Teaching English Language Journal
ISSN: 2538-5488 – E-ISSN: 2538-547X – http://tel.journal.org

Please cite this paper as follows:
Ozcelik, S., & Cosgun Ogeyik, M. (2026). The effects of written corrective feedback on L2 writing skills and grammatical accuracy through digital technology. *Teaching English Language, 20*(1), 139-177. https://doi.org/ 10.22132/tel.2025.477858.1691

Research Paper

The Effects of Written Corrective Feedback on L2 Writing Skills and Grammatical Accuracy Through Digital Technology*

Sibel Ozcelik
English Language Instructor, School of Foreign Languages, Istanbul Technical University
Muhlise Cosgun Ogeyik[1]
Prof., Department of English Language Teaching, Faculty of Education, Trakya University

Abstract
This study investigated the effectiveness of indirect written corrective feedback in improving overall writing performance. In this quasi-experimental study, the lessons were carried out on online platforms. Google Classroom, which allows the use of Google documents, was used to assign and collect the weekly writing tasks. The participants were a total of thirty-six Turkish university students enrolled in a General English preparatory class at the School of Foreign Languages at a Turkish University. While the experimental group received indirect feedback through error correction codes, the control group received direct feedback, which is a common and traditional way of providing written corrective feedback. The results indicated that the scores of both groups, regardless of the feedback type, significantly increased from the pre-test to the post-test. However, the findings showed no significant difference between the two groups. It was also seen that the first and the second post-tests were not significantly different in terms of the mean total scores, and both groups retained their gains in the second immediate post-test. Overall, the study results displayed that through written corrective feedback trails, either direct or indirect, the students'

[1] Corresponding author: muhlisecosgun@trakya.edu.tr

writing performance was developed. The findings also indicated that such trails were useful for improving the level of accuracy in writing performance. The study findings may have certain implications for writing courses in foreign language teaching process. It can be suggested that using corrective feedback in writing courses may boost students' enthusiasm for producing texts in the target language.
***Keywords**:* Written Corrective Feedback, Direct and Indirect Feedback, Comprehensive Corrective Feedback

Received: September 10, 2024
Accepted: April 11, 2025

1. Introduction

Second language (L2) writing which is a crucial skill helps students both express the self and communicate with the reader. In effective writing, the organization of ideas or information, the range and accuracy of grammar and vocabulary, and finally creating an appropriate style for the subject matter and the readers are key issues which make writing demanding for learners of English (Hedge, 2000). Comparing writing with speaking, for instance, Hedge (2000) states that in L2 writing, it is not possible to benefit from devices like gestures, facial expressions, and intonation. The writer has to find other ways to express herself or himself correctly. Similarly, considering speech as a more natural and passionate form of communication, Rousseau (2018) emphasizes that writing is a tool used to secure the spoken language.

In writing instruction, a teacher's responsibility is not only to assign a writing task and then collect the papers in and mark them, but also to help learners find topics, get ideas, organize their ideas into a text, get corrective feedback on language use and content, write several drafts, revise and edit their papers along with the given feedback (Scrivener, 2011). Corrective feedback has an important role in terms of increasing motivation and improving linguistic accuracy (Ellis, 2009). Generally speaking, corrective

feedback, whether direct or indirect, has been considered essential for writing development, and students have usually expected and liked to be corrected by their teachers (Hyland & Anan, 2006).

2. Literature Review

Corrective feedback has been discussed and investigated in numerous studies in the relevant literature for deliberating its impacts on writing skills development. Generally, corrective feedback in writing process is provided through two main methods as direct correction feedback (DCF) and indirect correction feedback (ICF) (Sheen & Ellis, 2011). In direct correction, the learner is provided with correct or reformulated forms. In indirect correction, the error is not explicitly corrected; instead, learners are expected to find out the correct form through various interventions such as by highlighting erroneous part, or simply by underlining the erroneous structure, or by using some codes referring to specific errors.

The relevant studies on written corrective feedback mostly examine the effectiveness of WCF on improving the overall text quality. Zohrabi and Khalili (2014) investigated the effectiveness of written corrective feedback on enhancing English for Specific Purposes (ESP) students' writing skills and determined that direct and reformulation-based strategies improved ESP students' writing ability in the short, medium, and long terms, while the indirect strategy was only effective in the short term. Bitchener et al. (2005) concluded that those who received a combination of DCF and individual conference with the teacher performed much better than the others who were not exposed to any types of CF. Van Beuningen et al. (2008; 2012) determined the short-term effects of DCF and ICF on improving accuracy, and only DCF has a significant effect in the long-term. Sheen et al. (2009) examined the differential effects of focused and unfocused CF treatments on

the accuracy of grammatical forms and stated the positive impacts of such treatments on grammatical accuracy over time in all the post-tests. Gholaminia et al. (2014) compared DCF with metalinguistic code correction and specified that the experimental group outperformed the control group who received DCF.

Karim and Nassaji (2018) measured the effects of comprehensive DCF, ICF (underlining only), ICF (underlining and metalinguistic cues) both on revision accuracy and new pieces of writing and concluded that the experimental groups performed better than the control group in revision tasks. Suzuki et al. (2019) investigated the interactional effects of direct and indirect WCF explicitness on revision accuracy and new pieces of writing and concluded that both types of WCF helped learners improve revision accuracy of the target structures in the short-term. Recently, some researchers have broadened the scope of their studies, and measured the effectiveness of WCF on new pieces of writing in the long term besides revision accuracy (Karim & Nassaji, 2019). In addition to the mentioned variables affecting corrective feedback, the role of corrective feedback timing is also important for getting positive impacts of the feedback trails (Javan Ameni, Sadeghi, & Alavinia, 2024; Shooshtari, Bayizade, & Jalilifar, 2024).

Although the majority of the research confirms the effectiveness of WCF, some studies claim the ineffective positions of WCF in improving accuracy and complexity. Truscott (1996, 2004, 2007) has repeatedly claimed that written corrective feedback is ineffective in developing writing accuracy. Truscott's (1996) other strong claim was that corrective feedback is harmful since it demotivates learners by increasing their stress levels. Similarly, Al-Rubai'ey and Nassaji (2013) who carried out research with two intact EFL classes who were provided with DCF and indirect metalinguistic CF reported that no significant differences across the two treatment groups were observed.

Liu (2008) conducted a quasi-experimental study to investigate the effectiveness of error feedback and emphasized that although direct correction was effective for reducing the errors in the immediate draft, there was no improvement in accuracy in a different text; Lui (2008) suggested that some mini-lessons or workshops would be necessary to improve the accuracy level and students' self-editing abilities.

As indicated in the relevant literature above, the research that has been carried out so far is not conclusive. While some studies favor direct feedback, others claim that indirect feedback has better results, especially in the long term. Many beneficial or impractical aspects of WFC may be set forth depending on the diverse nature of the applied groups regarding their linguistic and age levels, their cultural diversity, their personality, use of strategies and so on. Therefore, in this study, it was aimed to test how indirect feedback works on online platforms through distance education.

The study seeks answers to following research questions:

RQ 1: Is written corrective feedback effective in improving students' overall L2 writing performance?

RQ 2: Which corrective feedback is better? Direct or indirect written corrective feedback?

3. Methodology

The present study aims to examine the effectiveness of indirect written corrective feedback in improving overall writing performance of the participants on online platforms.

3.1 Research Method

In this study, a quasi-experimental study was designed with two groups of participants. Convenient sampling was used to conduct the research. While the experimental group received indirect written corrective feedback, the

control group received direct written feedback, which is a common and traditional way of providing written corrective feedback.

