



## The Journal of Language Teaching and Learning™

2026

Volume 16/Issue 1

Article 1

### The Impact of ChatGPT on L2 Writing Skill Development and Learners' Perceptions

Gülümser Efeoglu, Yıldız Technical University, İstanbul, Türkiye, gefeoglu@yildiz.edu.tr  
Erdem Talha Koç, Ministry of Education, Türkiye, erdemtalhakoc@gmail.com

#### Recommended Citations:

##### APA

Efeoglu, G. & Koç, E. T. (2026). The impact of ChatGPT on L2 writing skill development and learners' perceptions. *The Journal of Language Teaching and Learning*, 16(1), 1-14.

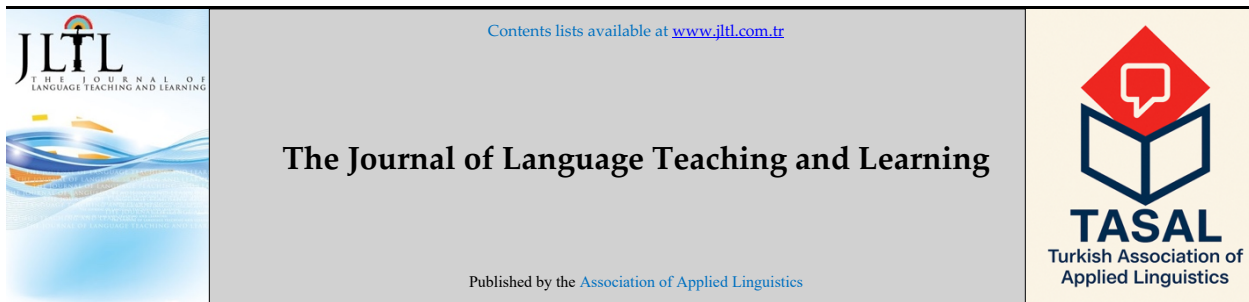
##### MLA

Efeoglu, Gülümser, and Erdem Talha Koç. "The Impact of ChatGPT on L2 Writing Skill Development and Learners' Perceptions." *The Journal of Language Teaching and Learning*, vol. 16, no. 1, 2026, 1-14.

The JLTl is freely available online at [www.jltl.com.tr](http://www.jltl.com.tr) with neither subscription nor membership required. Contributors are invited to review the Submission page and manuscript templates at [www.jltl.org/Submitonline](http://www.jltl.org/Submitonline)



As an online journal, the JLTl adopts a green-policy journal. Please print out and copy responsibly.



[www.jltl.com.tr](http://www.jltl.com.tr)

*The Journal of Language Teaching and Learning*, 2026(1), pp. 1-14

# The Impact of ChatGPT on L2 Writing Skill Development and Learners' Perceptions

Gülümser Efeoğlu<sup>1</sup>, Erdem Talha Koç<sup>2</sup>

## ARTICLE INFO

### Article History:

Received 15 Oct 2024

Revisions completed 7 Jul 2025

Published 31 Jan 2026

### Key Words:

English language teaching

L2 Writing

Written feedback

Technology-enhanced teaching

ChatGPT

## ABSTRACT

Following the pandemic, technology was inevitably incorporated into foreign language classrooms, which resulted in significant modifications to the way English is taught to non-native speakers. In particular, practitioners use a range of tools for various purposes when evaluating the written texts produced by L2 learners. The aim of the current study is to investigate whether written feedback generated via one of such tools, namely ChatGPT, could be used to foster L2 writing skill development and to figure out what L2 English learners consider about receiving feedback from ChatGPT instead of a human rater. For the purposes of this research, an embedded mixed design was administered to 46 participants who were recruited via a convenient sampling procedure. Five data collection tools, including a demographic form, an English proficiency test, pre and post writing proficiency tests, a semi-structured focus group interview, and an adapted questionnaire, were utilized to collect both quantitative and qualitative data. Throughout the intervention, participants received feedback for their essays without knowing the source of feedback. Thus, by employing several prompts prior to the study, feedback generated by AI was made indistinguishable from human. Meticulous analyses revealed that ChatGPT had positive impacts on L2 writing proficiency, whereas participants' perceptions as to the integration of it into assessment were divergent. Despite the fact that they haven't distinguished between ChatGPT and human rater feedback in many areas, like clarity and engagement, L2 English learners ultimately prefer human rater judgments due to motivational concerns and perceptions regarding the necessity and efficacy of the latter.

© Association of Applied Linguistics. All rights reserved

<sup>1</sup>Department of Foreign Languages Education, English Language Teaching, Yıldız Technical University, İstanbul, Türkiye, [gefeoglu@yildiz.edu.tr](mailto:gefeoglu@yildiz.edu.tr)

<sup>2</sup>Ministry of Education, Türkiye, [erdemtalhakoc@gmail.com](mailto:erdemtalhakoc@gmail.com)

Efeoğlu, G. & Koç, E. T. (2026). The impact of ChatGPT on L2 writing skill development and learners' perceptions. *The Journal of Language Teaching and Learning*, 16(1), 1-14.

