Computer Assisted Language Learning Electronic Journal (CALL-EJ), 25(3), 124-147, 2024

Comparing the Effects of Teacher Feedback, Automated Feedback, and Integrative Feedback on EFL Learners' Writing Accuracy and Writing Apprehension

Golnoush Haddadian, ghaddadian1@gsu.edu
Georgia State University, United States

Abstract

This study compared the effects of teacher feedback (TF) characterised as feedback provided by the teacher, automated feedback (AF) referred to feedback given by an Automated Essay Scoring system, and integrative feedback (IF) defined as feedback administered by both the teacher and an Automated Essay Scoring system on writing accuracy and writing apprehension (WA). The participants were 187 Iranian upper-intermediate EFL learners within the age range of 18 to 42 from both genders studying at nine language schools across Iran. Initially, these learners were divided into three experimental groups. Once the groups were in place, they were given a Test of English as a Foreign Language (TOEFL) independent sample task and a writing apprehension measure (WAM) as pretests. Then, one of the groups received TF (N = 64), another group was exposed to AF (N = 62), and the third group was provided with IF (N = 61). After the treatments, the three groups received a writing post-test and the WAM. The results of ANCOVA indicated significant differences among the effects of TF, AF, and IF on writing accuracy (effect size = .58) with the IF group outperforming the other two groups. However, there was no significant difference between the TF and AF groups. The results of Kruskal-Wallis showed that there were significant differences among the effects of TF, AF, and IF on WA (effect size = .83) with the IF group exhibiting more reduction in WA compared to the other groups. Additionally, there was a significant difference between the TF and AF groups, with the TF group displaying more reduction in WA compared to the AF group. Based on the results, it is suggested that EFL teachers employ both TF and AF in combination to enhance EFL learners' writing accuracy and reduce writing apprehension.

Keywords: teacher feedback, automated feedback, integrative feedback, witing accuracy, writing apprehension

Introduction

Writing is a pivotal language skill as a substantial amount of communication takes place in written mode to convey messages in various job-related, academic, and everyday life contexts (Kalman et al., 2023; Mirzaii & Aliabadi, 2013; Tsingos-Lucas et al., 2017). However, writing is a multifaceted (Shen & Bai, 2022), complex (Lei et al., 2023), and interactive language skill (Li & Zhang, 2023) which demands learners' cognitive (Wang & Han, 2022) and metacognitive attention (Sun, & Zhang, 2022). To master this skill, learners should pay attention to its various features such as correct spelling, punctuation, lexicon, grammar, content, organization, and relevance (Campbell & Batista, 2023; Xu et al., 2023). One of the vital components of writing is grammar. Therefore, learners are required to produce grammatically error-free writing (Baleghizadeh & Gordani, 2012; Shintani & Aubrey, 2016). As Foster and Skehan (1996) contend, accuracy is the extent to which language products are free from errors. Teachers use automated, conventional, or a combination of conventional and AF types to help learners improve their writing (i.e., integrative feedback) (Baleghizadeh & Gordani, 2012; Han & Sari, 2022; Stevenson, 2016).

The advancement of technology has given a paramount position to scoring systems in the realm of EFL writing (Haddadian & Haddadian, 2024; McCarthy et al., 2022). More specifically, the advent of technology has given rise to the emergence of Automated Essay Scoring (AES) systems to aid teachers in their pedagogical practices to conveniently provide timely corrective feedback (CF) to enhance learners' writing (Zhang & Hyland, 2018). AES systems draw on advanced technological affordances and language processing models to deliver feedback on various aspects of writing (Wang & Han, 2022; Zhang & Hyland, 2018). The results of previous investigations (e.g., Link et al., 2022; McCarthy et al., 2022; Ngo et al. 2022; Tian & Zhou, 2020; Wang & Han, 2022) have indicated the effectiveness of AES tools on writing improvement. However, such tools have some limitations which put constraints on their overall positive influence (Bai & Hu, 2017). Accordingly, teachers and researchers have adopted ways to integrate AES feedback with conventional feedback to address such restrictions (Stevenson, 2016). Apart from the limitations of these systems, such as inaccuracies in terms of analytic scoring categories, the results of available studies (e.g., Zhang & Hyland, 2018) have demonstrated that learners perceive and value teachers' feedback from a humanistic perspective as compared with feedback provided by computers. This humanistic perspective enmeshed with learners' value system can bear relevance to the affective and psychological dimensions of writing such as writing apprehension (WA) (Gaytan et al., 2022; Kelly et al., 2022; Perkins, 2022; Sun & Fan, 2022).

WA is defined as the learners' avoidance of the writing courses and writing process (Kelly & Gaytan, 2020). WA can adversely affect students' written products and their involvement in writing (Badrasawi et al., 2016). A review of extant empirical research shows that WA has so far been examined in relation to online classrooms and instructor behaviours (Gaytan et al., 2022), teacher efficacy and writing to learn (Perkins, 2022), instructor misbehaviours (Kelly, 2022), peer assessment (Rauf & Khan, 2022), WebQuest writing instruction program (Chuo, 2007), and Google drive versus face-to-face instruction (Marandi & Seyyedrezaie, 2017). As the results of previous studies reveal, there are associations between teachers' factors (e.g., Perkins, 2022) and behaviours and students' WA (e.g., Gaytan et al., 2022; Kelly, 2022). However, there is a dearth of investigations unravelling the effect of integrating teacher and computer-generated feedback on WA. Moreover, the previous comparative investigations (e.g., Dikli & Bleyle, 2014; Li et al., 2014; Shermis et al., 2004) have yielded conflicting results in terms of the effects of IF on writing performance. Thus, the present study was motivated in three ways. Firstly, the results of previous investigations in relation to comparing the effects of TF, AF, and IF are contradictory. Secondly, the area comparing the effects of TF, AF, and IF on writing accuracy is under-explored. Thirdly, the review of empirical research indicates that, few, if any study has compared the effects of TF, AF, and IF on EFL learners' WA. Accordingly, to fill the lacuna in the extant empirical literature, this study investigated any significant differences among the effects of TF, AF, and IF on EFL learners' writing accuracy and WA.

Literature Review

Writing Accuracy, CF, and AF

Writing accuracy is defined by Foster and Skehan (1996) as the extent to which learners' written products conform to the target language norms and are free from errors. To Foster and Skehan (1996), writing accuracy encompasses the ratio of error-free clauses over the number of independent clauses, sub-clausal units, and subordinate clauses. Due to its importance, writing accuracy has been subject to abundant investigations in relation to the use of AF (AF) (Han & Sari, 2022), teacher's CF (Baleghizadeh & Gordani, 2012), synchronous and asynchronous written CF (Shintani & Aubrey, 2016), CF in blended learning (Sarré et al., 2021), content and language integrated learning (Lahuerta, 2020), the use of Edmodo as a social learning network (Safdari, 2021), and blended learning via Google Classroom (Torabi, 2021).

CF is defined by Sheen and Ellis (2011) as the "feedback learners receive on the linguistic errors they make in their oral or written production" (p. 593). Baleghizadeh and Gordani (2012) state that writing ability and accuracy would be enhanced when enough feedback is included in the process of teaching and learning. This feedback can be shown in different forms such as TF, peer feedback, and computerized feedback, to name but a few. Shute (2008) contends that there are three types of feedback including Knowledge of Results (KR), Knowledge of Correct Response (KCR), and finally, the Elaborated Feedback (EF). In KR, with its roots in behaviourism, the error is identified but not corrected. The term "Error Flagging" is used to indicate the type of feedback in which the learner is provided with the error location. However, no additional information about the correct response is provided (Van der Kleij et al., 2015). In KCR, with its origins in cognitivism, revision is carried out based on students' responses to feedback. Ultimately, the correct response is provided by the teacher if the learner fails to successfully correct the error. Finally, the EF refers to the provision of "hints", "additional information", "extra study materials", and "explanation of correct answers" (Van der Kleij et al., 2015, p. 5).

