System 134 (2025) 103805

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

ChatGPT-generated versus human direct corrective feedback on L2 writing

Belén C. Muñoz Muñoz[a,*], Hossein Nassaji[b], Felipe I. Bello Carrillo[c]

[a] Universidad Del Bío-Bío, Chile
[b] University of Victoria, Canada
[c] Universidad de Concepción, Chile

ARTICLE INFO

Keywords:
ChatGPT
Written corrective feedback
EFL
Writing
Essay

ABSTRACT

The rapid advancement of AI is transforming language teaching, with tools like ChatGPT offering new ways to deliver written corrective feedback (WCF). While studies have long shown that teacher-delivered WCF is effective in promoting second or foreign language (L2) acquisition, the rise of AI-based tools raises questions about their potential to enhance or disrupt this process. The present study examined and compared the effectiveness of direct WCF provided by ChatGPT versus a human teacher in improving second-year English Pedagogy students' L2 essay writing, focusing on task response, cohesion and coherence, lexical resource, and grammatical range and accuracy. The study involved 44 students (aged 19–21) from a Chilean university, who were randomly assigned to receive corrective feedback from either ChatGPT (N = 20) or a trained human teacher (N = 22) over four sessions across two months. The findings indicate that while both sources of correction significantly enhance writing development, ChatGPT demonstrated overall superiority across all the assessed criteria. Thus, it can be suggested that this AI tool could be integrated into written language teaching as a complementary tool to support and promote WCF processes.

1. Introduction

Although Written Corrective Feedback (WCF) is empirically supported as a tool for promoting language acquisition (Li & Vuono, 2019; Muñoz & Sáez, 2019), its effectiveness remains debated, with factors such as learner differences, task characteristics, and feedback types potentially mediating its impact (Li et al., 2022; Nassaji, 2016). The integration of AI tools, including Automated Writing Evaluation (AWE) systems, introduces new challenges, although they offer benefits like consistency and multiple review opportunities (Koltovskaia, 2020, Nassaji, 2024a,2024b, 2025). Studies comparing human and AI-generated feedback have yielded mixed results: some favour human correction (Kaivanpanah et al., 2020), while others highlight the effectiveness of computational systems (Sistani & Tabatabaei, 2023). Among these, ChatGPT has generated significant interest, but empirical research on its role in language learning remains limited (Yan, 2023).

ChatGPT is a free AI-based language model, or GenAI chatbot, launched in November 2022. It was trained on large corpora from the internet and scanned books to statistically predict the most likely sequence of words in response to user prompts (OpenAI, 2023). GenAI tools like ChatGPT are increasingly integrated into education for their ability to generate, synthesize, and modify natural

* Corresponding author. Street 2, N° 1091, Dept. 8, Lomas de Bellavista, Concepción, Chile, Postal Code 4030000.
E-mail addresses: bmunoz@ubiobio.cl (B.C. Muñoz Muñoz), nassaji@uvic.ca (H. Nassaji), fbello2017@udec.cl (F.I. Bello Carrillo).

https://doi.org/10.1016/j.system.2025.103805
Received 27 January 2025; Received in revised form 31 July 2025; Accepted 4 August 2025
Available online 16 August 2025

language with high sophistication (Baktash & Dawodi, 2023). Recent studies have shown that ChatGPT can support various academic tasks, including essay, poem, and story writing, answering questions, and summarising or paraphrasing texts (Escalante et al., 2023).

ChatGPT has also shown potential to enhance educational processes more broadly. It has been associated with facilitating personalised and adaptable learning experiences (Baidoo-Anu & Owusu, 2023; Baskara & Mukarto, 2023; Farrokhnia et al., 2023; Xiao & Zhi, 2023; Zhang, 2023), improving the organisation of assessment processes (Deng & Lin, 2023; Rudolph et al., 2023), and addressing educational challenges (Kasneci et al., 2023). In language education, ChatGPT has been linked to the development of key linguistic skills and increased learner engagement (Baskara & Mukarto, 2023; Hong, 2023; Kohnke et al., 2023; Songsiengchai et al., 2023). Emerging research further highlights its potential to assist second language learners with writing tasks, including planning, revising, and idea generation (Li et al., 2024; Song & Song, 2023; Yan, 2023).

Specifically, within the field of WCF, a growing body of research has offered a good understanding of how ChatGPT-generated feedback compares to that of human teachers in L2 writing instruction. Evidence suggests that ChatGPT can match teacher feedback in areas such as vocabulary development, textual cohesion, and time efficiency (Cao & Zhong, 2023; Mizumoto & Eguchi, 2023; Song & Song, 2023) but remains limited in addressing complex aspects requiring cultural awareness and contextual understanding (Cao & Zhong, 2023; Song & Song, 2023). As a result, ChatGPT is increasingly viewed as a supplementary tool rather than a replacement for teacher feedback (Nassaji, 2024a). This perspective is particularly relevant in instructional settings, where ChatGPT may help reduce teacher workload while providing timely, personalised feedback to learners (Gul et al., 2016; Zou et al., 2023).

Nevertheless, this line of research remains in its infancy, largely due to the recent introduction of ChatGPT in educational contexts and the limited generalisability of existing findings. Moreover, diverse EFL settings require further investigation, as most studies to date have been conducted in Asian contexts (Li et al., 2024). This gap is particularly evident in the Chilean context, where, to our knowledge, no studies have yet examined the effectiveness of ChatGPT as a feedback source.

This study aims to compare the effectiveness of direct WCF provided by a human teacher and delivered by ChatGPT in supporting L2 writing. The primary goal is to assess and compare the impact of this type of feedback from both sources on improving students' writing, with a focus on four key aspects: task response, cohesion and coherence, lexical resource, and grammatical range and accuracy. Specifically, the study seeks to evaluate how the feedback from ChatGPT and a human teacher influences these aspects of writing and to identify which areas benefit most from each type of feedback. Direct WCF provides explicit corrections (e.g., supplying the correct form), making it easier to isolate and compare the effects of feedback from different sources. This also allows for a more controlled and consistent comparison between ChatGPT and human feedback, minimizing variation in feedback style or interpretation.

2. Literature review

2.1. Written corrective feedback

Written Corrective Feedback (WCF) refers to any written indication of errors in learners' texts, including markings, symbols, reformulations, and textual annotations, aimed at correcting linguistic inaccuracies at both local (grammar, lexis) and global (content, organisation) levels (Crosthwaite et al., 2022; Montgomery & Baker, 2007). In L2/FL contexts, WCF serves two main purposes: fostering the development of writing skills and enhancing second language acquisition (Ferris, 2012; Sheen, 2011). It represents a form of interaction between learners and someone more knowledgeable, which is pointed out as a vital practice to offer both assessment and guidance to support students' development (Cen & Zheng, 2024).

According to the literature, WCF is usually grouped into two broad categories: direct and indirect feedback (Ellis, 2009; Sheen, 2011; Nassaji, 2020). Direct feedback gives students the correct version of an error, while indirect feedback draws attention to the mistake without fixing it. Indirect correction might include underlining, circling, margin notes, or the use of error codes. Both types of feedback, direct and indirect, can be supported with metalinguistic explanations that clarify the nature of the error (Muñoz et al., 2023).

Theoretically, indirect feedback is believed to promote deeper cognitive engagement by encouraging learners to self-correct, thereby fostering greater long-term learning. In contrast, direct feedback may reduce cognitive load and lead to immediate improvement, particularly for learners who lack the linguistic knowledge to identify and correct their own errors (Muñoz et al., 2023; Muñoz & Sáez, 2019). Thus, each type of feedback can be theoretically justified based on different assumptions about how learners process and internalize corrective input.

The research on the effectiveness of different WCF strategies remains inconclusive. Some findings highlight the benefits of direct feedback, especially for rule-governed errors like article use and verb tense (Bitchener & Knoch, 2009; Ellis et al., 2008). Meanwhile, other studies favour indirect feedback, suggesting it leads to better long-term results by fostering cognitive processing and learner autonomy (Ferris, 2003; Muñoz & Sáez, 2019). Additionally, metalinguistic feedback has been shown to facilitate learners' understanding of errors (Bitchener, 2008), although greater explicitness does not always lead to better learning outcomes (Shintani & Ellis, 2013; Stefanou & Révész, 2015).

