers were predictors of argumentative essay quality.

Considering the research reported earlier, there is a scarcity of L2 writing research addressing the differential impacts of TF and AWE. Moreover, the existing research concentrating on the effects of cross-genre performances on CALF measures in EFL writings shows mixed findings. In light of the literature's scarcity of evidence, this study focuses on three important areas, including feedback (teacher vs. AWE), CALF measures, and writing genres (narrative vs. argumentative). It contributes to the literature by examining the differential impacts of TF and AWE on students' overall writing performance. Moreover, the potential interaction between the three above parameters is examined. The following research question was formulated to fulfill the study's objectives.

RQ1 Is there any significant difference between automated writing evaluation (AWE) and teacher-focused feedback (TF) in the global writing performance of EFL learners?

RQ2 Is there any significant difference between AWE and TF in syntactic complexity, accuracy, lexical diversity, and fluency (CALF) measures in EFL learners' narrative writings?

RQ3 Is there any significant difference between AWE and TF in CALF measures in EFL learners' argumentative writings?

Method

Design

This study featured a quasi-experimental design with non-random purposive sampling. Two intact classes were randomly assigned to the TF and AWE feedback groups. The

independent variable included the types of feedback (AWE vs. TF), and the dependent variables were the global writing performance as well as CALF measures (syntactic complexity, accuracy, lexical diversity, and fluency) in the participants' narrative and argumentative written productions.

Participants

A power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) suggested that a sample size of N = 47 was needed to ascertain medium effects (f = 0.39) in a mixed within-between-subject design. The data were collected from 53 adult EFL students (25 males and 28 females; ages 19–31, M = 23.31, SD = 2.3) from two classes enrolled in the General English course at an Iranian university. As a compulsory course, General English is taken by undergraduate students of all majors in Iran. The course aims to develop the students' four basic skills, amongst them writing. The treatment in this study thus constituted part of their curricular activities. The classes were held three times a week during a 15 week semester. The participants had experienced at least eight years of formal English learning, and their proficiency level at the time of data collection was intermediate, as determined by the Oxford Quick Placement Test (OQPT) scores (M = 34.3; SD = 1.7; $p > 0.05$).

Also, a writing test was administered, which required the participants to write a text on "My favorite sport" for 30 min. The compositions were assessed by the CEFR (Common European Framework of Reference) scale for A1 (lowest) to C2 (highest) using the website writeandimprove.com created by Cambridge English. The scores were then converted to numeric scores that ranged from 1 to 7 and compared (M = 4.7; SD = 1.03; F (2, 92) = 6.34, $p > 0.05$), suggesting their homogeneity in terms of L2 writing skill.

Prior to the treatment, all participants were informed orally about the study, and the procedures were explained. An informed consent form was also circulated, and the data of those who signed the form were included in the study. The two classes were allocated to one condition each, namely TF and AWE.

Instruments

Oxford quick placement test

Prior to treatment, to ensure the homogeneity of the participants, all participants took an Oxford quick placement test (OQPT) (2001) developed by Oxford University Press and Cambridge TESOL. The OQPT included 40 multiple-choice items, 25 items for vocabulary and 15 cloze items. The test took about 45 min to complete. The internal consistency of the test was also acceptable as indicated by a Cronbach's alpha coefficient of 0.79.

AWE software

In this study, a free version of Grammarly (https://app.grammarly.com/) was utilized. The free version of Grammarly provides feedback on spelling, grammar, punctuation, and clarity, as well as writing style, and conventions, including spacing, capitalization, and dialect-specific spelling. Grammarly serves as a user-friendly platform, instantly providing feedback for improvement once a paper is uploaded online. The uploaded paper appears on the left side of the screen with errors underlined in red (i.e., indirect

feedback), while direct feedback appears on the right side. Direct feedback contains the error type (e.g., grammar), possible error correction (e.g., the Internet for Internet) and suggestion (e.g., It appears that an article is missing before the word Internet. Consider adding the article.).

Pretest/post-test

The pre-test and post-test included a narrative and an argumentative written task where the participants were required to write a 300-word paragraph for each genre. The prompts were mainly related to daily routines, college life, education, and technology, as familiar subjects to students. The prompts for the pretest were "your first day at college" and "pros and cons of dorm life versus living at home" for narrative and argumentative genres, respectively. For the post-test, the participants wrote a narrative text on "your experience of online learning" and generated their arguments and counterarguments on "online learning versus face-to-face learning". During the task completion, the participants were not allowed to use external sources. It took about 60 min to complete the pretest and the post-test (30 min for each genre). Both writing prompts were on a similar topic to optimize comparison across genres.