Highlighting the importance of writing, Brandt (2005) notes that technological advancements, including email, text messaging, and instant messaging devices, have become quite common in writing instruction. Thus, to improve learners' writing, educators should employ different techniques for evaluating students' writing and giving feedback. Since assessment data plays a crucial role in guiding informed educational decision, assessment techniques, namely manual essay grading and computer-scored essay grading have been introduced for measuring students' ability (Haddadian et al., 2024; Riordan et al., 2000). These advancements have paved the way for providing quality feedback. According to Ranalli (2021), machine-learning techniques have provided vigorous tools for automated writing evaluation (AWE), which is available to L2 learning environments for both teachers and learners. In the context of AWE, AF refers to the feedback provided for the linguistic errors in a piece of writing by computer softwares

(Tian & Zhou, 2020; Zhang & Hyland, 2018). Warschauer and Grimes (2008) define AF as the deployment of "artificial intelligence to evaluate essays and generate feedback" (p. 22).

As the results of previous studies indicate, both conventional TF (Baleghizadeh & Gordani, 2012) and AF (Han & Sari, 2022) exert positive influences on writing accuracy. The use of AF partially addresses the concerns regarding the time-consuming nature of conventional TF (Dikli & Bleyle, 2014). However, AF systems can be used to complement more traditional ways of providing feedback (Wang & Han, 2022) to mitigate learners and teachers' workload (Lavolette et al., 2015). Several investigations have so far compared AF with traditional TF. Zhang and Hyland (2018) compared student engagement with TF and AF on L2 writing. The participants were two Chinese students in the first semester of their third year in a Chinese university. Zhang and Hyland identified the strengths and weaknesses of both types of feedback and revealed how engagement was a vital mediating factor in the effectiveness of feedback. As Zhang and Hyland (2018) maintain, AF possesses tangible merits over TF in relation to "timeliness, convenience, multiple drafting" (p. 11), and even promoting learner autonomy. However, as they noted, conventional TF has some advantages such as the provision of content and organization feedback and comprehensive feedback in terms of abbreviation and number usage, which fall out of the scope of AF systems. Moreover, they underscored the value attached to teachers' feedback by learners. As they contended learners view teachers' feedback as a human response to their written products compared to the use of machine-algorithms for highlighting their errors. Likewise, Stevenson (2016), through a critical interpretative synthesis of existing research, found that teachers drew on various creative ways of integrating AF in their classrooms. Moreover, Stevenson's findings showed that although students seemed to enjoy AF, there were several limitations with AF systems such as lack of students' knowledge for using and interpreting AF, and AF inaccuracies in terms of analytic scoring categories. One of the AES systems which has been widely employed to provide feedback on writing is the Criterion.

Criterion as an AES System

Automated CF is a computerized and software-based form of writing evaluation. CF can be provided via different tools such as ETS's e-Rater and Criterion system to analyse and evaluate a text by removing the barrier of human knowledge (Attali & Burstein, 2006). The Criterion system is one of the widely-used AES applications. According to Attali and Burstein (2006), by using e-rater, Criterion can evaluate various writing genres and topics at different levels. Criterion enjoying a high accuracy rate of feedback compared to humans (Warschauer & Ware, 2006) can evaluate the effectiveness of feedback on the accuracy of revisions. It also indicates the overall quality of writing and provides feedback for different dimensions of a composition. The provided feedback is visually keyed to specific sections of text. Criterion generates feedback on grammatical errors or provides more holistic assessment on aspects such as content or organization of writing (Ware & Warschauer, 2005).

Li et al. (2014) explored the holistic scores provided by Criterion in three university-level ESL writing courses. They probed the correlation between Criterion scores, teachers' grades, and analytic ratings. The participants comprised three writing instructors and 67 ESL learners. Li et al. concluded that Criterion was an effective tool to facilitate formative assessment as it provided learners with continuous feedback. Dikli and Bleyle (2014) compared the feedback provided by instructors to Criterion feedback. Participants were 14 advanced students from various linguistic backgrounds who received feedback from the instructor and the Criterion. The results showed differences between the efficacy of Criterion feedback and the feedback of instructors on L2 learners’ errors. Criterion missed or misidentified many errors made by the students. However, such errors were accurately identified by the instructor. In fact, the instructors provided both more and better quality-feedback on form compared to the AES system (Criterion). On the other hand, the findings of Li et al. showed that Criterion led to the enhancement of L2 learners’ writing accuracy. Moreover, learners were highly satisfied with the CF of Criterion. Most of the interviewed instructors valued the CF for grammar and mechanics, although some of them acknowledged the ineffectiveness of the machine feedback for writing organization and development.

Although the sustained body of AES research has concentrated on proving the validity and reliability of such systems and has reported high agreement levels between machine and human raters (e.g., Attali & Burstein, 2006; Rudner et al., 2006; Wang & Brown, 2007), several studies have focused on the instructional applications of AES systems. For instance, Shermis et al. (2004) investigated the impact of using Criterion on students’ writing development. Over a thousand high school students were randomly assigned to either an experimental group using Criterion or a control group which completed alternate writing assignments without using Criterion. No significant differences were found between the groups on a state writing exam at the end of the training. However, the group using Criterion showed a substantial increase in average essay length and a decrease in the number of errors. Such a decrease was more evident especially for errors in writing mechanics (spelling, capitalization, punctuation, and grammar). Grimes and Warschauer (2006) found that the U.S. high school students experienced increased motivation for practicing writing when MyAccess and Criterion were used. Furthermore, Shermis et al. (2008) examined the efficacy of AF on learners’ writing outcomes. They reported that the number of learners’ errors decreased as a result of AF. Notwithstanding the contributions of AF systems in general and Criterion in particular to enhancing writing performance, these systems have some limitations such as lack of the humanistic nature of TF, inaccuracies in scoring categories, and students’ insufficient knowledge for using and interpreting AF (Bai & Hu, 2017; Stevenson, 2016; Wang & Brown, 2007). Such limitations necessitate the combination of TF into these systems to partially obviate the problems of these systems and make the provided feedback as influential as possible. Closely related to the inclusion of the humanistic aspect of feedback in AF systems is writing apprehension as WA is directly related to learners' emotions.

Writing Apprehension

WA is conceptualized as an emotion-related construct pertinent to writing process, which is manifested in negative feelings associated with writing and can adversely influence individuals' writing performance (Daly & Wilson, 1983; Larson, 1985; Rankin-Brown, 2006; Thompson, 2007). WA is defined as the arousal of certain emotions when written products are to be evaluated (Larson, 1985). Thompson (2007) refers to WA as fear of the writing process. Rankin-Brown (2006) defined WA as negative and anxious feelings in a writing situation, which disrupt some parts of the writing process. Daly and Wilson (1983) believe that WA is a continuous dimension of everyone, which affects academic success, occupational decisions, self-esteem, and personality behaviours. As Daly and Miller (1975) state, messages written by a high level of apprehension are evaluated significantly lower in terms of quality than those encoded by low apprehension. In this respect, writing skill can be affected by apprehension. Accordingly, apprehension can influence the final writing production. According to Badrasawi et al. (2016), WA affects writing performance negatively and hinders students’ academic achievement.