Another important distinction in WCF is its focus or scope. Focused feedback targets a limited number of error types, while comprehensive feedback addresses most or all observed errors (Ellis et al., 2008). Studies comparing these approaches report mixed results. Focused WCF is often linked to better learning outcomes, as it reduces cognitive overload and allows students to concentrate on specific structures (Lee, 2019; Sheen, 2007). Conversely, some research favours comprehensive feedback for its greater ecological validity and closer alignment with classroom practice (Colpitts & Howard, 2018; Reynolds & Kao, 2022; Storch, 2010). According to Sheen et al. (2009), teachers tend to adopt a comprehensive approach, viewing it as the core purpose of written correction. However, this broader scope demands considerable time and effort (Han & Sari, 2024), a crucial factor in contexts where large class sizes limit individualised feedback opportunities (see Nassaji, 2016, 2020 for a detailed review of contextual factors).

This issue presents additional challenges for delivering feedback in EFL classrooms. The impracticality to systematically correct numerous texts may diminish the quality, efficacy, and consistency of feedback. Timely responses, which are generally considered more effective, are difficult to achieve in large classes (Bitchener & Knoch, 2008). This also restricts opportunities to provide iterative feedback across different writing assignments and to tailor responses, limiting the potential benefits of more personalised feedback (Hyland & Hyland, 2006). In addition, large class sizes make it difficult to ensure consistency due to the unfeasibility of addressing each student's errors in detail because of time restrictions (Lee, 2008). Nonetheless, technological developments in the field of AI have helped to mitigate some of these difficulties.

2.2. *An overwide of AI tools and WCF in language learning settings*

Recently, the rise of AI in language education has introduced additional dimensions to the ongoing debate in WCF studies. The integration of AI tools in educational settings, such as Automated Writing Evaluation (AWE) systems, has offered some advantages, like consistency and multiple review opportunities (Koltovskaia, 2020; Nassaji, 2024a, 2024b). Theoretically, AWE systems align with cognitive views of language learning, which emphasize the importance of noticing and repeated exposure to linguistic forms. However, studies comparing human versus AI-generated feedback have produced conflicting results, with some favouring human correction (Kaivanpanah et al., 2020), while others highlight the effectiveness of computational systems (Sistani & Tabatabaei, 2023). These mixed findings suggest that the effectiveness of AI-generated feedback may depend on contextual and learner-related variables that are not yet fully understood (Nassaji, 2025).

ChatGPT, an advanced technological chatbot, has become widely known all around the globe. This AI tool has been trained with extensive language corpora sourced from the internet and digitized books, allowing it to statistically determine the most probable word sequences in response to user-generated prompts (OpenAI, 2023). ChatGPT has demonstrated the ability to generate, synthesize, or modify natural language with a high level of sophistication (Baktash & Dawodi, 2023). In education, GenAI applications have been used to write essays, poems, and stories, answer questions, summarize, paraphrase, and synthesize texts (Escalante et al., 2023). In language teaching contexts, it has shown to promote language skills development and learners' autonomy, to improve students' lexical capabilities, to personalise language learning and adjust its responses to students' proficiency levels, to stimulate authentic interactions, to foster greater student engagement, among many other advantages (see Baskara & Mukarto, 2023; Hong, 2023; Kohnke et al., 2023; Songsiengchai et al., 2023).

ChatGPT has also demonstrated significant potential for teaching writing in a second language (Li et al., 2024; Song & Song, 2023; Yan, 2023), and for providing immediate and personalised feedback (Farrokhnia et al., 2023; Han & Li, 2024; Song & Song, 2023), characteristics that have been pointed out as fundamental in feedback effectiveness. The introduction of such an AI tool has also shown to reduce the amount of time invested in the time-consuming process of providing feedback and to foster consistency of revisions (Gul et al., 2016; Zou et al., 2023). These ChatGPT's advantages may have a significant impact on WCF treatments as they may foster their effectiveness, enhancing students' development. They also imply an essential contribution for teachers in charge of designing feedback

Table 1
Descriptive statistics and normality test across groups and time.

Test	Group	Descriptive statistics					Shapiro-Wilk		
		N	**Mean**	**SD**	**Skewness**	**Kurtosis**	**Statistic**	**df**	**Sig.**
T1	G1_Chat	20	1755	2.235	0.131	−0.821	0.938	20	0.217
	G2_Teacher	22	23.55	3.582	0.403	−0.572	0.959	22	0.467
T4	G1_Chat	20	25.63	2.038	−1.008	1.501	0.906	20	0.054
	G2_Teacher	22	23.82	3.850	0.499	0.372	0.934	22	0.152

Table 2
Wilcoxon test results for the differences between pre-test and post-test scores.

Group		Ranks	N	Mean rank	Sum of ranks	z	Sig.	η^2
G1 _Chat	post-test-pre-test	Negative ranks	20	10.5	210	−3.928	<0.001	0.772
		Positive ranks	0	0	0			
		Equal						
		Total	0					
G2_Teacher	post-test-pre-test	Negative ranks	11	6.0	66	−0.898	0.369	0.040
		Positive ranks	6	3.5	21			
		Equal	5					
		Total						

sessions as the possibility to save time and effort allows them to focus on other aspects of instruction (Kasneci et al., 2023).

Nonetheless, despite its potential, several limitations of ChatGPT have been identified. Scholars have noted that ChatGPT may produce biased, inappropriate, or discriminatory content due to its mono-cultural orientation (Hacker et al., 2023; Sallam, 2023), and it can generate inaccurate or misleading information (Borji, 2023; Nassaji, 2024a). Its algorithmic nature may also limit its ability to fully grasp complex concepts or cultural nuances (Baskara & Mukarto, 2023). Effective use of ChatGPT requires specific competencies, including AI literacy and prompt design skills (Strobelt et al., 2023). Ethical concerns have also been raised, particularly regarding potential risks to academic integrity, such as plagiarism and cheating (Farrokhnia et al., 2023). Moreover, Song and Song (2023) highlighted challenges related to the contextual accuracy of ChatGPT's responses, noting that students must apply higher-order thinking skills to critically evaluate the feedback. They also found that learners were cautious about relying too heavily on AI-generated feedback, fearing it could hinder independent critical thinking and creativity.

In conclusion, the integration of ChatGPT in L2 writing instruction has significantly transformed the landscape, particularly due to its potential to provide effective WCF, a key component for writing skills development (Zhang, 2023; Zhu et al., 2023). However, research findings have also shown some important limitations regarding its use. Therefore, it is essential to investigate both the advantages and pitfalls this AI system presents, taking into account the multiple variables that may influence its use as an effective feedback provider (Yan, 2023). Given the sudden emergence of this AI tool, research is still in its early stages; thus, as declared by Lundberg (2023, p.5), "the question remains whether AI-generated feedback, specifically ChatGPT feedback, has the same effects as teacher feedback."

2.3. ChatGPT- and human-generated feedback

The comparison between AI-generated and human feedback has emerged as a central focus in recent research on language learning and writing instruction. With tools like ChatGPT becoming widely used in educational contexts, researchers have begun to explore whether AI can match, or even surpass, teachers in providing effective WCF. This interest has led to an important but still unanswered question: should AI be seen as an alternative to teacher feedback, or as a tool that complements and supports it? This question is both theoretically and pedagogically important. Theoretically, it calls for a re-examination of the role of feedback in second language learning, particularly in terms of how feedback contributes to the learning process. Feedback is often seen not just as a way to correct errors, but as a tool to increase learner awareness, encourage engagement, and foster deeper understanding. As AI becomes more integrated into educational settings, researchers are now tasked with exploring whether machine-generated feedback can support learning in ways that are similar to human feedback or if it operates through fundamentally different processes. Pedagogically, the question has significant implications for classroom practice, teacher workload, and learner autonomy. If AI tools can reliably enhance or supplement teacher feedback, they may offer important learning opportunities and reduce the burden on teachers, especially in large or resource-limited classrooms. Rather than replacing educators, many scholars propose that AI has the potential to work alongside teachers, amplifying the quality and efficiency of feedback. AI, for example, can provide immediate, consistent feedback, which can be particularly helpful in large classrooms, and identify common patterns in students' errors to inform instruction. There are currently several studies re-examining the strengths and limitations of both AI and human feedback in L2 writing instruction, contributing to the ongoing discussion about the evolving role of AI in L2 writing instruction.

Escalante et al. (2023) conducted a pre- and post-test study with 48 university-level English learners in the Asia-Pacific region to compare the effectiveness of AI-generated and teacher-provided feedback. Students wrote 300-word academic texts and completed a

Table 3
Mann-Whitney *U* test results for the differences of pre-test and post-test scores between groups.

