The Effects of AI-generated Reformulation Text as a Form of Feedback on EFL Writing

Xinyue Guo

Note that some graphs/tables/images may be removed in order to comply with copyright restrictions.

MSc in Applied Linguistics and Second Language Acquisition, 2024

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	Xinyue Guo
Date:	26/8/2024

Acknowledgement

First and foremost, I would like to express my heartfelt gratitude to the incredible people in my department and at my college. The past year has been filled with moments of discovery, learning, and growth, all of which have been deeply enriched by the vibrant environment they cultivated. I am deeply grateful to my supervisor, Dr. Elizabeth Wonnacott, whose invaluable guidance and unwavering support have been pivotal. From the earliest stages, when my ideas were still just nascent thoughts, Liz encouraged me to explore uncharted territories. Throughout this journey, there were inevitable challenges and moments of doubt, but her consistent encouragement and insightful advice turned what could have been a daunting process into a deeply rewarding experience.

A special thanks to my teachers and friends from my alma mater. Their active engagement and willingness to help at every turn made all the difference. Without their kindness and steadfast support, this study wouldn't have been possible.

I would like to acknowledge the support of my family, for their love, patience, and understanding during the many hours I dedicated to this research. Their belief in me kept me motivated throughout this process and brought me here. As I move forward, I hope to continue facing obstacles with courage, to remain brave in the face of uncertainty, and to keep pushing forward, never giving up. Here's to a future filled with wonderful, beautiful, and worthy adventures.

Contents

Abstract

Exploring the ethical and effective use of generative AI tools in education, particularly in enhancing writing skills of second language (L2), is critical. Specifically, AI-generated reformulation texts can serve as an effective technique for written corrective feedback (WCF), helping to make students' essays sound more native-like while preserving their original ideas. Despite its promise, the reformulation technique has been underutilized and under-researched in L2 writing due to its demanding nature. This study aims to investigate the impact of AI-generated reformulation texts on improving L2 writing through a three-stage process involving initial writing, revision with AI feedback, and a final rewrite. Sixty university students of English as Foreign Language (EFL) were divided into experimental and control groups, with the former receiving AI feedback and the latter self-correcting their work. Statistical analyses revealed that AI-generated feedback did not significantly improve EFL students' argumentative writing scores compared to self-correction in the control group. However, the degree to which students accepted the AI-generated reformulations was a strong predictor of writing score improvements, particularly in organization, suggesting that adherence to AI feedback enhanced initial revisions. Additionally, participants in the experimental group noticed more language features than those in the control group, indicating that AI feedback improved learners' language awareness. The study also contributes to the literature by examining students' critical use of ChatGPT, focusing on their rejection of AI-generated changes. These rejections were influenced by internal, external, and AI-specific factors, highlighting learners' critical thinking when interacting with AI tools like ChatGPT.

Keywords: EFL, academic writing, reformulation, ChatGPT

List of acronyms and abbreviations

AI	Artificial Intelligence
AWCF	automated written corrective feedback
AWE	automated writing evaluation
AES	automated essay scoring
CCE	Cognitive Conflict Episode
DOP	depth of processing
EFL	English as Foreign Language
ESL	English as a second language
GenAI	Generative AI
L1	First Language
L2	Second Language
LRE	Language-Related Episodes
WCF	written corrective feedback

List of figures

List of tables

1. Introduction

1.1 GenAI and ChatGPT

Artificial Intelligence (AI) has become increasingly prevalent in our lives, bringing revolutions to almost every industry. By the early 1950s, many ideas, concepts, and programs related to AI had already been introduced (DuBose & Marshall, 2023). The term AI was first applied to a computer program named "Logic Theorist" that could prove mathematical theorems by selecting from a set of rules and applying them in a logical sequence like a human mathematician (Radanliev, 2024). Over the subsequent decades, AI's impact growing as it advanced to successfully perform more complex tasks from playing chess to computer programming, from language translation to detecting emotions (Fan et al., 2020; Nazari et al., 2021). On the other hand, certain aspects of human intelligence remain difficult frontiers for AI, such as artistic creativity and ethical judgment (Chomsky & Watumull, 2023; Radanliev, 2024). However, with the velocity of its development, AI has constantly pushed beyond the boundaries of "distinctly human faculties" (Chomaky & Watumull, 2023, p. 3)