CALF measures

The collected texts were analyzed for the three features of complexity, accuracy, lexical diversity, and fluency. Following Norris and Ortega (2009), Zhang (2018), and Chung and Ahn (2021), lexical diversity was operationalized by type/token ratio, and syntactic complexity was measured by the mean length of clauses/total number of clauses and the mean length of t-units/total t-units. Drawing on some previous studies (e.g., Chandler, 2003; Kim & Tracy-Ventura, 2013; Zhang, 2018), accuracy was operationalized by error-free T-units/total number of T-units. Fluency was also measured by the total number of words (Allaw & McDonough, 2019).

The rating was performed by two trained raters, including one of the researchers (Ph.D. with an average of 17 years of experience) and an EFL instructor who taught English for ten years. The data were coded twice for obtaining a global writing performance score and once again for assessing CALF measures. The rubric proposed by Jacob et al. (1981) was applied to obtain a global writing score, consisting of 50 discrete points for content and organizational aspects of writing and 50 points for formal aspects. Considering the purpose of this study, the raters assessed only the formal aspects of the written texts, yielding the final scores out of 50. For the assessment of both the global writing performance and CALF measures, the raters coded the data independently. The final scores were obtained by averaging the two rates' scores, yielding the acceptable Cohen's kappa inter-rater reliability indexes of 0.81, 0.83, 0.89, 0.79, and 0.81 for global writing performance, syntactic complexity, accuracy, lexical diversity, and fluency.

Data collection procedure

This study took four weeks, two 90 minute sessions per week. After getting homogenized on their general English proficiency level, the participants were randomly assigned to TF and AWE groups. Both groups received focused instruction on

narrative and argumentative genres before being engaged in composing texts. The AWE group was also trained on how to use AWE for text revision.

After administering the pretest in the first session, the teacher delivered the instruction on different aspects of a narrative essay using a sample in session 2. The instruction mainly discussed various elements in narratives, including setting, characters, relations, plot (sequence of events), moves, etc. In sessions 3 and 4, some prompts were provided, and both groups were engaged in composing in-class 300-word narrative texts on "a life lesson you have learned" and "a vacation you never forget", respectively. In session 5, the teacher delivered instruction on the argumentative genre. A typical argumentative essay was discussed. The position of the author on the topic and different steps taken for reasoning and supporting this stance were investigated, and the participants discussed how to generate their arguments and counterarguments on a certain topic. They composed argumentative texts in sessions 6 and 7 on "printed books or e-sources" and "social media pros and cons", respectively.

The writing task in each session was performed within 30 min, with no permission for the participants to consult the dictionary or the Internet. After completing the task in each session, the AWE group submitted their drafts to Grammarly. After logging into Grammarly and submitting their texts, they could receive immediate feedback. As mentioned earlier, the free version of the AWE tool was used, which mainly focuses on local-level revisions, such as grammar, vocabulary, punctuation, and sentence structure. Similarly, in the TF group, the teacher addressed local-level errors with a lesser focus on content and organizational aspects. Specifically, problematic grammar, words, and sentences were underlined, and suggestions for error corrections were provided by the teacher. The students could revise and rewrite their compositions several times until satisfied.

In the last session, the participants took a post-test, which was similar to the pretest in terms of time constraints and no permission to use supportive sources to compose the text. It required the participants to compose a narrative and an argumentative text using different prompts from the pretest. Table 1 shows the treatment procedure.