Test	Group	Mean rank	Sum of ranks	Mann-Whitney U	Wilcoxon W	Z	Sig.
Pre-test	G1_Chat	12.03	240.5	30,5	240.5	−4.77	<0.001
	G2_Teacher	30.11	662.5				
Post-test	G1_Chat	25.45	509	299	394	1.99	0.0468
	G2_Teacher	17.91	394				

Table 4
Friedman test results for each criterion between the G1_Chat and G2_Teacher groups.

Rank	Mean rank							
	G1_Chat				G2_Teacher			
	C1_TR	C2_CC	C3_LR	C4_GR	C1_TR	C2_CC	C3_LR	C4_GR
T1	1.23	1.18	1.15	1.15	2.43	2.55	2.3	2.2
T2	2.48	2.03	2.03	1.98	2.2	2.48	2.18	1.93
T3	2.75	3	3.13	3.2	2.23	2.05	2.55	2.75
T4	3.55	3.8	3.7	3.68	3.14	2.93	2.98	3.11
N	20				22			
Chi^2	39.175	50.597	50.869	51.486	9.268	6.880	6.942	14.337
Df	3				3			
Sig.	<0.001	<0.001	<0.001	<0.001	0.026	0.076	0.074	0.002

weekly questionnaire on feedback preferences. Using an analytic rubric, the researchers assessed the written tasks on content, coherence, language use, and use of sources and evidence. The results showed no significant differences in language learning between the two feedback groups. In terms of learners' preferences, the researchers found that half of the students favoured human feedback due to a preference for face-to-face interaction, while the other half preferred ChatGPT feedback for its clarity, detail, and readability. Based on these findings, the authors suggest a complementary approach that integrates both human and AI-generated feedback to leverage the strengths of each, potentially enhancing instructional efficiency and reducing teachers' feedback workload.

Cao and Zhong (2023) took the discussion further by including self-correction alongside AI and teacher feedback, offering a more detailed understanding of how different feedback sources affect lexicon, syntax, and cohesion. Forty-five EFL students of a master's program in translation and interpretation in China completed a Chinese-to-English translation task. Results showed that ChatGPT feedback was more effective in promoting lexical structures and referential cohesion, while teacher feedback and self-correction better supported overall translation quality and syntactic development. However, limitations were noted in ChatGPT's ability to detect errors in culturally sensitive translations, likely due to its English-centric orientation. Inconsistencies across student outputs also raised concerns about its reliability, particularly considering that ChatGPT remains in a developmental phase. Additionally, the study did not specify which version of ChatGPT was used, a notable omission given the tool's rapid evolution and the improvements observed in newer versions (Zhang, 2023).

Similarly, Dai et al. (2023) investigated the internal characteristics of AI-generated feedback by analysing 103 postgraduate student reports that had already received teacher comments. They focused on three aspects: the readability of ChatGPT feedback, its alignment with teacher feedback, and the presence of effective feedback elements. ChatGPT was found to generate more detailed feedback, and summarize student writing more fluidly and coherently. It also provided process-focused feedback in over half of the reports, an approach known to support deeper learning (Hattie & Timperley, 2007). However, ChatGPT's feedback did not consistently align with teacher evaluations, tending to be overly positive. This discrepancy was likely due to the lack of carefully designed prompts to guide its assessment. Although both ChatGPT and teachers provided task-level feedback, the findings suggest that ChatGPT should not be used as the sole feedback source. Nonetheless, they reinforce the view that ChatGPT can complement and enhance the teacher's role in the feedback process.

Mizumoto and Eguchi (2023) examined the reliability of ChatGPT for automated essay scoring (AES) by analysing 12,100 TOEFL11 essays written by speakers of 11 different first languages. Using the ChatGPT-based text-davinci-003 model, they instructed the tool to assign scores (0–9) based solely on linguistic features such as lexis, syntax, and cohesion, without offering explanations. The AI tool was instructed to assign scores from 0 to 9 without providing explanations for the assigned ratings. The findings revealed that this AI followed the same general patterns as the benchmark TOEFL11 level, suggesting that it can be effectively used as a practical tool for consistent and efficient scoring. The study also showed that the model significantly enhanced the prediction of benchmark levels, which may lead to better alignment with professional ratings. However, the researchers noted that ChatGPT still does not perfectly match human correction; therefore, it should be used as a supportive rather than an autonomous scoring tool.

Expanding the focus beyond scoring and feedback alignment, Song and Song (2023) conducted a mixed-method pre- and post-test study to examine the impact of ChatGPT on writing development and motivation among 50 EFL undergraduates in China. The experimental group, which used ChatGPT alongside instruction, showed notable improvements in organisation, coherence, grammar, and vocabulary, as well as increased motivation compared to the control group. Students appreciated the tool's accessibility, the chance to practise independently, and the immediate, personalised feedback it provided. However, some also raised concerns about potential overreliance on AI, limitations in contextual accuracy, and uncertainty regarding its long-term role in writing instruction. The contextual issue mirrors the concerns raised by Cao and Zhong (2023) regarding ChatGPT's limitations in addressing culturally specific or stylistically nuanced language.

Han and Li (2024) expanded existing research by examining the types and features of ChatGPT-supported feedback and how EFL students used it during text revision. The study involved 95 undergraduate students in a World Language Education programme in China. Two feedback types were explored: indirect coded feedback on 15 common errors (adapted from ChatGPT's direct responses) and holistic rhetorical feedback, which, though rarely provided by teachers, is essential for argument clarity, coherence, and idea organisation (Aljasir, 2021; Carter & Thirakunkovit, 2019). The findings suggest that combining AI with teacher input enhances feedback quality and depth, reduces future errors, and better supports writing development. This collaboration also eases teacher workload while offering students timely, personalised guidance. The study positions ChatGPT as a valuable co-facilitator in the feedback process, improving both instructional efficiency and learner engagement.

Adopting an instructor-centred perspective, Lin and Crosthwaite (2024) compared WCF provided by 25 EFL/ESL teachers with that generated by ChatGPT. Teachers predominantly used a mix of direct and indirect feedback, while ChatGPT relied mainly on metalinguistic explanations and reformulations, occasionally offering direct feedback and rarely using indirect strategies. However, ChatGPT often over-corrected, unnecessarily altering sentence or phrase structures and did not account for learners' proficiency, writing goals, or error types. Although more accurate than teachers at sentence-level corrections, ChatGPT focused primarily on local issues. In contrast, teachers addressed broader, global concerns through indirect WCF and followed more consistent patterns. The study simulated real classroom conditions by avoiding advanced prompting, allowing both AI and human participants to freely select feedback strategies—an aspect likely influencing the results.

Additionally, Wang et al. (2024) conducted a retrospective qualitative study using ChatGPT 3.5 to evaluate 50 argumentative texts from 42 sophomores at a university in Southern China, previously assessed by teachers. ChatGPT provided faster and more comprehensive linguistic feedback than teachers, though its accuracy declined with longer texts and frequent discourse markers. In contrast, teacher feedback—though slower and occasionally incomplete—reflected affective and personalised insights based on students' academic backgrounds. A key limitation of ChatGPT was its inability to assess argument quality effectively, often producing generic responses due to difficulties in analysing logical relationships. This highlights the need for a collaborative feedback model combining AI and teacher input. The researchers linked this issue to prompt quality, a concern echoed by Jacobsen and Weber (2025), who noted that "only the best prompt produced consistently high-quality feedback" (p. 2). These findings also align with Lin and Crosthwaite (2024), who reported that poorly designed prompts reduced ChatGPT's feedback effectiveness.

In conclusion, the reviewed studies offer a comprehensive view of how ChatGPT-generated feedback compares to that of human teachers in L2 writing instruction. While ChatGPT has demonstrated strong performance in areas like lexical enhancement, textual cohesion, and feedback efficiency, it continues to fall short in more complex dimensions such as contextual understanding, cultural sensitivity, and the affective qualities often present in teacher feedback. Several studies suggest that ChatGPT performs best when fed with strong prompts and used in collaboration with teachers; thus, rather than replacing teachers, this AI tool seems most effective when it complements teachers' work, supporting the revision process, easing workload, and giving students timely input. Nonetheless, this line of research is still in its infancy, and the generalisability of findings remains a challenge—particularly given the many EFL and L2 contexts that have yet to be thoroughly explored. This is especially true in the Chilean context, where, to our knowledge, no studies have addressed this issue.

To better understand how AI tools compare with traditional instructional methods, this study investigates the relative effectiveness of WCF provided by a human teacher and that generated by ChatGPT. The focus is on direct feedback and its contribution to the development of L2 writing skills, particularly in the areas of task response, cohesion and coherence, lexical resource, and grammatical range and accuracy. By examining learners' writing performance following feedback from these two sources, the study aims to shed light on the specific strengths of each, and the aspects of writing most positively influenced.