Many learners cannot effectively exhibit written communication skills due to WA (Autman & Kelly, 2017). While in the face-to face teaching contexts, teachers are capable of using immediate behaviours to offer clarity, such behaviour can be partially absent in online classrooms or online instruction (Kelly & Gaytan, 2020). Iksan and Abdul Halim's (2018) results indicated that, in online instruction, students were able to overcome their WA and enhance their writing performance as a result of writing in groups and drawing on their peers' feedback provided online. Such results point to the importance of the presence and involvement of teachers and peers in decreasing WA. Accordingly, the integration of computer-assisted instruction with teacher instruction can possibly assist learners in reducing their WA. Additionally, a review of previous studies on WA (e.g., Badrasawi et al., 2016; Chuo, 2007; Gaytan et al., 2022; Kelly, 2022; Marandi & Seyyedrezaie, 2017; Perkins, 2022; Rauf & Khan, 2022) reveals the non-existence of a study exploring the comparative effects of TF, AF, and IF on EFL learners' writing accuracy and WA.

As the results of previous studies indicate, AES systems have some restrictions, especially in regard to the lack of humanistic nature of such systems, which call for the integration of TF into these systems. Moreover, WA as an emotion-related aspect of the writing process can adversely affect writing performance. Furthermore, the available literature reveals that there is a gap in the literature comparing the effects of TF, AF, and IF on EFL learners' writing accuracy and WA. Thus, this study motivated by the void in the available literature and the interconnection of the emotions and writing explored any significant differences among the effects of TF, AF, and IF on EFL learners' writing accuracy and WA. In line with the research objectives, the following research questions were conceived:

Research Question 1: Are there any significant differences among the effects of TF, AF, and IF on EFL learners' writing accuracy?

Research Question 2: Are there any significant differences among the effects of TF, AF, and IF on EFL learners' WA?

Method

Participants

The initial participants comprised 283 EFL learners studying at the upper-intermediate level of language proficiency at nine private language schools in seven major cities in Iran. Persian was the mother tongue of all the learners. They had been learning English as a foreign language for two years (approximately 1200 hours of instruction). The participants were within the age range of 18 to 42 and from both genders (Females=145, Males=138). The researcher had to include learners within the age range of 18 to 42 as all the learners at the upper-intermediate level were within this age range. These learners were selected from among 102 available upper-intermediate classes via cluster sampling. Cluster sampling was used to gain access to various classes of learners at different language schools across the country. Moreover, since it was not possible to choose individual participants randomly from among the available learners across the language schools, the researcher had to resort to cluster sampling. Although random sampling of individual learners could have yielded more robust results and enhance the external validity of the findings, the researcher was not able to select the target participants based on a pure randomized manner. Out of the 102 classes, three classes at each school were selected. Thus, there were 27 classes in total, which consisted of 283 learners. These learners were studying in TOEFL preparation courses. The 283 learners were given an Oxford Quick Placement Test (OQPT) and 187 who scored within the range of 40 to 47 were selected as upper-intermediate learners in line with the placement scoring guidelines of OQPT. Apart from the learners, nine teachers also participated in this study to deliver the types of feedback. Four of the teachers were male and five were female. Their age ranged between 28 to 31 and their teaching experience fell within the range of 8 to 10 years. All teachers had prior experience of using AES systems such as Grammarly. However, to make sure that they had an adequate level of familiarity with Criterion a workshop was held. In the workshop, the teachers were provided with information on how to use the Criterion and interpret the scores.

Instruments

Oxford Quick Placement Test (OQPT)

OQPT is a reliable and valid measure developed by Oxford University Press. This test is designed to assess the English language ability of non-native speakers. It consists of 60 items and assesses learners' language performance in terms of grammar, vocabulary, and reading comprehension. Test takers' performance is measured based on their scores which display their level of language proficiency from beginners to high advanced: 1-17 (Beginner), 18-27 (Elementary), 28-36 (Lower-Intermediate), 37-47 (Upper-intermediate), 48-55 (Advanced), and 56-60 (high advanced). The results of this test are based on the Common European Framework of Reference (CEFR) scale. This test has proven to provide an accurate measure of English knowledge in a reliable and quick way (Wistner et al., 2009).

Writing Pretest and Posttest

Two topics, selected randomly from among 30 TOEFL independent writing tasks, were given to the participants as writing pretest and posttest. The learners were required to develop an essay within 300 to 400 words on the assigned topics in 30 minutes. The reason behind the selection of TOEFL writing tasks was that the present study was carried out in institutes in which TOEFL courses were held and all the participating learners were attending the classes to improve their TOEFL scores. Thus, the writing tasks were selected since they were in alignment with the course objectives.

Writing Accuracy Measure

Writing accuracy was computed in line with Foster and Skehan (1996). In so doing, initially the number of error-free clauses were counted. Next, the total number of independent clauses, sub-clausal units, and subordinate clauses were counted. Then, the number of error-free clauses was divided by the total number of independent clauses, sub-clausal units, and subordinate clauses. The yielded number was then multiplied by 100. To provide assurance concerning the consistency of the scores, two raters independently scored each writing and inter-rater reliability was computed. The inter-rater reliability index turned out to be .86 which is considered desirable.

Writing Apprehension Measure (WAM)

In order to measure students' WA, the WAM developed and validated by Autman and Kelly (2017) was used. Autman and Kelly shortened Daly and Miller's (1975) 26-item test to a six-item instrument. This measure was used in the current study since it had acceptable psychometric properties including desirable validity and reliability indices yielded by confirmatory factor analysis and Cronbach's Alpha, respectively. Moreover, since the number of participants who were asked to fill out the questionnaire was 187, this questionnaire was appropriate for the present research because a more extended questionnaire would be time-consuming and consequently pose feasibility problems for data collection. Additionally, as Autman and Kelly (2017) maintained this measure fits the present tech-savvy society and thus it was an appropriate measure for this study. Autman and Kelly updated Daly and Miller's (1975) 26-item instrument in terms of the wording of the questions and dropped items with overlapping variances (Stephens et al., 2020). This six-item test is based on a 7-likert point response scale ranging from strongly disagree (1) to strongly agree (7). Since reliability is sample dependent, this instrument was piloted on 30 non-participants, having similar characteristics to those of the main participants, and Cronbach's Alpha was run. The Cronbach's Alpha value turned out to be .73 which is considered desirable (Howitt & Cramer, 2014).

The Criterion Software

This web-based instructional software was developed by Educational Testing Service (ETS) as an online tool to evaluate and teach writing skills in the TOEFL test. Students, teachers, and administrations can all benefit from Criterion as it provides abundant opportunities for students to put writing into practice and get immediate diagnostic feedback to revise their writings. Teachers and administrations can also get an advantage as they can easily monitor what their students and system, as a whole, have done.

Data Collection

To recruit the participants for this study, the researcher initially sought the agreement of the CEOs of private language institutes in six cities in Iran. To this aim, the researcher posted an invitation via the ministry of education's official Telegram channel. Almost all private language schools around the country are present in this channel. Their participation was based on their written consent indicating that they were interested in participating in the present research. One condition for participation was that the institute was running TOEFL classes at the time of this research and was equipped with computers. Nine language schools gave their consent to take part in the study. Then, one orientation session and an introductory meeting was held for the principals and the CEOs to inform them about the implementation issues. This session's goal was to give an overview of the process to help the principals and CEOs better fit their institute's plans and educational goals so as to implement the project. Following that, 27 upper-intermediate classes consisting of 283 learners in the nine schools were selected and given an OQPT to assure the selection of a homogenized group of learners in terms of overall language proficiency. Prior to administering the test, the learners were given a consent form to sign. Based on the test results, 187 who scored within the range of 40 to 47 were selected. These learners continued their TOEFL preparation courses in the 27 classes. Before initiating the study, 27 teachers, who stayed in intact classes during this investigation, were asked to sign a consent form. The classroom teachers, teaching the 27 selected classes, were asked to fully participate in one informative workshop for two sessions. These sessions were held through a week, each lasting for 2 hours, to ensure that they gained adequate information on how to implement and how to give feedback based on Criterion. In the workshop, the first step was to introduce the Criterion platform and its features. Teachers were guided through the process of accessing the platform, creating accounts, and navigating the interface. Clear instructions were provided on how to upload student writing samples, select appropriate prompts, and initiate the scoring process. Emphasis was placed on the importance of selecting relevant prompts that align with learning objectives and language proficiency levels of the learners. Following the introduction to the Criterion platform, the next step involved a detailed explanation of the scoring criteria and rubrics. Teachers were presented with examples of different levels of writing proficiency and corresponding scores to illustrate the scoring scale. They were guided through each criterion, such as organization, development, language use, and mechanics, and shown how to identify strengths and areas for improvement in student writing. Practical tips were shared on how to provide constructive feedback based on the scores obtained, highlighting specific areas that students need to focus on for enhancement. By the end of the workshop, teachers were equipped with the knowledge and skills needed to effectively use the Criterion platform to assess and improve EFL learners' writing performance.