Table 1 The treatment procedure

Week/session	Treatment	TF group	AWE group
Session1	Pretest	✓	✓
Session 2	Instruction on the narrative genre	✓	✓
Sessions 3 and 4	Writing two narrative texts on "a life lesson you have learned" and "a vacation you never forget"	Received feedback from the teacher on their narrative texts	Submitted their narrative texts to AWE for feedback
Session 4	Instruction on the argumentative genre	✓	✓
Sessions 5 and 6	Writing two argumentative texts on "printed books or e-sources" and "social media pros and cons"	Received feedback from the teacher on their argumentative texts	Submitted their argumentative texts to AWE for feedback
Session 7	Post-test	✓	✓

Table 2 Results of paired samples *t*-test for the pretest and post-test of each group

Group	N	Pretest	Post-test	SD	Skewness	Kurtosis	F	t	Sig.
AWE	25	27.35	33.78	2.31	−0.21	0.4	14.37	5.23	0.000
TF	28	26.93	34.03	1.89	0.61	1.23	17.63	3.72	0.000

Data analysis

This study aimed at exploring whether exposure to AWE and TF differentially affected the global writing performance of EFL learners. Moreover, it examined the effects of these two types of feedback modes on CALF measures in EFL learners' writing across narrative and argumentative genres. A series of statistical analyses were performed using SPSS, version 22, to conduct within-group and between-group comparisons. The statistical tests included a series of paired-sample *t*-tests, analysis of covariance (ANCOVA), analysis of variance (ANOVA)s, and post-hoc pairwise comparisons using the Scheffe test.

To obtain an insight into the effectiveness of each feedback mode, the scores obtained for the global performance of each group in the pretest and post-test were compared. Accordingly, within-group comparisons using paired samples *t*-tests for each group were run to explore whether each of the feedback modes yielded improvement from the pretest to the post-test. Further, a one-way between-group ANCOVA was conducted to find out the difference between the two experimental groups. The pretest scores were regarded as covariates to control for any pre-existing difference between the two groups. Subsequently, a series of ANOVAs were conducted to examine the effects of each feedback type on CALF measures and the interaction between each feedback mode and text genre on the CALF measures. Finally, post-hoc paired comparisons were run to exactly locate the main effect of feedback modes on the CALF scores in two text genres.

Results

The effects of AWE and TF on the global writing performance of EFL learners

A preliminary screening was conducted to ensure no violation of normality, linearity, and homogeneity of variances. As shown in Table 2, the skewness ratios for the scores were within the legitimate range of ± 1.5, suggesting the normality of distributions. A paired-sample *t*-test was conducted to analyze whether there was any significant difference in terms of each group's total scores in the pretest and the post-test. As shown in Table 1, both groups showed an improvement in their writings in the post-test ($M = 33.78$ and 34.03 for the AWE and TF groups, respectively) compared with their pretest scores ($M = 27.35$ and 26.93 for the AWE and TF groups, respectively), with a significant difference at the 0.05 probability level ($p = 0.000$), suggesting that both types of feedback were effective in fostering students' overall writing performance.

A one-way ANCOVA was also run to examine the difference between the two experimental groups in the post-test scores. As Table 3 shows, there is no significant difference in the total scores of the two groups ($p > 0.05$, $F = 7.731$, $\eta 2 = 0.008$), indicating that exposure to both feedback modes resulted in the students' performing equally well in terms of global writing performance.

Table 3 Results of ANCOVA for the post-test of two groups

Source	Type III sum of squares	df	Mean square	F	Sig.	Partial η2
Covariate (pre-test)	276.821	1	276.821	9.531	0.07	0.006
Between-subjects	214.765	1	214.765	7.731	0.12	0.008
Within-subjects	657.096	51	31.42			

Table 4 Pared comparison of AWE and TF across CALF measures

Complexity measures	Group	N	M	SD	F	Sig.	Paired comparisons	Partial η2
Lexical diversity	AWE	25	19.30	1.12	13.24	0.000	AWE > TF	0.81
	TF	28	19.26	1.02				
syntactic complexity: mean length of clauses/ total clauses	AWE	25	1.29	0.95	45.05	0.01	AWE > TF	0.83
	TF	28	1.14	1.23				
Syntactic complexity: mean length of t-units/ total t-units	AWE	25	0.54	1.04	23.87	0.09	AWE = TF	0.79
	TF	28	0.54	1.01				
Accuracy	AWE	25	0.73	1.09	15.43	0.07	AWE = TF	0.71
	TF	28	0.72	0.46				
Fluency	AWE	25	256.5	0.67	7.61	0.00	TF > AWE	0.77
	TF	28	253.5	0.54				

A paired comparison of the two experimental groups across CALF measures (Table 4) showed that, in general, AWE performed better in lexical diversity ($p < 0.05$, F = 13.24, η2 = 0.81) and syntactic complexity, as measured by mean length of clauses/total clauses ($p < 0.05$, F = 45.05.4, η2 = 0.83). The performances of both g groups were equally well in syntactic complexity, measured by mean length of t-units/total t-units ($p > 0.05$, F = 23.87, η2 = 0.79) and grammatical accuracy ($p > 0.05$, F = 15.43, η2 = 0.71). The fluency scores of the TF group outweighed that of the AWE group ($p < 0.05$, F = 7.61, η2 = 0.77).

