Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 13, No. 1, March 2023

ISSN 2476-5880

Formative Assessment Feedback to Enhance the Writing Performance of Iranian IELTS Candidates: Blending Teacher and Automated Writing Evaluation

Mojtaba Mohammadi[1]*, Maryam Zarrabi[2], Jaber Kamali[3]

Received: 01 September 2022 Accepted: 11 December 2022

Abstract

With the incremental integration of technology in writing assessment, technology-generated feedback has found its way to take further steps toward replacing human corrective feedback and rating. Yet, further investigation is deemed necessary regarding its potential use either as a supplement to or replacement for human feedback. This study aims to investigate the effect of blending teacher and automated writing evaluation, as formative assessment feedback, on enhancing the writing performance among Iranian IELTS candidates. In this explanatory mixed-methods research, three groups of Iranian intermediate learners (N=31) completed six IELTS writing tasks during six consecutive weeks and received automated, teacher, and blended (automated + teacher) feedback modes respectively on different components of writing (task response, coherence and cohesion, lexical resource, grammatical range and accuracy). A structured written interview was also conducted to explore learners' perception (attitude, clarity, preference) of the mode of feedback they received. Findings revealed that students who received teacher-only and blended feedback performed better in writing. Also, the blended feedback group outperformed the others regarding *task response*, the teacher feedback group in *cohesion and coherence,* and the automated feedback group in *lexical resource*. The analysis of the interviews revealed that the majority of the learners confirmed the clarity of all feedback modes and learners' attitude about feedback modes was positive although they highly preferred the blended one. The findings suggest new ideas to facilitate learning and assessing writing and support the evidence that teachers can provide comprehensive, accurate, and continuous feedback as a means of formative assessment.

Keywords: Automated writing evaluation (AWE); Blended feedback; Formative assessment; IELTS writing; Learners' perception

1. Introduction

With the outbreak of the Covid-19 pandemic, the use of technology in language classrooms has witnessed incremental changes. One of these changes was the considerable attention to digital modes of L2 writing feedback (e.g., Gao & Ma, 2022; Jiang & Lu, 2022; Link et al., 2022; Ranalli & Yamashita, 2022; Shang, 2022). Automated writing evaluation

[1] ELT Department, West Tehran Branch, Islamic Azad University, Tehran, Iran.
Applied Linguistics Research Center, Roudehen Branch, Islamic Azad University, Roudehen, Iran
Email: mohammadi.mojtaba@wtiau.ac.ir; mojtabamohammadi@gmail.com
[2] English Department, Science and Research Branch, Islamic Azad University, Tehran, Iran. Email: m.zarrabi21@gmail.com
[3] Ibn Haldun University, Istanbul, Turkey. Email: jaber.kamali@ihu.edu.tr

(AWE), besides its purpose as a means of scoring, is a tool to provide formative assessment (Ranalli et al., 2017). Recent studies have acknowledged that AWE can be of service to learners to alleviate language-related challenges in writing more conveniently (e.g., Li et al., 2015). However, there are still many questions to raise, like, is AWE alone sufficient and efficient enough for developing writing ability? Do students have a positive perception of machine-generated feedback as opposed to their traditional teacher feedback? Which one do they prefer? Does AWE tap into the higher-order aspects of assessing writing components (e.g., coherence and task achievement)? These issues and others are also the critical concern of teachers in preparation courses for high-stakes international tests like IELTS (Pearson, 2018). According to Zhang and Hyland (2018), teacher feedback can address more error categories while AWE feedback mostly highlights, rather than corrects, students' errors. More comments can be provided by AWE which can, in turn, only raise learners' awareness (Ferris, 2002).

This study further expands the knowledge base of technology-integrated feedback in some ways. First, most of the studies use holistic scoring to measure writing performance, and analytic scoring of the components of writing which can be a source of formative assessment is overlooked (Biber et al, 2011). Second, the literature still lacks a convincing consonance with regard to the effectiveness of computer-generated feedback on learning; That is why some scholars, rather recently, (e.g., Stevenson & Phakiti, 2014; Zhang & Hyland, 2018) are inclined to suggest the integration of human and machine feedback to optimize the effect of feedback provision. It is deemed worthwhile to further investigate since little literature is available regarding the effectiveness of the integration of these two modes and their comparison with any one of them separately in practice. Next, from a methodological perspective, this study has approached the issue from a quantitative/qualitative perspective which can lead to a thorough understanding of the integration of both modes of feedback. Last, the majority of the studies have dealt with writing at either university or school levels and few studies investigated writing in exam classes. This study works on the IELTS writing task two and has a specific focus on opinion essays.

The present study investigates the effectiveness of implementing three feedback modes (automated, teacher, blended) on the learners' writing quality and examines their attitudes and preferences toward them and the clarity of these feedback modes in practice.

2. Review of Literature

2.1. Written Corrective Feedback (WCF)

Written corrective feedback (WCF), according to Sheen (2011), deals with different aspects of writing assessment such as assessing content, organization, rhetoric, linguistic accuracy, and mechanics. Scholars in Second Language Acquisition (SLA) have looked at the concept of corrective feedback from different theoretical perspectives – e.g., the Interaction Hypothesis (Long, 1983, 1996), the Input Hypothesis (Krashen, 1982), the Output Hypothesis (Swain, 1985, 1995), Noticing Hypothesis (Schmidt, 1994, 2001), and Skill-Learning theory (DeKeyser, 1998, 2007). In Sociocultural theory (SCT), there are also several major concepts related to Corrective Feedback and L2 acquisition: Mediation, regulation, internalization, and Zone of Proximal Development (ZPD). SCT researchers claim that different learners may require different types of corrective feedback and there is no ideal type to be applied to all learners. In the WCF typology proposed by Ellis (2009), there are 6 options for teachers to correct the errors: direct, indirect, metalinguistic, focused/unfocused, electronic, and reformulation. Electronic feedback is the one that introduces a new delivery mode that has been in the spotlight in recent years, especially after the outbreak of the pandemic. With the technological developments in recent years, this has transformed into newer aspects of feedback provisions as well as evaluation in which the process is integrated with computer-generated feedback.