One of the many definitions of technology-enhanced teaching is the integration of technology into education to facilitate the education process. Even though there are arguments about what exactly the definition entails (Kirkwood & Price, 2013), there is a consensus that it aims to improve learning outcomes. With the use of a number of instruments such as radios and audio devices, English Language Teaching (henceforth ELT) is no exception to this. In line with the advancement in technology, ELT methodologies kept evolving by making use of every piece of recent technology available to meet the needs of the learners.

With the realization of the potential computers offered these older methods have lost popularity and replaced with computers. This led to the emergence of Computer Assisted Language Teaching (CALL), which started in the 1960s and went through multiple stages such as behavioristic CALL, communicative CALL, and Integrative CALL (Warschauer & Healey, 1998). Behavioristic CALL started in the 1960s and focused on drills. Communicative CALL emerged in the late 1970s and early 1980s with the belief that the focus should be on the use of forms rather than the forms themselves. Finally, in the late 1980s and early 1990s, Integrative CALL emerged with an aim to integrate different skills through multimedia networks and technology into language learning (Warschauer & Healey, 1998).

Thanks to the widespread of the internet, and the increasing power of computers, further advancements took place not long after. Computers were not the only technological devices with increasing power. Mobile devices such as phones and tablets, thanks to new technological developments became more powerful; thus, they became on par with computers in terms of what you could do with them. This development led to Mobile Assisted Language Learning (MALL). As the name suggests, MALL utilizes mobile technologies for language learning, which makes it possible for learners to learn in their own time and pace, outside of the traditional classroom settings (Viberg et al., 2020). Likewise, Intelligent CALL (ICALL), appeared in the 1990s, which employs artificial intelligence (Gamper & Knapp, 2002).

ICALL can adapt materials to meet the learners' needs, increase interaction and offers a more personalized learning experience (Schulze & Heift, 2013). Especially in the post pandemic era, a number of artificial intelligence generated tools have become an indispensable part of foreign language learning practices including writing. Given the nascent stage of these applications and their in-class reflections, a more thorough examination is necessary to discern their relative merits and demerits in supporting foreign language instruction.

## 2. Literature Review

One of the foreign language skills that benefit from these advancements in technology is writing. Writing is a productive skill which has been scrutinized from various perspectives and, yet, poses a great challenge for acquirers. In particular, the field of second language (henceforth L2) writing has been dominated by four main theories; contrastive rhetoric, genre theories, cognitive models of composing and socio-cultural theory (Cumming, 2016, p. 65). Among these, socio-cultural theory, which suggests that one-to one interaction with instructor or collaborative writing, is essential to improve writing skills has dominated EFL writing predominantly (Cumming, 2016, p. 82).

### 2.1. Assessing Writing

These theories have also altered the writing assessment. Although both analytic and holistic/impressionistic scoring procedures have been utilized for a variety of purposes in assessing L2 writing, feedback within social constructivist framework plays a crucial role (Biber et al., 2011; Hyland & Hyland, 2006). Feedback can be direct or indirect. It can address a particular type of error or does not directly address it. It can be oral or written. In the lens of social constructivism, feedback can be given by peers or by the instructor as what is important is the interaction with feedback, the source. Learners should have a chance to interact with feedback and reach an understanding. However, both peer and instructor feedback have

constraints. As for peer feedback, the feedback peer provides may not be perceived as on par with the feedback an instructor would provide as learners view instructors as the more knowledgeable one (Tsui & Ng, 2000). Peers' feelings towards each other may also affect the feedback positively or negatively (Saito & Fujita, 2004). Lastly, peers mostly focus on the product and sentence level errors, rather than process (Storch, 2005). As for instructor feedback, providing feedback to every single learner in the classroom is quite time-consuming.

Regardless of the type of feedback given, giving feedback creates a burden on instructors and drains valuable class time, especially in crowded classrooms. That's where technology comes to aid. The first attempts to use technology to assess writing skills began in the 1960s with Project Essay Grade (PEG) (Warschauer & Grimes, 2008). This tool was trained on essays graded by human raters, and it would compare the essays to those it was trained on. Designed to help instructors with scoring high-stakes exams (Allen et al., 2016, p. 317), for example, Automated Essay Scoring (henceforth AES) tools are being used for TOEFL and GMAT (Stevenson, 2016). These tools mainly use the essays they were trained on to rate other essays. The main point of AES tools is to reliably score essays, thus reducing the workload for human raters and reliability of these tools has been proven by previous research (Shermis, 2014). Some of the known AES tools are e-rater by Educational Testing Service, IntelliMetric by Vantage Learning, and Intelligent Essay Assessor by Pearson.

## 2.2. Automated Writing Evaluation Tools and ChatGPT

Automated Writing Evaluation (henceforth AWE) tools, similar to AES tools, have a scoring engine. However, they also provide formative feedback to students and let them practice as much as needed, thus increasing learner autonomy (Grimes & Warschauer, 2010) and saving time for the instructors (Allen et al., 2016). The use of these resources in classrooms has increased as a result of these alternatives (Li et al., 2015), and helped learners to develop writing skills (Allen et al., 2016;

Stevenson, 2016). These tools have proven to be helpful, especially at initial stages of writing, mostly with grammar errors (Chen & Cheng, 2008; Li et al., 2015). A study conducted in China with PigAi (an AWE tool developed in China) revealed that students found AWE tool useful as it provided them feedback on grammar (Huang & Renandya, 2020).