One controversial area of AI, and the area which is also relevant to this dissertation, is its capacity in language use. The rise of statistical learning techniques and neural networks trained on large datasets has significantly advanced AI capabilities in perceiving and producing languages. Generative AI (GenAI), an AI technology that leverages Large Language Models (LLMs) to develop nuanced understandings of human language is functioned to automatically generates content in reaction to prompts in natural-language conversational interfaces (UNESCO, 2024). ChatGPT is one of the most accessible GenAI programs with a broad audience (UNESCO, 2024). Built on LLM GPT-3 and subsequently GPT-4, ChatGPT continues to impress users with its natural, interactive, and context-sensitive responses (DuBose & Marshall, 2023; Radanliev, 2024). Furthermore, ChatGPT has demonstrated its capacity to mimic human patterns of language use across a wide range of linguistic tasks (e.g., Cai et al., 2023). However, critics argue that despite its ability to mimic language patterns effectively, ChatGPT may

not achieve meaningful conversations, as it lacks moral thinking and can generate morally objectionable content (Chomsky & Watumull, 2023). Additionally, since the ChatGPT derives responses from internet-based sources, some of these responses, though seemingly credible, may be incorrect (DuBose & Marshall, 2023; Shankland, 2023).

Despite the ongoing debate, ChatGPT has already been integrated into various industries as a powerful language generation engine. This dissertation will explore its application in the higher education sector, focusing on its use in EFL learning for adult learners. The rationale for connecting ChatGPT with EFL learning in this dissertation is that, since over 90% of ChatGPT's training data is in English, its English content is of relatively high quality compared to other languages (UNESCO, 2024). Therefore, ChatGPT may be most effective as a resource for EFL learning rather than for other L2 languages.

1.2 EFL Academic Writing in Higher Education

EFL academic writing is important in higher education. In contemporary educational systems, student writing is utilized as a primary method in both formative and summative assessments to promote and evaluate critical thinking, the ability to construct arguments, the synthesis of information, knowledge acquisition, competency, and language proficiency (Behizadeh & Engelhard, 2011). Beyond these academic purposes, writing also fosters self-awareness, community participation, and personal enjoyment (Florio & Clark, 1982). This dissertation defines academic writing broadly, encompassing various text types for academic purposes. Academic writing is a problematic, affective, and complex process that is crucial for scientific careers (Rahimi & Zhang, 2018). It is challenging for both native and international students, with additional difficulties for EFL learners due to linguistic and educational barriers (Hanauer et al., 2019). As a productive skill, writing requires more cognitive effort than receptive language skills such as reading and listening. For EFL writing, students must not only generate sufficient content but also consider organization, grammar, vocabulary, and mechanics, which may cause huge burdens in their composition process. This

dissertation focuses on a particular text type: argumentative writing. As a critical aspect of academic literacy, argumentative writing requires various abilities including reasoning and critical thinking (Aull & Ross, 2020; Su et al., 2023). This type of writing is typically taught before scholarly writing such as paper writing, another typical type of academic writing, as it shares critical elements such as argumentation, synthesis of information, and language proficiency, laying the foundation for academic development (Su et al., 2023).

It is important to recognize that writing is not merely a final product but an ongoing process. Many researchers and educators have indicated that developing writing ability is a complex aspect of language learning that involves repeated and recursive steps (e.g., Gebhard, 2002; Sulistyo & Heriyawati, 2017). As Gebhard (2002) suggested, students need to engage in a process of creating and recreating their writing until they discover and clarify their intended message. Consequently, various writing approaches should be integrated into EFL academic writing activities to maximize the quality of students' compositions. Among these approaches, feedback plays a crucial role in improving students' writing throughout the process. The next section will discuss the role of feedback in enhancing EFL academic writing and introduce the potential of ChatGPT in providing effective feedback to EFL learners.