3.2 Participants

The participants in the present study were thirty-six university students enrolled in a General English preparatory class at a Turkish university. They were assigned into experimental and control groups through convenient sampling by the researchers. In the experimental group, there were 5 male and 13 female participants whose ages ranged from 19 to 23. On the other hand, there were 9 male and 9 female participants in the control group, and their ages were between 19 and 22. All participants in both groups had to take The Oxford Placement Test which is a computer-adaptive test for non-native speakers of English when they began their preparatory classes. According to the placement test scores, they were at the A2 level as defined in the Common European Framework of Reference for Languages. The participant details are displayed in Table 1.

Table 1.
Detailed information about the participants

Participants	Age	Male	Female	Level	L1
Experimental Group	19-23	5	13	A2	Turkish
Control Group	19-22	9	9	A2	Turkish

To see if the test scores differed across groups, a t-test was run, considering that the residuals of the placement test data were normally distributed ($SW = .044$, $df = 36$, $p > .05$) as suggested by Kéry and Hatfield (2003) and Field (2018). The results of the t-test are presented below.

Table 2.
Placement Test Comparison

Group	n	M	SD	T	df	p
1	18	28.44	5.27	-1.48	34	.148
2	18	31.22	5.97			

As seen in the table, the mean placement score of control group was 28.44 (SD = 5.27) and that of experimental group was 31.22 (SD = 5.97). The t-test results showed that the difference between the mean placement scores of the feedback groups was not statistically significant ($t = -1.48$, $df = 34$, $p > .05$).

Table 3.
Comparison of Pre-test Scores

Variable	Group	*n*	*M*	*SD*	*Mdn*	*t/Z*	*df*	*P*
Task Achievement*	1	18	2.44	0.68	2.50	-1.454	-	.171
	2	18	2.75	0.35	2.75			
Coherence & Cohesion	1	18	2.56	0.48	2.50	-0.297	34	.768
	2	18	2.61	0.63	2.75			
Vocabulary	1	18	3.08	0.67	3.00	-0.147	34	.884
	2	18	3.11	0.47	3.00			
Grammar	1	18	2.69	0.84	2.50	-0.071	34	.716
	2	18	2.61	0.47	2.50			
Total	1	18	10.78	1.90	11.00	-0.361	34	.605
	2	18	11.08	1.60	11.00			

*: Mann-Whitney U (Z Score Reported)

As displayed in the table, no statistically significant difference was observed among the feedback groups in terms of the pre-test scores ($p > .05$).

3.3 Data Collection Tools

In this study, pre and post-tests were administered to collect the quantitative data. While the experimental group was given indirect written corrective feedback through error correction codes, the control group was given direct written corrective feedback during the treatment process which lasted seven weeks. As the first data collection instrument, the participants in both experimental and control groups were given a pre-test to measure their existing writing performance. The same test was later administered to all participants in both groups as a post-test after the treatment.

A writing rubric developed by the institution was used for the evaluation of the participants' writing papers. The writing rubric included four main categories: Task Achievement; Coherence / Cohesion; Range and Accuracy of Vocabulary for Purpose; Range and Accuracy of Grammar for Purpose. Each category in the rubric was graded out of 5 points, which made a total score of 20 points. The total number of points was then converted into a score out of 100. The pre-tests and post-tests were assessed twice, first by the researchers and then by another colleague to make sure that the results were accurate and consistent in order to ensure reliability. Therefore, it can be suggested that the scoring rubric helped reduce subjectivity (Moscal & Leydens, 2000).

In addition, a follow-up post-test that the participants were asked to write on a different topic was implemented. The aim was to collect the data that allows the researchers to compare the two post-test results and see the short-term transfer effects of the corrective feedback treatment on a new piece of writing. According to Larsen-Hall (2016), Cronbach's Alpha reliability coefficient should be used in cases where the aim is to see if data coming from different judges agree or not. For this reason, total scores and each analytical item within the rubric were matched for Rater 1 and Rater 2 to

calculate a Cronbach's Alpha value for each pair. The results are given below.

Table 4.
Interrater Reliability Analyses

Pretest	αPretest	αPosttest	αPosttest2
Task Achievement	.915	.946	.956
Coherence & Cohesion	.887	.929	.959
Vocabulary	.952	.904	.923
Grammar	.941	.939	.950
Total	.955	.971	.978

As shown in the table, the alpha values obtained in the analysis were between .89 and .98, indicating reliable scoring for all parts of the rubric as well as the total scores.

3.4 Procedure

In this quasi-experimental study, the lessons were carried out with the help of online platforms. Google Classroom, which allows the use of Google documents, was used to assign and collect the weekly writing tasks. The same platform was also used to provide feedback for the writing tasks. The treatment period was completed in seven weeks. During the first week, the participants in both groups were asked to write an opinion paragraph which was evaluated as the pre-test to see their initial writing performance. When selecting writing topics, it was important to keep in mind that the topics should not have been too specialized or require much research, but they should have been the ones about which all participants may have some ideas. The purpose of the pre-test was to determine if the participants could form an opinion paragraph by first expressing their opinions with a topic sentence and then supporting their ideas with details and examples, as well as provide

coherence and cohesion in the paragraph by using a variety of cohesive devices such as linking words, sequencers, referencing, and so on. Subsequently, papers from both groups were scored by the researchers and another colleague. However, the participants were not provided with the feedback immediately after the pre-test was conducted in order to familiarize them with the feedback procedure and the Error Correction Codes.

The participants in both groups had reading lessons in addition to writing lessons. Each week, a total of six hours were allocated for reading and writing lessons. The participants read two reading texts relating to the topic on which they were supposed to write each week. Q-Skills for Success Reading and Writing 2 (3rd ed.) by Oxford University Press was used for the reading lessons. The writing topics were therefore created in accordance with the themes of the reading book's units. As Hirvela (2004) emphasized, reading is closely connected to writing as it provides a rich source of input in terms of the organizational pattern, language, and style, all of which help generate a written text. In this study, it was aimed to enrich students' knowledge of vocabulary through the reading texts. Besides, it was a stage that started students' thinking processes on the related writing topic. In addition to reading texts, supplementary writing materials were used to teach how to form an opinion paragraph and draw attention to the important points to consider while writing. During the treatment, only the opinion paragraph genre was addressed.

In Week 1, the Error Correction Codes were introduced to the experimental group, and the participants were informed about the way the feedback procedure was going to take place. As the use of Error Correction Codes as a way of providing written corrective feedback was completely new for the participants, the instructor elaborated on what each code referred to in the following weeks to make sure that they were all clearly understood. In the

same week, after the completion of reading and writing lessons, the first writing task was assigned to both groups, and the participants completed the task during the lesson under the supervision of the instructor.

In Week 2, the instructor returned the students' tasks from Week 1. The experimental group had indirect feedback through codes, and their errors were also highlighted in the text, as illustrated by (1). On the other hand, the control group had direct feedback and their errors were explicitly corrected by the instructor, as illustrated by (2).

(1) I think before the pandemic, we are **(VT)** so free and fearless.

(2) Before the pandemic, nobody ~~wears~~ **wore /would wear** masks because we ~~don't need to~~ **didn't need to**.

In addition to the feedback for grammatical and lexical errors, the instructor gave some overall feedback for organizational aspects of writing like task achievement and coherence and cohesion in the margins or at the bottom of the text. The participants in both groups were given feedback on paragraph structure, expanding supporting points, or building coherence and cohesion, as illustrated by (3), (4), and (5).