It is worth noting even though AES tools were designed to help with scoring and AWE tools designed to provide feedback on various aspects of essays, the distinction between them is getting blurry as AES tools start to offer feedback and AWE tools offer essay scoring. One example of this overlap would be Grammarly. It offers feedback on the content and form of the writing in addition to its scoring engine which scores the text. Therefore, from now on, we will use AWE to refer to automated tools which aim at providing both feedback and score for written texts.

The use of AWE tools offers some advantages. One possible advantage of using AWE tools is that it saves time for the teachers. In her literature review Stevenson (2016) found that the most common reason to use AWE tools was to save time. The time was saved thanks to the help AWE tools offer on sentence level (grammar, punctuation and spelling). The basic feedback these tools offer is appreciated by students with low proficiency levels. One of students who used PigAi in Huang and Renandya's study (2020) stated that PigAi could point out the basic grammar errors thus he could avoid repeating them. Another benefit of using AWE seems to be that the time these tools save can be used by teachers to focus on other aspects of L2 writing. In a writing course, students used Criterion, and instructors of the course reported that the use of Criterion offered instant feedback and helped students with sentence level errors, which, allowed them to focus on bigger issues like thesis argument development (Schroeder et al., 2008). Last but not least, the AWE tools foster learner autonomy by making it possible for students to practice writing and receive feedback (Allen et al., p. 320).

There are also some criticisms for the use of AWE tools. One of the criticisms is that these tools are not suitable for English language learners as the developers of these tools trained these tools using

data from native speakers of the language (Warschauer & Ware, 2006), which means the errors these tools detect and the feedback these tools give may not be suitable for non-native speakers of English. What type of errors these tools detect is another problem as these tools are tuned to detect certain types of errors (McNamara et al., 2015) which leads to overlook of the other types of errors. These tools can also be tricked by writing in a certain way which resembles the essays the machine was trained on to get higher scores (Harrington & Moran, 2012; Powers et al., 2001) or in some cases these tools may fail to score essays. One such case was exemplified by Liu and Kunnan's study (2016) which tested how Writetolearn would perform compared to human raters. After Writetolearn failed to score some essays, they contacted the developers and learned that if the essay is not similar enough to the ones Writetolearn was trained on, the tool fails to score it. This raises questions such as if it is worth using these tools in non-native classroom settings.

Another issue with the use of AWE tools is that most of the existing research on these tools is either directly conducted or funded by the developers themselves (Liu & Kunnan, 2016; Stevenson, 2016; Chen & Cheng, 2008). The generic feedback provided by these tools does not consider individual differences and needs (Warschauer & Grimes, 2008). Harrington and Moran (2001) took a closer look at AWE tools from a different perspective which is still relevant. They claimed that we write to influence the readers, and writing to a machine would desensitize us as we would know there is no human being reading our writings. They stated that testing 4000 students using AWE tools would cost them 3,900,000\$ including the infrastructure, which was half the salary budget of their faculty. The last point they made was although these tools were marketed as a means to lighten the burden of the instructors, they would only make sure the problem of overcrowded classrooms persisted.

In light of the above discussion, we recognize several potential advantages of ChatGPT over traditional AWE tools, prompting us to explore its use as an AWE tool. Unlike traditional AWE

systems, ChatGPT is designed to generate human-like text. Its training on a diverse range of sources, including data created by human trainers and publicly available internet data, might give it some advantages. Firstly, this diverse training potentially makes ChatGPT better at detecting social, cultural and contextual aspects of the written products, which is a limitation often observed in AWE tools (Vojak et al., 2011). The second advantage is the clarity of the feedback it will provide, as it is designed to generate human-like text, the feedback it will provide should be clearer, in theory, at least. The last reason we decided to use ChatGPT is the interaction. Unlike the AWE tools, one can interact with the feedback given by ChatGPT and ask to clarify it or even ask it to focus on any aspect of the writing to assess. In a recent study, Tsai et al. (2024) demonstrated that the use of ChatGPT to revise English compositions had positive impacts on EFL learners' vocabulary, grammar, and organization of their text development. Adding ChatGPT's ability to remember the previous conversations and form its answers accordingly takes interaction to a whole new level which is never seen before in any AWE tool. One last advantage ChatGPT has over the AWE tools is that unlike most of these tools, it, at least the 3.5 version, does not require any kind of payment or subscription to use.

When we look at ChatGPT through the lens of social constructivism, we see it could have advantages over other forms of feedback. The feedback it provides can focus on what is asked, which makes it a better alternative than peer feedback in terms of knowledge. It has no feelings, which is another benefit compared to peer feedback. There is also interaction with the feedback by asking questions, which lowers the workload of instructors. In conclusion, implementing ChatGPT into the classroom might enhance the L2 writing process by giving learners meaningful feedback they can interact with, and giving instructors the time which they can use to focus on other aspects of L2 writing (Straume & Anson, 2022).

Concerns related to originality and reliability of ChatGPT in its use for educational purposes overshadows its advantages (Rudolph et al., 2023). Yet, a meticulous scrutiny is required to detect

where and how to integrate ChatGPT in education settings. We aim to address this by integrating ChatGPT into a EFL writing classroom setting for a semester. Stevenson (2016) observed that AWE tools are often seen as sophisticated editing tools rather than tools designed to enhance students' overall writing capabilities. Through our study, we aim to investigate whether ChatGPT can serve as a tool to assist English language learners in refining their overall writing skills and EFL teachers in assessing and scoring writing. Taking these into consideration, we aim to answer the following research questions:

- 1- Does feedback either from human rater or ChatGPT have any effects on L2 writing proficiency?
- 2- What do EFL students think about receiving feedback for their essays from ChatGPT?