Next, the 27 classes were divided randomly and equally into three groups receiving TF, AF, and IF, respectively. The first group which received TF consisted of 64 learners, the second group which comprised 62 learners was exposed to AF, and the third group which contained 61 learners was treated with IF. Each group was studying in three classes. Moreover, each teacher taught one of the classes in each of the groups. Following the grouping, the learners in the three groups were given the writing pretest and the WAM. After that, the treatment in each group started.

As for the TF group, the participants were required to write an essay from the TOEFL independent writing task every other session. The teacher checked their writings and provided them with feedback in line with Shute’s (2008) three types of feedback. These types included Knowledge of Results (KR), Knowledge of Correct Response (KCR), and finally, the Elaborated Feedback (EF) for maximum efficiency as recommended by Van der Kleij et al. (2015). Shute's model was adopted in this study to provide learners with extensive feedback from the teacher. Accordingly, the instructors in TF group employed all three types of feedback to provide more elaborate feedback expected to result in enhanced learning outcomes. In the AF group, similar to the TF group, the learners were required to write one sample essay from the TOEFL independent writing task for every two sessions, but no TF was provided for this group. However, in one introductory session, the teachers were asked to provide information in regard to working with Criterion and obtaining the output. Moreover, the learners were required to send the revised draft of their writings along with Criterion reports to the teacher via email. As for the IF group, both types of feedback were provided. In so doing, the learners wrote an essay for each two sessions and initially submitted that to the Criterion. Next, they sent their revised draft along with the Criterion report to the teacher. The teacher was then required to provide the three feedback types in line with Shute (2008). To provide assurance that consistency was observed during the implementation of treatments, the researcher held three separate sessions with the participating teachers for each group via Telegram and made sure that the teachers were providing the feedback types appropriately. In so doing, each teacher was asked to explain how they were giving feedback for the writing assignments. During the sessions, it was found that all teachers were on track and were carrying out the feedback types in line with the objectives of each group.

The whole course of the data collection lasted 12 weeks. Every week two sessions were held. Overall, the learners wrote 12 essays and received the feedback type associated with each group. At the end of treatment, the three groups received writing posttest and the WAM. The writing accuracy pretest and posttest scores were fed into SPSS 26 to address the first research question. Similarly, the WAM pretest and posttest scores of the three groups were also calculated and inserted into SPP to examine the second research question.

Data Analysis

To address the first research question, initially the writing accuracy pretest scores of the three groups were compared via running a One-way ANOVA. However, since there were significant differences among the means of the writing accuracy scores, the pretest scores were considered as covariate and a One-way ANCOVA was run. As for the second research question, a similar statistical procedure was adapted, but since the assumptions of ANCOVA were not met, the gain scores were computed for the WA pretest and posttest scores and a Kruskal-Wallis test was run.

Ethical Considerations

To observe ethical considerations, the CEOs of the language institutes, the participating teachers, and the learners were asked to fill out consent forms. In the consent forms, the CEOS were provided with sufficient information regarding the objectives of the study and how the treatment types were implemented. The teachers were also given enough information about the study objectives and were asked to express their willingness to partake in the study. The learners were informed that their participation in the study was voluntary and they had the chance to withdraw from the study at any stage they wished.

Results

Research Question 1

To address the first research question, the researcher decided to run a one-way ANOVA to make sure that the three groups were not statistically different in terms of pretest writing accuracy scores. Table 1 displays the results of descriptive statistics for the pretest writing accuracy scores.

Table 1

Descriptive Statistics of Pretest Writing Accuracy Scores

	N	Range	Minimum	Maximum	Mean	SD	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Accuracy Pretest TF	64	23.00	34.00	57.00	44.203	6.01633	36.196	.323	.299	-.746	.590
Accuracy Pretest AF	62	26.00	32.00	58.00	42.338	7.34834	53.998	.553	.304	-.853	.599
Accuracy Pretest IF	61	22.00	35.00	57.00	46.245	6.31574	39.889	.124	.306	-1.023	.604
Valid N (listwise)	61										

As indicated in the above table, the score means for the TF, AF, and IF groups are 44.20, 42.33, and 46.24, respectively. This indicates that there might be statistically significant differences among the groups in terms of pretest writing accuracy scores. Table 2 shows the results of one-way ANOVA on the pretest writing accuracy scores.

Table 2

Results of One-Way ANOVA for Pretest Writing Accuracy Scores

Pretest Accuracy					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	469.629	2	234.815	5.423	.005
Within Groups	7967.558	184	43.302		
Total	8437.187	186			

As presented in Table 2, the sig value equals 0.005, which is lower than .01, indicating that the three groups are statistically different in terms of pretest writing accuracy scores. Such differences in the scores could be attributed to the possible contextual variations existing in the institutes from which the data were collected. Such contextual variations could include the over-emphasis of some of the institutes on writing practices in general and writing accuracy in particular. However, since this study intended to provide results which can have a high generalizability power, such differences at the beginning were compensated for by considering the pretest scores as covariate. Thus, the researcher decided to consider the pretest writing accuracy scores as a covariate and run ANCOVA to address the first research question. Before running ANCOVA the assumptions of normality, linearity, homogeneity of regression slopes, and homogeneity of variances in ANCOVA is essential for ensuring the validity and reliability of the results obtained. By assessing these assumptions, researchers can identify any potential issues that may impact the interpretation of treatment effects and make informed decisions about the appropriateness of using ANCOVA for their analysis.

Initially, the normality assumption for the covariate (pretest scores) and the dependent variable was checked. Deviations from normality can lead to biased estimates and inaccurate conclusions. Checking for normality helps ensure that the results of the ANCOVA analysis are valid and reliable. Table 3 displays the results of descriptive statistics and skewness and kurtosis values for the writing accuracy pretest and posttest scores.

