# Application of Chat-GPT in Iranian EFL Learners' Writing Complexity, Accuracy, and Fluency

**Shiva Farmoudi[1], Masoud Yazdanimogaddam[2*], Ali Rabi[3]**

[1]Ph.D. Candidate, Department of English, ST.C., Islamic Azad University, Tehran, Iran
* Professor, Department of English, Ga. C., Islamic Azad University, Gramsar, Iran
*Massood.yazdanimoghaddam@gmail.com*
[3]Assistant Professor, Department of English, ST.C., Islamic Azad University, Tehran, Iran
*Ali.rabi22@gmail.com*

**Abstract**
Due to the importance and prevalence of the application of ChatGPT in various domains, the current study set out to investigate the effect of feedback provided by this AI on the complexity, accuracy, and fluency components of EFL learners' writing. To this end, 60 undergraduate EFL students were selected via convenience sampling and were homogenized using the Oxford Proficiency Test. They did a writing as the pretest and were assigned randomly to one experimental group (n=30) and one control group (n=30). In the intervention phase, the experimental group was engaged in writing activities, using Chat-GPT over 10 weeks, with each session lasting for 60 minutes. The participants interacted with Chat-GPT to receive feedback for writing, grammar practice exercises, and vocabulary expansion activities designed to enhance their complexity, accuracy, and fluency in written English. The control group received the usual instruction. In the posttest, both groups did another writing. Paired samples t-tests were used to find out the within-group difference between the pretest and posttest, and independent samples t-tests were used to find out the between-group difference between the control and the experimental groups. The result showed that there was a significant difference both in the writing group and between groups in complexity, accuracy, and fluency in the writing of EFL learners due to using ChatGPT. The findings can have implications for EFL learners, teachers, and curriculum developers.
**Keywords:** Accuracy, ChatGPT, Complexity, EFL writing, Fluency

## INTRODUCTION

The emergence of advanced generative artificial intelligence, particularly Chat-GPT (Generative Pretrained Transformer), has ignited a wave of interest, debates, and controversies around educational environments, especially in the field of education and writing, which is an inextricable academic requirement (Arif et al., 2023; Chen, 2023; Lim et al., 2023). Chat-GPT is defined as "... a machine-learning system that independently learns from data and can generate complex and seemingly intelligent writing after being trained on an extensive dataset of text" (van Dis et al., 2023, p. 224). Advanced by Open-AI, it is a natural language processing model featuring a human-like conversational AI interface referred to as a chatbot (George & George, 2023; Tlili et al., 2023). With this chatbot, learners can submit text instructions, known as prompts or queries, and the application rapidly generates text responses using its pre-trained large data corpus (Pavlik, 2023). By the end of January 2023, ChatGPT had achieved the record for the fastest-growing user base of any program, reaching over 100 million active users within just two months of its launch in late November 2022 (Haleem et al., 2023).

However, despite the acknowledged human-like capabilities of Chat-GPT, it has raised significant concerns regarding ethical implications (Chen, 2023; Liebrenz et al., 2023; Salvagno et al., 2023), particularly when utilized for composing research papers. One of the reasons for this concern is that Chat-GPT has considerable effects on writing (Halaweh, 2023; Kumar, 2023), especially regarding issues of writing complexity, accuracy, and fluency (CAF), which is integral to writing quality in any

field, with EFL as no exception. Unfortunately, the relatively recent emergence of Chat-GPT technology means that limited research has been conducted on its deployment in the CAF context, notably in the realm of academic writing. While some articles have addressed Chat-GPT, a review of the literature focusing on its application in education reveals that only a few tackle the issues related to academic papers (Rudolph et al., 2023). Therefore, the relationships of AI with writing`s complexity, accuracy, and fluency are not well comprehended. Engaging in discussions and debates around this topic is strongly encouraged to enhance our understanding of how this new technology may influence the writing landscape, particularly regarding the controversies linked to its application in this area.

Additionally, while some individuals argue that this groundbreaking technology may ultimately cause more harm than benefits in academic circles, others contend that Chat-GPT could enhance writing productivity if applied wisely and ethically (van Dis et al., 2023). It is evident that Chat-GPT is a permanent fixture in the writing adventure, making it crucial to assess and discuss the value of this innovative technology along with its various implications according to van Dis et al. to validate its worth. Therefore, this paper aims to contribute to the ongoing discussions. It is posited that, although there are numerous challenges and concerns to address, Chat-GPT and its potentials in improving writing complexity, accuracy, and fluency (Halaweh, 2023; Chen, 2023) remains a worthy concern to be considered. This awareness can improve the awareness of the users that they need to adopt this new technology with care, while firmly upholding academic integrity, honesty, and transparency.

**Writing Complexity, Accuracy, Fluency**

Since the 1990s, these three concepts have been significantly presented in L2 studies, often occurring together, primarily as dependent variables. This means they are attributes of L2 learners' performance evaluated to examine the impact of various factors. These elements have been assessed to explore the influence of age on L2 achievement (Alghizzi, 2017; Nosratinia & Razavi, 2016), the effects of different types of instructions (Fathi & Rahimi, 2021; Teng & Huang, 2022), individual differences (Wigglesworth & Storch, 2009), as well as the implications of the learning environment or task design (Soleimani et al., 2017). In recent years, the cognitive shift in L2 research has led to CAF being investigated as a significant point in its own right (e.g., Lennon 2000; Ortega, 2003; Riggenbach 2000; Robinson, 2001; Segalowitz 2000; Towell & Dewaele, 2005). In several of these studies, CAF is revealed as the main byproduct of the psycholinguistic processes and mechanisms that underlie the learning, representation, and processing of L2 systems.