(3) Before going to the job interview, firstly you should research about the company.
You should start your paragraph with a topic sentence.

(4) For example, technological inventions make people lazy.
Why and how? You should expand your idea by giving examples.

(5) *Well- done! It is good to see that you expand your ideas by giving some specific examples. And you use linking words effectively. Please check the use of 'because' again.*

The participants in both groups were then asked to review their papers with corrections, and they were given 30 minutes to write their second drafts during the lesson.

At the beginning of the study, the experimental group had difficulty while correcting their errors on their own, and they needed guidance, especially for their lexical errors. Therefore, both experimental and control groups were taught how to use online dictionaries to find out the correct forms of the words and some collocations. In the following weeks, the same procedure for each writing task was implemented. In total, the participants submitted six writing tasks, except for the pre and post-tests. For all the tasks, they were required to write the first and second drafts. The instructor administered the same instructional materials and conditions to both groups.

In Week 6, the first post-test was administered. In addition, in the final week, the participants took a writing exam, and the writing exam results were used as the second post-test with a different topic. Both post-tests were carried out under the supervision of the instructor. The participants' activities during the exam were logged, and before the evaluation of the writing exam, the authenticity of the students' papers was checked through Turnitin, a plagiarism detection software.

3.5 Data Analysis

For data analysis, the assumptions of each parametric test were checked initially. However, a number of studies show that regression models such as ANOVA's, ANCOVA's or t-tests are robust to minor violations of those assumptions (e.g. Blanca, et al., 2017; Oljenik & Algina, 1984; Shielzeth et al., 2020). Even so, parametric results were reported along with their

nonparametric counterparts in the case of assumption violations to avoid erroneous conclusions as recommended by Mizumoto and Takeuchi (2018).

Comparing the scores related to the whole group regardless of the feedback type they received necessitated a paired-samples t-test (i.e. dependent t-test). A paired-samples t-test assumes that the differences between pre and post scores, not the scores themselves, are normally distributed (Field, 2018). Therefore, the differences were checked for distribution normality initially.

Table 5.
Normality Tests for Differences in the Paired Comparisons

Score	*SW*	*df*	*p*
Total	.969	36	.387
Task Achievement	.925	36	**.018**
Coherence & Cohesion	.920	36	**.012**
Vocabulary	.929	36	**.023**
Grammar	.950	36	**.000**

As seen in the table, Shapiro-Wilk Tests indicated that the paired differences were normally distributed only in the total scores ($p > .05$). In Task Achievement, Coherence & Cohesion, Vocabulary and Grammar, the distributions of the differences were seen to be non-normal ($p < .05$). To check the severity of deviations from normality, Q-Q plots (Figures 1, 2, 3, 4) were investigated.