The effects of AWE and TF on CALF measures in narrative and argumentative writing

While no significant difference was observed between the two groups in their global writing performance, there was an interaction between the type of genres and the CALF measures. Thus, the analysis centered on assessing the CALF measures across the narrative and argumentative writings of the two groups.

Table 5 shows the results of a series of ANOVAs run to compare the two groups' narrative and argumentative writings on lexical diversity and syntactic complexity measures. In terms of lexical diversity, the trend was TF-NA = AWE-NA > AWE-AG > TF-AG (t = 16.79, $p < 0.05$, Partial η2 = 0.85), suggesting that the narrative texts produced by either group were associated with higher diversity of lexical items.

Concerning syntactic complexity, as measured by the mean length of clauses divided by total clauses and the mean length of t-units/total t-units, it was found that these measures were significantly affected by genre (F = 56.4, $p < 0.05$) with a large effect size (Partial η2 = 0.89 and 0.91). The post hoc paired comparisons demonstrated the pattern

Table 5 Group comparison on syntactic complexity and lexical diversity

Complexity measures	Group	N	M	SD	F	Sig.	Paired comparisons	Partial η2
Lexical diversity: Type/token ratio	AWE*-NA	25	19.34	1.03	16.79	0.000	TF-NA = AWE-NA > AWE-AG > TF-AG	0.85
	AWE-AG		19.27	1.14				
	TF-NA	28	19.29	0.97				
	TF-AG		19.24	1.21				
syntactic complexity: Mean length of clauses/total clauses	AWE-NA	25	1.23	1.08	56.4	0.019	TF-AG = AWE-AG > AWE-NA > TF-NA	0.89
	AWE-AG		1.35	0.67				
	TF-NA	28	1.21	0.78				
	TF-AG		1.37	0.77				
Syntactic complexity: Mean length of t-units/total t-units	AWE-NA	25	0.51	1.04	32.98	0.000	AWE-AG > TF-AG > TF-NA > AWE-NA	0.91
	AWE-AG		0.58	1.21				
	TF-NA	28	0.53	1.08				
	TF-AG		0.56	0.86				

AWE Argumentative; *TF* Teacher feedback; *AG* Argumentative; *NA* Narrative

Table 6 Group comparison on grammatical accuracy

Accuracy measures	Group	N	M	SD	F	Sig.	Paired comparisons	Partial η2
Error-free t-units/total number of t-units	AWE-NA	25	0.77	0.16	32.06	0.000	AWE-NA = TF-NA > TF-AG > AWE-AG	0.83
	AWE-AG		0.70	0.15				
	TF-NA	28	0.74	0.12				
	TF-AG		0.70	0.12				

Table 7 Group comparison on fluency

Fluency measures	Group	N	M	SD	F	Sig.	Paired comparisons	Partial η2
Total number of words	AWE-NA	25	261	1.21	1.24	0.00	AWE-NA > TF-NA > AWE-AG > TF-AG	0.87
	AWE-AG		252	1.03				
	TF-NA	28	258	1.20				
	TF-AG		249	1.09				

TF-AG = AWE-AG > AWE-NA > TF-NA, and AWE-AG > TF-AG > TF-NA > AWE-NA, suggesting the higher syntactic complexity in argumentative texts as compared to narratives under both feedback conditions.

As shown by Table 6, the grammatical accuracy measure was also affected by genre (AWE-NA = TF-NA > TF-AG > AWE-AG). Narrative texts in both TF and AWE groups outweighed in terms of accuracy, followed by argumentative texts in the TF group and argumentative texts in the AWE group ($F = 32.06$; $p = 0.000$). A partial et squared value of 0.83 demonstrates a quite large effect size (Cohen's criterion = 0.14). It can be concluded that narrative genres written by EFL writers under AWE and TF conditions did not show more or fewer errors. Still, argumentative genres that received the instructor's feedback were more grammatically accurate than the argumentative essays subjected to automatic machine feedback.