2.2. Automated Writing Evaluation (AWE)

Pedagogical computer-based programs for feedback provision known as Automated Writing Evaluation (AWE), is a type of formative, frequent, and process-oriented assessment which emphasizes the active role and engagement of the learners (Zhang & Hyland, 2018). Initially introduced in the 1960s, AWE refers to a network-based teaching system that can contribute to essay evaluation and scoring by providing immediate diagnostic holistic scores and feedback on writing samples. Originally, it was used mainly for high-stakes tests to provide summative scores for assessment. A review of the literature demonstrates that a myriad of studies has been done regarding different modes of providing writing feedback. Using computer-generated feedback, as one of these modes, has garnered the attention of so many scholars and practitioners in recent years. Several studies have hailed the favorable outcome of adopting technology to develop the writing quality and the behavior of the learners (e.g., Burstein et al., 2004; Chen & Cheng, 2008; Crossley et al., 2016; Dikli, 2006; Herrington & Moran, 2001; Li et al., 2015; Liao, 2016; Luo & Liu, 2017; Philips, 2007; Potter & Wilson, 2021; Waer, 2021; Warschauer & Ware, 2006; Zhu et al., 2020) while some others have criticized or felt suspicious regarding the efficiency of technology alone as the scoring instrument for writing (e.g., Anson, 2006; Freitag Ericsson, 2006; Fuet al., 2022; Herrington & Moran, 2012; Huang & Renandya, 2018; Lang et al., 2019; Patterson, 2005). The proponents (e.g., Luo & Liu, 2017; Warschauer & Ware, 2006; Zhu et al., 2020) enumerate such positive factors as improving revision behavior, giving more instruction time to teachers, providing immediate feedback at a suitable moment and in multiple drafting, giving more opportunity for learners to revise the drafts, and fostering learners' autonomy. The opponents (e.g., Fu et al., 2022; Huang & Renandya, 2018; Lang et al., 2019) enlisted other factors as being a disproportionate replacement for the human, offering only unilateral feedback provision, inability to provide a meaningful score, lacking human inferencing skill, adopting only a formalist approach by tapping into the lower-order aspects of writing, and acting more in developing testing strategies rather than writing skill.

In recent years, however, literature tends toward the feedback provided by both computers and humans (Zhang & Hyland, 2018). In this regard, scholars have investigated the efficiency of the provision of these two modes of feedback in the development of writing (sub)skills. To explore the effectiveness of AWE software besides human feedback in facilitating writing, Chen and Cheng (2008) investigated three L2 university writing courses in Taiwan using software called *My Access*. Results showed that regarding idea development AWE feedback performed unsatisfactorily, and human feedback was more useful for students with respect to global items. Therefore, as AWE seemed to be 'limited in its higher-order aspects of writing', Attali et al. (2012, p. 127) proposed a 'division of labor, that is, teacher feedback should focus on the global errors and AWE on the local errors of writing. This also complies with the principles Zamel (1985) proposed for the process-writing pedagogy.

In a quasi-experimental study, Wilson and Czik (2016) assigned language learners to two different conditions, combined (Teacher + PEG Writing) and teacher-only feedback. The results revealed that although the students received the same amount of feedback in both conditions, they received feedback on higher-level writing skills only in the combined feedback condition by teachers. Also, a 16-week study was performed by Zhang and Hyland (2018) on two Chinese students of English investigating the effect of AWE feedback modes on the development of their writing ability. The positive and negative points of teacher-only and AWE feedback types were identified using students' texts in multiple drafts and interviews. Results indicated an almost similar level of student engagement with both feedback types and suggest that the integration of two types of feedback in writing classrooms could be more effective. Furthermore, a quantitative study, conducted by Parra and Calero (2019), investigated the effectiveness of utilizing free AWE tools (*Grammark*® and *Grammarly*®) on writing performance in a Training Program. These tools were implemented as complementary treatments for the teachers' provision of feedback which ultimately led to students' significant improvement in writing. In a multiple case study, Koltovskaia (2020) investigated students' engagement using *Grammarly* automated written corrective feedback (AWCF) while revising their final drafts. Behavioral, cognitive, and affective dimensions of students' engagement were analyzed through the analysis of the students' screencasts (*QuickTime*), and their views were collected as a result of a stimulated recall and semi-structured interview. Findings indicated different levels of engagement. However, both students showed moderate changes in their drafts. The results of a study by Link et al. (2022) indicated that using AWE feedback, as a complement to teacher feedback, had no significant effect on the quantity of higher-level writing feedback by teachers but using teacher-only feedback increased the chance for the students to receive more lower-level feedback from their teacher. However, students who received AWE feedback showed more improvements and retention in writing accuracy.