1.3 EFL Writing Feedback

Feedback is broadly regarded as essential for both fostering and reinforcing learning, and its importance is similarly acknowledged in the area of L2 writing (Hyland K. & Hyland F., 2006; Vygotsky, 1978). Over the past decades, changes in writing pedagogy and research have transformed feedback practices, with a shift from summative feedback—such as scoring and rubrics evaluating writing as a product—to formative feedback that focuses on future writing and the development of writing processes (Hyland & Hyland, 2006). In recent years, with GenAI technologies like ChatGPT revolutionizing EFL writing pedagogy and research, innovative methods for various types of feedback have flourished (e.g., Escalante et al., 2023; Shi & Aryadoust, 2024;

Storey, 2023). Although feedback plays a central role in EFL writing, teachers frequently feel that they are not fully harnessing its potential (Hyland & Hyland, 2006). Additionally, in the research literature, a universally agreed taxonomy of feedback strategies in EFL writing is still absent.

Among the various feedback strategies, reformulation stands out yet receives less attention than other modes of feedback such as form-focused written corrective feedback (WCF). Cohen (1983) defined reformulation as a technique where a native speaker rewrites a learner's essay, maintaining the original ideas while enhancing its native-like quality (p. 4). Learners are then expected to compare their original text with the reformulated version in terms of vocabulary, syntax, cohesion, and rhetorical functions (Cohen, 1983, p. 5).Research suggested that reformulated writing can be an effective tool for L2 pedagogy in promoting the noticing and improving students' essays, particularly in cohesion (Lázaro-Ibarrola, 2009; Hylland, 2007; Sulistyo & Heriyawati, 2017). Student responses were largely positive, as reformulations offered them a comprehensive example of a more native-like way to express their ideas (Yang & Zhang, 2010). Yet, it was also found that with errors beyond sentence level remained unnoticed with reformulation as feedback (Sulistyo & Heriyawati, 2017).

One issue of using reformulation in EFL writing classrooms and EFL writing research is the difficulty of finding an educated native speaker or qualified teacher to reformulate the students' essays. Even when available, reformulation is time-consuming and labour-intensive (Lim & Phua, 2019). New advanced writing tools, powered by GenAI and available on mobile devices, might be a promising solution.

However, it is important to proceed the application of ChatGPT in EFL academic writing with caution. First, the efficacy of ChatGPT as an AWE tool is questionable due to its training on general internet text rather than specific domains (Escalante et al., 2023). Additionally, ChatGPT has known limitations, including a tendency to produce untruthful or malicious content (OpenAI, 2023). Third, there are concerns regarding the appropriate and legal use of ChatGPT, particularly in relation to academic ethics,

plagiarism, and cheating on assignments (Gao et al., 2023). These limitations of ChatGPT suggested that a need to teach learners to approach GenAI-produced output critically.

Despite concerns about the potential misuse of ChatGPT's functions, the existence of this language generation model is both irreversible and inevitable (Tsai et el., 2024). Rather than resisting or banning its use in education, it may be more beneficial to integrate ChatGPT into the classroom as a tool for EFL learners to enhance their writing skills. This dissertation aims to contribute to the exploration of how ChatGPT can be effectively integrated into EFL writing by enhancing traditional reformulation feedback, while also exploring how students critically evaluate AI-generated content. Specifically, this study will examine the effects of ChatGPT on improving Chinese university EFL learners' argumentative writing by providing AI-generated reformulation feedback.

1.4 Structure of the dissertation

This dissertation comprises 6 chapters. Chapter 2 reviews pertinent literature that investigates the current application of ChatGPT in EFL academic writing and reformulation as a mode of feedback. Chapter 3 outlines the methodology of this study. Chapter 4 presents the results of the data collected. Chapter 5 discusses the findings from the data and provides implications. Chapter 6 concludes the study.