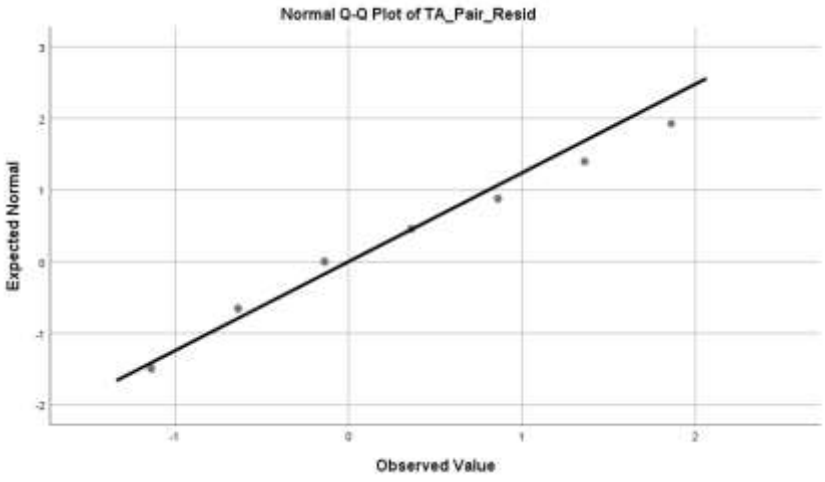

Figure 1. *Paired Residuals of Task Achievement*

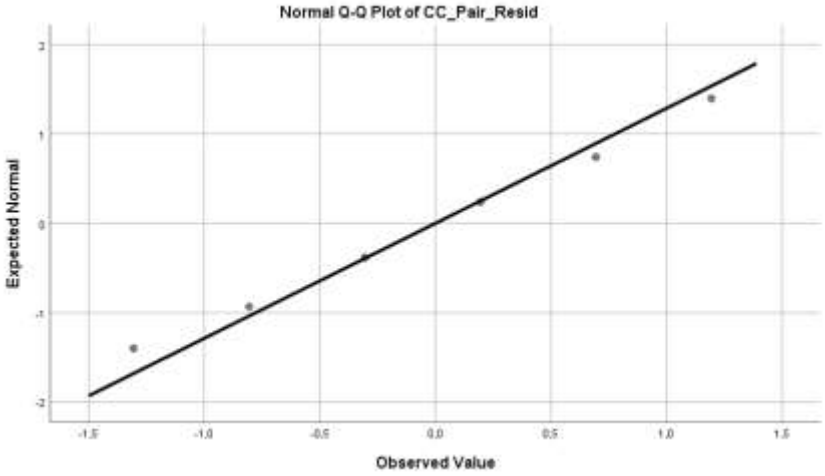

Figure 2. *Paired Residuals of Coherence & Cohesion*

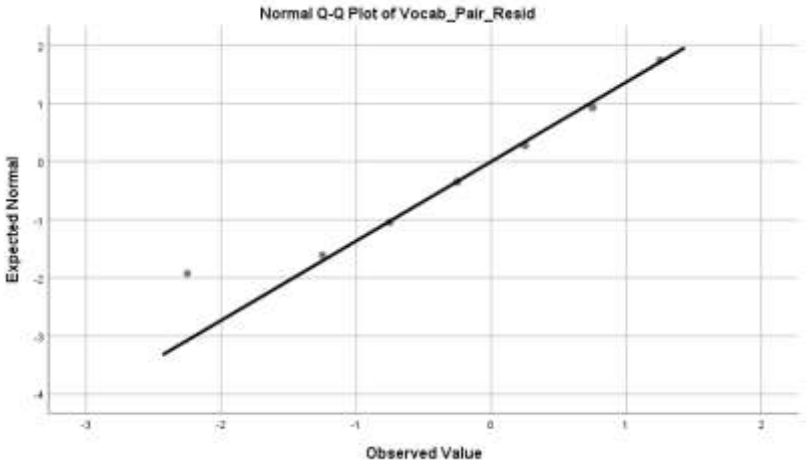

Figure 3. *Paired Residuals of Vocabulary*

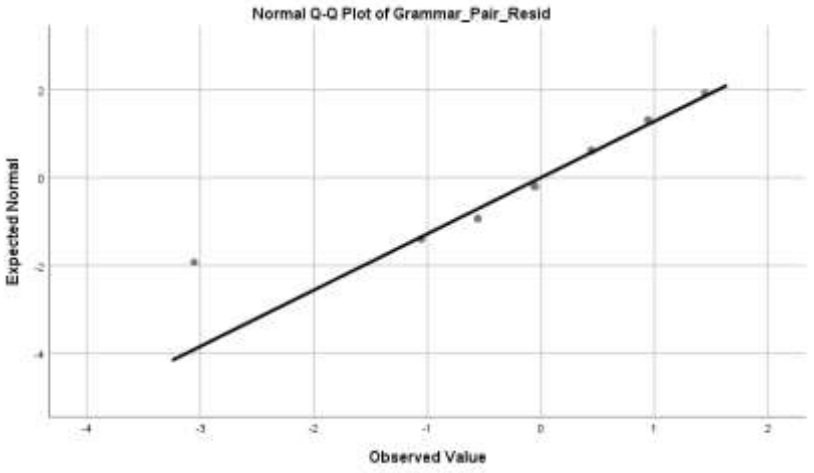

Figure 4. *Paired Residuals of Grammar*

As seen in the figures, the distributions approximated normality, which was sufficient for most parametric analyses. However, a visible outlier was present in Vocabulary and Grammar scores. Although removing outliers is common practice in educational science and applied linguistics, Nicklin and Plonsky (2020) warn against "blindly removing" (p. 26) them unless there is a confirmable error in measurement since such practice results in the loss of legitimate data even though winsorizing and log e transformation produce valid results without any loss of legitimate data points. Also taking into account the robustness of regression models to deviations from normality, parametric results were reported and compared with their nonparametric alternatives for those variables as suggested by Mizumoto and Takeuchi (2018) without any data removal since there was no evidence that the outlier was due to erroneous measurement.

Since the comparison of the pre-test and post-test scores for each written corrective feedback group necessitated an ANCOVA model, the assumptions of ANCOVA, namely approximate normality in residuals, variance homogeneity, homogeneity of regression slopes, linearity and absence of a difference in the pretest scores were initially tested (Field, 2018; Kéry & Hatfield).

3.5.1 Normality of Residual Distributions

Due to its particular power in smaller samples (e.g. $n > 50$), Shapiro-Wilk tests were run to test the residual distribution normality assumption (Larson-Hall, 2016; Ricci, 2005). The results are presented below in different tables due to the large number of variables investigated.

Table 6.
Normality Tests for Residuals in the ANCOVA Model

Score	*SW*	*df*	*p*
Task Achievement	.895	36	**.003**
Coherence & Cohesion	.953	36	.130
Vocabulary	.960	36	.219
Grammar	.973	36	.519
Total	.960	36	.220

As shown in the table, the residuals of the Coherence & Cohesion, Vocabulary, Grammar and Total scores were normally distributed in the ANCOVA model ($p > .05$). However, the residuals of Task Achievement were seen to be non-normally distributed ($p < .05$). Nonetheless, an investigation of the Q-Q plots for the residuals of Task Achievement showed that the distribution approximated normality (Figure 5), which met the requirement for an ANCOVA model. Even so, the analysis regarding Task Achievement were performed both parametrically and non-parametrically in order to check if the results confirmed one another.

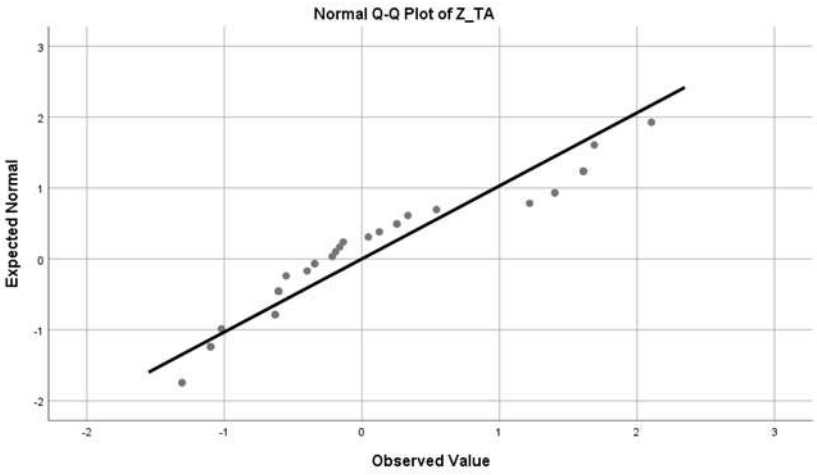

Figure 5. *Residual Distribution of Task Achievement*

3.5.2 Variance Homogeneity

The F test results regarding the assumption of variance homogeneity are given below.

Table 7.
Variance Homogeneity Tests

Score	F	df1	df2	p
Task Achievement	1.776	1	34	.191
Coherence & Cohesion	1.012	1	34	.321
Vocabulary	2.271	1	34	.141
Grammar	.233	1	34	.632
Total	.725	1	34	.400

As seen in the tables, all the F tests produced non-significant p values, indicating variance homogeneity in all ANCOVA models tested in this thesis ($p > .05$).

3.5.3 Homogeneity of Regression Slopes

According to Larsen-Hall (2016), one of the ways to test the homogeneity of regression slopes in ANCOVA models is to test the interaction effect between the grouping variable and the covariate. The results of the interaction tests are presented below:

Table 8.
Regression Slopes Homogeneity Tests

Score	*F*	*df*	*p*
Task Achievement	0.253	1	.504
Coherence & Cohesion	0.666	1	.268
Vocabulary	0.019	1	.829
Grammar	0.166	1	.519
Total	0.624	1	.731

According to the results, the interaction effects between the grouping variable and each covariate were not statistically significant, indicating homogeneity of regression slopes for all ANCOVA models built in this study ($p > .05$).

3.5.4 Linearity

Deviations from linearity between covariates and dependent variables were also tested to see if they met the assumptions of ANCOVA. The results are tabulated below:

Table 9.
Linearity Tests

Score	*F*	*df*	*p*
Task Achievement	0.687	4	.606
Coherence & Cohesion	0.436	4	.729
Vocabulary	0.710	4	.098
Grammar	3.216	4	**.026**
Total	0.382	4	.820

As seen in the results, only the Grammar scores significantly deviated from linearity in terms of the relationship between the dependent variable and the covariate ($p < .05$) For this reason, non-parametric analyses were also conducted for this variable, too.

3.5.5 Absence of Differences between Pre-test Scores

To test the last assumption of ANCOVA, the pre-test scores of the written corrective feedback groups were compared using t-tests or Mann-Whitney U tests depending on the residual distribution of the variables. The results are tabulated below:

Table 10.
Comparison of Pre-test Scores

Variable	Group	*n*	*M*	*SD*	*Mdn*	*t/Z*	*df*	*P*
Task Achievement*	1	18	2.44	0.68	2.50	-1.454	-	.171
	2	18	2.75	0.35	2.75			
Coherence & Cohesion	1	18	2.56	0.48	2.50	-0.297	34	.768
	2	18	2.61	0.63	2.75			

Vocabulary	1	18	3.08	0.67	3.00	-0.147	34	.884
	2	18	3.11	0.47	3.00			
Grammar	1	18	2.69	0.84	2.50	-0.071	34	.716
	2	18	2.61	0.47	2.50			
Total	1	18	10.78	1.90	11.00	-0.361	34	.605
	2	18	11.08	1.60	11.00			

*: Mann-Whitney U (Z Score Reported)

As seen in the table, no statistically significant difference was observed among the feedback groups in terms of the pretest scores ($p > .05$).

Based on the assumption tests, it was seen that ANCOVA models could be used to test the effectiveness of the written corrective feedback intervention. However, since Task Achievement had a slight deviation from normality and Grammar had a slight deviation from linearity, these variables were analyzed in both parametric and non-parametric terms to see if the presence of outliers in the data sets caused different results. Even though ANCOVA models are generally robust to minor deviations from normality as well as linearity and Type I error rate is not seriously affected in minor deviations (Oljenik & Algina, 1984), nonparametric ANCOVA models are known to be stronger than parametric ANCOVA models in serious deviations, producing fewer Type I errors (Rheinheimer & Penfield, 2001). For this reason, the Task Achievement and Grammar variables were analysed by both ANCOVA and Quade's Test, which is one of the non-parametric alternatives to ANCOVA (Cangür, Şungur & Ankaralı, 2018; Oljenik & Algina, 1985).