The literature concerning the effects of various modes of feedback on writing performance (e.g., Chen & Cheng, 2008; Link et al., 2020; Wilson & Czik, 2016; Zhang & Hyland, 2018) has evidenced diverse findings. Also, it was reported in the literature that there is a positive attitude toward the formative assessment and its effectiveness in learning (Ghazizadeh & Motallebzadeh, 2017; Hazim Jawad, 2020); however, little research has been done on the effects of using various modes of feedback (teacher, automated, and blended) on students' writing improvements, learners' perception towards these modes, and the formative nature of feedback modes. Overall, applying an effective method of feedback can offer a more positive learning experience to students but it is still unclear whether and how different modes of feedback can help students' writing development or which mode can be more effective.

This study, therefore, aims to contribute to our understanding of whether and how three modes of feedback can be efficient or can improve Iranian EFL learners' writing accuracy. Moreover, it seeks to understand which mode of feedback the learners prefer and what are the reasons behind these preferences. The present study is guided by these questions:

(1) Is there any significant difference among learners in the three feedback groups with regard to their writing quality in the essay-writing tasks?

(2) How do learners progress in their writing quality across essay-writing tasks in each group?

(3) Is there any significant difference among learners in the three feedback groups with regard to their writing quality in the components of writing (task response, coherence and cohesion, lexical resource, grammatical range and accuracy)?

(4) How do learners feel about the mode of feedback they received in terms of attitude, clarity, and preference?

3. Method

3.1. Research Design

An explanatory sequential mixed-methods research (MMR) design was used to investigate the effectiveness of three modes of feedback in the quantitative phase and then to examine the learners' attitudes and preference towards these modes and their clarity in the qualitative phase. This design is used "when researchers want to further explain their quantitative findings through more in-depth qualitative data and analysis" (Riazi, 2016, p. 114)

3.2. Participants

The participants volunteered to take part in the project based on public notice in social media. They were all IELTS candidates who have planned to take part in the actual IELTS exam for a variety of purposes such as immigration, education, or employment. Out of 55 people who registered for the project, 31 were selected based on the writing proficiency level in their previous performance in IELTS real or mock exam. Having converted their scores to the Common European Framework of Reference, they were from B1 to B2 - from limited user to competent user - based on IELTS official preparation material (e.g., Hashemi & Thomas, 2011). All the participants were Persian native speakers and their ages ranged from 20 to 40. They were randomly divided into three approximately equal groups receiving different feedback modes (teacher, automated, and blended).

3.3. Instruments

There were two instruments for collecting data: *writing tasks* and *a structured interview* and an instrument for rating and providing formative feedback. Six writing task topics were assigned to participants in three groups to write an essay of not less than 250 words. The topics were all opinion essays and the participants were requested to write the next one as they received feedback on the previous one.

The structured interview was formed based on three issues raised in the fourth research question - attitude, clarity, and preference - each of which was inquired by one question. The interview was held online in written mode after completing the writing tasks and receiving and addressing the related feedback. This type of interview was adopted as we were constrained in terms of time (due to the Pandemic), the accessibility of the participants, and the straightforward and efficient analysis of the data. The interview included three written questions inquiring about how they felt about the feedback provided, how the feedback was engaging and clear to them, and whether they prefer the feedback mode(s) they received.

The Virtual Writing Tutor website was also the instrument used for scoring and providing formative feedback. It is a free online website helping writers to score their essays. It has several sections to check grammar, punctuation, and vocabulary, and gives the writer the

CEFR level. Essays are scored based on their type including opinion, argument, film analysis, and literary critique. This study used the criteria for the opinion essay section in which the website provides the user with the score and the formative feedback on four aspects of writing and their subskills separately. The criteria were *writing quality* which rates and gives feedback on the cohesion, dynamism, provocativeness, cliches, and exclamation marks; *essay structure and content* which deals with the organization of the essay and the degree to which the title and the content developed are related; *vocabulary* centers around the topic related words, and the general and academic vocabulary profiles traced within the text; and *language accuracy* which deals with the score and formative feedback which rates the grammatical accuracy of the text. In this study, these factors were considered equivalent to the four assessment criteria in IELTS writing band descriptor: *coherence and cohesion*, *task response*, *lexical resource*, and *grammatical range and accuracy*.

3.4. Data collection procedures and analysis

Data collection employed both quantitative and qualitative methods with the priority given to the former. The data collection consisted of six essay writing tasks and structured written interviews. The quantitative data were used to explore how the students work with different modes of feedback and the qualitative data are collected subsequently to understand students' views of these modes: QUAN (qual). The quantitative phase of the project started with sending a call for participation in social media within the discourse communities of language learners inviting them to be volunteered in writing the essays. We received 55 requests out of which 31 were selected considering the level of proficiency from B1 to B2. They were randomly assigned to three groups where they received three modes of feedback (automated, teacher, blended - teacher + automated). Formative assessment feedback was adopted in this study and the participants were provided with the chance to receive the feedback, do the revisions, and send back the final version. Therefore, having been given the first topic, the participants wrote the first task and sent it back to the researchers. The essays of the students with different modes of feedback were rated and sent back to the participants for revision. The teacher feedback group tasks were rated by an IELTS writing expert and the automated feedback group received the feedback using the *Virtual Writing Tutor* website. This first task was considered as the pretest to measure their initial capability of them at the outset of the project. The same procedure was separately followed for any one of the five topics which totally took around five months. The sixth task was regarded as the post-test of the study. Moreover, the participants' results from tasks 2 to 5 were also collected separately to probe into the effect of different modes of formative feedback on the degree of their gradual development from one task to the next.

In the qualitative phase, the participants were requested to answer a structured written interview inquiring about different aspects of their attitudes toward the feedback they received. The first two groups receiving automated and teacher feedback were to answer a five-item interview while the third group's interview included six items. The extra item here refers to the comparison between the types of feedback since they received both of them.