Table 7 shows the ANOVA results for writing fluency as measured by the total number of words. It was found that the narratives produced by the AWE group were the most

fluent, followed by narratives by the TF group, argumentative essays by the AWE group, and argumentative texts by the TF group ($F = 1.24$; $p < 0.05$) with a large effect size (Partial $\eta 2 = 0.87$). The pattern was thus AWE-NA > TF-NA > AWE-AG > TF-AG, indicating the overall better fluency of narrative texts as compared with argumentative essays.

Discussion

This study aimed to investigate how the use of TF and AWE modes could affect the students' global writing performance and CALF measures in an EFL environment. Overall, both types of feedback were found to positively affect the students' global writing performance and CALF measures in L2 writing. After the employment of both feedback modes, the students' writings demonstrated a significant improvement in terms of overall writing performance as well as CALF measures as compared with their compositions prior to using the feedback. AWE's contribution to improving writing was reported in some previous studies, too (e.g., Ranalli, 2018; Shang, 2019; Link et al., 2022). According to Wang et al., (2013), the detailed and diagnostic feedback provided by AWE enables improving the quality of L2 writing across several drafts of the same text. It promotes noticing, provides direct metalinguistic explanations, gives students an awareness of their lapses, and results in self-directed learning (Barrot & Gabinete, 2019). This study also found evidence for the efficacy of teacher-focused feedback in developing a high-quality written product, which is consistent with the findings of some previous studies (Cheng, Zhang, & Yan, 2021; Zhang and Zhang, 2022). Teacher feedback enables real communication by monitoring errors, delivering feedback, and discussing misconceptions, yielding more effective revisions (Shang, 2019). Such real-time mutual interactional discourses do not normally occur on an AWE platform.

Regarding CALF measures, the written products of the AWE group were characterized by higher lexical diversity and syntactic complexity (as measured by the mean length of clauses/total clauses). On the other hand, the TF group's writings yielded better fluency scores at the expense of lexical diversity and syntactic complexity. It seems likely that the teacher offered a wide range of common and frequently-used lexical and syntactic structures with lesser sophistication. On the other hand, AWE supported the use of more complex language as it allowed the students to access samples of writings, lexical suggestions, and web-based dictionaries, hence more syntactic and lexical variation. Both feedback modes yielded similar levels of accuracy, suggesting that the detailed and diagnostic feedback offered to both groups enhanced students' metalinguistic awareness of grammatical structures and lexical usage and enabled the production of more accurate writings. This finding is inconsistent with that of Dikli (2013) and Dikli and Bleyle (2014), who found that, unlike the instructor, the automated platform did not identify a large number of errors and failed to contribute to students' improvement in their linguistic accuracy.

The analysis of narrative and argumentative writings showed that there was an interaction between the use of feedback types and text genres in CALF measures. Overall, the argumentative genre yielded higher scores in syntactic complexity. On the other hand, the narrative genre resulted in significant improvement in lexical diversity, accuracy, and fluency at the expense of syntactic complexity. The cognition hypothesis may explain the higher syntactic complexity of the argumentative texts (Robinson, 2001; 2003), which

states that more complex tasks push learners toward greater complexity of language production to meet the task demands placed on learners. The development of an argumentative draft, which demands justification, reasoning, and causal relations, is normally more complex than composing a narrative draft, which requires the simple conveying of information. According to Biber and Conrad (2009), the communicative and functional demands placed by different tasks on learners vary, resulting in the use of different lexical items. The finding of this study in terms of better scores of argumentative essays in syntactic complexity measures is congruent with the findings of Lu (2011). In terms of lexical diversity, narrative drafts included a higher lexical diversity than argumentative texts, where the students tended to rely on a few formulaic expressions common in the argumentative discourse. The greater lexical diversity in the narrative genre is also reported in some studies (Olinghouse & Wilson, 2013; Yoon & Polio, 2017; Chung & Ahn, 2020).

Regarding accuracy, the finding that the narrative genre yielded more accurate drafts is consistent with the findings of Way et al. (2000) but contradicts those of Yoon and Polio's (2017) study, which found no genre effect on accuracy. The narrative genre normally occurs in daily routines and represents a rather familiar genre as compared to argumentative discourse. According to Skehan (2009), tasks that are familiar and possess a clear structure, such as presenting personal information or recounting events, yield more accuracy in oral and written performances than tasks involving factual information. Previously acquired schemas are instantaneously activated by familiar topics, and the relevant schemas free up cognitive resources for other functions by reducing the need to process new information (Sweller, 1994).