The study addresses the following two research questions.

1. How does ChatGPT-provided direct WCF compare to human-provided direct WCF in improving EFL students' essays at a Chilean university?
2. Which aspects of the essays benefit most from ChatGPT versus human feedback?

3. Methods

The proposed study adopts a quasi-experimental pre-test-post-test design. Participants were 44 first-year English Pedagogy students (26 females, 18 males) aged 19–21 from a Chilean university, enrolled in a mandatory course on English language didactics as part of their pre-service teacher education program. Although learners' additional exposure to English outside the classroom could not be fully controlled, their parallel courses addressed different topics, and such variation reflects the ecological validity of the research.

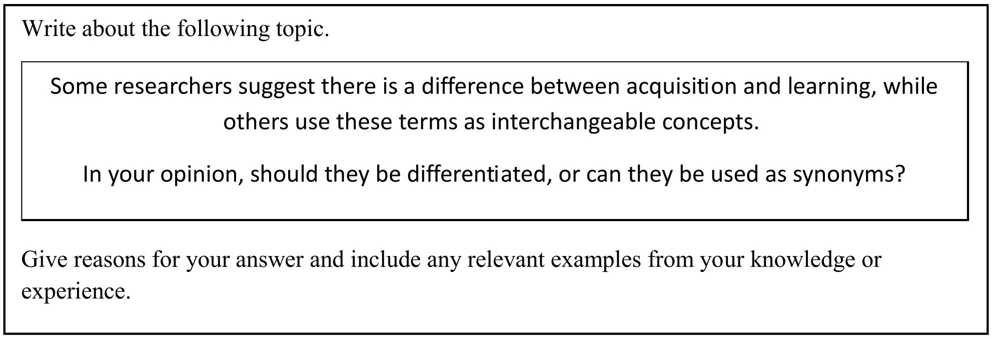
All participants were classified as between high A2 and low B1 levels according to the Common European Framework of Reference for Languages (CEFR). This classification was based on the results of the annual institutional placement test administered at the start of the academic year. The test, routinely used by the institution, evaluates grammar, vocabulary, reading, and writing skills to place students into appropriate language courses. The range reflects the reality of classroom composition in educational contexts, where students within the same level band (A2–B1) often display individual variation in specific language abilities. For this study, each section was randomly assigned an experimental condition: one group (N = 20) received corrective feedback from ChatGPT (G1_Chat), while the other (N = 22) received feedback from a human teacher (G2_Teacher).

The instructor responsible for G2_Teacher feedback was a hired teacher with eight years of experience in pre-service education, a master's degree in English teaching, and expertise as an international examiner and research collaborator. He was specifically trained to provide targeted feedback (see Section 2.3). Ethical procedures followed the University's Bioethics and Biosafety Committee guidelines, ensuring voluntary participation and the right to withdraw at any time.

3.1. *Writing tasks*

The writing tasks comprised four opinion essays based on the IELTS Writing Task 2 with the first and fourth essays serving as the pre-test and post-test, respectively. The IELTS tasks were chosen for the following reasons: a) IELTS is commonly used by many Chilean universities to assess the linguistic proficiency of English Pedagogy students before graduation; b) The students in the sample are familiar with this writing tasks of IELTS as part of their regular activities, which helps them prepare for international exams, c) a specific genre tends to promote a more homogeneous use of lexical and grammatical resources.

The prompts covered a range of academic topics addressed in the course to make sure that students possess the necessary knowledge to effectively express their opinions. The overall topics included: language acquisition theories, teaching methods, acquisition versus learning, and emotional variables in language learning. The instructions and number of words required followed those of the original IELTS exam. A sample of one of the essay prompts is presented below.



Write about the following topic.

Some researchers suggest there is a difference between acquisition and learning, while others use these terms as interchangeable concepts.

In your opinion, should they be differentiated, or can they be used as synonyms?

Give reasons for your answer and include any relevant examples from your knowledge or experience.

To validate the writing prompts, six academics from different English Pedagogy Programs completed a Likert scale assessment on the quality of the tasks and test prompts, the number of tasks, and the equivalence test. Kendall's W was used to evaluate the level of expert agreement. Results indicated moderate but significant agreement among these experts ($W = 0.423$; $p = 0.007$). Based on their comments, prompts with lower agreement values were revised. The final prompts were then piloted with a group of 10 students with similar characteristics in order to identify potential issues such as time, comprehension of instructions, and genre characteristics. These participants were not included in the final sample.

The IELTS rubric was used to assess the writing tasks because it is a standardized, globally consistent instrument. The rubric includes four criteria: task response, cohesion and coherence, lexical resource and grammatical range and accuracy. Each criterion is rated on a scale from 0 to 9; the final task score corresponds to the average of these four ratings.

3.2. Data collection and procedures

The study was conducted over 4 sessions across 2 months. In session 1, students in both the ChatGPT and the teacher groups completed Writing Task 1. They received a pen-and-paper form containing the prompt and were given 40 min to complete the task. Students were not allowed to write outside the classroom as they usually use tools such as translators or AI tools to improve their writing, hindering proper composition. Before handing in their texts, students were encouraged to read through their writing and make surface-level corrections as needed, such as for spelling or grammar. They were also encouraged to align their writing with the rubric criteria, which had been shared with them in advance.

In session 2, two weeks later, they received their written tasks back with the rubric assessing the criteria. Each group received WCF according to its experimental condition. ChatGPT-generated feedback was transferred to a paper rubric for students, eliminating any potential influence of the feedback medium. Students were given 10 min to review their writings and the rubric. They were encouraged to compare their original texts and the feedback received in order to understand the comments given and the level achieved in each criterion. After this, the instructor of the course collected all the texts and rubrics before assigning a new writing task. This process was repeated in sessions 3 and 4. Re-writing was not considered in this study in order to avoid decontextualized tasks; students in the researched context are not used to rewriting texts as part of their teaching process.

The intervention was integrated into the course curriculum to prevent it from being perceived as an external activity, which may have helped control mediating variables such as attitude, motivation, and engagement that could affect feedback effectiveness.

Session	Time	Task
1	● 40 min	● Task 1 (T1)
2	● 10 min review ● 40 min	● Corrective feedback Provision (T1) ● Task 2 (T2)
3	● 10 min review ● 40 min	● Corrective feedback Provision (T2) ● Task 3 (T3)
4	● 10 min review ● 40 min	● Corrective feedback Provision (T3) ● Task 4 (T4)

3.3. Feedback provision

The feedback provided was based on the IELTS Task 2 rubric, which evaluates writing across four key criteria: task response, cohesion and coherence, lexical resource, and grammatical range and accuracy.

For the G2_Teacher group, feedback was delivered by an experienced EFL teacher, well-trained in IELTS evaluation standards. The teacher assessed the students' written texts by assigning a band score for each of the four criteria, offering general comments, and providing direct written corrective feedback (WCF). This WCF strategy was selected for two main reasons. First, it is one of the preferred WCF strategies among Chilean learners (Muñoz et al., 2023), making it highly relevant and pedagogically authentic. Second, it aligns closely with the type of feedback ChatGPT is capable of generating. Unlike indirect or metalinguistic feedback, which require learners to interpret or infer corrections, direct feedback offers explicit revisions, something that AI systems like ChatGPT can readily provide. As noted earlier, this comparability allows for a more equitable evaluation of how feedback from a human teacher and an AI tool affects learner outcomes, since both sources deliver corrections in a similarly straightforward format (see Appendix A).

The teacher received the same instructions as those given to ChatGPT to ensure both feedback sources were comparable (see Appendix B). The only difference between the two conditions was the format of feedback presentation: the teacher marked corrections directly above the errors within the students' texts, while ChatGPT presented its feedback in a structured two-column table format, listing the original errors alongside the corrected versions. The researchers conducted a series of test trials prior to data collection to ensure alignment between the feedback provided by ChatGPT and that provided by the human teacher, particularly in terms of correction type, level of explicitness, and adherence to the four IELTS criteria.

The G1_Chat group received feedback generated by ChatGPT, specifically using the GPT-4o model, the latest version developed by Open AI. This model accepts text, image, and video inputs with a responsive time as low as 232 ms (OpenAI, 2023). ChatGPT was instructed to provide feedback on these same four rubric criteria to enable comparison. As ChatGPT requires specific prompts, it was trained using the procedure outlined by Escalante et al. (2023) as follows.

- ChatGPT was assigned the role of an English language teacher.
- ChatGPT was instructed to focus on the four rubric criteria of the IELTS rubric: task response, cohesion and coherence, lexical resource, and grammatical range and accuracy.
- ChatGPT was asked to generate a table identifying linguistic errors, with one column specifying the error and a second column providing a suggestion for correction.