In order to compare the first post-test and the second one with a different topic, their differences were initially computed and checked for normality as suggested by Field (2018). The results are given below:

Table 11.
Normality Tests for the Differences in the Paired Post-test Comparisons

Score	*SW*	*df*	*p*
Total	.945	36	.072
Task Achievement	.896	36	**.003**
Coherence & Cohesion	.902	36	**.004**
Vocabulary	.886	36	**.001**
Grammar	.938	36	**.045**

As seen in the table, only the total score residuals were normally distributed in the paired posttest comparisons ($p > .05$). The residuals of Task Achievement, Coherence & Cohesion, Vocabulary and Grammar scores were non-normally distributed ($p < .05$). The severity of deviations from normality for those variables were checked on their Q-Q plots (Figures 6, 7, 8, 9).

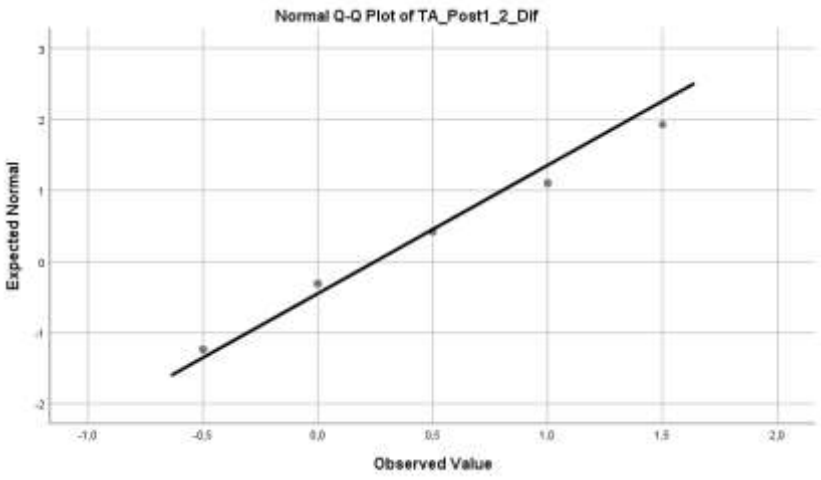

Figure 6. *Paired Residuals of Task Achievement Post-tests*

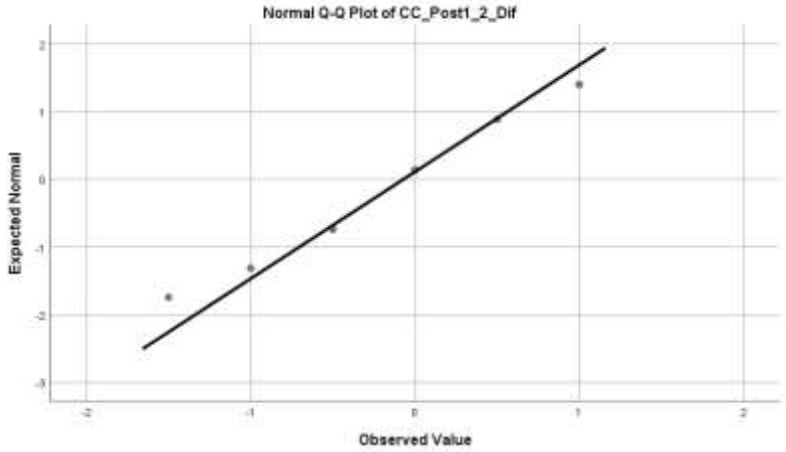

Figure 7. *Paired Residuals of Coherence & Cohesion Post-tests*

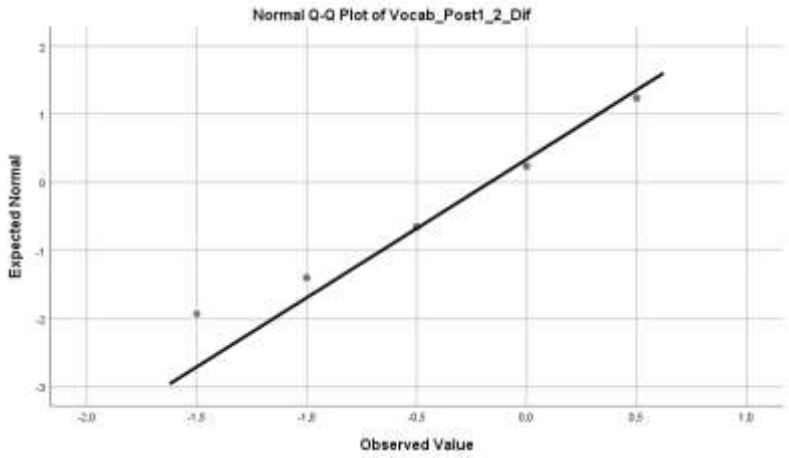

Figure 8. *Paired Residuals of Vocabulary Post-tests*

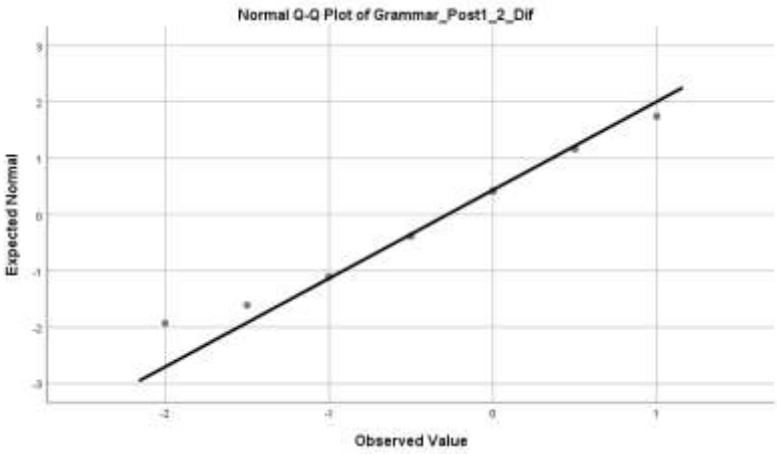

Figure 9. *Paired Residuals of Grammar Post-tests*

The investigation of the Q-Q plots, as seen in the figures, showed the distributions in the post-test residuals approximated normality. Despite the existence of a few outliers in the data, they were kept as is due to the absence of evidence regarding measurement error (Nicklin & Plonsky, 2020). Therefore, paired samples t-tests were run on the data and reported along with Wilcoxon Signed Rank Test results, or their non-parametric alternatives (Field, 2018; Mizumoto & Takeuchi, 2018).

3.5.6 Pre-test and Post-test Comparisons of the Whole Group

The pre-test and post-test results of the whole group of participants, regardless of the type of feedback they received, were compared initially. The results are presented below.

Table 12.