2. Literature Review

This section will first review the literature on the effect of reformulation feedback and factors influencing the effects of reformulation. Following this, this section will consider previous research on using ChatGPT as the reformulator for improving EFL writing. In the end, the research gap and research questions will be addressed.

2.1 The Effect of Reformulation Feedback

2.1.1 Theoretical background

As introduced earlier, reformulation, proposed by Levenston (1978) and developed further by Cohen (1989), is feedback that involves having a native (or native-like) speaker rewrite the learner's essay while preserving the original ideas. Reformulations can be viewed as an extended form of written recast, offering implicit feedback and demonstrating how ideas or content can be expressed in a more native-like way (Adams, 2003; Yang & Zhang, 2010). For the implementation, reformulation feedback is usually involved in multi-stage writing activities, encompassing three key stages: (1) the composing stage, where learners respond to writing prompts; (2) the noticing or comparison stage (also referred to as the revision stage), in which learners compare their original text with a reformulated version, identifying differences between the two; and (3) the rewriting stage, where learners rewrite or revise their original writing text based on what they have learnt during the noticing stage.

Research on the reformulation feedback mode agrees that its effectiveness is theoretically based on two fundamental SLA hypotheses: output hypothesis and noticing hypothesis (e.g., Chen & Wu, 2022; Lázaro-Ibarrola, 2009; Qi & Lapkin, 2001). The Output Hypothesis suggests that producing output is a key step towards language learning, transiting from language understanding to language use, as it requires learners to engage more deeply and exert greater mental effort than merely receiving input (Swain, 1995). According to Swain, output is not just a result of language production, but most importantly, a trigger that makes learners aware of their problems and pushes them to sort them out from the language production process. Correspondingly, composing as a

form of written output, plays this triggering role for producers (i.e. learners) to notice their difficulties and limitations of their knowledge. Reformulation feedback provides solutions to help them find better ways to express their ideas, solve these problems and correct errors by reformulating their output in an authentic way. Furthermore, output facilitates leaners' L2 knowledge as it is a trigger for noticing, which relates to the following hypothesis to be introduced. Schmidt's Noticing Hypothesis (Schmidt, 1990) states that L2 learners acquire the language only when they consciously notice the target-like form within the comprehensible input (Krashen, 1982). Reformulation text is one kind of such input. By comparing their own compositions with the reformulation texts, L2 learners' are provided with opportunities to notice not only features in the reformulation feedback, but also the linguistic forms they have produced, that is, to "notice their own output" (Lázaro-Ibarrola, 2009, p. 193) in the comparison stage.

2.1.2 Cognitive perspective

In addition to the aforementioned theoretical foundations, empirical research has also provided evidence to support the effectiveness of reformulation feedback (e.g., Chen & Wu, 2022; Coyle et al., 2020; Kim & Bowles, 2019; Yang & Zhang, 2010; Milla & García Mayo, 2024). As noted previously, reformulation provides changes needed to be noticed for improving L2 learners' written output. To unveil the process of how students consider these changes, many studies delve into the cognitive process of students' learning experiences with reformulation. For example, Tocalli-Beller and Swain (2005) focused on the role of cognitive conflict when L2 learners compared their original writing against reformulated one. The study involved 12 adolescent French immersion students who participated in a multi-stage task. The process began with students writing a text based on a visual or auditory stimulus, which was then reformulated by a native speaker. Students were asked to compare their original text to the reformulated version, followed by the stimulated recall where they watched recordings of their noticing stage and discussed the changes they made. The final stage involved rewriting their original text individually. Language-related episodes, including Cognitive Conflict Episodes (CCEs), where students explicitly questioned, disagreed with, or expressed uncertainty

about the changes made in the reformulated text were coded. By calculating the percentage of matching correct post-test changes generated by CCEs, they found that over 60% of these corrective expressions matches reformulation texts, indicating that students' could learn from reformulation. The researchers also provided a qualitative analysis of specific CCEs, highlighting how the cognitive conflict during these episodes led to changes in students' understanding and application of language rules, which was reflected in their post-test writing. They concluded that the conflict and disagreement in the comparison stage help learners to re-examine their language use and clarify their thoughts, thereby supporting their individual rewriting. With a small sample, we may need to proceed with caution on the generalization of the results. But their study provides evidence that engaging reformulation feedback can be helpful from the cognitive perspective.