Table 3

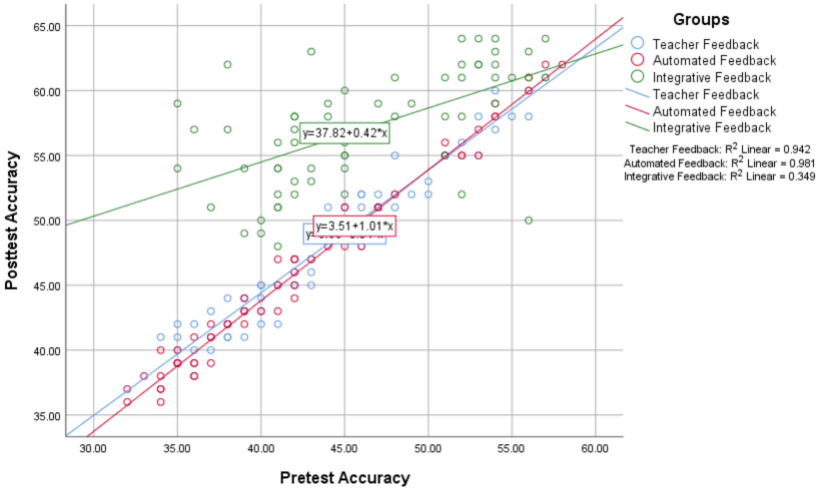
Descriptive Statistics and Skewness and Kurtosis Values of Writing Accuracy Pretest and Posttest

	N	Minimum	Maximum	Mean	SD	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Pretest Accuracy	187	32.00	58.00	44.2513	6.73507	45.361	.153	.178	-.288	.354
Posttest Accuracy	187	36.00	64.00	50.4973	7.64070	58.380	-.059	.178	-.116	.354
Valid N (listwise)	187									

As Table 3 exhibits, the skewness and kurtosis ratios for the writing accuracy pretest and posttest scores lay within the range of +/- 1.96. This indicates that the data sets were normally distributed (Tabachnick & Fidell, 2007). The second assumption was linearity. Violations of this assumption can lead to biased estimates of the treatment effects in ANCOVA. To check the linearity assumption, the scatterplot of the variables was inspected (Figure 1). As seen in Figure 1, lines for the variables are in the form of a straight diagonal line, which indicates that the assumption of linearity was not violated (Pallant, 2011). The third assumption was the homogeneity of regression slopes. If this assumption is violated, it suggests that the effect of the covariate on the dependent variable differs between groups, which can impact the interpretation of treatment effects in ANCOVA.

Figure 1

Scatterplot of accuracy pretest and posttest scores



To check the homogeneity of regression slopes, the table for Tests of Between-Subjects Effects was consulted. The respective results are presented in Table 4.

Table 4

Tests of Between-Subjects Effects for Accuracy Pretest and Posttest Scores

Dependent Variable: Posttest Accuracy

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	9892.284a	5	1978.457	370.527	.000
Intercept	983.163	1	983.163	184.127	.000
Groups	881.530	2	440.765	82.547	.093
WApre	4836.719	1	4836.719	905.823	.121
Groups * WApre	539.763	2	269.882	50.544	.178
Error	966.464	181	5.340		
Total	487705.000	187			
Corrected Total	10858.749	186			

a. R Squared = .911 (Adjusted R Squared = .909)

The significant value corresponding to Groups * WApre equals .178 which is greater than 0.05 (Table 4). This indicates that the assumption of the homogeneity of regression slopes was warranted. The last assumption was the homogeneity of variances. Violations of this assumption can lead to inflated Type I error rates and inaccurate conclusions in ANCOVA. Table 5 displays the results of Levene's test of variances (Table 5).

Table 5

Levene's Test of Equality of Error Variances for Accuracy Scores

Dependent Variable: Posttest Accuracy

F	df1	df2	Sig.
17.063	2	184	.210

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + WApre + Groups

Based on the results of the Levene's test, variances in the dependent and covariate variable were equal. Therefore, the assumption of homogeneity of variances was met (F = 17.06, p = .21 > .05). Having assured that all the assumptions were met, the researcher examined the main ANCOVA output (Table 6).

Table 6

Results of ANCOVA for the Accuracy Pretest and Posttest Scores

Dependent Variable: Posttest Accuracy

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	9352.521a	3	3117.507	378.763	.000	.861
Intercept	850.035	1	850.035	103.275	.000	.361
WApre	5253.465	1	5253.465	638.273	.012	.777
Groups	2157.062	2	1078.531	131.037	.001	.589
Error	1506.228	183	8.231			
Total	487705.000	187				
Corrected Total	10858.749	186				

a. R Squared = .861 (Adjusted R Squared = .859)

As Table 6 shows, the sig value corresponding to the groups turned out to be smaller than the critical value (*p*= .001<.001). This shows that there were significant differences among the performances of the three groups in terms of accuracy scores. The partial eta squared turned out to be .58, which is an indication of a large effect size showing that this large magnitude of effect in the context of this study is attributed to the treatment types (Cohen, 1988). Table 7 displays the estimated marginal means for the three groups' accuracy scores.

Table 7

Estimated Marginal Means for Accuracy Scores

Dependent Variable: Posttest Accuracy

Groups	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
TF	48.430a	.359	47.722	49.137
AF	47.730a	.370	47.001	48.460
IF	55.479a	.373	54.743	56.214

a. Covariates appearing in the model are evaluated at the following values: Pretest Accuracy = 44.2513.

Table 8 presents the pairwise comparisons among the three groups' writing accuracy scores. There is a significant difference between the IF and TF groups (*p* = .00 < 0.001), with the IF group outperforming the TF group (Mean difference = 7.04). Likewise, there is a significant difference between the IF and AF groups (*p* =. 00 < 0.001), with the IF group outperforming the AF group (Mean difference = 7.74). However, there is not a significant difference between the TF and AF groups (*p* = .176 > 0.05, Mean difference = .699).

Table 8

Pairwise Comparisons Among the Three Groups' Accuracy Scores

Pairwise Comparisons

Dependent Variable: Posttest Accuracy

(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.b	95% Confidence Interval for Differenceb Lower Bound	Upper Bound
TF	AF	.699	.515	.176	-.316	1.715
	IF	-7.049*	.518	.000	-8.070	-6.028
AF	TF	-.699	.515	.176	-1.715	.316
	IF	-7.748*	.532	.000	-8.799	-6.698
IF	TF	7.049*	.518	.000	6.028	8.070
	AF	7.748*	.532	.000	6.698	8.799

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Research Question 2

To examine the second research question, the researcher decided to run a one-way ANOVA to make sure that the three groups were not statistically different in terms of pretest WAM scores. Table 9 displays the results of descriptive statistics for the pretest WAM scores.

Table 9

Descriptive Statistics of Pretest WAM Scores

	N	Range	Minimum	Maximum	Mean	SD	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
App-Pretest TF	64	11.00	28.00	39.00	33.51	2.7079	7.333	.153	.299	-.788	.590
App-Pretest AF	62	8.00	28.00	36.00	32.59	1.778	3.163	-.011	.304	-.232	.599
App Pretest IF	61	9.00	31.00	40.00	35.11	2.523	6.370	.342	.306	-.524	.604
Valid N (listwise)	61										

As presented in the above table, the score means for the TF, AF, and IF groups are 33.51, 32.59, and 35.11, respectively. Such variations could be attributed to the possible diversity in the individual characteristics of the participants, which was not feasible to be controlled in the context of this study. The mean differences indicate that there might be statistically significant differences among the groups in terms of pretest WAM scores. Table 10 shows the results of one-way ANOVA on the pretest WAM scores.

Table 10

Results of One-Way ANOVA for WAM Scores

Pretest Apprehension

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	199.531	2	99.765	17.700	.000

Within Groups	1037.100	184	5.636
Total	1236.631	186	

As seen in Table 10, the sig value turns out to be 0.00, which is smaller than .001. This shows the three groups are statistically different in terms of pretest WAM scores. Thus, the researcher decided to consider the pretest WAM scores as a covariate and run ANCOVA to address the second research question. ANCOVA was run to be statistically on the safe side as there were significant differences among the pretest mean scores of the three groups.

Table 11 presents the results of descriptive statistics and skewness and kurtosis values for the WAM pretest and posttest scores.

Table 11
Descriptive Statistics and Skewness and Kurtosis Values of WAM Pretest and Posttest

	N	Minimum	Maximum	Mean	SD	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Pretest-App	187	28.00	40.00	33.7326	2.57848	6.649	2.12	.178	-.244	.354
Posttest-App	187	18.00	35.00	27.2406	4.82610	23.291	-.215	.178	-1.952	.354
Valid N (listwise)	187									

As Table 11 indicates, the skewness and kurtosis ratios for the WAM pretest and posttest scores did not fall within the range of +/- 1.96. This demonstrates that the data sets violated the normality assumption (Tabachnick & Fidell, 2007). Thus, the researcher computed the gain scores for each group. Table 12 depicts the descriptive statistics for the pretest and posttest WAM scores for the three groups.