Pre-test and Post-test Comparisons of the Whole Group (N = 36)

Variable	$M_{pretest}$	$SD_{pretest}$	$M_{posttest}$	$SD_{posttest}$	*t*	*df*	*p*	*D*
Total	10.93	1.74	15.18	2.46	-11.039	35	**.000**	2.00
Task Achievement	2.60	0.56	3.74	0.75	-8.465	35	**.000**	1.72
Coherence & Cohesion	2.58	0.55	3.89	0.75	-10.080	35	**.000**	1.99
Vocabulary	3.10	0.56	3.85	0.62	-6.148	35	**.000**	1.27
Grammar	2.65	0.67	3.71	0.65	-8.102	35	**.000**	1.61

According to the paired samples t-test results, there were statistically significant differences with large effects in all pretest and posttest results ($p < .001$), $d > 0.80$). Since the differences between the pretest and posttest scores of Task Achievement (Z = -4,968, $p < .001$, r = 0.59), Coherence & Cohesion (Z = -4,899, $p < .001$, r = 0.58), Vocabulary (Z = -4,531, $p < .001$, r = 0.53) and Grammar (Z = -4,531, $p < .001$, r = 0.53) scores slightly deviated from normality, Wilcoxon Signed Rank Tests were also run on these variables and it was seen that the non-parametric results were in complete agreement with the parametric ones producing statistically significant results with large effects. Therefore, it was concluded that the scores related to all variables significantly increased from the pretest to the posttest.

3.5.7 Pre-test and Post-test Results Based on Feedback Groups

The participants were also divided by their respective feedback groups and compared by means of ANOVA's and Quade's tests to see if one group outperformed the other or not. The descriptive results regarding the scores divided by feedback groups are presented below.

Table 13.
Pre-test and Post-test Results Based on Feedback Groups

Variable	Group	*N*	$M_{pretest}$	$SD_{pretest}$	$Mdn_{pretest}$	$M_{posttest}$	$SD_{posttest}$	$Mdn_{posttest}$
Total	1	181	10.78	1.90	11.00	14.75	2.49	13.50
	2	181	11.08	1.60	11.00	15.61	2.41	15.50
Task Achievement	1	181	2.44	0.68	2.50	3.58	0.69	3.50
	2	181	2.75	0.35	2.75	3.89	0.80	3.75
Coherence & Cohesion	1	181	2.56	0.48	2.50	3.81	0.75	4.00
	2	181	2.61	0.63	2.75	3.97	0.76	4.00
Vocabulary	1	181	3.08	0.65	3.00	3.78	0.71	3.50
	2	181	3.11	0.47	3.00	3.92	0.52	4.00
Grammar	1	181	2.69	0.84	2.50	3.58	0.62	3.50
	2	181	2.61	0.47	2.50	3.83	0.66	4.00

An investigation of the pre-test scores divided by feedback groups showed that slight differences were present in the pre-test scores related to all the variables studied. However, it was previously confirmed in Table 11 as a part of ANCOVA assumption checks that none of those differences were statistically significant ($p > .05$). Similarly, slight differences were visible in the post-test scores of two feedback groups. Descriptive results showed that the mean post-test scores of Group 2 were slightly higher than those of Group 1 in all variables, by 0.86 point in Total scores, 0.31 point in Task Achievement, 0.16 point in Coherence & Cohesion, 0.14 point in Vocabulary

and 0.25 point in Grammar. In order to see if these differences were statistically significant, covariance analyses were run for each of the variables. The results are given below in their separate respective tables.

Table 14.
Group Comparison for Total Scores (ANCOVA)

Source	*Type III SS*	*df*	*MS*	*F*	*p*	η_p^2
Corrected Model	44.069	2	22.035	4.354	**.021**	0.21
Intercept	64.020	1	64.020	12.65	**.001**	0.28
Pretest	37.396	1	37.396	7.389	**.010**	0.18
Feedback Group	4.111	1	4.111	0.812	.374	0.02
Error	167.007	33	5.061			
Total	8507.25	36				
Corrected Total	211.076	35				

R^2 = .21, Adjusted R^2 = .16

ANCOVA results indicated that the Total pre-test scores had a statistically significant effect on the total post-test scores with a large effect ($F_{(1, 33)}$ = 7.389, $p < .05$, η_p^2 = 0.18). When the total pre-test scores were controlled for, it was seen that there was no difference in the Total post-test scores of Group 1 and Group 2 ($F_{(1, 33)}$ = 0.812, $p > .05$, η_p^2 = 0.02).

Table 15.
Group Comparison for Task Achievement Scores (ANCOVA)

Source	*Type III SS*	*df*	*MS*	*F*	*p*	η_p^2
Corrected Model	1.784	2	0.892	1.639	.209	0.09
Intercept	12.396	1	12.396	22.778	**.000**	0.41
Pretest	0.944	1	0.944	1.735	.197	0.05

Feedback Group	0.373	1	0.373	0.686	.413	0.02
Error	17.959	33	0.544			
Total	522.25	36				
Corrected Total	19.743	35				

R^2 = .09, Adjusted R^2 = .04

In Task Achievement, it was seen that the pre-test did not have a statistically significant effect on the post-test scores ($F_{(1, 33)}$ = 1.735, p > .05, η_p^2 = 0.05). The post-test scores of the feedback groups were not found to be significantly different, either ($F_{(1, 33)}$ = 0.686, p > .05, η_p^2 = 0.02). Since the residuals of Task Achievement scores slightly deviated from normality, a Quade's Test was also run for confirmatory purposes and it was seen that nonparametric results confirmed their parametric counterparts regarding the absence of a statistically significant difference between the Task Achievement post-test scores of Group 1 and Group 2 ($F_{(1, 34)}$ = 0.548, p = .464).

Table 16.
Group Comparison for Coherence & Cohesion Scores (ANCOVA)

Source	*Type III SS*	*Df*	*MS*	*F*	*p*	η_p^2
Corrected Model	2.139	2	1.069	2.026	.148	0.11
Intercept	12.099	1	12.099	22.923	**.000**	0.41
Pretest	1.889	1	1.889	3.578	.067	0.10
Feedback Group	0.184	1	0.184	0.350	.558	0.01
Error	17.417	33	0.528			
Total	564.000	36				
Corrected Total	19.556	35				

R^2 = .11, Adjusted R^2 = .06

The analysis of Coherence & Cohesion scores revealed that the pre-test scores did not have a significant effect on the post-test scores ($F_{(1, 33)} = 3.578$, $p > .05$, $\eta_p^2 = 0.10$). The post-test scores did not significantly differ between feedback groups, either ($F_{(1, 33)} = 0.350$, $p > .05$, $\eta_p^2 = 0.01$).

Table 17.
Group Comparison for Vocabulary Scores (ANCOVA)

Source	*Type III SS*	*Df*	*MS*	*F*	*p*	η_p^2
Corrected Model	0.867	2	0.434	1.141	.332	0.07
Intercept	10.357	1	10.357	27.25	**.000**	0.45
Pretest	0.694	1	0.694	1.825	.186	0.05
Feedback Group	0.156	1	0.156	0.412	.526	0.01
Error	12.542	33	0.380			
Total	546.25	36				
Corrected Total	13.41	35				

$R^2 = .07$, Adjusted $R^2 = .01$

According to the results, the effect of the Vocabulary pre-test scores on the post-test scores were not statistically significant ($F_{(1, 33)} = 1.825$, $p > .05$, $\eta_p^2 = 0.05$). Furthermore, the differences in the Vocabulary post-test scores were not significant between the feedback groups ($F_{(1, 33)} = 0.412$, $p > .05$, $\eta_p^2 = 0.01$).