In conclusion, this varied collection of studies reveals that complexity, accuracy, and fluency are recognizable elements of L2 proficiency and performance, which can manifest differently depending on various L2 usage conditions, and may be developed in distinct ways by different types of learners, under various learning environments, and even different types of modern technologies.

**Research Questions**

1. Does Chat-GPT-generated feedback have any significant effect on Iranian EFL learners` writing complexity?
2. Does Chat-GPT-generated feedback have any significant effect on Iranian EFL learners` writing accuracy?
3. Does Chat-GPT-generated feedback have any significant effect on Iranian EFL learners` writing fluency?
4. Is there a significant difference between the experimental and control groups in terms of complexity?
5. Is there any significant difference between the experimental and control groups in terms of accuracy?
6. Is there any significant difference between the experimental and control groups in terms of fluency?

**Method**

**participants**

The participants in this study consisted of 60 undergraduate male and female English learners, aged between 20 and 30 years. They were all Iranian residents of Qom and Tehran. They were selected from a pool of 75 EFL students based on their performance on the Quick Oxford Placement Test (QOPT), which ascertained their English proficiency level to be intermediate. Participants were selected using a convenience sampling method due to practical constraints and logistical challenges that limited the use of other sampling techniques (Emerson, 2021; Golzar et al., 2022).

**Desing**

The design used in this study was a quasi-experimental pretest/posttest in which the participants were selected via convenience sampling and were assigned to an experimental and a control group randomly. The independent variable was the feedback provided by ChatGPT, and the dependent variables were three components of complexity, accuracy, and fluency in EFL learners' writing.

**Data Collection Procedures**

In the course of conducting the current study, the following steps were taken. The students took a placement test on general English proficiency. All 75 students were asked to conduct a Quick Oxford Placement Test (QOPT), which assessed their general knowledge on vocabulary and grammar. Based on their performance, 60 students who scored one standard deviation above and below the mean were chosen as the participants of this study. In the next phase, they were divided into one experimental group (n = 30) and one control group (n = 30). The process of assigning to these groups was random. Specifically, the following steps were taken for data collection.

1. Pre-test Assessment: Before the start of the study, all participants completed a pre-test assessment to be used as the benchmark to analyze the complexity, accuracy, and fluency as the main variables in the current study. The topic was 'Twenty thousand people are killed in road accidents each year. What are the causes of this national tragedy in Iran?'

2. Intervention Phase: The experimental group was engaged in writing activities, using Chat-GPT over weeks (60 minutes, 10 sessions). The participants interacted with Chat-GPT to receive prompts for writing, grammar practice exercises, and vocabulary expansion activities designed to enhance their complexity, accuracy, and fluency in written English. The experimental group was prepared through a short training program on using ChatGPT in the writing classroom. Yuan and Ellis (2002) have recommended training for learners before using a digital tool in the classroom, and before Chat-GPT is used for generating ideas and samples. The participants in the experimental group were provided with a topic for each session to write and revise or edit with the help of ChatGPT. They were informed that each writing task had to be around 300 words, and the whole activity was supposed to take place in around 60 minutes. These requirements were presented to them as the general guidelines rather than strict rules, so that they could feel comfortable completing them

3. The control group participants were engaged in similar writing activities as those in the experimental group, but without access to Chat-GPT or any other AI assistance as part of their mainstream assignments.

4. The participants' writings in the pretest were analyzed by employing the Profile of Larsen-Freeman (2006) to estimate their writing complexity, accuracy, and fluency.

5. Post-Test Assessment: After 10 weeks of the intervention period, all participants from both groups completed a post-test assessment similar to the pre-test assessment tasks.

5. Essays were analyzed for the dependent variables. To assess the quality of the participants' written output, measures of fluency, complexity, and accuracy were utilized. These measurements were primarily consistent with those applied in previous research in the literature (e.g., Ellis & Yuan, 2004; Storch, 2005; Wigglesworth & Storch, 2009). For the purpose of this analysis, all written materials were initially coded for T-units and clauses. To assess complexity and accuracy, the essays needed to be examined for clauses, differentiating between independent and dependent clauses. Some researchers argue over the methodology for coding clauses, especially dependent clauses. In this research, a

dependent clause was identified as containing either a finite or non-finite verb along with at least one additional clause element, such as a subject, object, complement, or adverbial.

Another aspect evaluated was complexity, which indicated the writers' readiness to engage with and utilize a variety of syntactic structures, progressing from basic coordination to more intricate forms that involved subordination and embedding (Storch, 2007). Complexity was quantified by the ratio of clauses to T-units, a reliable metric according to Foster and Skehan (1996), which correlates effectively with other complexity measures. Additionally, complexity was examined through the ratio of dependent clauses to total clauses (DC/C), which assesses the level of embedding present in the text (Wolfe-Quintero et al., 1998).