ChatGPT was instructed to identify inaccuracies, taking into account the students' proficiency levels and instructional setting, in the areas of spelling, capitalization, punctuation, singular and plural nouns, verb tense, subject-verb agreement, word form, awkward phrasing, prepositions, articles, and sentence fragments and run-ons. Feedback was provided for each essay in a single session by the same researchers. A new chat session was created for each of the writing texts to prevent any retention bias that could have introduced potential variation in WCF response (Lin & Crosthwaite, 2024).

3.4. *Data analysis*

Descriptive statistics for G1_Chat and G2_Teacher scores on the pre-test (T1) and post-test (T4), along with Shapiro-Wilk test results, are presented in Table 1. Results indicated that both groups were normally distributed ($p > 0.05$). However, Levene's test for equality of error variances indicated a lack of homogeneity for the pre-test ($p = 0.048$). Consequently, non-parametric tests were chosen for the analysis: Wilcoxon Signed-Rank and Mann-Whitney U tests. The analysis was conducted in Python, primarily utilizing SciPy and statsmodels libraries.

4. Results

A Wilcoxon Signed-Rank test was conducted to examine differences in writing performance between the pre-test and post-test. Results, shown in Table 2, indicate a significant improvement in writing scores for learners in G1_Chat ($z = -3.928$, $p < 0.005$) but no significant difference in G2_Teacher ($z = -0.898$, $p = 0.369$). The sum of the difference scores and the mean rank suggests that this significant difference favours the post-test results.

Eta squared value[1] (η^2) was calculated to assess the intervention's effect size. Results showed a large effect size for G1_Chat group ($\eta^2 = 0.772$), and a small effect size for G2_Teacher group ($\eta^2 = 0.040$). This indicates that the G1_Chat group experienced a more substantial impact from ChatGPT feedback on post-test scores compared to the human teacher feedback for the G2_Teacher group.

Mann-Whitney *U* test was run to underscore the statistical significance of mean rank differences between the G1_Chat and the G2_Teacher groups in both pre-test and post-test scores. Table 3 reveals that, prior to the intervention, the groups had a statistically significant difference in performance with a U statistic of 30.5 and a z-value of -4.77 ($p < 0.001$). The negative z-value indicates that G1_Chat had lower ranks than G2_Teacher, suggesting that this last group had higher initial scores. It is important to note that this difference emerged despite the fact that all participants were classified as being between high A2 and low B1 levels according to the CEFR. This could be because the CEFR levels represent broad proficiency bands, which may not capture fine-grained differences in individual learner ability. It is, therefore, possible that the groups differed more in actual performance than the CEFR classification suggests. Additionally, the pre-test instrument may have been more sensitive to specific language skills measured, providing a more precise measure of these initial differences.

However, following the intervention, a significant difference was observed in post-test scores, with a U statistic of 299.0 and a z-value of 1.99 ($p = 0.046$), indicating a statistically significant improvement in the performance of the G1_Chat group. The positive z-value suggests that G1_Chat ranked higher than G2_Teacher in the post-test. Importantly, although G2_Teacher had significantly higher performance prior to the intervention, G1_Chat not only caught up but also significantly outperformed G2_Teacher by the end of the study. This finding highlights the effectiveness of the chat-based intervention in promoting learning gains, particularly in writing

[1] Norouzian and Plonsky (2018) establish the cutoff values for eta squared as follows: small size ($\eta^2 = 0.01$), medium size ($\eta^2 = 0.06$), and large size (η^2 =0.14).

proficiency. Despite the initial advantage held by the G2_Teacher group, the G1_Chat group demonstrated the most substantial improvement, suggesting that ChatGPT-based feedback had a greater positive impact on learner outcomes.

The Friedman test was used to evaluate whether there were significant differences in each rubric criterion's scores within the same group throughout the treatment. Results in Table 4 show that the G1_Chat group significantly improved across all four criteria (TR, CC, LR, and GR), whereas the G2_Teacher group showed significant improvement only on criteria 1 (TR) and 4 (GR).

5. Discussion

This study compared the effectiveness of direct written corrective feedback (WCF) from ChatGPT and human teachers in supporting L2 writing development. It examined whether AI-driven feedback could serve as an alternative or complement to traditional feedback. By evaluating participants' writing after receiving feedback from both sources, the study offers insights into AI's potential role in L2 writing instruction.

The first research question was: How does ChatGPT-provided direct WCF compare to human-provided direct WCF in improving EFL students' essays at a Chilean university?

The findings showed that both traditional teacher feedback and ChatGPT-based feedback positively supported L2 writing development. However, ChatGPT feedback produced a larger effect size, indicating a stronger impact on learner improvement, although no statistically significant differences were found, learners who received AI-based feedback showed greater overall progress. This aligns with Song and Song (2023), who observed notable writing improvements attributed to ChatGPT use. The results contribute to growing evidence supporting AI tools as effective complements to traditional teaching, a view also shared by Cao and Zhong (2023) and Mizumoto and Eguchi (2023). Several factors may explain these findings.

ChatGPT may also be effective because it is based on a large database of language patterns, grammar rules, and writing practices, enhancing the clarity and specificity of its feedback. However, contrary evidence exists. Wang et al. (2024) found that teacher feedback was more personalised and affective than ChatGPT's when assessing argumentative texts. This discrepancy may reflect contextual differences: in Wang et al.'s study, teachers knew their students and offered formative comments informed by learners' linguistic development. In contrast, the teacher in the present study had no prior relationship with the participants, having been hired solely to provide feedback on their texts. This difference in teacher-student relationship may have influenced how human feedback was perceived and taken up, possibly contributing to the stronger impact of AI-generated feedback in our context. These findings highlight the need for further research into how teacher-student dynamics affect the effectiveness of feedback provision and uptake, particularly when AI tools are introduced as part of the process.

The present findings, which suggest the superiority of ChatGPT-generated feedback, contrast with those of Cao and Zhong (2023), who found teacher feedback to be more effective in improving writing quality. This discrepancy may stem from differences in task type. While our participants wrote opinion essays, Cao and Zhong's participants completed translation tasks from Chinese to English. These tasks involve distinct cognitive and linguistic demands: opinion essays require generating original content, expressing viewpoints, and organising persuasive arguments, whereas translation involves accurately rendering existing content while preserving meaning, tone, and structure. As such, translation feedback often targets specific linguistic issues, enabling direct correction, while opinion writing demands more abstract revisions, such as refining arguments or clarifying ideas, which may be harder to implement. The observed differences in feedback effectiveness may reflect how these task characteristics interact with the type of feedback provided. This supports the view that task type significantly influences the effectiveness of WCF (Nassaji, 2016). Hence, it can be suggested that WCF generated by ChatGPT is effective for assessing written essays, which is in line with what Su et al. (2023) propose about the effectiveness of this tool in improving learners' performance when writing argumentative essays. Thus, the need to explore which other types of texts benefit from ChatGPT-based feedback remains.

Cultural awareness is also a crucial consideration in feedback on translation tasks. ChatGPT's mono-cultural orientation may limit its ability to interpret culturally nuanced meanings essential to accurate translation (Hacker et al., 2023; Sallam, 2023). Zhai (2022) warns that this limitation can reinforce stereotypes and discriminatory representations. This aligns with the *garbage-in, garbage-out* principle (Farrokhnia et al., 2023), as ChatGPT's data-driven algorithms may replicate historical biases rather than promote equity. As Baskara and Mukarto (2023) note, "irrespective of its advanced capabilities, ChatGPT, like any AI tool, cannot fully grasp and interpret subtle cultural intricacies" (p. 109). This poses a challenge, particularly in translation tasks, where cultural competence influences lexical and syntactic choices (Cao & Zhong, 2023). As AI-generated feedback becomes more common, future research should explore how cultural bias may shape learner writing. Educators must also guide students in critically engaging with AI feedback.

The second research question was: Which aspects of the essays benefit most from ChatGPT versus human feedback? Specifically, it explored how different elements of L2 opinion essay writing, such as task response, cohesion and coherence, lexical resource, and grammatical range and accuracy, are affected by feedback from ChatGPT versus that from a human teacher.