### **3. Methodology**

#### *3.1. Research Design*

In line with the research questions a mixed method design is employed. Specifically, an embedded experimental design has been conducted since it allows “to enhance a quantitative experimental study by including a secondary qualitative component to explore the procedures or process of the experiment” (Clark & Creswell, 2014, p. 391). In this design both quantitative and qualitative means have been utilized subsequently to triangulate data. Moreover, participants have been classified as control and experimental groups. The control group received feedback from human raters while the other received it from ChatGPT. In order not to bias data participants have not been informed about the source of their feedback until the end of the study.

#### *3.2. Setting and Participants*

The study was conducted at a state university in Istanbul. 46 participants were recruited via a convenient sampling procedure. Participants were all freshmen students who enrolled in English Language Teaching program and were all taking

Writing course. They were divided into two groups as control and experimental ones randomly. Their English proficiency was accepted as C1 according to CEFR scheme and to ensure their proficiency level Oxford Quick Placement Test was administered. The study was lasted for ten weeks from February to June in 2023. After receiving approval from the Institutional Review Board with an issue number 2023.06 written consents were obtained from the participants by informing them that identities would be kept anonymous to ensure privacy.

#### *3.3. Procedure*

As a result of the earthquake that occurred in February in Türkiye, the Higher Education Institutions switched to online education. Hence, we decided to use the combination of Google Forms, Google Classroom and Zoom to conduct the study. We used Google Forms to create the assignments, as it has timestamp function which could help us see when the participants had started working on the assignments. It was used to make sure they would not go beyond the time limit or complete assignments at a different time (i.e. Google Classroom to share the assignment links as assignments and to give feedback to their works as private comments). Participants were using Zoom to connect to Writing class. Hence, researchers used Zoom to communicate with them at the end of their Writing class and shared the link on the Google classroom, which they had already joined, and they would start working on the task.

Participants completed the English Proficiency Test and Demographic Form in the second week of the term. The following week, the pre-test for writing proficiency was administered. Then, till the end of the semester, participants in both groups followed the same syllabus producing three different types of essays such as cause and effect, process or argumentative. Topics ranged from technology and education to social media. Participants were required to write at least 250 words about the topic in 40 minutes during the class. For each essay, they received feedback and prepared the final draft without knowing the source of feedback. The same rubric that was used to score

participants' writing proficiency was employed as the basis for feedback provided to participants in both groups. The rubric was prepared in line with the course objectives and five criteria were determined: Introduction and Thesis; Organization and Structure; Evidence and Support; Grammar, Spelling and Mechanics; Conclusion and Overall Impressions. Analytic scoring was believed to make both types of feedback look like more similar than different since raters (both ChatGPT and HRs) are restricted to most commonly preferred criteria (Brown, 2004; Huges, 2020).

More specifically, researchers instructed ChatGPT to give feedback to the uploaded materials by sticking to the pre-determined rubric. Thus, ChatGPT group received written feedback for each subcategory specified in the rubric meanwhile human rater (HR) group received the feedback in form of a table which included the same five categories along with written feedback. In order not to bias the scoring procedure with holistic and impressionistic assessment under two distinct conditions one being ChatGPT while the other HR, analytic scoring procedure was selected. Scoring of all the essays done by the two researchers. To ensure the rater reliability, researchers, after designing the rubric scored a couple of essays and agreed upon the details. Whenever they gave different scores to one of the five criteria, the average of the two scores was given as the final score for that criterion.

At the end of the first task, participants who submitted the assignment were divided into two groups randomly. The first group received feedback from ChatGPT, which was unbeknownst to them. They received written feedback, which was actually provided by asking ChatGPT to provide feedback to their writings based on five criteria and use a tone which would resemble a teacher sending feedback to his/her student's writing. Before the exact prompt we used, the researchers asked ChatGPT which prompt to use. Here is the prompt used to get written feedback:

*Researcher A:* I want you to give feedback to essays which will be provided by me. What prompt should I use to get the best results?

*ChatGPT:* To get the best results when asking for feedback on essays, you can use the following prompt format:

"Please provide feedback on the following essay on [topic]. Here is the text: [Insert essay here] Please address your feedback as if you are directly communicating with the student.

Replace [topic] with the subject of the essay, and [Insert essay here] with the full text of the essay you would like to receive feedback on. This format will help provide context and ensure that the AI understands your request clearly."

*Researcher A:* Please provide feedback on the following essay on [topic] by using the following criteria: 1. Introduction and Thesis Statement 2. Organization and Structure 3. Evidence and Support 4. Grammar, Spelling, and Mechanics 5. Conclusion and Overall Impression. Here is the text: [Insert essay here]

The same prompt was copied and pasted for each essay being evaluated. Responses provided by ChatGPT were downloaded as they were in order not to interfere with the original content generated by AI. Hence, participants received GPT feedback copied and pasted on a word document without knowing that it was taken from GPT. HR feedback was also provided in the same format on a word document to mask the source of the feedback.