In accordance with Wigglesworth and Storch (2009), fluency was assessed by calculating the average number of words, T-units, and clauses per text. To assess accuracy, two metrics were utilized: the ratio of error-free T-units to total T-units (EFT/T) and the ratio of error-free clauses to the overall number of clauses (EFC/C). These ratios were presented as percentages in line with Wigglesworth and Storch. Errors addressed in this study comprised syntactical mistakes (such as issues with word order and missing components) and morphological errors (including verb tense, subject-verb agreement, incorrect usage of articles and prepositions, and errors in word forms). Lexical errors (errors in word choice) were considered only when they hindered understanding. All spelling and punctuation errors were taken into consideration.

**Data Analysis**

An independent sample t-test was used to find out the between-group differences between the control and the experimental group in the pre-test and post-test. To find out within-group differences, paired samples t-tests were used for the experimental group for the dependent variables of complexity, accuracy, and fluency to compare their mean scores in the pre-test and post-test. Because a number of separate analyses were used, it is suggested that we set a higher alpha level to reduce the chance of a Type 1 error (i.e., finding a significant result when there isn't really one). The most common way of doing this is to apply what is known as a Bonferroni adjustment. In its simplest form, this involves dividing your original alpha level of .05 by the number of analyses that you intend to do (see Tabachnick & Fidell, 2001, p. 349). In this case, we have three dependent variables to investigate; therefore, we would divide .05 by 3, giving a new alpha level of .017. We will consider the results significant only if the probability value (Sig.) is less than .017.

**RESULTS**

Before reporting the main statistics, all 75 participants were asked to sit a Quick Oxford Placement Test (QOPT), which assessed their general knowledge on vocabulary and grammar. Based on their performance, 60 students who scored one standard deviation below and one standard deviation above the mean were selected as the participants in the current study. In the next part, the participants were assigned to two groups of 30 participants for the experimental group to receive the treatment and 30 for the control group to have their mainstream, usual education. The result of the proficiency test is presented in Table 1.

Table 1

*Descriptive Statistics for the Homogeneity of the Participants*

| | **N** | **Min** | **Max** | **Mean** | **SD** | **Variance** | **Skewness** | | **Kurtosis** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Statistic** | **Statistic** | **Statistic** | **Statistic** | **Statistic** | **Statistic** | **Statistic** | **Std. Error** | **Statistic** | **Std. Error** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| OPT | 75 | 20.00 | 50.00 | 38.29 | 5.947 | 35.37 | -.74 | .27 | .44 | .54 |
| Valid N (listwise) | 75 | | | | | | | | | |

As can be seen in Table 1, the minimum score was 20 and the maximum is 50, with the mean score of 38.29. The standard deviation is 5.94. The value of skewness is -.74, which falls between -1.96 and +1.96, and for kurtosis, it is .27, which falls between -2 and +2, indicating that the data is normally distributed.

**Within-Group Differences between Pretest and Posttest**

The first research question was 'Does Chat-GPT-generated feedback have any significant effect on Iranian EFL learners' writing complexity?' To answer this question, a paired sample t-test was run to see if there was a statistically significant within-group difference between the pretest and posttest of the experimental group for the complexity component in their writing. For this purpose, the number of t-units was determined and used as the final score for each writing task. The result is presented in Table 2.

Table 2

*Descriptive Statistics of the T-unit Related to Complexity in Pre-and Posttest*

| | N | Minimum | Maximum | Mean | Std. Deviation | Variance | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Complexity pretest | 30 | 10.00 | 17.00 | 13.16 | 1.662 | 2.76 | .296 | .42 | -.300 | .83 |
| Complexity posttest | 30 | 11.00 | 17.00 | 14.86 | 1.502 | 2.25 | -.542 | .42 | -.070 | .83 |
| Valid N (listwise) | 30 | | | | | | | | | |

According to the data analysis presented in Table 2, the mean number of t-units for the pretest was 13.16 with a standard deviation of 1.66, and for the posttest, it was 14.86 with a standard deviation of 1.50. The skewness for the pretest was .29, which is between -1.96 and +1.96, and the kurtosis was -.3, which is between -2 and +2. For the post test, the skewness was -.54, falling between -1.96 and +1.96, and the kurtosis was -.7, falling between -2 and +2. It indicated that both pretest and posttest met the normal distribution, despite negative values for skewness and kurtosis in the posttest, which could be attributed to the effect of the intervention. The result of the paired sample t-test is presented in Table 3

**Table 3**

*Paired Samples Test for the Experimental Group in Pre- and Posttest for Complexity*

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | Complexity pretest – complexity posttest | -1.700 | 2.479 | .452 | -2.625 | -.774 | -3.75 | 29 | .001 |

According to the result presented in Table 3, the result of the paired samples t-test showed that there was a statistically significant decrease in complexity components scores from pre-treatment (M=13.16, SD=1.66) to post-treatment (M=14.86, SD=1.50, t (29) =3.75, p = .001. Therefore, it can be stated that the first null hypothesis is rejected because ChatGPT did affect the writing of EFL learners in terms of improving their complexity, measured in terms of the t-unit.