The data collected from the quantitative and qualitative phases were analyzed to answer the research questions posed at the beginning of the study.

4. Results

4.1 Research Question One

To answer the first research question; that is, "Is there any significant difference among learners in three feedback groups with regard to their writing quality in the essay-writing tasks?", one-way between-groups analysis of variance (ANOVA) was performed on the gain scores of writing ability (the deviation score) in three feedback groups (*blended feedback, teacher feedback, and automated feedback)* after having examined the necessary assumptions. The results showed that feedback modes significantly affected students' writing scores. Table 1 and Table 2 show that the blended feedback group outperformed the automated feedback group, yet no significant difference was founded between the blended feedback group and the teacher feedback group. In addition, the teacher feedback group outdid the automated feedback group.

Table 1

Descriptive Statistics of Writing Gain over Feedback Groups

Group	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
				Lower Bound	Upper Bound
Blended	1.41	1.04	0.31	0.71	2.11
Teacher	1.40	0.70	0.22	0.90	1.90
Automated	0.45	0.60	0.19	0.02	0.88
Total	1.10	0.91	0.16	0.76	1.43

Table 2

Bonferroni Post-hoc Test of Writing Gain over Feedback Groups

Group		Mean Difference	Std. Error	Sig.	95% CI	
					Lower Bound	Upper Bound
Blended	Teacher	.01	.36	1.00	-.90	.91
	Automated	.96	.36	.04	.05	1.86
Teacher	Blended	-.01	.36	1.00	-.91	.90
	Automated	.95	.36	.04	.02	1.88
Automated	Blended	-.96	.36	.04	-1.86	-.05
	Teacher	-.95	.36	.04	-1.88	-.02

4.2 Research Question Two

Regarding the second research question, that is, "how do learners progress in their writing ability across tasks in each group?", three one-way repeated-measures ANOVAs, with one within-group variable of time (topic as a within-subjects variable in this study had six levels representing those six writing tasks) was conducted on three groups' writing performance scores in those six topics.

Before conducting the first repeated-measures ANOVA for the blended feedback group, the sphericity assumption of this test was checked and *Mauchly*'s test of sphericity showed that a violation occurred, X^2 (14) = 26.25, *p* = .02, so we reported Greenhouse-Geisser corrected *df.* One-way repeated-measures ANOVA revealed that time had a large significant effect on students writing performance over those six writing tasks given in terms of topic 1 to topic 6, *F* (2.63, 26.37) = 7.13, *p* = .00, sphericity not assumed, partial eta squared of .41 (Cohen, 1988).

To shed more analytic light on the students' writing performance over time, from the first writing task (topic 1) to the last writing task (topic 6), *Repeated* contrasts were utilized. The *repeated* contrasts were exploited to compare students' writing performance in each task with its next adjacent counterpart. Table 3 shows a large statistically significant difference between students' writing performance just from topic 5 to topic 6, *F* (1, 15) = 20.25, *p* = .00, with partial eta squared of .67 (Cohen, 1988). Nonetheless, there was no statistically significant difference between other pairs of topics.

Table 3
Tests of Repeated and Polynomial Within-subjects Contrasts for blended feedback Group

Source		Type III Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Topic	Level 1 vs. Level 2	.20	1.00	.20	.41	.54	.04
	Level 2 vs. Level 3	.09	1.00	.09	.23	.64	.02
	Level 3 vs. Level 4	1.84	1.00	1.84	2.87	.12	.22
	Level 4 vs. Level 5	.20	1.00	.20	.80	.39	.07
	Level 5 vs. Level 6	7.36	1.00	7.36	20.25	.00	.67

Note: Level stands for topic

Also, prior to conducting the second repeated-measures ANOVA for the teacher feedback group, the sphericity assumption of this test was checked and *Mauchly*'s test of sphericity showed that no violation arose, X^2 (14) = 21.97, *p* = .09. One-way repeated-measures ANOVA revealed that time had a large significant effect on students writing performance in this group over those six writing tasks given from topic 1 to topic 6, *F* (5, 45) = 5.25, *p* = .00, sphericity assumed, partial eta squared of .33 (Cohen, 1988).

As can be seen in Table 4, there was a large statistically significant difference between students' writing performance just from topic 4 to topic 5, *F* (1, 9) = 8.3, *p* = .00, with a partial eta squared of .5 (Cohen, 1988). However, there was no statistically significant difference between other pairs of topics.

Table 4

Tests of Repeated and Polynomial Within-subjects Contrasts for Teacher Feedback Group

Source		Type III Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Topic	Level 1 vs. Level 2	2.0	1.0	2.0	3.2	0.1	0.3
	Level 2 vs. Level 3	0.4	1.0	0.4	1.2	0.3	0.1
	Level 3 vs. Level 4	0.9	1.0	0.9	0.6	0.5	0.1
	Level 4 vs. Level 5	5.6	1.0	5.6	8.3	0.0	0.5
	Level 5 vs. Level 6	0.9	1.0	0.9	0.8	0.4	0.1

Note: Level stands for topic

Pertaining to the last repeated-measures ANOVA for the automated feedback group, the sphericity assumption of this test was again tested and *Mauchly*'s test of sphericity showed that no violation occurred, X^2 (14) = 13.53, *p* = .51. One-way repeated-measures ANOVA showed that time regarding this feedback did not have any significant effect (very small effect as it was manifested in partial eta squared) on students writing performance over those six writing tasks given from topic 1 to topic 6, *F* (5, 45) = 0.88, *p* = .55, sphericity assumed, partial eta squared of .08 (Cohen, 1988).