As previously noted, both ChatGPT-generated and teacher-provided feedback supported overall writing development. However, ChatGPT feedback proved more effective by the end of the intervention. As shown in Table 4, all four IELTS rubric components improved significantly in the G1_Chat group, while only task response and grammatical range and accuracy reached significance in the

G2_Teacher group. This aligns with Song and Song (2023), who found that ChatGPT supports improvements in organisation, coherence, grammar, and vocabulary. Notably, the lack of significant gains in cohesion/coherence and lexical resource in the teacher feedback group prompts further reflection. These results support earlier findings that link ChatGPT to positive outcomes in lexical development, which in turn can enhance textual cohesion. In particular, our results reinforce Cao and Zhong's (2023) conclusion that ChatGPT is especially effective in promoting lexical complexity. This strength may stem from its vast and continuously updated lexical database, which allows for precise and focused feedback, potentially surpassing what a single teacher can offer in terms of lexical input and scaffolding.

In addition to its impact on vocabulary and grammatical accuracy, ChatGPT's influence may extend to cohesion and coherence, as the overall writing organisation may benefit from the improvement of these linguistic features (Song & Song, 2023). These findings are particularly relevant, as they contribute to the growing body of research examining how ChatGPT can be integrated into feedback practices. Given the stronger effects observed for AI-generated feedback in this context, further research is needed to explore how such tools might be effectively incorporated into hybrid feedback models where teacher feedback is complemented by AI-generated comments, particularly in varied instructional settings.

Building on these insights, the results suggest that ChatGPT supports both low-order (e.g., lexical resource, grammatical range) and high-order (e.g., task response, cohesion, coherence) writing skills, as assessed by the IELTS rubric. This contrasts with Farrokhnia et al. (2023), who, based on a SWOT analysis, identified ChatGPT's failure to develop higher-order thinking skills as a limitation. However, their claim appears based on secondary sources, whereas the present findings are empirically grounded.

Nevertheless, it is important to acknowledge that the current findings differ from previous studies that did not report superior performance by ChatGPT. For example, Lin and Crosthwaite (2024) found that ChatGPT-generated feedback was inconsistent, often over-corrected sentence structures unnecessarily, and failed to account for learners' proficiency levels. One possible explanation lies in the quality of the prompts used: aiming to replicate authentic classroom conditions, the researchers did not provide precise or guided prompts, allowing both ChatGPT and teachers to choose any feedback strategy. This lack of standardisation likely contributed to variability in feedback and complicated direct comparisons between the two sources.

While such methodological differences warrant attention, another important point to consider is the difference between immediate writing performance and long-term language learning. While AI may enhance L2 writing performance by improving grammar, vocabulary, and overall text quality, its contributions to long-term L2 learning remain a subject of ongoing debate (Nassaji, 2025). From a language acquisition and cognitive science perspective, Nassaji (2025) draws a distinction between writing as a surface-level activity focused on textual accuracy and fluency, and language learning as a deeper, multifaceted process involving the internalization of linguistic structures and the development of communicative competence. He contends that excessive reliance on AI tools, while beneficial for producing polished texts, may inadvertently encourage passive engagement with language. This can diminish learners' opportunities to actively resolve linguistic challenges, engage in self-monitoring, and reflect on language use, cognitive processes widely recognized as essential for fostering long-term language development and autonomous learning. Therefore, although AI-assisted writing may yield immediate performance gains, its pedagogical value must be carefully evaluated in relation to its impact on sustained L2 learning outcomes (Nassaji, 2024a).

Finally, it is important to recognise that the effectiveness of AI-generated feedback may also be influenced by various contextual and methodological factors. For instance, whether feedback is embedded in a required course that contributes to students' final grades or delivered as part of a voluntary task can shape learners' engagement and motivation, as suggested in prior research on teacher-provided feedback. Similarly, the mode of delivery, ranging from extended one-on-one sessions to brief written comments, may affect how feedback is interpreted and applied. These variations can influence learners' ability to benefit from AI-generated feedback and, ultimately, impact writing outcomes. As Nassaji (2016) notes, differences in feedback implementation and assessment procedures are key methodological variables that may account for varied results across studies, including those involving AI tools. While such issues lie beyond the scope of the present research, they highlight valuable considerations for the design of future WCF interventions.

6. Conclusion, implications and limitations

This study aimed to compare the effectiveness of direct WCF provided by ChatGPT versus a human teacher in improving second-year English Pedagogy students' L2 opinion essay writing. The findings suggest that both feedback sources improved writing, but ChatGPT outperformed the teacher across all assessed areas: task response, cohesion and coherence, lexical resource, and grammatical range and accuracy. Thus, it can be concluded that ChatGPT-feedback supports both low- and high-order writing skills, making it a valuable tool beyond surface-level correction. These findings may be attributed to its ability to produce personalised feedback, which can make it more familiar to students and foster a more impactful learning experience, even in the absence of a teacher-student relationship that typically allows for tailored support. Additionally, ChatGPT appears to provide more balanced feedback by equally addressing positive and negative aspects of students' writing, an aspect pointed out as essential for effective feedback uptake (Farrokhnia et al., 2023). Nevertheless, it is still premature to advocate for this AI capacity as further empirical evidence is needed (Nassaji, 2024a).

Despite these advantages, some considerations must be acknowledged. First, ChatGPT's effectiveness may depend on the type of writing task. As existing research remains inconclusive, the present findings cannot be generalised to other genres, though they highlight a valuable direction for future inquiry. Second, ChatGPT's feedback is not free from cultural bias, requiring users to critically assess its responses to address context-specific nuances. This issue is particularly important in educational settings historically underrepresented in AI training data, where cultural realities may not be adequately reflected.

In conclusion, the advantages demonstrated by ChatGPT position it as a valuable complementary tool to support teachers in providing corrective feedback on written texts. Integrating ChatGPT into the feedback process could help address the significant time and effort typically required (Dai et al., 2023). Given that time constraints are a major challenge for delivering meaningful and timely feedback (Gul et al., 2016; Zou et al., 2023), the current findings suggest that ChatGPT may help alleviate this burden (Farrokhnia et al., 2023). Its time-saving potential could enable teachers to assign more writing tasks and offer more frequent, consistent feedback—an especially important advantage in EFL contexts, where large class sizes often restrict opportunities for detailed, personalised responses.

As Nassaji (2025) suggested, it is essential to educate both language teachers and students on the effective use of ChatGPT and other AI tools to support language learning and foster AI literacy. Using ChatGPT as a complementary feedback source requires clear and purposeful prompts to generate relevant responses (Jacobsen & Weber, 2025). Additionally, engaging with AI-generated feedback demands higher-order thinking skills to critically assess its content and suitability for specific learning contexts, especially given the tool's known limitations (Cao & Zhong, 2023; Dai et al., 2023; Wang et al., 2024). Until more empirical evidence becomes available, a cautious and informed approach to AI integration in education is recommended.

As with any research, certain limitations must be acknowledged. The brief period of feedback exposure may have influenced the outcomes, suggesting the need for longitudinal studies to examine the long-term effects of AI versus teacher feedback across multiple writing cycles. Additionally, as the study was embedded in a course, external exposure to English could not be fully controlled, potentially affecting student performance. A mixed-method approach could have enriched the findings; for instance, exploring learners' perceptions of ChatGPT or conducting a content analysis of teacher and AI-generated feedback might have offered deeper insight into feedback uptake and learner engagement. We also acknowledge that despite efforts to make ChatGPT's responses consistent by using the same prompt, device, and chat session, some degree of variability in its output is inherent and beyond the researchers' control. While steps were taken to minimize these inconsistencies, this could still represent a limitation of the study and should be considered when interpreting the findings. Finally, the generalisability of the results is limited, as only opinion essays were examined. Future studies should explore other writing genres to broaden the applicability of the findings.

Declaration of interest statement

We have nothing to declare.

Data statement

Research data are unavailable due to confidentiality, in accordance with the ethical guidelines of the institution.

Funding sources

This work was supported by *the National Agency for Research and Development, ANID Chile, Fondecyt Iniciación [11230546]; FDD-*UBB*, Chile [FDD2024-24]*.