Eta square was calculated via t squared/t t-squared /t plus the number of participants minus one. The value turned out to be 14/0625/43/0625 = .32. To interpret the eta squared values, the following guidelines can be used (Cohen, 1988): .01=small effect, .06=moderate effect, .14=large effect. Given our eta squared value of .32, it can be concluded that there was a large effect, with a substantial difference in the Fear of Statistics scores obtained before and after the intervention.

The second research question stated, 'Does Chat-GPT-generated feedback have any significant effect on Iranian EFL learners` writing accuracy?' For this purpose, the writing samples were assessed in terms of accuracy, which included lexical, grammatical, and punctuation accuracy. To quantify this part, each case of inaccuracy was tabulated. To see the impact of the feedback provided by Chat-GPT, the mean scores of these mistakes in the writing samples of the experimental group before and after the treatment were calculated. The result of the descriptive statistic is presented in Table 4.

Table 4

*Descriptive Statistics on the Accuracy of Writing on Pre-test and Post-test*

| | N | Minimum | Maximum | Mean | Std. Deviation | Variance | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Accuracy pretest | 30 | 22.00 | 42.00 | 31.80 | 5.040 | 25.40 | .114 | .42 | -.841 | .83 |
| Accuracy posttest | 30 | 20.00 | 40.00 | 29.73 | 5.03 | 25.30 | .125 | .42 | -.913 | .83 |
| Valid N (listwise) | 30 | | | | | | | | | |

As presented in Table 4, the mean for the number of mistakes as an indication of accuracy for the pretest was 31.80, and the standard deviation was 5.04. For the posttest, the mean score was 29.73 and the standard deviation was 5.03. The skewness values were .114 and .125 for the pretest and posttest, respectively, which were between -1.96 and +1.96. The value for kurtosis was -.84 and -.91 for the pretest and posttest, receptively, which fell between -2 and +2, indicating that the assumption for normal distribution was met. The result of the paired sample t-test is presented in Table 5.

**Table 5**

*Paired Samples Test on the Accuracy of Writing in Pretest and Posttest*

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | Accuracy pretest – accuracy posttest | 2.066 | .691 | .126 | 1.808 | 2.324 | 16.37 | 29 | .000 |

According to the result presented in Table 5, the result of the pared samples t-test showed that there was a statistically significant decrease in complexity components scores from pre-treatment (M= 31.80, SD=5.04) to post-treatment (M= 29.73, SD=5.3, t (29) = 16.37, p = .000) Therefore, it can be stated that second null hypothesis was rejected because chat-GPT did affect the writing of EFL learners in terms of improving their accuracy measured in terms of the t-unit. Eta square for this component was calculated to be .9, which, according to Cohen's guideline, can be considered a big effect size.

The third research question stated, 'Does Chat-GPT-generated feedback have any significant effect on Iranian EFL learners` writing fluency?' To assess writing fluency, several factors had to be taken into account. The most important one emphasized in the literature is the speed of writing, measured by words or characters per minute. This can be quantified by setting the timer to record the number produced within a certain time limit. However, putting a premium on the mere number of words without considering other equally important features of the writing does not sound logical. The next factor, which is subjective, is the ease of writing, indicated by the writer's ability to generate text without excessive pauses or revisions during the initial drafting phase, which can be measured by eye-teaching software to find out which aspect of the text required more time. This can be followed by the think-aloud protocol to partially find out what the writers have gone through in that certain part of the writing. Of course, the think-aloud protocol is not without its shortcomings (Alhadreti & Mayhew, 2018). The next feature for fluency is the overall flow and coherence of the text, evident in the logical progression of ideas and smooth transitions between sentences and paragraphs. The last factor is the appropriate use of vocabulary and sentence structures, which further contributes to the overall impression of writing fluency. Taking into account the purpose of the current study, two criteria were considered for the fluency: the time limit within which the assigned writing tasks were produced and the quality of writing in terms of coherence and logical progression of ideas and smooth transitions between sentences and paragraphs. Since the instruction kept the word number the same, the only factor was the time in which the writing task was completed. For the second case, the coherence and smooth transition of ideas and moves, the scoring of the two raters was used. The descriptive statistic for the fluency component is presented in Table 6.

Table 6
*Descriptive Statistics on the Time for Writing in Pre-test and Post-test as an Indication of Fluency*

| | N | Minimum | Maximum | Mean | Std. Deviation | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Fluency time pre | 30 | 35.00 | 58.00 | 45.93 | 5.64 | .210 | .42 | -.187 | .833 |
| Fluency time post | 30 | 35.00 | 55.00 | 44.53 | 5.33 | .337 | .42 | -.189 | .833 |
| Valid N (listwise) | 30 | | | | | | | | |

According to the data analysis presented in Table 6, the mean length of time spent on completing the writing task in the pretest was 45.93 minutes with the standard deviation of 5.64, skewness of .21 (which is within -.196 and +1.96) and the kurtosis of -.18 (which is within the range -2 and +2). For the posttest, the mean length of time for completing the writing task was 44.53 with the standard deviation of 5.33, skewness of .33 (which is within -.196 and +1.96), and kurtosis of -.18 (which is within the range -2 and +2). As can be seen, the difference between the mean length in pretest and posttest is one and a half minutes. The result of the paired samples t-test presented in Table 7 shows whether this difference is statistically significant.