4.3 Research Question Three

Concerning the third research question; that is, "is there any significant difference among learners in three feedback groups with regard to the writing components?", a three-group MANOVA was used to measure the potential effects of feedback modes on different components of writing. More specifically, this MANOVA was conducted to investigate the effects of different feedbacks on four writing components of *task response, coherence and cohesion, lexical resource, and grammatical range and accuracy*, considered as four main dependent variables (DVs). Having measured those four DVs, those three feedback modes, that is, blended feedback, teacher feedback, and automated feedback (the independent variable with three levels) were compared to see whether they were different with regard to different writing components.

The results of the three-group MAMOVA illustrated that there was a significant difference between three feedback groups, that is, blended feedback, teacher feedback, and automated feedback, on those four components of writing, *F* Wilk's Lambda (8, 50) = .11, *p* = .00, partial eta squared of 0.99, showing a large effect. Hence, it can be argued that feedback mode did have a statistically significant large holistic effect on performance in those four components of writing.

Four univariate *F* tests (embedded in the three-group MANOVA) were utilized to have a better portrait of the feedback group's performance in each of the four DVs separately. Univariate *F* tests (see Table 5) for DVs showed that there were large group differences on three out of four components of writing, that is, *task response, coherence and cohesion,* and *lexical resources*.

Table 5

Tests of Between-Subjects Effects

Source		Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Group	TR	24.04	2	12.02	11.54	.00	.45
	CC	26.56	2	13.28	12.92	.00	.48
	LR	8.79	2	4.39	8.45	.00	.38
	GRA	.43	2	.22	.11	.90	.01
Error	TR	29.17	28	1.04			
	CC	28.78	28	1.03			
	LR	14.55	28	.52			
	GRA	57.94	28	2.07			
Total	TR	1205.50	31				
	CC	1244.50	31				
	LR	1858.25	31				
	GRA	1549.50	31				

Note: TR = Task Response, CC = Coherence and Cohesion, LR = Lexical Resource, GRA = Grammatical Range and Accuracy.

Detailed comparisons of different groups' performance in different components of writing were presented in Table 6. Firstly, regarding *task response*, it can be seen that the teacher feedback group (*M* = 7.30, *SD* = 1.16) outperformed the blended feedback group (*M* =5.86, *SD* = .95) and also this group (teacher feedback group) outdid the automated group (*M* = 5.15, *SD* = 0.94), without observing any other differences. Secondly, pertaining to *coherence and cohesion*, interestingly, both teacher (*M* = 7.30, *SD* = 0.95) and the blended feedback group (*M* = 6.27, *SD* = 0.85) outperformed the automated only group (*M* = 5.00 *SD* = 1.22). With regard to *lexical resource*, a reverse trend was seen given that the automated-only group (*M* = 8.45 *SD* = 0.50) outdid both the teacher (*M* = 7.20, *SD* = 0.79) and blended feedback group (*M* = 7.45, *SD* = 0.82). Also, no other differences could be mentioned in this part.

Table 6

Bonferroni Post-hoc Test of Different Groups' Performance in Different Components of Writing

DV	Feedback		Mean Difference	Std. Error	Sig.	95% CI	
						Lower Bound	Upper Bound
TR	Blended	Teacher	-1.44*	.45	.01	-2.57	-.30
		Automated	.71	.45	.36	-.42	1.85
	Teacher	Blended	1.44*	.45	.01	.30	2.57
		Automated	2.15*	.46	.00	.99	3.31
	Automated	Blended	-.71	.45	.36	-1.85	.42
		Teacher	-2.15*	.46	.00	-3.31	-.99
CC	Blended	Teacher	-1.03	.44	.08	-2.16	.10
		Automated	1.27*	.44	.02	.14	2.40
	Teacher	Blended	1.03	.44	.08	-.10	2.16
		Automated	2.30*	.45	.00	1.15	3.45

	Automated	Blended	-1.27*	.44	.02	-2.40	-.14
		Teacher	-2.30*	.45	.00	-3.45	-1.15
LR	Blended	Teacher	.25	.31	1.00	-.55	1.06
		Automated	-1.00*	.31	.01	-1.80	-.19
	Teacher	Blended	-.25	.31	1.00	-1.06	.55
		Automated	-1.25*	.32	.00	-2.07	-.43
	Automated	Blended	1.00*	.31	.01	.19	1.80
		Teacher	1.25*	.32	.00	.43	2.07
GR	Blended	Teacher	-.28	.63	1.00	-1.88	1.32
		Automated	-.08	.63	1.00	-1.68	1.52
	Teacher	Blended	.28	.63	1.00	-1.32	1.88
		Automated	.20	.64	1.00	-1.44	1.84
	Automated	Blended	.08	.63	1.00	-1.52	1.68
		Teacher	-.20	.64	1.00	-1.84	1.44

Note: * shows a significant difference

4.4 Research Question Four

To address the fourth research question, that is, “How do learners feel about the mode of feedback they received in terms of attitude, clarity, and preference?”, learners’ responses to the structured written interview were analyzed and reported regarding the three aspects – i.e., attitude, clarity, and preference. The learners’ views and preferences towards different modes of feedback were considered as evidence and support of the results of the quantitative phase. In this thematic analysis, we used the bottom-up deductive approach (Riazi, 2016) to analyze the interview data by surveying the transcript of the interviews and extracting the codes and common themes/patterns from the text without having an a priori theoretical perspective. The codes were extracted by two researchers and in case there was a controversy, the third researcher rechecked the coding. Coding was based on three levels of open, axial, and selective coding which are explained in more detail in every section below.