With respect to fluency, the total number of words in narratives outweighed those of the argumentative texts. This is not surprising as the narrative genre is generally a less demanding genre that involves a simple and frequent lexicon as compared to the argumentative genre, which is characterized by the use of a less frequent complex lexicon (Yoon & Polio, 2017). It is argued that students require less planning time in performing a narrative text than other genres (De Smedt, Van Keer & Merchie, 2016), producing more fluent language. This is partially supported by Schleppegrell (2004) who argued that the development in mastering genres in one's native language progresses from personal genres (e.g., narratives) to analytic genres (e.g., argumentative). Similarly, Ruth et al. (2007) asserted that the development of writing skills does not follow the same path. In monolingual students, narrative structures are mastered by age ten, while argumentative structures are acquired far later.

Conclusion and pedagogical implications

The findings of this study provide support for the contribution of teacher assessment and automated evaluation platforms in the development of L2 writing. The effectiveness of each type of feedback can be determined with reference to the type of writing task, the course's purpose, and the students' proficiency level. Like any other technological tool, AWE is fallible, and decisions on the selection and use of certain AWE tools should be made with caution, continuously evaluating these tools' performance across various EFL contexts. Various studies encouraged the use of AWE as supplementary to teacher

feedback (Jiang et al., 2020; Link et al., 2022). AWE can be used in numerous ways, including employing it as a text editor, a scaffold for teachers, and an interface promoting collaborative written tasks (Stevenson, 2016).

The limitations of the study should be acknowledged. Due to feasibility concerns, both feedback types in this study addressed lower-order local-level errors rather than global and content-oriented ones. Focusing on errors at content and organizational levels, using other AWE platforms, and employing various genres may yield different outcomes in terms of writing quality, which remains an area of research for future studies. A further limitation of this study relates to the use of limited CALF measures. Future studies are recommended to consider a variety of CALF measures and add more depth and breadth to their investigation.

Author contributions
ZFA wrote the main manuscript text. SE reviewed the manuscript.

Funding
Not applicable.

Availability of data and materials
The data that support the findings of this study are available upon request.

Code availability
Not applicable.

Declarations

Ethics approval and consent to participate
Written informed consent was obtained from all participants before the data collection.

Competing interests
The author declares that there is no conflict of interest regarding the publication of this article.

Received: 25 December 2022 Accepted: 15 June 2023
Published online: 08 September 2023

References

Ahmadi, A., & Parhizgar, S. (2017). Coherence errors in iranian EFL learners' writing: A rhetorical structure theory approach. *Journal of Language Horizons, 1*(1), 9–37. https://doi.org/10.22051/lghor.2017.8588.1011

Allaw, E., & McDonough, K. (2019). The effect of task sequencing on second language written lexical complexity, accuracy, and fluency. *System*. https://doi.org/10.1016/j.system.2019.06.008

Barrot, J., & Gabinete, M. (2019). Complexity, accuracy, and fluency in the argumentative writing of ESL and EFL learners. *International Review of Applied Linguistics in Language Teaching, 59*(2), 209–232. https://doi.org/10.1515/iral-2017-0012

Berman, R. A. (2008). The psycholinguistics of developing text construction. *Journal of Child Language, 35*, 735–771. https://doi.org/10.1017/S0305000908008787

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9

Cheng, X., Zhang, L. J., & Yan, Q. (2021). Exploring teacher written feedback in EFL writing classrooms: Beliefs and practices in interaction. *Language Teaching Research*. https://doi.org/10.1177/13621688211057665

Chung, E. S., & Ahn, S. (2021). The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2020.1871029

De Smedt, F., Van Keer, H., & Merchie, E. (2016). Student, teacher and class-level correlates of Flemish late elementary school children's writing performance. *Reading and Writing, 29*, 833–868. https://doi.org/10.1007/s11145-015-9590-z

Dikli, S. (2013). The nature of automated essay scoring feedback. *CALICO Journal, 28*(1), 99–134. https://doi.org/10.11139/cj.28.1.99-134