Table 7
*Paired Samples Test of the Time for Writing in Pretest and Posttest as an Indication of Fluency*

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | Fluency time pre – fluency time post | 1.400 | 1.886 | .344 | .695 | 2.104 | 4.06 | 29 | .000 |

According to the result presented in Table 7, the result of the paired-samples t-test showed that there was a statistically significant decrease in fluency component scores from pre-treatment (M= 45.93, SD= 5.64) to post-treatment (M= 44.53, SD=5.33, t (29) = 4.06, p = .000. Therefore, it can be stated that third null hypothesis was rejected because chat-GPT did affect the writing of EFL learners in terms of improving their fluency measured in terms of the t-unit. Eta square for this was calculated to be .36, which, according to Cohen's guideline, can be considered a big effect size.

**Between-Group Difference between the Experimental and Control Groups**

To have a between-group comparison between the experimental and the control group on the effect of Chat-GPT on the writing quality in terms of the variables of complexity, accuracy, and fluency, an independent samples t-test was run. The result is presented in the following parts. The fourth research question was 'Is there a significant difference between the experimental and control group in terms of complexity?' The descriptive statistic for complexity in the pretest is presented in Table 8.

Table 8

*Group Statistics for the Complexity in the Writing Tasks in the Pretest between the Control and the Experimental Groups*

| | group | N | Mean | Std. Deviation | Std. Error Mean | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Complexity pre | control | 30 | 13.50 | 1.592 | .290 | -.192 | -.333 |
| | experimental | 30 | 13.16 | 1.662 | .303 | .296 | -.300 |

According to the data analysis presented in Table 8, the mean score for the control group in the pretest was 13.50 with a standard deviation of 1.5. For the experimental group, the mean score was 13.16 with a standard deviation of 1.6. The values for skewness are .19 and .29, respectively, indicating the normality of distribution. The result of the independent samples t-test presented in Table 9 shows whether this difference is significant.

Table 9

*Independent Samples Test for the Complexity in the Writing Tasks in the Pretest between the Control and the Experimental Groups*

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| Complexity pre | Equal variances assumed | .034 | .854 | .793 | 58 | .431 | .333 | .420 | -.507 | 1.174 |
| | Equal variances not assumed. | | | .793 | 57.89 | .431 | .333 | .420 | -.507 | 1.174 |

According to Table 9, the value for Levene's test of equality of variance is 85, which is larger than .05, indicating that the assumption of equality of the variances has been observed, and therefore, the first row should be used. As can be seen, there was no significant difference in scores for the control (M=13.50, SD=1.59) and experimental group [M=13.16, SD=1.66; t (58) =.79, p=.43]. Therefore, it can be assumed that the possible between-group difference between the control and the experimental group in the posttest can be attributed to the effect of the treatment. The results are presented in Tables 10 and 11

Table 10

*Group Statistics for the Complexity in the Writing Tasks in the Posttest between the Control and the Experimental Groups*

| | group | N | Mean | Std. Deviation | Std. Error Mean | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Complexity post | control | 30 | 13.93 | 1.311 | .239 | -.066 | .100 |
| | experimental | 30 | 14.86 | 1.502 | .274 | -.542 | -.070 |

As can be seen in Table 10, the mean score for the control group for the complexity in their posttest was 13.93 with a standard deviation of 1.31, and for the experimental group, it was 14.86 and 1.50, respectively. The values for skewness are -.6 6, and -.54, indicating normality of distribution. The result of the independent samples t-test is shown in Table 11.

Table 11

*Independent Samples Test for the Complexity in the Writing Tasks in the Posttest between the Control and the Experimental Groups*

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| Complexity post | Equal variances assumed | .522 | .473 | -2.5 | 58 | .013 | -.933 | .364 | -1.662 | -.204 |
| | Equal variances not assumed. | | | -2.5 | 56.95 | .013 | -.933 | .364 | -1.662 | -.204 |

According to the results of the data analysis presented in Table 11, the value for Levene's test is .47, which is larger than .05; indicating that equality of variances has been observed, and therefore, the first row should be used for analyzing the data. The result shows that there was a significant difference between the control group (M = 13.93, SD = 1.31) and experimental group (M = 14.86, SD = 1.50, t (58) = -.25, sig = .013). This indicates that the feedback incorporated from Chat-GPT into writing tasks made a significant difference in improving the complexity of the writing.

The Fifth research question was 'Is there any significant difference between the experimental and control group in terms of accuracy?' The results of the possible difference between these two groups in the pretest are presented in Tables 12 and 13

Table 12

*Group Statistics for Accuracy between Control and Experimental Groups in Pretest*

| | group | N | Mean | Std. Deviation | Std. Error Mean | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| | control | 30 | 32.03 | 4.097 | .748 | .34 | -.45 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Accuracy pretest | experimental | 30 | 31.80 | 5.040 | .920 | .11 | -.84 |

The result of the data analysis presented in Table 12 shows that the mean score for the control group for accuracy in the pretest was 32.03, with a standard deviation of 4.09 for the experimental group; the mean score was 31.80, with a standard deviation of 5.04. The values for skewness are .34 and .11, indicating normality of distribution. The result of the independent samples t-test presented in Table 13 shows whether this difference is significant.