Similar findings were found in the study by Kim and Bowles (2019). Their findings also supported that reformulations provide more opportunities for deeper cognitive processing than direct corrections, where errors were explicitly marked and corrected. The study involved 22 high-intermediate ESL learners, primarily L1 Mandarin speakers, enrolled in an academic writing course at a Midwestern U.S. university. The participants completed two argumentative writing tasks on topics relevant to their academic studies, each requiring a minimum of 500 words. One week later, they returned to review the feedback, either reformulation or direct correction, while performing a think-aloud task, verbalizing their thoughts as they compared their original text with the feedback. The tasks assessed the depth of cognitive processing prompted by each type of feedback. The depth of processing (DOP) was categorized to High DOP, which involved activities like hypothesizing about language rules, applying learned rules, or spending considerable cognitive effort to understand the feedback, and Low DOP, which involved merely acknowledging the feedback without further elaboration or cognitive engagement. The study found that reformulations prompted significantly more high DOP responses compared to direct corrections. Specifically, 33.8% of the errors

corrected through reformulation were processed at a high DOP, whereas only 22.2% of the errors corrected through direct corrections reached high DOP. However, their study did not measure participants writing performance with a post-test, so there is no evidence whether the reformulation, and the deeper processing, could lead to better writing performance. Besides, while think-aloud protocols provided valuable insights into the cognitive processes, there is a potential issue of reactivity, where the act of verbalizing thoughts might have influenced how participants processed the feedback (see, Sachs & Polio, 2007). Overall, previous research provides evidence that reformulation can benefit learners with deep cognitive processing, though the direct evidence of the effect of reformulation on improving writing remains vague.

2.1.3 Effect on writing performance

A significant portion of the literature on reformulation feedback involves its comparison with other feedback modes. These studies offer a comprehensive, in-depth, and objective perspective on the effects of reformulation feedback.

Many researchers have compared reformulation with other feedback modes, finding mixed results for the effectiveness of reformulation in improving written output. For example, Sachs and Polio (2007) compared the effectiveness of reformulations with written error corrections on accuracy of English as a second language (ESL) learners' writing using a three-stage composition-comparison-revision task. Following a pilot study with a sample size of 15 participants, the second study was conducted with 54 participants who were randomly assigned to one of the four groups: error correction, reformulation, reformulation with think-aloud, or a control group that received no feedback. Over a three-day sequence, participants wrote an essay, received feedback according to their group, and then revised their text without access to the feedback. As for analysis, similar to Kim and Bowles (2019), the researchers used the think-alound protocols to the depth of processing with the reformulation feedback, indicated by the use of metalanguage and providing reasons for changes in the revision. The accuracy was measured by evaluating how many T-units (an independent clause and all its dependent

clauses) were correctly revised in the participants' original texts during the second revision, without having the feedback in front of them. Each T-unit in both the original and revised texts was assessed to determine whether it contained errors and whether those errors were corrected in the second revision. The results indicated that participants in the reformulation condition corrected errors in 70% of T-units, compared to 55.2% in the control group. However, direct error correction was significantly more effective, with participants correcting 87.6% of their T-units. Statistical analysis confirmed the significance of these differences. These findings suggest that while reformulation is not as effective as direct error correction in improving grammatical accuracy, it still leads to better accuracy compared to the control group that did not receive any feedback. Furthermore, the study found that reformulation encourages deep cognitive processing as learners actively engage with and reflect on the feedback through the think-aloud protocols. However, this deeper engagement does not always result in more accurate revisions, as stated earlier. This contrast suggested that while reformulation can foster noticing and critical thinking about language, it may require additional support or explicit instruction to be as effective as direct error correction on improving accuracy of learners' writing.