4.4.1 Attitude. The first question of the interview inquires about the attitude of the participants in all three groups. The results indicated a sweeping positive attitude of the learners toward the feedback they received. The teacher feedback group used such modifying words and phrases as *excellent, pleased, positive, enlightening, constructive, and useful* showing a positive attitude. The automated feedback group also revealed their positive feelings by mentioning their feedback as *precise, motivating, fruitful, enjoyable, comprehensive,* and *very detailed and accurate*. However, the blended group, as they had experienced both modes of feedback, mentioned positive and negative aspects of the automated feedback. Out of 69 codes underlined in the open coding level and 11 codes in the axial coding level, three codes were formed in the selective coding level. On the negative side, they mentioned such issues as administrative (e.g., *having lengthy and time-consuming details)*, affective (*[believing that] the comments were interesting at the beginning but a bit mechanical, repetitive, and boring at the end)*, and functional (e.g., *fail[ing] to understand ... task response)*. Nevertheless, on the positive side, the emerged themes showing their attitude were mainly concerned with only affective (e.g., an *amazing experience)* and functional (e.g., *it detected their writing faults accurately using different colors, the marks given by it were not biased, dogmatic, or based on personal judgment,* and *satisfied with regard [to] the vocabulary feedback as it suggested many* *academic terms together with the percentages of accuracy or usage of words as well*). All in all, the feedback modes showed no significant difference in terms of learners' attitudes.

4.4.2. Clarity. The second question of the interview dealt with the extent to which the feedback was clear and sufficient enough for the participants. The majority of learners reported an appropriate level of clarity irrespective of the mode they received. Concerning teacher-only feedback, three themes finally emerged after going through different levels of coding to explain the clarity of the feedback: sufficiency (e.g., *very good and self-explanatory*), efficiency (e.g., *help to organize ideas & support them logically*, and brevity (e.g., *straightforward and to the point and brief*). However, three participants in the teacher feedback group believed that the explanation presented was not sufficient for different reasons such as not assigning a time limit for the tasks, expecting further study notes, and lack of clarity which might mean that the teacher was rather parsimonious in using further explanation in the feedback notes.

In the automated feedback group, they underlined factors of sufficiency (e.g., *it was subtle to most parts*) and efficiency (e.g., *One part that showed great clarity was the grammatical errors and mistakes [in grammatical range and accuracy] part,* and *cohesion and coherence part was not often clear*).

Interestingly enough those who had received two modes of feedback (blended), referred to both factors of sufficiency and efficiency of the provided feedback, considering them as complementary. A few instances of the interview excerpts are below:

Instance for sufficiency

- *I suppose automated feedback is more detailed than teacher-based feedback, however; it is a little complicated and is not as straightforward as teacher-based feedback.*

Instances of efficiency

- *There were positive points in both and I guess the collaboration of them could be better.*
- *Automated feedback was more useful in LR (lexical resource) and GRA (grammatical range and accuracy) while teacher feedback was more practical and individualized with regard to CC (cohesion and coherence) and TR (task response).*

4.4.3. Preference. As for the third interview question, the participants of the blended feedback group were asked if the feedback mode(s) they received were of their preference and would like to receive it in the future. In the blended feedback group, they generally agreed that integrating both is marginally better than receiving either of them. They mentioned a few positive features of the automated feedback mode, such as *providing detailed information*, *setting predetermined criteria*, [being] *easier to administer, providing the solution*, and *good for improving lexical and grammatical proficiency,* and a negative feature, i.e., *complicated and contradictory responses*. Their view toward teacher feedback was also contradictory and can be summarized as *short responses*, *more precise in rating*, *more practical*, *lack of suggested solutions for the errors*, and *good for improving cohesion and coherence and task response*. In conclusion, most learners' preference was an integration of the two modes. These are a few excerpts from the interviews:

- *well, both, as they complement each other*
- *I hope to receive both because I think both are effective.*
- *both were useful in different aspects.*

5. Discussion

This study has attempted to investigate how computer-mediated and teacher-led feedback modes or their integration can be effective in writing ability in general and the components of writing in specific. Besides, the study analyzed the learners' attitudes and preference toward feedback modes and the clarity of the feedback provision.

In response to the first research question, the findings manifested that the blended feedback group outperformed the automated one. While the teacher feedback group did better than the automated one, no significant difference was seen between teacher and blended feedback groups. We speculate that this is partly due to two reasons: first, the traditional teacher-fronted mindset Iranian learners are grown up with in which the teacher is the sole source of knowledge, second, in this study, teachers did not have AWE software at their disposal, otherwise, they might regularly use it and form a more trusting relationship with technology. Such results confirm the findings by Zhang and Hyland (2018) and Wilson and Czik (2016) in terms of the benefits of the combination of both teacher and automated feedback. The result is also supporting the claim by Stevenson and Phakiti (2014) for the well-integration of automated feedback into the classroom. Some studies also reiterated the positive impact of the AWE tool as a supplementary to the teacher commentary in the process of rating (e.g., Chen & Cheng, 2008; Dikli & Bleyle, 2014; Grimes & Warschauer, 2010; Parra & Calero, 2019; Waer, 2021; Ware, 2011; Warschauer, 2010; Weigle, 2013). Furthermore, these findings are in line with those of Huang and Renandya (2018) and Liao (2016) which indicated that the adoption of automated-only feedback may not necessarily result in the improvement of the students' writing quality. Furthermore, in their critical review of research in computer-mediated feedback, Stevenson and Phakiti (2014) claimed that in between-group designs, the computer-mediated feedback group showed either mixed effects (found in some texts, not in others) or no effects on the quality of writing. However, this is not in line with the study by Li et al. (2015) which supported the effectiveness of automated feedback.