Table 13

*Independent Samples Test for Accuracy between Control and Experimental Groups in Pretest*

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| accuracy pretest | Equal variances assumed | 2.537 | .117 | .197 | 58 | .845 | .233 | 1.186 | -2.140 | 2.6074 |
| | Equal variances not assumed. | | | .197 | 55.67 | .845 | .233 | 1.186 | -2.142 | 2.6095 |

As can be seen in Table 15, the level of significance for Levene's is .11, which is larger than .05, indicating that the assumption of equality of variances has been met. The result of the independent samples t-test shows that there is not a significant difference for accuracy in the pretest between the control group (M = 32.03, SD = 4.09) and the experimental group (M = 31.8, SD = 5.04); t (58) = .19, sig = .84.). The between group difference in the post-test is presented in Table 14 and 15.

Table 14

*Group Statistics for Accuracy between Control and Experimental Groups in Posttest*

| | group | N | Mean | Std. Deviation | Std. Error Mean | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Accuracy post | control | 30 | 32.63 | 3.624 | .661 | .54 | -.40 |
| | experimental | 30 | 29.73 | 5.030 | .918 | .12 | ٩١.- |

As can be seen in Table 14, the mean score for the control group in the posttest in terms of accuracy was 32.63, and the standard deviation was 3.62. For the experimental group, the mean score was 29.73 and the standard deviation was 5.03. The values for skewness are .54 and .12, indicating the normality of the distribution. The result of the independent samples t-test is presented in Table 15.

Table 15

*Independent Samples Test for Accuracy between Control and Experimental Groups in Posttest*

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| Accuracy post | Equal variances assumed | 5.65 | .021 | 2.56 | 58 | .013 | 2.90 | 1.132 | .634 | 5.165 |
| | Equal variances not assumed. | | | 2.56 | 52.718 | .013 | 2.90 | 1.132 | .629 | 5.170 |

According to the result of the data analysis in Table 15, the level of significance for Levene's test is .02, which is smaller than .05, indicating that the assumption of equality of variables has been violated and therefore, the second row should be used for reporting the result. The result of the independent sample t-test showed a significant difference between the control (M = 32.63, SD = 3.62) and the experimental group (M = 29.73, SD = 5.03), t (28) = 2.56, Sig. = .013) on accuracy in the posttest. This indicates that the between-group difference in the posttest can be attributed to the effect of the treatment, which was the feedback provided by ChatGPT on the accuracy of the writing.

The sixth research question was 'Is there any significant difference between the experimental and control group in terms of fluency?' For this purpose, similar to what was done for complexity and accuracy, two independent sample t-tests were run; one for the pretest and the other for the posttest. The results of the pretest are presented in Tables 16 and 17.

Table 16

*Group Statistics for Fluency between the Control and Experimental Groups in the Pretest*

| | group | N | Mean | Std. Deviation | Std. Error Mean | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Fluency rater pre | control | 30 | 14.63 | 1.376 | .251 | .12 | .24 |
| | experimental | 29 | 14.13 | 1.552 | .288 | -.61 | .47 |

According to Table 16, the mean score for the control group in the pretest was 14.63 with a standard deviation of 1.37. For the experimental group, the mean score was 14.13 and the standard deviation was 1.55. The values for skewness were .12 and .61, indicating normality of distribution. The result of the independent samples t-test presented in Table 17 shows whether this difference is significant.

Table 17

*Independent Samples Test for Fluency between the Control and Experimental Group in Pretest*

| | Levene's Test for Equality of Variances | t-test for Equality of Means |
|---|---|---|

| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Lower | Upper |
| Fluency rater Pre | Equal variances assumed | .197 | .658 | 1.298 | 57 | .199 | .495 | .381 | -.268 | 1.259 |
| | Equal variances not assumed. | | | 1.29 | 55.6 | .201 | .495 | .382 | -.270 | 1.261 |

The result of the data analysis presented in Table 17 shows that the level of significance for Levene's test was .65, which is greater than .05, indicating that the assumption of equality of variables has been observed. The result of the independent samples t-test shows that there was not a significant difference between the control group (M 14.63, SD = 1.37) and the experimental group (M = 14.13, SD = 1.55), t (47) = 1.2, Sig. = .19.) The result of the independent samples t-test on the fluency in the posttest between the control and the experimental group is presented in Table 18 and Table 19.

Table 18

*Group Statistics on Fluency in the Posttest between the Control Group and the Experimental Group*

| | group | N | Mean | Std. Deviation | Std. Error Mean | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Fluency post-rater | control | 30 | 14.30 | 1.368 | .249 | .19 | .-84 |
| | experimental | 30 | 15.23 | 1.406 | .256 | -.68 | .61 |

Data analysis presented in Table 18 shows that the mean score for the control group in fluency in the posttest was 14.3, and the standard deviation was 1.36; for the experimental group, the mean score was 15.23, and the standard deviation was 1.40. The values for skewness are .19 and -.68 indicating normality of distribution. The result of the independent samples t-test presented in Table 19 shows whether this difference is significant.