In studies comparing reformulations with models, another popular feedback mode, findings have also been mixed. A model text is a native speaker's version of a written text, which serves as an example of well-formed, target-like language. Unlike reformulations that closely mirror a student's original text but improve its language to be more native-like, a model text provides a completely independent example of how a native speaker might express ideas based on the same writing prompt. For example, the study by Milla and García Mayo (2024) compared the impact of reformulations and models on the writing performance of primary school children learning English as a foreign language, specifically exploring how these feedback methods influence students' ability to revise their writing when working either individually or collaboratively. The study involved 39 students aged 11–12 from two 6th-year primary education classes in a

semi-public school in northern Spain. They were required to complete a writing task based on a series of pictures, and after their initial drafts, one week later they revised their work using the feedback provided either individually or collaboratively according to their group. Four weeks later after the second draft, participants did the revision again which was intentionally implemented to examine whether their revisions reflected a retention of the feedback provided. The effectiveness of the feedback was evaluated by analyzing the features that the students noticed and incorporated into their revised drafts. They found that models tend to lead to greater noticing and incorporation of vocabulary, while reformulations were more helpful to grammar and spelling features, consistent with other studies (e.g., Coyle et al., 2018; Coyle & Roca de Larios, 2014; Kang, 2020).

Studies comparing reformulation feedback with other feedback modes provide a more comprehensive, nuanced, and objective understanding of the effects of reformulation. These findings suggest that while reformulation feedback has its advantages and disadvantages, and there is inconsistency regarding which specific aspects of writing it most effectively enhances.

2.2 Factors Influencing the Effects of Reformulation

The potential benefit of writing feedback is influenced by a number of factors including the nature of the feedback technique, the timing of the intervention, and the learners' language proficiency (Coyle et al., 2020; Chen & Wu, 2022; Qi & Lapkin, 2001). In the previous section, the nature of the reformulation was discussed. This section will review the literature on the other two factors.

2.2.1 Learners' proficiency

Cohen's (1983) claim that reformulation may benefit ‘‘learners at intermediate levels and above’’ and ‘‘may have its greatest impact among advanced students’’ (p. 5). Previous literature has provided evidence that learners' language proficiency could influence the effectiveness of reformulation on writing improvement. For example, the study by Qi and Lapkin (2001) involved two advanced Mandarin-speaking adults who

completed a single narrative writing task based on a picture-story prompt, conducted in three stages: composing the narrative (Stage 1), comparing their original text with a reformulated version (Stage 2), and revising the original narrative (Stage 3). Language-Related Episodes (LREs) were measured through think-aloud protocols during the composing and reformulation stages, capturing moments where participants focused on language use, including vocabulary, grammar, or syntax. The task was set in the context of narrative writing, and the LREs were analyzed to assess the effectiveness of noticing and language processing. Wu, the higher-proficiency participant, produced more LREs and resolved a greater proportion correctly, indicating that higher language proficiency enhanced the ability to notice and effectively use reformulation feedback. Considering that only two participants were involved, this finding is only anecdotal evidence and based on individual base. Similar evidence suggests that learners with higher proficiency outperform those with lower proficiency in research on reformulation feedback (e.g., Hanaoka, 2006; Lapkin et al., 2002). However, since no statistical analysis was conducted in these studies, caution is needed when generalizing these findings.

Only a few studies provided valuable statistical evidence on the role of language proficiency in language learners' engaging reformulation feedback. For example, Hanaoka (2007) involved 37 Japanese college students divided into two proficiency groups: advanced and intermediate. Participants completed a four-stage narrative writing task, which included initial writing, comparison with native-speaker models, and two subsequent revisions. LREs were measured through note-taking, focusing on what learners noticed during the writing and comparison stages. The study found that higher proficiency learners noticed significantly more language features during the comparison stage than lower proficiency learners, and they incorporated these features more effectively into their revisions. This suggests that language proficiency plays a crucial role in the ability to notice and incorporate feedback during the writing process.