Regarding the second research question, it was revealed that time had a large significant effect on students' writing as a significant difference was found between students' writing in the blended feedback group from topics five to six. Moreover, a significant difference was investigated in the teacher feedback group between students' writing from topics four to five. However, there was not seen any significant difference in the automated group. These findings indicate that blended feedback needs a longer time to take effect and enrich the quality of the students' writing while this took a shorter time in teacher feedback. It can be attributed to the degree of the students' adaptability in handling both modes of feedback at the same time.

Respecting the third research question, the results indicated that teacher feedback could significantly improve the students' *task response* compared to the other two groups. While teacher and blended feedback showed a significant effect on improving *cohesion and coherence*, automated feedback could help learners have better *lexical resources*. Furthermore, there was no significant difference among groups with respect to *grammatical range and accuracy*. Regarding the results, it can be inferred that teacher feedback is more useful and clearer in providing explanations on how to write an essay that covers completely all the requirements of the task in comparison with automated one which was likely to be more mechanistic concerning the *task response*. In terms of *cohesion and coherence*, the teacher lies in the heart of both groups as the teacher feedback seems to be more illuminating to improve the quality of writing.

Automated feedback could provide the students with the suggested substitution words which can be the reason behind its improvement among the students. However, this improvement can be contributed to the mechanical use of the lexical resources and expressions suggested by the program rather than developing the quality of their writing skill itself. In other words, there is little known whether the learners have fostered the metacognitive skills to detect the problematic areas and correct them successfully (Stevenson & Phakiti, 2014). This is in line with the result of Luo and Liu (2017), where vocabulary was improved as a result of automated feedback provision of the written text. The literature is, however, mostly rich with the positive influence of the AWE feedback on a number of linguistic components, more specifically grammatical accuracy or the related components such as spelling and punctuation. Waer (2021) found that students performed significantly better in grammatical knowledge when received automated feedback. The study by Crossley et al. (2016) further reported that automated feedback could enhance text cohesion and coherence. Moreover, several other studies (e.g., Dikli, 2006; Milton, 2006; Warschauer & Ware, 2006) also indicated that by receiving too much detailed feedback and explanation provided by the AWE, students can improve learners' grammatical structures and spelling, and achieve greater improvement in writing accuracy.

In terms of the language level, it can be mentioned that teacher-only and blended feedback modes resulted in a significant improvement in the higher-level aspect of IELTS writing (i.e., cohesion & coherence and task response) while AWE feedback could significantly enhance the quality of lower-level aspect of writing (i.e., lexical resource). This confirms the results by Attali et al. (2012), Link et al. (2022), and Hyland and Hyland (2006) but is against Cheville (2004) which concluded that AWE feedback can address content development of writing.

Regarding their attitudes toward feedback modes, students use such modifying words and phrases as *excellent*, *pleased*, *positive and enlightening*, *constructive*, *motivating*, *precise and comprehensive*, … for all three modes of feedback showing that they highly appreciate the act of providing feedback. A large number of studies in the literature underlined the positive perception of the learners towards automated feedback (e.g., Dikli & Bleyle, 2014; Grimes & Warschauer, 2008, 2010; Li et al., 2014; Rich, 2012). However, there are few studies reporting learners' negative attitudes toward AWE feedback (e.g., Chen & Cheng, 2008; Lai, 2010).

With respect to clarity of teacher feedback, three themes of efficiency, sufficiency, and brevity emerged. While participants in the teacher feedback group characterized all three major themes concerning feedback clarity, the automated and integrated feedback groups acknowledged the first two. Learners in the teacher feedback group raised the need for further study notes and lack of clarity which might mean that the teacher was rather parsimonious in using further explanation compared to the long and detailed explanation of automated feedback. The reason is certainly rooted in the workloads and limited time of the teachers.

With regard to students' preference, they generally agreed that having both modes of feedback is marginally better than receiving either of them which indicates the fact that they have found some advantages in each of them. A closer look at the interview data provides evidence that human feedback lends itself well to a higher-order aspect of assessing writing, i.e., coherence, content, and organization, while automated feedback address lower-order sub-skills of writing, i.e., structure and lexical resources. This confirms the results of Link et al., (2022) which demonstrated the development of lower-level skills as a result of teacher-only feedback, and Yang et al. (2006) which revealed that teacher feedback alone is not preferred by the learners.

6. Conclusion

In a traditionally-designed educational context like the one in this study, the teacher, as a human agent in the language class, is considered to be the center of education and source of learning. Therefore, it seems that it is not the right time for purely automated feedback, and integration of teacher and automated feedback is more appealing for this transient period of the shift from traditional to more modern technologically-enhanced assessment. The findings of this study indicated that AWE and teacher feedback integration, not a replacement, might be more effective. It can contribute to teachers and practitioners to fine-tune their feedback and designate more space and time for lesson presentation or to formative support to develop the learners' rhetorical knowledge in writing. Teacher educators and trainers are also advised to develop teacher professional development courses enriched with both AWE and human feedback modes. Furthermore, formative assessment feedback, devoid of the modes of provision, can be a formative tool for teaching, learning, and assessing learners' continuous development even though they may tap to lower- or higher-level aspects of writing development. It is worth mentioning that the improvement in the quality of writing in this study may not be exclusively attributable to the use of either teacher or automated feedback but to the learners' revising skills, instructional factors, and even developmental factors (Stevenson & Phakiti, 2014). Also, the results could be different if other AWE tools had been used since they may focus on different aspects of language. Moreover, the scoring rubrics other than the IELTS task 2 band descriptor might lead to diverse results. Building upon the findings of this study, future research might further explore other aspects of blending both teacher and automated feedback modes in a variety of writing tasks within EFL and/or ESL contexts.