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

Fathi, J., & Rahimi, M. (2022). Examining the impact of flipped classroom on writing complexity, accuracy, and fluency: A case of EFL students. *Computer Assisted Language Learning, 35*(7), 1668–1706. https://doi.org/10.1080/09588221.2020.1825097

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavior, and biomedical sciences. *Behavior Research Methods Instruments & Computers, 39*, 175–191. https://doi.org/10.3758/BF03193146
Huisman, B., Saab, N., Broek, P., & Driel, J. V. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education, 44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896
Hyland, K. (2009). *Teaching and researching writing*. New York: Routledge.
Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
Jiang, L., & Yu, S. (2021). Understanding changes in EFL teachers' feedback practice during COVID-19: Implications for teacher feedback literacy at a time of crisis. *Asia-Pacific Education Researcher, 30*, 509–518. https://doi.org/10.1007/s40299-021-00583-9
Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System, 93*, 102302. https://doi.org/10.1016/j.system.2020.102302
Kim, Y., & Tracy-Ventura, N. (2013). The role of task repetition in L2 performance development: What needs to be repeated during task-based interaction? *System, 41*, 829–840. https://doi.org/10.1016/j.system.2013.08.005
Lewis, S. B. (2018). Human versus automated essay scoring: A critical review. *Arab World English Journal (AWEJ)*, *9*(2), https://doi.org/10.2139/ssrn.3201916. Available at SSRN: https://ssrn.com/abstract=3201916 or.
Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System, 99*, 102505. https://doi.org/10.1016/j.system.2021.102505
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
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*(3), 1–30. https://doi.org/10.1080/09588221.2020.1743323
Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science, 12*(1), 311–351. https://doi.org/10.1111/tops.12455
Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*, 36–62. https://doi.org/10.5054/tq.2011.240859
Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003
Lv, X., Ren, W., & Xie, Y. (2021). The effects of online feedback on ESL/EFL writing: A meta-analysis. *Asia-Pacific Education Researcher, 30*, 643–653. https://doi.org/10.1007/s40299-021-00594-6
Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics, 30*, 555–578. https://doi.org/10.1093/applin/amp044
Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing: An Interdisciplinary Journal, 26*, 45–65. https://doi.org/10.1007/s11145-012-9392-5
Pourdana, N., & Asghari, S. (2021). Different dimensions of teacher and peer assessment of EFL learners' writing: Descriptive and narrative genres in focus. *Language Testing in Asia, 11*(6), 1–22. https://doi.org/10.1186/s40468-021-00122-9
Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of chinese EFL learners' writing performance. *Journal of Second Language Writing, 33*, 3–17. https://doi.org/10.1016/j.jslw.2016.06.001
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
Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 285–316). Cambridge: Cambridge University Press.
Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language learning. *Second Language Studies, 21*(2), 45–105.
Ruth, A., Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes, 43*, 79–120. https://doi.org/10.1080/01638530709336894
Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. New York: Routledge.
Shang, H. F. (2022). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments, 30*(1), 4–16. https://doi.org/10.1080/10494820.2019.1629601
Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532. https://doi.org/10.1093/applin/amp047
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
Swales, J. (1990). *Genre analysis: English for academic and research settings*. Cambridge: Cambridge University Press.
Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295–312. https://doi.org/10.1016/0959-4752(94)90003-5
Sybing, R. (2021). Examining dialogic opportunities in teacher-student interaction: An ethnographic observation of the language classroom. *Learning Culture and Social Interaction, 28*, 100492. https://doi.org/10.1016/j.lcsi.2021.100492
Thi, N. K., & Nikolov, M. (2022). How teacher and grammarly feedback complement one another in Myanmar EFL students' writing. *Asia-Pacific Education Researcher, 31*, 767–779. https://doi.org/10.1007/s40299-021-00625-2
Wang, Y., Shang, H., & 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

Way, P. D., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal, 84*, 171–184. https://doi.org/10.1111/0026-7902.00060

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

Yoon, H., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly, 51*(2), 275–301. https://doi.org/10.1002/tesq.296

Zabihi, R., Mousavi, S. H., & Salehian, A. (2020). The differential role of domain-specific anxiety in learners' narrative and argumentative L2 written task performances. *Current Psychology, 39*, 1438–1444. https://doi.org/10.1007/s12144-018-9850-6

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

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

Publisher's Note