Both groups received written feedback three times in total. The week after the posttest for writing proficiency (12th week) a focus group interview was conducted, in which students were informed about the true source of the feedback they had been receiving throughout the semester. During the focus group interview they were asked about their opinions on the feedback ChatGPT provided and their opinions on the use of ChatGPT as a feedback source in addition to several other questions. Finally, they were required to fill out a questionnaire that targets their perception of written feedback that they had received.

The Table 1 below provides a brief summary of data collection procedure and intervention.

Table 1

## A Synopsis of Data Collection

Order	Data Collection Tools	Time
1.	Demographic Form	2 <sup>nd</sup> week
2.	English Proficiency Test	2 <sup>nd</sup> week
3. A)	Pre Writing Proficiency Test	3 <sup>rd</sup> week
	a. Summarizing Information	
	b. A Free Writing Task: Argumentation	
	Intervention: In-class Tasks and Written Feedback	4 <sup>th</sup> – 11 <sup>th</sup>
	a. Cause and Effect: First Draft → Feedback → Final Draft	weeks
	b. Process: First Draft → Feedback → Final Draft	
	c. Argumentative: First Draft → Feedback → Final Draft	
3. B)	Post Writing Proficiency Test	11 <sup>th</sup> week
	a. Summarizing Information	
	b. A Free Writing Task: Argumentation	
	Participants were informed that some received feedback from ChapGPT	11 <sup>th</sup> week
4.	Semi-structured Focus Group Interviews	12 <sup>th</sup> week
5.	5-point Likert Scale Questionnaire Adapted from (Huang & Renandya, 2020)	12 <sup>th</sup> week

## 3.3. Data collection tools

For the purposes of this study five data collection tools were administered. The first one was the Demographic Form which was used to elicit detailed information as to participants' language background, computer literacy skills, etc. The second tool was the Oxford Quick Placement test which was used to test participants' English proficiency level. This plays a crucial role since difference in proficiency might result in discrepancies in the overall gain of feedback (López et al., 2017). Since the overall proficiency test was missing a writing component, another instrument was utilized.

The third tool was Pre-post writing proficiency tests which were administered both at the beginning and at the end of the study to understand whether the feedback given had any impacts on overall L2 English writing proficiency. Each test involved two writing tasks. The first one was geared towards summarizing the information presented via prompts while the second task was a free writing in which they argue for or against an idea. They were scored analytically based on a pre-determined rubric. Likewise, post writing proficiency was carried out at the end of the semester. Everything was just the same as the pre writing proficiency tests except for the topics. All participants were informed

that some had received feedback from ChatGPT at the end of the semester.

The fourth instrument was semi-structured focus group interviews administered at the end of the study, and the aim was to gain deeper insight into the participants' experiences of both human rater and AI generated feedback for their L2 English writing performance. Lastly, they were requested to fill out a questionnaire which was adapted from Huang and Renandya (2020). It consisted of 16 items on a 5-point Likert-scale (i.e. 8 for ChatGPT and 8 for human-rater related items). To validate the reliability of the questionnaire, we ran a reliability test and got the score of 0.913, which is above the required threshold of 0.70 for internal consistency reliability.

## 4. Results

Out of 46 participants, 12 of them either dropped out as they did not complete at least one of the given tasks. Out of the remaining 34 participants, 17 of them were in the ChatGPT group and 17 were in the human rater group. In order to answer the first research question about the probable effects of ChatGPT originated feedback on L2 writing proficiency, scores of pre and post writing proficiency tests were scrutinized quantitatively on SPSS 23. To ensure group homogeneity, a Levine test was conducted and it was found that groups



were homogenous with p value of 0.9362. After ensuring the homogeneity of the groups, we proceeded with paired sample t-tests to see whether the feedback provided by ChatGPT had led to any improvements.

Prior to the intervention, participants of the human rater group functioning as the control group had a mean score of  $M=83.7647$ ,  $SD= 10.43749$ , and after the intervention  $M= 87.0588$ ,  $SD= 7.28415$ . A paired sample t-test revealed there was no significant difference between two time periods  $t(16)= -.964$ ,  $p=.349$ , 95% CI [56, 96]. Although the effect size was medium ( $d=.712$ ), it was not statistically significant, implying any observed differences in the control group's scores may be due to chance.

Likewise, for ChatGPT group a paired samples t-test was run. The results signaled a significant difference  $t(16)=-2.266$ ,  $p=.038$ , 95% CI [72, 100]. Test mean scores for pre and post tests were  $M=81.4118$ ,  $SD=10.57745$  and  $M=85.1765$ ,  $SD= 9.97644$  respectively. The effect size was medium ( $d=.0549$ ). Thus, the intervention, namely the feedback provided by ChatGPT resulted in a statistically significant difference on experimental group's L2 writing performance.

In brief, the quantitative findings demonstrated that human raters and ChatGPT originated feedback might have divergent impacts on L2 writing proficiency. To triangulate evidence, qualitative data were collected via focus group interviews a week after the post test.