Table 12
Descriptive Statistics for Pretest Posttest WAM Scores of the Three Groups

	N	Range	Minimum	Maximum	Mean	SD	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Pretest-App TF	64	11.00	28.00	39.00	33.515	2.7079	7.333	.153	.299	-.888	.590
Pretest - App AF	62	8.00	28.00	36.00	32.596	1.7783	3.163	-.011	.304	-.232	.599
Pretest-App IF	61	9.00	31.00	40.00	35.114	2.5238	6.370	.342	.306	-.524	.604
Posttest-App TF	64	13.00	20.00	33.00	28.250	3.1922	10.190	-.489	.299	-.661	.590
Posttest-App AF	62	8.00	27.00	35.00	31.822	1.6647	2.771	-.039	.304	-.159	.599
Posttest-App IF	61	11.00	18.00	29.00	21.524	1.6391	2.687	1.539	.306	6.441	.604
Valid N (listwise)	61										

As seen in the above Table, the means of WAM scores for the TF, AF, and IF groups have decreased on posttest, indicating that there has been some reduction in WAM posttest scores compared with the pretest. Thus, to compute the gain scores, the pretest scores were subtracted from the posttest scores. Table 13 displays the descriptive statistics and skewness and kurtosis values for the WAM gain scores.

Table 13

Descriptive Statistics for the WAM Gain Scores

	N	Range	Minimum	Maximum	Mean	SD	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
TF Gain Scores	64	11.00	1.00	12.00	5.2656	2.3854	5.690	1.982	.299	1.858	.590
AF Gain Scores	62	4.00	-1.00	3.00	.7742	1.0467	1.096	-1.236	.304	-.096	.599
IF Gain Scores	61	18.00	3.00	21.00	13.590	2.8541	8.146	-2.482	.306	2.536	.604
Valid N (listwise)	61										

According to the information in Table 13, the skewness and kurtosis ratios for the WAM scores did not fall within the range of +/- 1.96. This is an indication of the violation of normality assumption (Tabachnick & Fidell, 2007). Therefore, the researcher ran the non-parametric test of Kruskal-Wallis to find any significant differences among the gain scores for the three groups. Table 14 shows the respective results.

Table 14

Independent-Samples Kruskal-Wallis Test Summary

Total N	187
Test Statistic	157.827a
Degree Of Freedom	2
Asymptotic Sig. (2-sided test)	.000

a. The test statistic is adjusted for ties.

As demonstrated in Table 14, the sig value turned out to be .00, which is lower than 0.001, indicating that there are significant differences among the gain scores of the three groups' WAM scores. The effect size calculated based on the eta2[H] = (H - k + 1)/(n - k) formula, where H is the value obtained in the Kruskal-Wallis test; k is the number of groups; and n is the total number of observations, turned out to be 0.83 which is an indication of a large effect size. This large effect size shows that the treatment types have had substantial effects on the participants' WA (Cohen, 1988). Table 15 shows the results of pairwise comparisons.

Table 15

Pairwise Comparisons of Groups

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.a
Automated Feedback- TF	60.689	9.591	6.328	.000	.000
Automated Feedback-IF	-121.942	9.707	-12.563	.000	.000
TF -IF	-61.253	9.631	-6.360	.000	.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

As seen in Table 15, there is a significant difference between the IF and TF groups (*p* = .00 < 0.001), with the IF group showing more reduction in WAM scores than the TF group. Likewise, there is a significant difference between the IF and AF groups (*p* =. 00 < 0.001), with the IF group displaying more reduction in WAM than the AF group. Similarly, there is a significant difference between the TF and AF groups (*p* = .00 < 0.001), with the TF group exhibiting more reduction in WAM scores compared with the AF group.

Discussion

This study compared the effects of TF, AF, and IF on EFL learners' writing accuracy and WA. The results of statistical analysis indicated significant differences among the effects of TF, AF, and IF on EFL learners' writing accuracy with the IF group outperforming the TF and AF groups. However, there was not a significant difference between the TF and AF groups. The results also indicated that there were significant differences among the effects of TF, AF, and IF on EFL learners' WA with the IF group exhibiting more reduction in WA compared to the other two groups. Additionally, there was a significant difference between the TF and AF groups, with the TF group displaying more reduction in WAM scores compared to the AF group.

Writing Accuracy

The more effectiveness of IF than AF and TF on writing accuracy can be justified based on the merits of each feedback type. In other words, the merits of the two types of feedback in tandem have proved more beneficial for learners than each feedback type in isolation. As Zhang and Hyland (2018) found "timeliness, convenience, multiple drafting" (p. 11), and fostering learner autonomy were the advantages of AF over TF. On the other hand, more comprehensive feedback can be delivered via TF. Therefore, during the implementation of the two feedback types, these advantages have been capitalized on and have contributed to writing accuracy development. Along the same lines, previous empirical research has revealed that both conventional TF (Baleghizadeh & Gordani, 2012), and AF (Han & Sari, 2022) positively impact writing accuracy. Thus, when both types of feedback are used in combination, they can prove more effective. In essence, IF combines the strengths of both TF and AF, providing learners with a more comprehensive and diverse set of feedback sources (Stevenson, 2016). The combination of human expertise and machine precision in IF may have led to a more effective feedback process, allowing learners to receive tailored guidance from both sources (McCarthy et al., 2022; Wang & Han, 2022; Zhang & Hyland, 2018). The integration of teacher and automated feedback likely facilitated a more holistic approach to addressing writing errors and promoting language development, ultimately resulting in improved writing accuracy for the EFL learners in the IF group.

The lack of a significant difference between AF and TF in improving writing accuracy are inconsistent with the findings of Dikli and Bleyle (2014). Dikli and Bleyle found differences between the efficacy of Criterion feedback and TF on L2 learners' errors. This result can be explained based on the model of feedback (i.e., Shute, 2008) adopted in the present study. Adopting Shute's model, in this study the researcher used comprehensive feedback to incorporate KR, KCR, and EF for maximum efficiency. Put it another way, although learners in the AF group had the chances of convenient use of Criterion for revising their writings several times, the TF group received extensive feedback from the teacher, which has similarly enhanced their writing accuracy. Moreover, Criterion might have missed or misidentified some errors as pointed out by

Dikli and Bleyle (2014). The lack of a significant difference between the TF and AF groups in terms of writing accuracy can be attributed to the nature of the feedback provided in both groups leading to comparable outcomes. Teacher feedback and automated feedback may have addressed similar types of errors or provided similar suggestions for improvement, resulting in overlapping effects on writing accuracy. It is also possible that individual learner characteristics, such as learning preferences, influenced how they benefited from the different types of feedback. Overall, the lack of a significant difference between the TF and AF groups highlights the importance of considering the specific characteristics and implementation of feedback strategies in language learning research.

Writing Apprehension

The significant reduction of WA in the IF group compared to AF and TF groups can be explained based on the cumulative instructional support provided by both feedback types used in the IF group. Contrary to the TF group, and the AF group, who had only the teacher, or the Criterion at their disposal to overcome their WA, the IF group concomitantly enjoyed the instructional affordances of the Criterion and the humanistic potential of the instructor for support. Accordingly, the humanistic nature of the TF along with the speed of the Criterion in providing feedback (Zhang & Hyland, 2018) might be the factors responsible for more reduction in learners' WAM in comparison with the other two groups.