CRediT authorship contribution statement

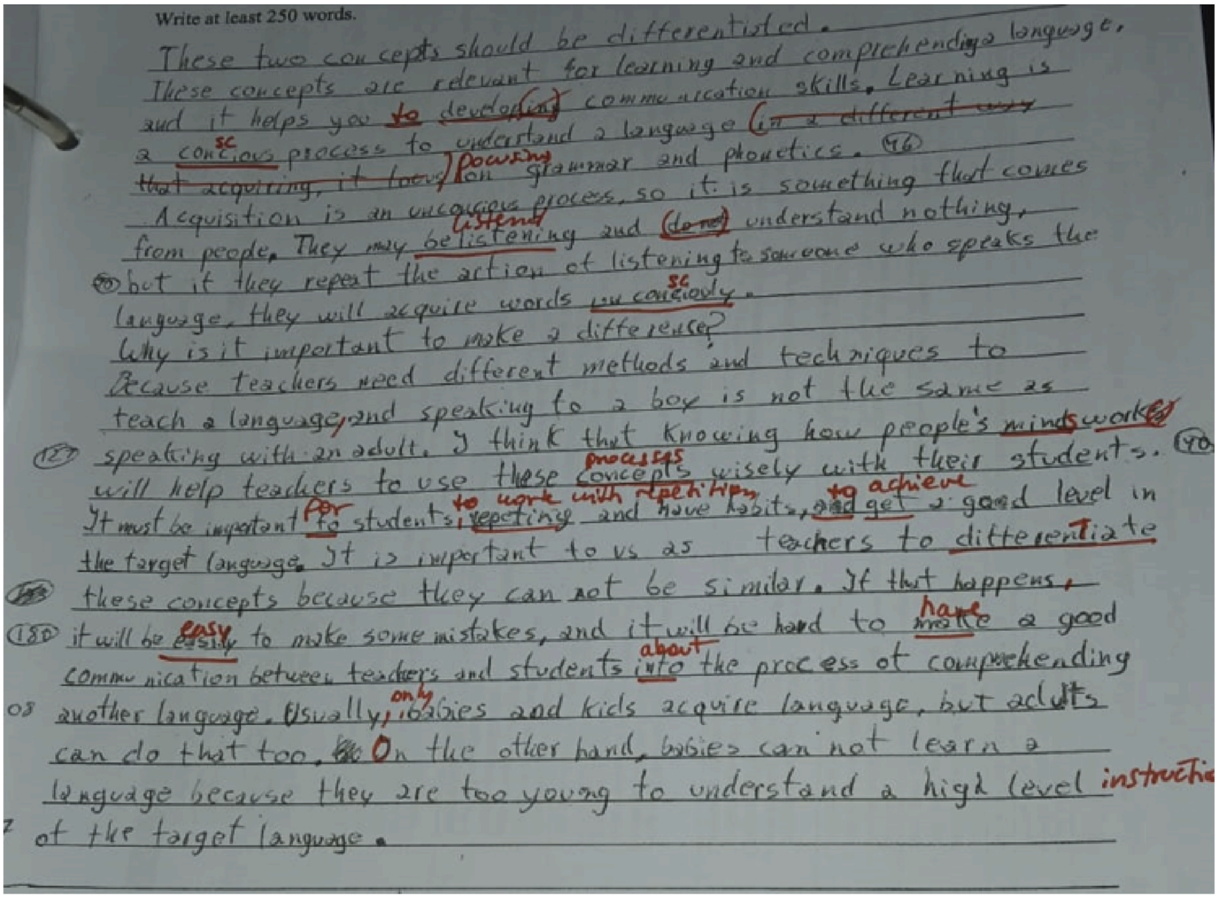
Belén C. Muñoz Muñoz: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Hossein Nassaji:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Felipe I. Bello Carrillo:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Acknowledgements

Research Groups in Neurosciences and Artificial Intelligence at the service of Psychosocial Wellbeing and Learning (NIABA in Spanish), Universidad del Bío Bío, Code GI2310037.

Appendix A. Feedback provision

Teachers' feedback_ Direct WCF



Write at least 250 words.

These two concepts should be differentisted.
These concepts are relevant for learning and comprehending a language,
and it helps you to develo(ping) communication skills. Learning is
a con sc cious process to understand a language (~~in a different way~~
~~that acquiring, it focus~~) focusing on grammar and phonetics. (46)
Acquisition is an unconscious process, so it is something that comes
from people. They may listend be listening and (done) understand nothing,
(80) but if they repeat the action of listening to someone who speaks the
language, they will acquire words unconsc iously.
Why is it important to make a difference?
Because teachers need different methods and techniques to
teach a language, and speaking to a boy is not the same as
(127) speaking with an adult. I think that knowing how people's minds worked
will help teachers to use these processes concepts wisely with their students. (140)
It must be important for to students, to work with repetition repeting and have habits, and to achieve get a good level in
the target language. It is important to us as teachers to differentiate
(15) these concepts because they can not be similar. If that happens,
(180) it will be easy easily to make some mistakes, and it will be hard to have make a good
communication between teachers and students about into the process of comprehending
08 another language. Usually, only babies and kids acquire language, but adults
can do that too. On the other hand, babies can not learn a
language because they are too young to understand a high level instructio
z of the target language.

Teacher's comments

1. "The opinion that guides the essay is not clearly stated. The ideas of paragraphs are not properly linked. Mind word choice, spelling and sentence order."
2. "The body paragraphs do not directly support the opinion. Include linking expressions to start your paragraphs. There are word choice and punctuation mistakes."

ChatGPT's feedback_Direct correction

Error	Correction
"exist differents method"	**"exist different methods!**
"feel interesting about"	**"feel interesting in"**
"how speak"	**"how to speak"**
"talk him in english"	**"talk to him in English"**

ChatGPT's comments

1. "The essay effectively communicates a personal and engaging perspective on English teaching methods, emphasizing the value of making learning fun and interactive. To enhance the effectiveness of future essays, focusing on grammatical accuracy, and refining transitions."
2. "The essay provides a well-argued perspective on the differences between acquisition and learning. To improve further, the writer should focus on refining grammatical accuracy and correct spelling. These improvements would make the argument clearer."

Appendix B. Instructions for feedback provision

> You are a professional language teacher specializing in English, with expertise in providing feedback on the writing of English language learners. You will work with adapted prompts from Task 2 of the IELTS exam. Your students, who are in their second year of an English language teacher training program in a country where English is learned as a foreign language, have a proficiency level between A2 and B1. Using the IELTS Writing Task 2 rubric, you will assess each student's work, assigning them to one of the bands for each criterion. You will also provide general comments using simple language. Additionally, you must create a table to detail linguistic errors, including spelling, capitalization, punctuation, singular and plural nouns, verb tense, subject-verb agreement, word form, awkward phrasing, prepositions, articles, and sentence fragments and run-ons. This table should have two columns: one for the error and one for the correction. This is the writing instruction and prompt.

References

Aljasir, N. (2021). Matches or mismatches? Exploring shifts in individuals' beliefs about written corrective feedback as students and teachers-to-be. *Journal of Teaching and Teacher Education, 9*(1), 1–10. https://doi.org/10.12785/jtte/090101

Baidoo-Anu, B., & Owusu, M. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN*. https://doi.org/10.2139/ssrn.4337484

Baktash, J., & Dawodi, M. (2023). GPT-4: A review on advancements and opportunities in natural language processing. *arXiv*. https://arxiv.org/abs/2305.03195.

Baskara, F., & Mukarto, F. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics, 7*(2), 343–358.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*, 102–118. https://doi.org/10.1016/j.jslw.2007.11.004

Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research, 12*(3), 409–431. https://doi.org/10.1177/1362168808089924

Bitchener, J., & Knoch, U. (2009). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics, 31*(2), 193–214. https://doi.org/10.1093/applin/amp016

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv*. https://doi.org/10.48550/arXiv.2302.03494

Cao, S., & Zhong, L. (2023). Exploring the effectiveness of ChatGPT-based feedback compared with teacher feedback and self-feedback: Evidence from Chinese to English translation. *arXiv*. https://doi.org/10.48550/arXiv.2309.01645

Carter, T., & Thirakunkovit, S. (2019). A comparison of L1 and ESL written feedback preferences: Pedagogical applications and theoretical implications. *Journal of Response to Writing, 5*(2), 139–174. https://scholarsarchive.byu.edu/journalrw/vol5/iss2/7.

Cen, Y., & Zheng, Y. (2024). The motivational aspect of feedback: A meta-analysis on the effect of different feedback practices on L2 learners' writing motivation. *Assessing Writing, 59*, Article 100802. https://doi.org/10.1016/j.asw.2023.100802

Colpitts, B., & Howard, L. (2018). A comparison of focused and unfocused corrective feedback in Japanese EFL writing classes. *Lingua Posnaniensis, 60*(1), 7–16. https://doi.org/10.2478/linpo-2018-0001

Crosthwaite, P., Ningrum, S., & Lee, I. (2022). Research trends in L2 written corrective feedback: A bibliometric analysis of three decades of scopus-indexed research on L2 WCF. *Journal of Second Language Writing, 58*, Article 100934. https://doi.org/10.1016/j.jslw.2022.100934

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *EdArXiv*. https://doi.org/10.35542/osf.io/hcgzj

Deng, J., & Lin, Y. (2023). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems, 2*(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal, 62*(2), 97–107. https://doi.org/10.1093/elt/ccn023

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*(3), 353–371. https://doi.org/10.1016/j.system.2008.02.001

Escalante, J., Pack, A., & Barret, A. (2023). AI-Generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(57). https://doi.org/10.1186/s41239-023-00425-2

Farrokhnia, M., Banihashem, S., Norooz, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education & Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Ferris, D. (2003). Responding to writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 119–140). Cambridge University Press. https://doi.org/10.1017/CBO9781139524810.