Table 19

*Independent Samples Test on the Fluency in the Posttest between the Control Group and the Experimental Group*

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| Fluency post-rater | Equal variances assumed | .18 | .67 | -2.6 | 58 | .012 | -.933 | .358 | -1.65 | -.21 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Equal variances not assumed. | -2.6 | 57.9 | .012 | -.933 | .358 | -1.65 | -.21 |

According to the result of the data analysis presented in Table 19, the level of significance for Levene's test was .67, which is greater than .05, indicating that the assumption of equality of variances has been observed. The result of the independent samples t-test shows a significant difference between the control group (M = 14.30, SD = 1.36) and the experimental group (M 15.23, SD = 1.40), t (58) = -2.6, Sig. = .012, on the fluency in the posttest.

**DISCUSSION**

The result of the current study showed that using Chat-GPT can have a positive effect on the writing task of EFL learners. This positive effect was seen both between groups, the control and the experimental groups and within groups in pretest and posttest, indicating that Chat-GPT can be a beneficial and viable instructional tool in assisting EFL learners in their writing skills. The findings of the current study gain support from Amini and Susanti (2024), Tsai et al. (2023), Xu et al. (2024), and Yuan et al. (2024), who deployed Chat GPT for EFL learners' descriptive writing and found a significant effect between the pretest and posttest scores of the participants. They suggested that this tool be used as an auxiliary tool to tackle part of the writing difficulties that EFL learners go through. According to them, the beneficial effect of Chat-GPT in writing a descriptive essay does not come from the mere feedback on the mechanics of writing but from generating novel ideas, on which this form of AI is quite handy for most of the EFL learners. This is in line with the proposition made by Brown (2012) that EFL learners can cultivate a profound level of coherence in their writing by effectively developing ideas and material, a point that seems to be assisted optimally by Chat-GPT for any other form of AI.

The other support comes from Mabrouk (2023), who investigated the effect of using Chat-GPT on improving EFL learners' writing and motivation. The rationale for inclusion of motivation in this study was the notion that it did play an important role in different aspects of language and especially writing. This study found that while Chat-GPT is perceived as a useful tool by both EFL learners and teachers in assisting the learners in their writing ability and reducing their anxiety, the challenges associated with using it need to be acknowledged and dealt with accordingly.

It seems that the beneficial effect of Chat-GPT cannot be restricted to EFL writing only. Ouis (2023) studied the potential effect of this AI on improving the ESP students' writing qualitatively. The finding in this study supports the findings in the current study. The main advantage in this study came from the immense variety that AI could provide EFL learners with in finding the correct content, improving their background information which is useful for their writing in their respective field. This along with the result of the current study is in accordance with Mun (2024) who investigated the utilization of AI-assisted feedback by EFL college learners in their English writing. He specifically, aimed to find out how writing assisted by Chat-GPT improved EFL learners' writing skills in comparison to peer feedback. Additionally, the study attempted to understand the learners' perceptions regarding the use of Chat-GPT in editing English writing. The findings showed that the participants who received Cha-GPT generated feedback on different aspects of writing progressed significantly compared with the control group in terms of the holistic scores about content and organization. However, no significant difference was found in terms of the word count which can be interpreted as the fluency. Therefore, as far as the fluency was concerned the findings of the current study depart from what Mun found.

The other support for the findings of the current study comes from Song and Song (2023) who studied the effect of Chat-GPT on improving the writing skill and motivation of EFL learners. The quantitative analysis in this study showed notable gains in both writing abilities and motivation for students who received AI-supported instruction, in contrast to the control group. The experimental

group exhibits improved skills in several areas of writing, such as organization, coherence, grammar, and vocabulary which is in line with the findings in the current study. The qualitative insights revealed varied opinions, from acknowledging AI's innovative teaching role and its positive effects on writing skills and motivation to concerns about its contextual accuracy and the risk of over-dependence. The participants in this study raised concern about the long-term effects and sustainability of AI-driven instruction, stressing the importance of continual development and adaptation of AI tools. In other words, for the effect of AI to have long-term effect, it seems that needs to be acknowledged as an important and inextricable part of the education in general and EFL in particular.

**References**


Alghizzi, T. M. (2017). Complexity, accuracy, and fluency (CAF) development in L2 writing: the effects of proficiency level, learning environment, text type, and time among Saudi EFL learners.

Alhadreti, O., & Mayhew, P. (2018, April). Rethinking thinking aloud: A comparison of three think-aloud protocols. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12).

Amini, F., & Susanti, A. (2024). Chat GPT: Enhancing Students' Writing Skills for EFL Students in Descriptive Text. *International Journal of Research and Innovation in Social Science*, *8*(10), 2273-2285.

Arif, T. B., Munaf, U., & Ul-Haque, I. (2023). The future of medical education and research: Is ChatGPT a blessing or blight in disguise? Medical Education Online, 28(1), 1-2. https://doi.org/10.1080/10872981.2023.2181052.

Brown, H. D. (2012). T*eaching By Principles: An Interactive Approach to Teaching Pedagogy* (2nd Ed.). New York: Pearson Education, Inc.

Chen, T.-J. (2023). ChatGPT and other artificial intelligence applications speed up scientific writing. Journal of the Chinese Medical Association, 86(4), 351-353. https://doi.org/10.1097/jcma.0000000000000900.

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

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in second Language acquisition*, *26*(1), 59-84.