Previous literature has established evidence that for reformulation to have a positive effect, learners need to possess a sufficient level of language proficiency to

effectively engage with the feedback. Consequently, this dissertation will focus on the population of university EFL students majoring in English. This approach is intended to better control and explore the role of AI-generated reformulation when comparing its effects with traditional reformulation within this proficiency range.

2.2.2 Feasibility

One big concern of reformulation is its feasibility in classroom settings due to its time-consuming nature and task difficulty for the teacher. As Lázaro-Ibarrola (2009) claimed that, though there is some evidence on its effect on improving students' writing, reformulation will only have a real impact on students' learning if it is used as a "regular correction strategy" (p. 207). To test the effectiveness and feasibility of reformulation on EFL classroom, Lázaro-Ibarrola (2009) tested the validity of reformulation feedback in classroom settings by comparing it with self-correction through a structured writing-correction-rewriting task involving 16 Spanish EFL students but only 2 students' noticing process was recorded for analysis due to the time constraints. According to the researcher, these two participants were aware of the requirement that they had to explain what they had noticed to the teacher what which might motivate them for the noticing session then the learners than usual. But still, c3eeed4onsistent with previous literature, the results showed that reformulation led to the detection of a higher percentage of errors (79.23%) compared to self-correction (44.80%), highlighting its effectiveness in helping students' noticing. Most intriguingly, the research conducted a survey with the teacher in the classroom which the study was conducted and two selected participants. The teacher expressed satisfaction with the outcomes but highlighted significant challenges, particularly the time-consuming nature of reformulation and uncertainty about the accuracy of her reformulations. These concerns led her to state that she would not use reformulation regularly in her classroom. Students, on the other hand, reported that they felt they had learned more through reformulation than through traditional correction methods, although they felt more lost and unsure during self-correction tasks. The study concluded that while reformulation had a positive impact on error detection and was generally well-received by students, its practicality in the classroom was limited due to

the considerable time and effort required from teachers, suggesting that adaptations would be necessary for regular classroom use. To address this issue, the researcher suggested, as other authors have previously done, that teachers could use this strategy by having the whole class work on just one reformulated text (Allwright et al., 1988). However, in this case, the advantage of providing personalized feedback that adheres to students' original ideas would be undermined.

This dissertation innovatively proposes generative AI, specifically ChatGPT, as a solution to the feasibility challenges of reformulation feedback due to its fast response and the large language model it is based. The next section will review the relevant literature on these issues.

2.3 ChatGPT as a Reformulator

Writing assistance programs powered by AI, including automated written corrective feedback (AWCF), automated writing evaluation (AWE), and automated essay scoring (AES) have gained popularity in language learning and teaching (see, Shi & Aryadoust, 2024). ChatGPT in particular has shown its strengths as an agent for writing feedback by providing personalized, efficient and real-time feedback (Imran & Almusharraf, 2023; Ranalli & Yamashita, 2022). Relating to the current dissertation, ideally, ChatGPT has potentials to do the reformulation tasks of an instructor who typically reads and gives comments and suggestions to students on how to better improve their writing skills with a faster response.

To our knowledge, only the study by Tsai et al (2024) used ChatGPT to generate reformulation feedback. In their study, 44 EFL college English majors were first asked to write a short essay individually using Microsoft Word in class without any external resources or assistance, but the specific details about the topic and the genre of the required writing task were not provided. For the revision stage, the participants, who had little prior experience with ChatGPT, received instruction and training on how to use ChatGPT for reformulations. The instructor guided students by first demonstrating basic operations of ChatGPT, and then providing a sample for reformulation using seven