Ferris, D. (2012). Written corrective feedback in second language acquisition and writing studies. *Language Teaching, 45*(4), 446–459. https://doi.org/10.1017/S0261444812000250

Gul, R., Tharani, A., Lakhani, A., Rizvi, N., & Ali, S. (2016). Teachers' perceptions and practices of written feedback in higher education. *World Journal of Education, 6*(3), 10–20. https://doi.org/10.5430/wje.v6n3p10

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. *arXiv*. https://doi.org/10.48550/arXiv.2302.02337

Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System, 126*, Article 103502. https://doi.org/10.1016/j.system.2024.103502

Han, T., & Sari, E. (2024). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning, 37*(4), 961–985. https://doi.org/10.1080/09588221.2022.2067179

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hong, W. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation, 5*(1), 37–45. https://doi.org/10.61414/jeti.v5i1.103

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Jacobsen, L., & Weber, K. (2025). The promises and pitfalls of LLMs as feedback providers: A study of prompt engineering and the quality of AI-driven feedback. *AI, 6*(2), 1–17. https://doi.org/10.3390/ai6020035

Kaivanpanah, S., Alavi, M., & Meschi, R. (2020). L2 writers' processing of teacher vs. computer-generated feedback. *Journal of English Language Teaching and Learning, 12*(26), 175–215. https://doi.org/10.22034/elt.2020.11472

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kohnke, L., Moorhouse, B., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback. *Assessing Writing, 44*, Article 100450. https://doi.org/10.1016/j.asw.2020.100450

Lee, I. (2008). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*(3), 144–164. https://doi.org/10.1016/j.jslw.2007.12.001

Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching, 52*(4), 524–536. https://doi.org/10.1017/S0261444819000247

Li, S., Hiver, P., & Papi, M. (2022). Individual differences in second language acquisition: Theory, research, and practice. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of SLA and individual differences* (pp. 3–34). Routledge.

Li, B., Lowell, V., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. *Computers and Education: Artificial Intelligence, 7*, Article 100266. https://doi.org/10.1016/j.caeai.2024.100266

Li, S., & Vuono, A. (2019). Twenty-five years of research on oral and written corrective feedback in *System*. *System, 84*, 93–109. https://doi.org/10.1016/j.system.2019.05.006

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System, 127*, Article 103529. https://doi.org/10.1016/j.system.2024.103529

Lundberg, S. (2023). *ChatGPT vs. teacher feedback provision. An investigation of the efficacy of ChatGPT feedback provision on written production across proficiency levels* [Master' thesis, Stockholm University] DiVA https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-228001</div.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Montgomery, J., & Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing, 16*(2), 82–99. https://doi.org/10.1016/j.jslw.2007.04.002

Muñoz, B., Ortiz, M., & Sáez, K. (2023). Preferencias y opiniones de estudiantes de un Programa de Pedagogía en Inglés con distinto nivel de competencia lingüística acerca del tratamiento de los errores en la escritura en LE: Estudio de caso en una universidad chilena. *Literatura y Lingüística, 47*, 279–306. https://doi.org/10.29344/0717621X.47.2736

Muñoz, B., & Sáez, K. (2019). Indirect written corrective feedback in the treatment of subject-verb agreement in third person singular among students of english as a fl, El feedback correctivo escrito indirecto en el tratamiento de la concordancia sujeto-verbo en tercera persona singular entre estudiantes de inglés como LE. *Alpha, 49*, 275–290. http://doi.org/10.32735/s0718-2201201900049755.

Nassaji, H. (2016). Anniversary article: Interactional feedback in second language teaching and learning: A synthesis and analysis of current research. *Language Teaching Research, 20*(4), 535–562. https://doi.org/10.1177/13621688166449.

Nassaji, H. (2020). Assessing the effectiveness of interactional feedback for L2 acquisition: Issues and challenges. *Language Teaching, 53*(1), 3–28. https://doi.org/10.1017/S0261444819000375

Nassaji, H. (2025, June). *Keynote speech. AI and L2 Acquisition: A Tool to Enhance Writing or Facilitate Learning? [Conference presentation]. The 7th National Forum on Instructed Second Language Acquisition*. Dalian, China: Dalian University of Technology.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research, 34*(2), 257–271. https://doi.org/10.1177/0267658316684904

OpenAI. (2023). GPT-4 system card. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Reynolds, B., & Kao, C. (2022). A research synthesis of unfocused feedback studies in the L2 writing classroom: Implications for future research. *Journal of Language and Education, 8*(4), 5–13. https://doi.org/10.17323/jle.2022.16516

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). https://doi.org/10.37074/jalt.2023.6.1.9

Sallam, M. (2023). The Utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv*. https://doi.org/10.1101/2023.02.19.23286155

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *Tesol Quarterly, 41*(2), 255–283. https://doi.org/10.1002/j.1545-7249.2007.tb00059.x

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

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*(4), 556–569. https://doi.org/10.1016/j.system.2009.09.002

Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and ME on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing, 22*(3), 286–306. https://doi.org/10.1016/j.jslw.2013.03.011

Sistani, H., & Tabatabaei, O. (2023). Effects of teacher vs Grammarly feedback on Iranian EFL learners' writing skill. *International Journal of Foreign Language Teaching and Research, 11*(45), 75–87. https://doi.org/10.30495/jfl.2023.703271

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*. https://doi.org/10.3389/fpsyg.2023.1260843

Songsiengchai, N., Sereerat, B., & Watananimitgul, W. (2023). Leveraging artificial intelligence (AI): Chat GPT for effective English language learning among Thai students. *English Language Teaching, 16*(11), 68–79. https://doi.org/10.5539/elt.v16n11p68

Stefanou, C., & Révész, A. (2015). Direct written corrective feedback, learner differences, and the acquisition of second language article use for generic and specific plural reference. *The Modern Language Journal, 99*(2), 263–282. https://doi.org/10.1111/modl.12212

Storch, N. (2010). Critical feedback on written corrective feedback research. *International Journal of English Studies, 10*(2), 29–46. https://doi.org/10.6018/ijes/2010/2/119181

Strobelt, H., Webson, A., Sanh, V., Hoover, B., Beyer, J., Pfster, H., & Rush, A. (2023). Interactive and visual prompt engineering for ad-hoc task adaption with large language models. *IEEE Transactions on Visualization and Computer Graphics, 29*(1), 1146–1156. https://doi.org/10.1109/TVCG.2022.3209479

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, Article 100752. https://doi.org/10.1016/j.asw.2023.100752

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity, 51*, Article 101440. https://doi.org/10.1016/j.tsc.2023.101440

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages, 8*(3). https://doi.org/10.3390/languages8030212

Nassaji, H. (2024, October). *Corrective Feedback in the Age of AI: Discoveries, Methodological Challenges, and Future Directions*. Chengdu, China: The 8th High-End Forum on Second Language Acquisition Research [Keynote speech].

Nassaji, H. (2024, November). *The AI Paradox: Enhancing L2 Writing or Facilitating L2 Learning?* Qingdao. China: Ocean University of China [Keynote speech]. The 2024 International Seminar on Interdisciplinary Research in SLA.

Yan, X. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies, 28*, 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSNR*. https://doi.org/10.2139/ssrn.4312418

Zhang, B. (2023). *Preparing educators and students for ChatGPT and AI technology in higher education: Benefits, limitations, strategies, and implications of ChatGPT & AI technologies*. ResearchGate. https://doi.org/10.13140/RG.2.2.32105.98404

Zhu, C., Sun, M., Luo, J., Li, T., & Wang, M. (2023). How to harness the potential of ChatGPT in education? *Knowledge Management & E-learning, 15*(2), 133–152. https://doi.org/10.34105/j.kmel.2023.15.008

Zou, D., Xie, H., & Wang, F. (2023). Effects of technology enhanced peer, teacher and self-feedback on students' collaborative writing, critical thinking tendency and engagement in learning. *Journal of Computing in Higher Education, 35*(1), 166–185. https://doi.org/10.1007/s12528-022-09337-y