The significant difference between the TF and AF groups, with the TF group displaying more reduction in WAM compared to the AF group, substantiate the results of previous studies in which the associations between teachers' factors and behaviours and WA have been documented. For instance, Gaytan et al. (2022) found that teacher's behaviour is related to shaping learners' WA. Moreover, Perkins (2022) showed associations between teacher efficacy and WA. In this study, the teachers in the TF group are likely to have displayed more efficacy and attuned behaviour in the provision of comprehensive feedback; something which was absent in the AF group as they were only exposed to Criterion feedback.

Conclusion

This study aimed to divulge any significant differences among the effects of TF, AF, and IF on EFL learners' writing accuracy and WA. The results of ANCOVA run to address the first research question indicated significant differences among the effects of TF, AF, and IF on writing accuracy (effect size = .58) with the IF group outperforming the other two groups. However, there was no significant difference between the TF and AF groups. The results of Kruskal-Wallis carried out to examine the second research question showed that there were significant differences among the effects of TF, AF, and IF on WA (effect size = .83) with the IF group exhibiting more reduction in WA compared to the other groups. Additionally, there was a significant difference between the TF and AF groups, with the TF group displaying more reduction in WA compared to the AF group. Based on the results, EFL teachers are encouraged to integrate AF with TF in their instructional practices to enhance EFL learners' writing accuracy and decrease their WA. In so doing, teachers should provide comprehensive TF to complement the limitations and shortcomings of AES systems. Moreover, learners can be provided with awareness over the possible adverse effects that AES systems can exsert on their WA. Learners can be encouraged to seek TF in tandem with AF to accommodate the role of affective factors in their writing development. The superiority of IF in improving writing accuracy suggests that a combination of TF and AF can be a powerful approach to enhancing EFL learners' writing skills. Educators may consider integrating both human expertise and machine precision in providing feedback to students, as this holistic approach can offer a more comprehensive and tailored feedback experience. By leveraging the strengths of both TF and AF, language instructors can create a more effective feedback environment that supports learners in addressing writing errors and improving their overall writing accuracy. Implementing IF strategies in language classrooms may lead to more significant gains in writing proficiency among EFL learners. Furthermore, the significant reduction in WA observed in the integrative feedback IF group compared to the TF and AF groups underscores the potential benefits of combining different feedback sources in addressing learners' emotional responses to writing tasks. Reducing WA is crucial for promoting learners' confidence and motivation in writing activities, which are essential components of language learning success. Language teachers can use IF approaches to not only enhance writing accuracy but also alleviate learners' anxiety and apprehension towards writing tasks. By providing a supportive and comprehensive feedback system that combines human guidance with technological assistance, teachers can create a more positive learning environment that fosters EFL learners' writing skills development and overall language proficiency. However, a balance should be struck between the effectiveness of IF and the potential benefits of TF and AF should be acknowledged. Overall, language teachers should adopt a more informed and nuanced approach to feedback integration in language learning contexts.

The findings in this study revealed the more effectiveness of IF in comparison with TF and AF in enhancing writing accuracy and decreasing WA. Nonetheless, such findings cannot be conclusive given the contradictory results of previous studies. Thus, the replication of this study in other settings and with other AESs can provide a more comprehensive picture of the role of IF in contributing to writing accuracy and WA. The replication of this study with other AESs is of empirical value since there is the possibility of variations in results across different AES systems. This study was carried out in the Iranian EFL context and in private language institutes. Since contextual variations can play a role in the results, this study can be replicated in contexts other than EFL settings and different educational contexts such as state schools and universities. Additionally, qualitative data collection instruments such as observations and interviews can provide more insights in regard to the comparative role of different feedback types in influencing writing accuracy and WA. Such data collection methods and instruments could complement the quantitative findings and contribute to a richer understanding of the topic.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3),1–31. https://ejournals.bc.edu/index.php/jtla/article/view/1650/1492

Autman, H., & Kelly, S. (2017). Reexaming the writing apprehension measure. *Business and Professional Communication Quarterly, 80*(4), 516-529. https://doi.org/10.1177/ 2329490617691968

Badrasawi, K. J., Zubairi, A., & Idrus, F. (2016). Exploring the relationship between writing apprehension and writing performance: A qualitative study. *International Education Studies, 9*(8), 134-143. https://files.eric.ed.gov/fulltext/EJ1110238.pdf

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

Baleghizadeh, S., & Gordani, Y. (2012). Academic writing and grammatical accuracy: The role of corrective feedback. *Gist: Education and Learning Research Journal*, (6), 159-176. https://dialnet.unirioja.es/servlet/articulo?codigo=4707905

Brandt, D. (2005). Writing for a living: Literacy and the knowledge economy. *Written Communication, 22*(2), 166-197. https://doi.org/10.1177/0741088305275218

Campbell, C. W., & Batista, B. (2023). To peer or not to peer: A controlled peer-editing intervention measuring writing self-efficacy in South Korean higher education. *International Journal of Educational Research Open*, *4*, 100218. https://doi.org/10.1016/j.ijedro.2022.100218

Chuo, T. W. I. (2007). The Effects of the WebQuest writing instruction program on EFL learners' writing performance, writing apprehension, and perception. *Tesl-ej*, *11*(3), n3. https://files.eric.ed.gov/fulltext/EJ1065001.pdf

Cohen, J.W. (1988). *Statistical power analysis for the behavioral sciences (2nd ed).* Lawrence Erlbaum Associates.

Daly, J. A., & Miller, M. D. (1975). The empirical development of an instrument to measure writing apprehension. *Research in the Teaching of English*, *9*(3), 242–249. http://www.jstor.org/stable/40170632

Daly, J. A., & Wilson, D. A. (1983). Writing apprehension, self-esteem, and personality. *Research in the Teaching of English*, *17*(4), 327-41.

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

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, *18*(3), 299-324. https://doi.org/ 10.1017/S0272263100015047

Gaytan, J., Kelly, S., & Brown, W. S. (2022). Writing apprehension in the online classroom: The limits of instructor behaviors. *Business and Professional Communication Quarterly*, *85*(4), 376-394. https://doi.org/10.1177/23294906211041088

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment, 8*(6). https://ejournals.bc.edu/index.php/jtla/article/view/1625/1469

Haddadian, G., & Haddadian, N. (2024). Innovative Use of Grammarly Feedback for Improving EFL Learners' Speaking: Learners' Perceptions and Transformative Engagement

Experiences in Focus. *The Journal of Applied Instructional Design*, 13(2). https://doi.org/10.59668/1269.15640

Haddadian, G., Radmanesh, S., & Haddadian, N. (2024). Construction and validation of a Computerized Formative Assessment Literacy (CFAL) questionnaire for language teachers: an exploratory sequential mixed-methods investigation. *Language Testing in Asia*, *14*(1), 1-24. https://doi.org/10.1186/s40468-024-00303-2

Han, T., & Sari, E. (2022). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning*, 1-24. https://doi.org/10.1080/09588221.2022.2067179

Howitt, D., & Cramer, D. (2014). *Introduction to research methods in psychology* (4th ed.). Harlow, England: Pearson.

Iksan, H., & Abdul Halim, H. (2018). The effect of e-feedback via wikis on ESL students' L2 writing anxiety level. *Malaysian Online Journal of Educational Sciences, 6*(3), 30-48. https://www.learntechlib.org/p/192237/

Kalman, J., Méndez-Arreola, R., & Valdivia, P. (2023). Conceptualizing everyday writing. In H., Rosalind, *The Routledge international handbook of research on writing* (pp. 317-333). Routledge.