The recordings of interviews were transcribed verbatim and thematic analysis was conducted as it is suitable for identifying, analyzing and describing patterns in a dataset (Braun & Clarke, 2006). Intra-rater reliability was sustained by multiple readings of the raw data. An important discovery was that all participants believed the source of all feedback was the researchers, revealing how ChatGPT was good at imitating human-like written feedback. Detailed thematic analysis revealed some distinct patterns, pointing strengths and limitations of each source of feedback. In particular, three arguments prevailed the thematic analysis: affect, perceived value of feedback, preferred type of feedback.

Unlike human-raters who read and rate the essays, ChatGPT has no emotions. The reactions to the absence of emotions in feedback were mixed. Some claimed the lack of emotions were rather a good thing as a human rater may fail to be neutral. For instance, one of the participants stated:

*I might have been offended by a lexical choice of a human rater but when I receive feedback from ChatGPT it sounds more objective because it does not know me, thus I feel less threatened (GPT-2).*

On the other hand, some claimed that emotions are important and indispensable part of communication, one participant told:

*I believe the evaluation of a human is better as both parties can understand each other's feelings (HR-1).*

This and other comments suggest that participants have mixed feelings as to the availability of emotions in feedback and ChatGPT offers a good alternative for those who prefer feedback with no reference to emotions.

Another key theme was perceived value of the feedback. Participants took a more pragmatic approach, stating ChatGPT offered good feedback so it could be used as long as it is practical. One participant said:

*If the feedback is useful, the source does not matter (GPT-4).*

Even though participants found the feedback provided by ChatGPT detailed and adequate to use, the participants objected to the idea of receiving feedback from it for prolonged periods of time. Main reason of this objection was motivation. Participants stated that getting feedback from ChatGPT instead of the instructor, would decrease their motivation as they would start to question whether it's worth to write. One participant mentioned:

*If the instructor is going to use ChatGPT all the time, I can ask ChatGPT to write the essays for me. What is*

*the point of wasting time if the instructor is not going to read (HR-8)?*

Thus, interacting with the audience as one of the purposes of writing tasks was missing in AWE generated feedback. Learners write not only to improve their writing skills but also to arouse a feeling and stimulate some sort of a change on the reader. In brief, participants highlighted that for the sake of objectivity AWE generated feedback is better while in the long run they strive for a human reaction in L2 writing.

When asked about the preferred form of feedback the participants answered diversely. Some claimed irrespective of the source (i.e. human rater or AWE) written feedback is better as it is permanent and accessible whenever needed. For instance, one of the participants answered:

*I like written feedback more because I know that it stays there forever and I can use it whenever I want (HR-6).*

On the other hand, those who preferred the oral feedback claimed that as oral feedback is

immediate, one can talk about the details and understand the feedback better:

*I prefer oral feedback since I can negotiate the content of the feedback with the instructor. If I have questions, I can ask for clarification (HR-4).*

Lastly, when asked to rank AWE feedback, human-rater feedback, and peer feedback, AWE feedback was always placed in the middle. When asked about the reason, it is revealed that participants believed instructors were more knowledgeable, thus they would provide better feedback, meanwhile the feedback peers give would be less stressful but more general. Still, the fact that none of the participants chose AWE feedback as the first option indicates that participants favored human-rater more than AWE tools.

Lastly, the questionnaire that was administered after revealing the source of feedback was analyzed quantitatively by calculating means and standard deviations for each construct. The results are available in Table 2 below.

Table 2  
*Results for Perceptions of Feedback on a 5-point Likert Scale*

Construct	Sample Item	HR		GPT	
		Mean	SD	Mean	SD
Engagement	I can understand feedback in XX	4.35	0.81	4.28	0.6
Perceived Effectiveness	XX can enhance my writing performance	4.55	0.99	4.15	1.08
Clarity	The feedback in XX is clear	3.26	0.91	3.4	0.68
Motivational Influence	I feel motivated to improve my writing skills after receiving feedback in XX	3.64	1.18	3.85	1.22
Perceived Necessity	It is necessary to receive feedback from XX	4.65	1.22	3.15	1.33

Participants seem to hold similar beliefs in terms of engagement and motivational difference between HR and ChatGPT, and they do not distinguish one from the other. As for clarity, means for both groups are comparatively low when compared to other constructs although there is not much difference between feedback types. This might result from the nature of the written feedback that was based on analytic scoring. As there were predetermined

criteria and the feedback was limited to detailed explanations on these criteria, participants could have had difficulties in interpreting the feedback.

On the other hand, constructs targeting perceptions suggested more diversity for the type of feedback. In particular, participants rated HR feedback as more effective than GPT, which overlaps with the qualitative findings. Furthermore, the most dramatic discrepancy was observed in

perceived necessity. This implies that although participants consider both almost equivalent in many respects they still accredit HR feedback more than GPT. Hence, based on the difference in perceived necessity, it might be concluded that despite all positive assets of AWE, feedback participants still posit that written feedback from human raters is needed.

## 5. Discussion

This study was guided by two research questions investing whether AWE generated, namely ChatGPT, written feedback had any influence on L2 English writing skill development and how L2 English learners perceive the use of ChatGPT in receiving feedback.