Table 18.
Group Comparison for Grammar Scores (ANCOVA)

Source	*Type III SS*	*df*	*MS*	*F*	*p*	η_p^2
Corrected Model	2.011	2	1.006	2.618	.088	0.14
Intercept	17.900	1	17.900	46.600	**.000**	0.59
Pretest	1.449	1	1.449	3.772	.061	0.10

Feedback Group	0.679	1	0.679	1.767	.193	0.05
Error	12.676	33	0.384			
Total	509.750	36				
Corrected Total	14.687	35				

R^2 = .14, Adjusted R^2 = .09

In Grammar scores, it was seen that the pre-test did not have a statistically significant effect on the post-test (F(1, 33) = 3.772, p > .05, ηp2 = 0.10). The post-test scores did not differ according to the feedback group after controlling for the pre-test, either (F(1, 33) = 1.767, p > .05, ηp2 = 0.05). Since the Grammar scores deviated from linearity, the results were compared with the Quade's Test results, which also confirmed that there was no statistically significant difference between Group 1 and Group 2 in the Grammar post-test scores (F(1, 34) = 1.370, p = .250).

3.5.8 Comparisons of Post-test Scores

Apart from the post-test which was identical to the pre-test, a second post-test which was of the same genre with a different writing prompt was also collected from the participants. The pairwise comparisons related to these two post-tests are presented below. Both parametric and nonparametric pairwise comparison results were reported for Task Achievement, Coherence & Cohesion, Vocabulary and Grammar due to the deviations from normality in their pre-test and post-test differences.

Table 19.
Post-test 1 and Post-test 2 Comparisons (N = 36)

Variable	$M_{posttest1}$	$SD_{posttest1}$	$M_{posttest2}$	$SD_{posttest2}$	*t*	*df*	*p*	*D*

Total	15.18	2.46	14.92	2.29	0.940	35	.354	0.11
Task Achievement	3.74	0.75	3.99	0.74	-2.707	35	**.010**	0.34
Coherence & Cohesion	3.89	0.75	3.82	0.72	0.657	35	.515	0.10
Vocabulary	3.85	0.62	3.68	0.51	2.029	35	.050	0.30
Grammar	3.70	0.65	3.43	0.62	2.615	35	**.013**	0.43

As seen in the table, the first and the second post-tests were not significantly different in terms of the mean Total scores ($t = 0.940$, $df = 35$, $p > .05$) and the mean scores of Coherence & Cohesion ($t = 0.657$, $df = 35$, $p > .05$, $Z = -0.459$, $p = .646$, $r = 0.05$) and Vocabulary ($t = 2.029$, $df = 35$, $p = .050$, $Z = -1.949$, $p = .051$, $r = 0.23$). On the other hand, Task Achievement ($t = -2.707$, $df = 35$, $p = .01$, $Z = -2.588$, $p = .010$, $r = 0.31$) and Grammar ($t = 2.615$, $df = 35$, $p < .05$, $Z = -2.023$, $p = .043$, $r = 0.24$) scores were significantly different in the first and the second post-tests. It was seen that both parametric and nonparametric analyses yielded the same conclusions regarding the pairwise comparisons.

4. Results and Discussion

In order to seek legitimate responses to the first research question, the results of the pretest and posttest were compared. Initially, the pre-test and post-test results of the whole group of participants, regardless of the type of feedback they received, were compared. As a result, the scores of the whole group of participants related to all variables including Task Achievement, Coherence and Cohesion, Grammar and Vocabulary significantly increased from the pre-test to the post-test. Therefore, the findings demonstrated that both direct

and indirect written corrective feedback seem to have facilitated improvement in revision task.

ANCOVA results showed that the total pre-test scores had a statistically significant effect on the total post-test scores with a large effect. When the total pre-test scores were reviewed, it was seen that there was no difference in the total post-test scores of the control group and the experimental group. Moreover, it was seen that similar results were achieved for all variables. It could be argued that the small sample size might be the main reason for this result. In addition, this result could partly be due to the fact that the corrections were provided via an online platform in distance education context, and the participants who received direct correction were able to access their first drafts with corrections while revising their texts. For the participants, the possibility of copying and pasting the corrections would have been prevented. If the participants had been allowed to review their corrections only for a short period of time, the results might have been different and some differences between the feedback groups may have been observed.

Moreover, this research was completed in a relatively short time period, which lasted seven weeks. None of the participants in the experimental group had received indirect written corrective feedback with error correction codes before. Therefore, the indirect corrective feedback group might need more training and time to make sense of error correction codes (Ferris, 2004). In conclusion, this result adds to the previous research which found both types of written corrective feedback effective in developing revision accuracy (e.g. Ferris & Roberts, 2001; Van Beuningen et al., 2008; Van Beuningen et al., 2012; Al-Rubai'ey and Nassaji, 2013; Lopez et al., 2018; Karim & Nassaji, 2018; Suzuki et al., 2019).

The scores for task achievement and grammar were significantly different in the first and the second post-tests. It revealed that both parametric and nonparametric analyses yielded the same results with regard to the pairwise comparisons. The mean total scores showed that the whole group of participants retained their gains in the second post-test, which was on a new piece of writing. However, significant differences between the first and second post-tests in task achievement and grammar should be noted. While the participants performed better in achieving the task in the second post-test, they did not perform well in grammar as they did in the first post-test. The task effect may be the reason why the participants performed better in achieving the task in the second post-test. Moreover, since the second post-test scores were gathered from their writing quiz results, the low score in grammar may be associated with exam anxiety. Through the final weeks of the module, an apparent decrease in the participants' motivation was also observed.

The results of the present indicated that both direct and indirect written corrective feedback had positive effects in improving the students' overall writing performance both in revision task and in new piece of writing. In this respect, the study results are consistent with the findings of the study by Al-Rubai'ey and Nassaji (2013) who reported that no significant differences were observed across the two treatment groups who were exposed to direct and indirect corrective feedback. The findings of the present study are also in line with the study results by Suzuki et al. (2019) who concluded that both types of corrective feedback helped learners improve revision accuracy of the target structures in the short-term. The research findings demonstrated that written corrective feedback was effective in facilitating accuracy development both in revision and in new pieces of writing; those findings are

in line with the other study findings, for instance, by Van Beuningen et al., (2012), Lopez et al., (2018), and Suzuki et al., (2019).

On the other hand, the results of this present study in terms of the efficiency of direct and indirect corrective feedback in both experimental and control groups are not consistent with the other studies of the researchers who claimed that only experimental groups benefitted from the corrective feedback: for instance, Sheen et al. (2009), Van Beuningen et al. (2012), Gholaminia et al. (2014), Karim and Nassaji (2018) concluded that the experimental groups in their studies improved in grammatical accuracy over time in all the post-tests.

To sum up, the present study demonstrated that the students' writing performance were developed through written corrective feedback, either direct or indirect, due to its positive effects in improving students' overall writing performance. The distinctive nature of writing leads to challenges in learning and teaching writing. Writing does not only rely on the mechanics of language, but also has a creative side. As regards the research findings, it may be suggested that corrective feedback trails can be designed in writing courses for improving the level of accuracy in writing; and before designing the course, it is essential to take learners' preferences and expectations into consideration, as it may influence their success and motivation (Storch, 2010; Faraji et al., 2023).