Kelly, S., & Gaytan, J. (2020). The effect of instructors' immediate behaviors and clarity on student writing apprehension. *Business and Professional Communication Quarterly, 83*(1), 96-109. https://doi.org/10.1177/2329490619868822

Kelly, S., Violanti, M., Denton, E., & Berry, I. (2022). Instructor misbehaviors as predictors of students' writing apprehension. *Communication Quarterly*, 1-19. https://doi.org/10.1080/01463373.2022.2077123

Lahuerta, A. (2020). Analysis of accuracy in the writing of EFL students enrolled on CLIL and non-CLIL programmes: The impact of grade and gender. *The Language Learning Journal*, *48*(2), 121-132. https://doi.org/10.1080/09571736.2017.1303745

Larson, R. (1985). Emotional scenarios in the writing process: An examination of young writers' affective experiences. In M. Rose (Ed.), *When a writer can't write* (pp. 19-42). The Guilford Press.

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language, Learning & Technology*, *19*(2). https://cupola.gettysburg.edu/cgi/viewcontent.cgi?article=1001&context=lrc

Lei, L., Wen, J., & Yang, X. (2023). A large-scale longitudinal study of syntactic complexity development in EFL writing: A mixed-effects model approach. *Journal of Second Language Writing*, *59*, 100962. https://doi.org/10.1016/j.jslw.2022.100962

Li, M., & Zhang, M. (2023). Collaborative writing in L2 classrooms: A research agenda. *Language Teaching*, *56*(1), 94-112. https://doi.org/10.1017/S0261444821000318

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

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

Marandi, S. S., & Seyyedrezaie, M. S. (2017). The multi-course comparison of the effectiveness of two EFL writing environments: Google drive versus face-to-face on Iranian EFL learners' writing performance and writing apprehension. *CALL-EJ*, *18*(1), 9-21.

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

Mirzaii, M., & Aliabadi, R. B. (2013). Direct and indirect written corrective feedback in the context of genre-based instruction on job application letter writing. *Journal of Writing Research*, *5*(2), 191-213. https://doi.org/10.17239/jowr-2013.05.02.2

Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2022). The effectiveness of automated writing evaluation in EFL/ESL writing: a three-level meta-analysis. *Interactive Learning Environments*, 1-18. https://doi.org/10.1080/10494820.2022.2096642

Pallant, J. S. (2011). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows (4th ed).* Allen & Unwin, Crows Nest.

Perkins, M. A. (2022). The relationship between teacher efficacy, writing apprehension, and writing to learn using structural equation modeling. *The Journal of Writing Analytics*, *6*, 15-57. https://doi.org/10.37514/JWA-J.2022.6.1.03

Pourfeiz, J. (2022). Willingness to write and writing performance of EFL students: Pursuit of relevance. *Eurasian Journal of Language Teaching & Linguistic Studies (EAJLTLS), 2*(1).

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

Rankin-Brown, M. (2006). *Addressing writing apprehension in adult English language learners*. Paper presented at the Proceedings of the CATESOL state conference.

Rauf, A., & Khan, W. A. (2022). Effect of peer assessment on English writing mechanics and writing apprehension of undergraduate students. *Pakistan Journal of Educational Research and Evaluation (PJERE), 9*(2). http://111.68.103.26/journals/index.php/PJERE/article/view/5303/2514

Riordan, D. A., Riordan, M. P., & Sullivan, M. C. (2000). Writing across the accounting curriculum: An experiment. *Business Communication Quarterly, 63*(3), 49-59. https://doi.org/10.1177/108056990006300305

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*(4), 1-21. https://files.eric.ed.gov/fulltext/EJ843850.pdf

Safdari, M. (2021). Contributions of Edmodo social learning network to Iranian EFL learners' writing accuracy. *CALL-EJ*, *22*(1), 355-373.

Sarré, C., Grosbois, M., & Brudermann, C. (2021). Fostering accuracy in L2 writing: Impact of different types of corrective feedback in an experimental blended learning EFL course. *Computer Assisted Language Learning*, *34*(5-6), 707-729. https://doi.org/10.1080/09588221.2019.1635164

Sheen, Y., & Ellis, R. (2011). Corrective feedback in language teaching. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 593- 610). Routledge.

Shen, B., & Bai, B. (2022). Chinese university students' self-regulated writing strategy use and EFL writing performance: influences of self-efficacy, gender, and major. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2020-0103

Shermis, M.D., Burstein, J.C., & Bliss, L. (2004). *The impact of automated essay scoring on high stakes writing assessments.* Paper presented at the Annual Meetings of the National Council on Measurement in Education, San Diego, CA.

Shermis, M.D., Garvan, C.W., & Diao, Y. (2008). *The impact of automated essay scoring on writing outcomes.* Paper presented at the Paper presented at the Annual Meetings of the National Council on Measurement in Education. https://files.eric.ed.gov/fulltext/ED501148.pdf

Shintani, N., & Aubrey, S. (2016). The effectiveness of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment. *The Modern Language Journal*, *100*(1), 296-319. https://doi.org/10.1111/modl.12317

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62. https://doi.org/10.1093/applin/17.1.38

Stephens, L. E., Kim, G., Fogle, E. L. V., Kleinbort, T., Duffy, L. N., Powell, G. M., ... & Gremillion, J. P. (2020). Reducing writing apprehension in undergraduate parks, recreation, and tourism management students. *SCHOLE: A Journal of Leisure Studies and Recreation Education*, 1-15. https://doi.org/10.1080/1937156X.2020.1860668

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

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

Sun, Q., & Zhang, L. J. (2022). Examining the effects of English as a foreign language student-writers' metacognitive experiences on their writing performance. *Current Psychology*, 1-16. https://doi.org/10.1007/s12144-022-03416-0

Tabachnick, B.G. & Fidell, L.S. (2007). *Using multivariate statistics*. Pearson Education.

Thompson, R. A. (2007). Emotion Regulation: Conceptual Foundations. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 3–24). The Guilford Press. https://doi.org/10.4236/am.2013.412233

Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System*, *91*, 102247. https://doi.org/10.1016/j.system.2020.102247

Torabi, S. (2021). Blended learning (B-learning) via Google Classroom (GC) and Iranian EFL learners' writing accuracy: Effects and percepts. *CALL-EJ*, *22*(3), 183-199.

Tsingos-Lucas, C., Bosnic-Anticevich, S., Schneider, C. R., & Smith, L. (2017). Using reflective writing as a predictor of academic success in different assessment formats. *American Journal of Pharmaceutical Education*, *81*(1), 8. https://doi.org/10.5688/ajpe8118

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475-511. https://doi.org/10.3102/0034654314564881

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6*(2). https://ejournals.bc.edu/index.php/jtla/article/view/1632/1476

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*. https://doi.org/10.3389/fpsyg.2022.909802

Ware, P. D., & Warschauer, M. (2005). Hybrid literacy texts and practices in technology-intensive environments. *International Journal of Educational Research, 43*(7-8), 432-445. https://doi.org/10.1016/j.ijer.2006.07.008

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

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

Wistner, B., Sakai, H., & Abe, M. (2009). An analysis of the Oxford Placement Test and the Michigan English Placement Test as L2 proficiency tests. *Bulletin of the Faculty of Letters, Hosei University, 58*(2), 33-44. file:///C:/Users/acer/Downloads/bun58_brian,sakai,abe%20(1).pdf

Xu, T. S., Zhang, L. J., & Gaffney, J. S. (2023). A multidimensional approach to assessing the effects of task complexity on L2 students' argumentative writing. *Assessing Writing*, *55*, 100690. https://doi.org/10.1016/j.asw.2022.100690

Yu, S., Xu, H., Jiang, L., & Chan, I. K. I. (2020). Understanding Macau novice secondary teachers' beliefs and practices of EFL writing instruction: A complexity theory perspective. *Journal of Second Language Writing*, *48*, 100728. https://doi.org/10.1016/j.jslw.2020.100728

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