A quick glance at the literature reveals that AWE feedback is indeed useful (Zhai & Ma, 2023) especially for grammar (Liao, 2015). It helps instructors save time and focus on other aspects of L2 writing (Carvalho et al., 2022). Contrary to many tools which are deemed to be highly practical in providing sentential and lexical level feedback, we wanted ChatGPT to focus on higher order skills such as organization, style, and tone. A comparison of the pre-post test scores of the control and experimental groups reveals that the feedback provided by ChatGPT, which we used as an alternative AWE, had positive effects on L2 writing proficiency. Moreover, non-availability of such impact in the experimental group suggests that the content of the feedback matters more than feedback itself. Although the same analytic rubric was employed in assessment, detailed content-based explanations generated by Chat GPT were found to be superior. Hence, it might be concluded that the content of the feedback overweighed the significance of its source in the current study.

In this study, even though both the control and experimental groups received written feedback based on the same (i.e. an analytic scoring oriented) rubric, ChatGPT feedback was more influential in L2 writing skill development. Although written feedback is indispensable part of L2 writing it does not guarantee success in acquiring L2 writing skills. Thus, practitioners should pay more attention to the

content of the feedback than its type (written vs. oral) and origin (HR vs. ChatGPT).

Regarding the participant perceptions, the qualitative data analysis revealed that the feedback ChatGPT provided was perceived to be useful. Unlike some of the feedback provided by the other AWE tools (Li et al., 2015), no one considered the feedback ChatGPT provided to be generic and impractical. In addition, one of the problems with the use of AWE tools is that they offer no social interaction (Lai, 2010). Even though some participants considered lack of emotions, which are typically embedded in the HR feedback, as one of the fallacies of AWE feedback; others believed that it makes the feedback sound more objective. Besides, motivation was found out to be another significant construct. In particular, the findings indicated that participants would feel demotivated if they had received written feedback from ChatGPT constantly. Participants noted that they want their progress to be traced by their teachers but not machines. In a way, the sense of approval appears to be the driving force behind L2 writing practice motivating the learners to tailor their L2 writing skills. Furthermore, HR feedback was not only perceived to be motivating but also essential and required for L2 writing skill development. However, the findings underlined that participants do not distinguish between two types of feedback in terms of clarity of the feedback content, perceived value, and engagement. In brief, it could be concluded that despite the hesitations related to motivation and perceived necessity ChatGPT might be a handful tool in helping the instructors with their workload by providing meaningful feedback in L2 writing classes.

## 6. Conclusion

As stated in the literature review, one of the downsides of the AWE tools was that they were designed using the data provided by the native speakers. ChatGPT, thanks to its source of information being the whole internet, eliminates this limitation. Thus, it can detect non-native mistakes easier, which is something some AWE tools fail to do. Another plus for the ChatGPT is that

it can alter the form of the feedback it gives. It can be asked to give it as if an instructor was e-mailing the feedback (which is what we did). This alteration makes it possible for the learners not to dismiss the feedback immediately as it is tailored for them. Furthermore, the comparison of two groups in this study demonstrated that ChatGPT generated feedback influences L2 writing skill development positively.

Taking all these pros into consideration, practitioners might employ AWE tools such as ChatGPT for the purposes of giving written feedback in foreign language learning classes. Besides, participants have not distinguished HR feedback from ChatGPT in many respects such as clarity and engagement. Thus, teachers might employ GPT for initial screening of student essays by providing clear prompts in line with their assessment criteria. This will also help them reduce heavy workload and save time for detailed evaluation of student essays. However, the motivational concerns and perceptions as to the necessity and effectiveness of the latter suggest that L2 English learners yearn for human rater judgements in the long run. Therefore, instead of replacing teachers completely, ChatGPT might be used to assist EFL teachers in reinforcing L2 writing. The current study was restricted in terms of genre (i.e. essay writing), frequency (i.e. three tasks), and scoring (i.e. an analytic rubric). However, ChatGPT evolved a lot ever since we conducted the experiment. Now it offers a creative writing coach which can give feedback on your writing. Likewise, ChatGPT editor lets you give ChatGPT certain instructions to redefine it and then share it with others, even to those without ChatGPT plus subscription. It makes it possible for instructors to

train the ChatGPT in the way they wish and share their version with students so they can share their writings with ChatGPT and get feedback just like their instructor would give. Training ChatGPT on feedback and making students use it would be an interesting study. Hence, all these new features are probable areas of research which call a closer scrutinization as to their integration in teaching English as a foreign language.

### **Declarations**

The participants were protected in accordance with the ethics guidelines and the details were described in the section on participants.

### **Competing Interests**

We, the authors, declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### **Data availability**

Data is available upon reasonable request to the corresponding author.

### **Acknowledgements**

A part of this research was presented at 13th International ELT Research Conference, Ankara, Türkiye. Besides, we would like to thank two anonymous reviewers for their constructive feedback.