Acknolwedgments

Our sincere appreciation goes to all the participants of the study who voluntarily attended the process of data collection.

Declaration of Conflicting Interests

We declare no potential conflicts of interest with respect to this study.

Funding

The authors received no financial support for this research

References

Anson, C. (2006). Can't touch this: Reflections on the servitude of computers as readers. In P. Freitag Ericsson, & R. Haswell (Eds.), *Machine scoring of student essays* (pp. 38-56). Utah State University Press.

Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125–41.

Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Research Report Series*, *64*(1), i-99. https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing.* Routledge.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine, 25*(3), 27-36.

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

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *The English Journal, 93*(4), 47-52.

Crossley, S.A., Kyle, K. & McNamara, D.S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227–1237.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 1–35.

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.

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal, 63*(2), 97-107.

Ferris, D. (2002). *Treatment of error in second language student writing*. University of Michigan Press.

Ferris, D.R. (2004). The "grammar correction" debate in L2 writing: Where are we, and where do we go from here? (And what do we do in the meantime?). *Journal of Second Language Writing*, *13*(1), 49-62.

Ferris, D. (2010). Second language writing research and written corrective feedback in SLA intersections and practical applications. *Studies in Second Language Acquisition*, *32*(2), 181-201.

Freitag Ericsson, P. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. Freitag Ericsson, & R. Haswell (Eds.), *Machine scoring of student essays* (pp. 28-37). Utah State University Press.

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

Gao, J., & Ma, S. (2022). Instructor feedback on free writing and automated corrective feedback in drills: Intensity and efficacy. *Language Teaching Research, 26*(5), 986-1009.

Ghazizadeh, F., & Motallebzadeh, K. (2017). The impact of diagnostic formative assessment on listening comprehension ability and self-regulation. *International Journal of Language Testing, 7*(2), 178-194.

Grimes, D., & Warschauer, M. (2008). Learning with laptops: A multi-method case study. *Journal of Educational Computing Research, 38*(3), 305–332.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Language, and Assessment, 8* (6), 4-43.

Hazim Jawad, A. (2020). Examination of Iraqi EFL teachers' attitudes, intentions, and practices regarding formative assessment. *International Journal of Language Testing, 10*(2), 145-166.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480-499.

Herrington, A., & Moran, C. (2012). Writing to a machine is not writing at all. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 219-232). Hampton Press.

Huang, S., & Renandya, W. A. (2018). Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in Language Learning and Teaching, 14*(1), 15-26.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101.

Jiang, L., & Yu, S. (2022). Appropriating automated feedback in L2 writing: Experiences of Chinese EFL student writers. *Computer Assisted Language Learning, 35*(7), 1329-1353.

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

Lai, Y. H. (2010). Which do students prefer to evaluate their essays: Peers or computer program? *British Journal of Educational Technology, 41*(3), 432–454

Lang, F., Li, S., & Zhang, S. (2019). Research on reliability and validity of mobile networks-based automated writing evaluation. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC), 10*(1), 18-31.

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.

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1-18.

Liao, H. (2016). Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal, 70*(3), 308-319.

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

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W.R. Ritchie & T. J. Bhatia (Eds.) *Handbook of second language acquisition* (pp. 413-468). Academic Press.

Luo, Y., & Liu, Y. (2017). Comparison between peer feedback and |automated feedback in college English writing: A case study. *Open Journal of Modern Linguistics, 7*(4), 197-215.

Milton, J. (2006). Resource-rich web-based feedback: Helping learners become independent writers. In F. Hyland & K. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 123-139). Cambridge University Press.

Parra G., L., & Calero S., X. (2019). Automated writing evaluation tools in the improvement of writing skills. *International Journal of Instruction, 12*(2), 209-226.

Patterson, N. (2005). Computerized writing assessment: Technology gone wrong. *Voices from the Middle, 13*(2), 56-57.

Pearson, W. S. (2018). Written corrective feedback in IELTS writing task 2: Teachers' priorities, practices, and beliefs. *The Electronic Journal for English as a Second Language, 21*(4), 1-32.

Philips, S. M. (2007). *Automated essay scoring: A literature review.* Society for the Advancement of Excellence Education.

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

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

Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of delivery. *Language Learning & Technology, 26*(1), 1-25.

Riazi, A. M. (2016). *The Routledge encyclopedia of research methods in applied linguistics*. Routledge.

Rich, C. S. (2012). The impact of online automated writing evaluation: A case study from Dalian. *Chinese Journal of Applied Linguistics, 35*(1), 83–99.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129-158.

Shang, H. F. (2022). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments, 30*(1), 4-16.

Sheen, Y. (2011). *Corrective feedback, individual differences, and second language Learning*. Springer

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 125-144). Oxford University Press.

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

Ware, P. (2012). Computer-generated feedback on student writing. *TESOL Quarterly, 45*(4), 769-774.

Warschauer, M. (2010). New tools for teaching writing. *Language Learning & Technology, 14*(1), 3-8.

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

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M.D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions (pp. 35-54)*. Routledge.

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.

Yang, M., Richard, B., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179-200.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19*(1), 79–101.

Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90-102.

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