Emerson, R. W. (2021). Convenience sampling revisited: Embracing its limitations through thoughtful study design. *Journal of visual impairment & blindness*, *115*(1), 76-77.

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

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. Studies in Second Language Acquisition, 18, 299–323.

George, A. S., & George, A. H. (2023). A review of ChatGPT AI's impact on several business sectors. Partners Universal International Innovation Journal, 1(1), 9-23. https://doi.org/10.5281/zenodo.7644359

Golzar, J., Noor, S., & Tajik, O. (2022). Convenience sampling. *International Journal of Education & Language Studies*, *1*(2), 72-77.

Halaweh, M. (2023). ChatGPT in education: *Strategies for responsible implementation . Contemporary Educational*. https://doi.org/10.30935/cedtech/13036

Haleem, A., Javaid, M., & Singh, R. P. (2023). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2(4), 1-8. https://doi.org/10.1016/j.tbench.2023.100089

Kumar, A. H. (2023). Analysis of ChatGPT tool to assess the potential of its utility for academic writing in biomedical domain. *Biology, Engineering, Medicine and Science Reports, 9* ,(١)24-30. https://doi.org/10.5530/bems.9.1.5

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied linguistics*, *27*(4), 590-619.

Lennon, P. (2000). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417.

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). *Generating scholarly content with ChatGPT: Ethical challenges for medical publishing*. The Lancet Digital

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. The International Journal of Management Education, 21(2), 1-13. https://doi.org/10.1016/j.ijme.2023.100790

Mabrouk, N. (2024). *Exploring the Impact of Using Chat GPT on Learners' Motivation in the Writing Classroom the Case of Second Year Students at Larbi Tébessi University–Tebessa_* (Doctoral dissertation, University of Martyr Sheikh Larbi Tebessi Tebessa).

Mun, C. Y. (2024). EFL learners' English writing feedback and their perception of using ChatGPT. *Journal of English Teaching through Movies and Media*, *25*(2), 26-39.

Nosratinia, M., & Razavi, F. (2016). Writing complexity, accuracy, and fluency among EFL learners: Inspecting their interaction with learners' degree of creativity. *Theory and practice in language studies*, *6*(5), 1043.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492-518.

Ouis, H. (2023). The effects of Chat GPT technology use on enhancing ESP students' writing proficiency. The case of master one students at the faculty of economic, commercial, and management sciences at Chadli Bendjedid University, ELTARF. *إفاق للعلوم, 8*(5), 74-82.

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. Journalism & Mass Communication Educator, 78(1), 84-93. https://doi.org/10.1177/10776958221149577

Riggenbach, H. (2000). *Perspectives on fluency*. Ann Arbor, Michigan: University of Michigan Press.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditiona assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), 342 363. https://doi.org/10.37074/jalt.2023.6.1.9

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*, *27*(1), 1-5. https://doi.org/10.1186/s13054-023-043802

Segalowitz, N. (2000). Automaticity and attentional skill in +uent performance. In H. Riggenbach (Ed.). *Perspectives on f!uency* (pp. 25–42). Ann Arbor: The University of Michigan Press.

Skehan, P. (1996). Second language acquisition research and task-based instruction. In J. Willis & D. Willis (Eds.), Challenge and change in language teaching (pp.17–30). Oxford: Heinemann.

Soleimani, M., Modirkhamene, S., & Sadeghi, K. (2017). Peer-mediated vs. individual writing: Measuring fluency, complexity, and accuracy in writing. *Innovation in Language Learning and Teaching*, *11*(1), 86-100.

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843.

Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of second language writing*, *14*(3), 153-173.

Storch, N. (2007). Writing Tasks: The Effects of Collaboration. *Investigating tasks in formal language learning/Multilingual Matters*.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th edn). New York: HarperCollins

Teng, M. F., & Huang, J. (2023). The effects of incorporating metacognitive strategies instruction into collaborative writing on writing complexity, accuracy, and fluency. *Asia Pacific Journal of Education*, *43*(4), 1071-1090.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10(1), 1-24. https://doi.org/10.1186/s40561- 023-00237-x

Towell, R., & Dewaele, J. M. (2005). The Role of Psycholinguistic Factors in the Development of Fluency Amongst Advanced Learners of. *Focus on French as a foreign language: Multidisciplinary approaches*, *10*, 210.

Tsai, C. Y., Lin, Y. T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and information technologies*, 1-19.

van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. Nature, 614, 224-226. https://doi.org/10.1038/d41586-023-00288-7

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing*, *26*(3), 445-466.

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language testing*, *26*(3), 445-466.

Wolfe-Quintero, K. (1998). The connection between verbs and argument structures: Native speaker production of the double object dative. *Applied Psycholinguistics*, *19*(2), 225-257.

Xu, J., Gao, J., Gong, X., Shadiev, R., & Li, Y. (2024). The impact of using ChatGPT on EFL students' writing. In *2024 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 126-128). IEEE.

Yuan, F., & Ellis, R. (2002). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied linguistics*, *24*(1), 1-27.

Yuan, Y., Li, H., & Sawaengdist, A. (2024). The impact of ChatGPT on learners in English academic writing: Opportunities and challenges in education. *Language Learning in Higher Education*, *14*(1), 41-56.