5. Conclusion

The present study investigated the effectiveness of indirect written corrective feedback on students' overall writing performance. It was seen that, regardless of the feedback type, there were statistically significant differences with large effects in all pre-test and post-test results. The scores related to all variables significantly increased from the pre-test to the post-test. Some

conclusions can be drawn from the study findings: there were no significant effects of indirect written corrective feedback given through error correction codes, as both feedback groups improved almost to the same degree in the revision of the same text; in the immediate post-test, it was seen that the students in both groups could transfer their gains to a new piece of writing; using error correction codes might not be effective for all kinds of errors because the students had difficulty while dealing with some error correction codes, especially when the meaning is unclear. A combination of direct and indirect written corrective feedback types could be more useful for students. In conclusion, the type of written corrective feedback is not the only factor to consider regarding the development of writing performance. The study has some limitations: this research was completed in a relatively short time period, which lasted seven weeks; the participants in the experimental group were not exposed indirect written corrective feedback with error correction codes before; and the study was administered in an online learning context and the participants were not familiar with the distance education context; and before designing the course students' opinions and preferences should have been determined.

Note

* This study is derived from the master's thesis titled "*The Effects of Written Corrective Feedback on L2 Writing Skill and Grammatical Accuracy*" completed by Sibel Ozcelik under the supervision of Prof. Dr. Muhlise Cosgun Ogeyik at Trakya University in 2022.

References

Al-Rubai'ey, F., & Nassaji, H. (2013). Direct and indirect metalinguistic feedback: A matter of suitability rather than superiority. In M. Mahmoud, & R. Al Mahrooqi (Eds.), *Issues in TEFL in the Arab World* (pp. 203-231). Muscat, Oman: Sultan Qaboos University Press.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*, 191-205.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema, 29*(4), 552-557.

Cangür, Ş., Sungur, M., & Ankaralı, H. (2018). The Methods Used in Nonparametric Covariance Analysis. *Duzce Medical Journal, 20*(1), 1-6. doi:10.18678/dtfd.424774

Ellis, R. (2009). Corrective Feedback and Teacher Development. *L2 Journal, 1*(1), 2-18.

Faraji, M. N., Dowlatabadi, H. R., Moelaie, B., Rahmanpanah, R. (2023). Investigating ESP Needs Analysis in Iranian Universities: The Case of Social Sciences Students. Teaching English Language, 17(2), 103-133.

Ferris, D. (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*, 49-62.

Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*, 161-184.

Field, A. (2018). *Discovery statistics using IBM SPSS Statistics.* London, UK: Sage Publications LTD.

Gholaminia, I., Gholaminia, A., & Marzban, A. (2014). An investigation of meta-linguistic corrective feedback in writing performance. *Procedia - Social and Behavioral Sciences, 116*, 316-320.

Hartshorn, K. J., Evans, N. W., Merril, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of Dynamic Corrective Feedback on ESL Writing Accuracy. *TESOL Quarterly, 44*(1), 84-109.

Hedge, T. (2000). *Teaching and Learning in the Language Classroom.* New York: Oxford University Press.

Hirvela, A. R. (2004). *Connecting Reading & Writing in Second Language Writing Instruction.* Ann Arbor, MI: The University of Michigan Press.

Hyland, K., & Anan, E. (2006). Teachers’ perceptions of error: The effects of first language and experience. *System, 34*(4), 509-519.

Javan Amani, Z., Sadeghi, K., & Alavinia, P. (2024). Combined effect of the phase of explicit instruction and feedback timing on task performance and grammar acquisition. Teaching English Language, 18(2), 69-92.

Karim, K., & Nassaji, H. (2018). The revision and transfer effects of direct and indirect comprehensive corrective feedback on ESL students' writing. *Language Teaching Research, 24*(4), 519-539.

Karim, K., & Nassaji, H. (2019). The effects of written corrective feedback: a critical synthesis of past and present research. *Instructed Second Language Acquisition, 3*(1), 28-52.

Kéry, M., & Hatfield, J. S. (2003). Normality of raw data in general linear models: The most widespread myth in statistics. *Bulletin of the Ecological Society of America, 84*(2), 92-94.

Larson-Hall, J. (2016). *A guide to doing research in second language acquisition with SPSS and R.* New York, NY: Routledge.

Liu, Y. (2008). The Effects of Error Feedback in Second Language Writing. *Arizona Working Papers in SLA & Teaching, 15*, 65-79.

Lopez, M. B., Steendam, E. V., Speelman, D., & Buyse, K. (2018). The Differential Effects of Comprehensive Feedback Forms in the Second Language Writing Class. *Language Learning, 68*(3), 813-850.

Mizumoto, A., & Takeuchi, O. (2018). Modeling a prototypical use of language learning strategies: Decision tree-based methods in multiple contexts. In R. L. Oxford, & C. M. Amerstorfer (Eds.), *Language Learning strategies and individual learner characteristics: Situation strategy use in diverse contexts* (pp. 99-122). New York, NY: Bloomsbury.

Moscal, B. M., & Leydens, J. A. (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research and Evaluation, 7*(10). doi:https://doi.org/10.7275/q7rm-gg74

Nicklin , C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics, 40*, 26-55. doi:https://doi.org/10.1017/S0267190520000057

Olejnik, S. F., & Algina , J. (1984). Parametric ANCOVA and the rank transform ANCOVA When the data are conditionally non-normal and heteroscedastic. *Journal of Behavioral and Educational Statistics, 9*(2), 129-149.

Olejnik, S. F., & Algina, J. (1985). A Review of Nonparametric Alternatives To Analysis of Covariance. *Evaluation Review, 9*(1), 51-83. doi:10.1177/0193841X8500900104

Ricci, V. (2005). *Fitting Distributions with R. Vienna: R Foundation for Statistical Computing.* Retrieved from https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf

Rousseau, J. -J. (2018). Essay on the Origin of Languages. In J. Rousseau, & V. Gourevitch (Ed.), *Rousseau: The Discourses and Other Early Political Writings* (pp. 257-310). Cambridge: Cambridge University Press.

Scrivener, J. (2011). *Learning Teaching: The Essential Guide to English Language Teaching.* Oxford, UK: Macmillan Education.

Sheen, Y., & Ellis, R. (2011). Corrective Feedback in Language Teaching. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning.* New York: Routledge.

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*, 556-569.

Shielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*(9), 1141-1152.

Shooshtari, Z., Bavizade, Kh., & Jalilifar, A. R. (2024). Incorporating feedback timing into learning and retention of vocabulary and morpho-syntax: Focus on Iranian EFL learners in a computer-mediated communication context. Teaching English Language, 18(1), 59-97.

Suzuki, W., Nassaji, H., & Sato, K. (2019). The effects of feedback explicitness and type of target structure on accuracy in revision and new pieces of writing. *System, 81*, 135-145.

Truscott, J. (1996). The Case Against Grammar Correction in L2 Writing Classes. *Language Learning, 46*(2), 327-369.

Truscott, J. (2004). Evidence and conjecture on the effects of correction: a response to Chandler. *Journal of Second Language Writing, 13*, 337-343.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*, 255-272.

Van Beuningen, C. G., De Jong, N. H., & Kuiken , F. (2012). Evidence on the Effectiveness of Comprehensive Error Correction in Second Language Writing. *Language Learning, 62*, 1-41.

Van Beuningen, C. G., De Jong, N. H., & Kuiken, F. (2008). The Effect of Direct and Indirect Corrective Feedback on L2 Learners' Written Accuracy. *International Journal of Applied Linguistics, 156*, 279-296.

ZamelStorch, N. (2010). Critical feedback on written corrective feedback research. *IJES, 10*, 29-46.

Zohrabi, M., & Khalili, A., (2024). A cross-cultural study into the utility of diverse written corrective feedback strategies in medicine students' ESP writing courses. *International Journal of Society, Culture & Language, 12(2), 150-169.* https://doi.org/10.22034/ijscl.2024.2025600.3436