## References

- Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 316–329). The Guilford Press
- Biber, D., Nekrasova, T. and Horn, B. (2011), The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. *ETS Research Report Series*, i-99. <https://doi.org/10.1002/j.2333-8504.2011.tb02241.x>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. Pearson Longman.
- Carvalho, L., Martinez-Maldonado, R., Tsai, Y.-S., Markauskaite, L., & De Laat, M. (2022). How can we design for learning in an AI world? *Computers and Education: Artificial Intelligence*, 3, 1–9. <https://doi.org/10.1016/j.caeai.2022.100053>
- Chen, C. & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning and Technology*. 12(2), 94-112.
- Clark, V. L. P., & Creswell, J. W. (2014). *Understanding research: A consumer's guide, Enhanced Pearson eText with loose-leaf version--Access Card Package*. Pearson
- Cumming, A. (2016). Theoretical orientations to L2 writing. In Manchon R, Matsuda PK (Eds.) *Handbook of second and foreign language writing*. Walter de Gruyter, pp. 65–88. <https://doi.org/10.1515/9781614511335-006>
- Gamper, J., & Knapp, J. (2002). A Review of Intelligent CALL Systems. *Computer Assisted Language Learning*, 15(4), 329–342. <https://doi.org/10.1076/call.15.4.329.8270>
- Grimes, D. & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment*, 8(6), 1-42.
- Herrington, A., & Moran, C. (2012). Writing to a machine is not writing at all. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 219–232). Hampton Press
- Huang, S., & Renandya, W. A. (2020). Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in Language Learning and Teaching*, 14(1), 15–26. <https://doi.org/10.1080/17501229.2018.1471083>
- Hughes, A. (2020). *Testing for language teachers*. Cambridge University Press.
- Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(2), 83-101. <https://doi.org/10.1017/S0261444806003399>
- Kirkwood, A., & Price, L. (2013). Technology-enhanced learning and teaching in higher education: what is “enhanced” and how do we know? A critical literature review. *Learning, Media and Technology*, 39(1), 6–36. <https://doi.org/10.1080/17439884.2013.770>
- Lai, Y. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology*, 41(3), 432–454. <https://doi.org/10.1111/j.1467-8535.2009.00959.x>
- Liao, H. C. (2015). Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal*, 70(3), 308–319. <https://doi.org/10.1093/elt/ccv058>
- López, M. B., Steendam, E.V., & Buyse, K. (2017). Comprehensive corrective feedback on low and high proficiency writers: Examining attitudes and preferences. *International Journal of Applied Linguistics*, 168, 91-128. <https://doi.org/10.1075/itl.168.1.04bon>
- Liu, S., & Kunnan, A. J. (2016). Investigating the Application of Automated Writing Evaluation to Chinese Undergraduate English Majors: A Case Study of WriteToLearn. *CALICO Journal*, 33(1), 71–91. <https://doi.org/10.1558/cj.v33i1.26380>
- McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35–59. <https://doi.org/10.1016/j.asw.2014.09.002>
- Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping E-Rater: Challenging the validity of automated essay scoring. *ETS Research Report Series*, 2001(1), i-44.
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1), 342–363. <https://doi.org/10.37074/jalt.2023.6.1.9>
- Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research*, 8(1), 31-54. <https://doi.org/10.1191/1362168804lr133oa>
- Schulze, M. & Heift, T. (2013) Intelligent CALL. In M. Thomas, H. Reinders and M. Warschauer (Eds.) *Contemporary computer-assisted language learning* (pp. 249-265). Bloomsbury.

- Schroeder, J., Grohe, B., & Pogue, R. (2008). The Impact of Criterion Writing Evaluation Technology on Criminal Justice Student Writing Skills. *Journal of Criminal Justice Education*, 19(3), 432–445. <https://doi:10.1080/10511250802476269>
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76. <https://doi:10.1016/j.asw.2013.04.001>
- Stevenson, M. (2016). A Critical Interpretative Synthesis: The Integration of Automated Writing Evaluation into Classroom Writing Instruction. *Computers and Composition*, 42, 1–16. <https://doi:10.1016/j.compcom.2016.05.001>
- Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, 14(3), 153–173. <https://doi:10.1016/j.jslw.2005.05.002>
- Straume, I., & Anson, C. (2022). Amazement and Trepidation: Implications of AI-Based Natural Language Production for the Teaching of Writing. *Journal of Academic Writing*, 12(1), 1–9. <https://doi.org/10.18552/joaw.v12i1.820>
- Tsai, CY., Lin, YT. & Brown, I.K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and Information Technologies*, 1-19. <https://doi.org/10.1007/s10639-024-12722-y>
- Tsui, A. B. M., and Ng, M. (2000) Do Secondary L2 Writers Benefit from Peer Comments? *Journal of Second Language Writing*, 9, 147-170. [http://dx.doi.org/10.1016/S1060-3743\(00\)00022-9](http://dx.doi.org/10.1016/S1060-3743(00)00022-9)
- Viberg, O., Wasson, B., & Kukulska-Hulme, A. (2020). Mobile-assisted language learning through learning analytics for self-regulated learning (MALLAS): A conceptual framework. *Australasian Journal of Educational Technology*, 36(6), 34-52. <https://doi.org/10.14742/ajet.6494>
- Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New Spaces and Old Places: An Analysis of Writing Assessment Software. *Computers and Composition*, 28(2), 97–111. <https://doi:10.1016/j.compcom.2011.04.004>
- Warschauer, M., & Grimes, D. (2008). Automated Writing Assessment in the Classroom. *Pedagogies: An International Journal*, 3(1), 22–36. <https://doi:10.1080/15544800701771580>
- Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, 31, 57-71.
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10 (2), 157–180. <http://dx.doi.org/10.1191/1362168806lr190oa>
- Zhai, N., & Ma, X. (2023). The Effectiveness of Automated Writing Evaluation on Writing Quality: A Meta-Analysis. *Journal of Educational Computing Research*, 61(4), 875-900. <https://doi.org/10.1177/07356331221127300>