prompt strategies for adjusting grammar, tone, intent, audience, style, emotion, and domain. Following this, students revised their essays with the assistance of ChatGPT within a 30-minute time limit, making final edits themselves and submitting the revised versions as a post-test. Additionally, participants were required to provide a revision report, indicating the specific prompts they used and the changes made in the revised texts compared to the original texts. This step was intended to reinforce students' learning through noticing. However, it is a pity that these revision reports were not analyzed for noticing. Instead, the noticing process was assessed through a qualitative survey regarding the usefulness of ChatGPT in the writing process. The survey results indicated that 86.5% of learners noticed a more diverse and precise use of vocabulary, while a significant 98.1% observed corrections in grammatical errors. Additionally, 19.2% of learners reported that sentence restructuring improved the fluency of their writing, and 30.7% noted that the revised sentence structures were more concise and easier to understand. Although a triangulation of the analysis of participants' revision reports and a content analysis of the original and revised texts would help provide more objective evidence than relying solely on the survey report, their study provides support to the positive effects of AI-generated reformulation as feedback on noticing. To examine the effects of AI-generated reformulation feedback on enhancing EFL writing performance, 44 original essays and 44 ChatGPT-assisted revised essays provided by participants were evaluated by two independent graders using a randomized crossover design to reduce grading bias. Scores from the original and revised essays were paired for a before-and-after comparison. The results indicated that ChatGPT-assisted revisions led to significantly higher scores among EFL college English majors. Notable improvements were observed across all four dimensions of writing quality assessment, with the most substantial gains in vocabulary, followed by grammar, organization, and content.

Their study used a rigorous experimental design with a mixed methods to provide evidence that ChatGPT-assisted revisions help improve both noticing process and the

writing outcomes. However, it remains unknown the effect of AI-generated reformulation on improving EFL argumentative writing.

2.4 Research Gap

After reviewing the relevant literature, several gaps have been identified.

First, the effectiveness of reformulation in improving EFL academic writing, particularly regarding which specific aspects of writing it influences, remains unclear and mixed. Additionally, much of the existing research on reformulation relies on case studies, which may limit the generalizability of the findings. There is a need for more empirical evidence to rigorously examine the effects of reformulation feedback.

Second, while AI, particularly ChatGPT, is increasingly being used as a promising writing assistant, there is limited research on its effectiveness in generating reformulation feedback. Few studies have investigated the extent to which AI-generated reformulation can enhance noticing and improve writing outcomes.

Third, there is a scarcity of research examining students' critical use of AI-powered tools, In the context of EFL academic writing, no research has explored students' evaluations of AI-generated reformulation output, which is crucial for understanding how students critically engage with and utilize ChatGPT in their learning processes.

To address these gaps, this dissertation aims to examine the effects of ChatGPT to provide reformulation feedback with university English major students in China, focusing on language improvement in different aspects. Learners' noticing and use of AI-generated reformulations in the writing process will also be investigated.

2.5 Research Questions

Specifically, this study will be guided by the following research questions:

1) To what extent does AI-generated reformulation text as feedback improve EFL learners' writing?

2) What factors significantly predict EFL writing score gains?

3) What do EFL learners notice when comparing their essays with AI-generated reformulation texts?

4) How do EFL learners use the AI-generated reformulation texts in the revision?

3. Methodology

This chapter summarizes the current study's methodology, including the research questions and hypothesis, participants, instruments, procedure of the task and methods for data analysis.

3.1 Research Questions

The research questions of the current study and their respective hypotheses are as follows:

RQ1. To what extent does AI-generated reformulation text as feedback improve EFL learners' writing?

The hypothesis for this RQ is that intervention group will show significantly better improvements in overall writing scores and in specific aspects of writing (content, organization, grammar, and vocabulary) over time compared to the control group.

RQ2. What factors significantly predict EFL writing score gains?

a. higher language proficiency will be associated with higher writing score gains.

b. higher AI-acceptance will be associated with higher writing score gains.

RQ3. What do EFL learners notice when comparing their essays with AI-generated reformulation texts?

This RQ has no hypotheses and aims only to identify patterns that emerged from the noticing/revision stage of the writing task.

RQ4. How do EFL learners use the AI-generated reformulation texts in the revision?

This RQ has no hypotheses and aims only to identify patterns of students' acceptance or rejection of the AI-generated reformulation texts in the revision.

3.2 Participants

The EFL learner population comprised university English majors in northern China. They were young adults (18-30 years old), with Mandarin as their L1. The population was chosen because they are more familiar with English academic writing compared to other EFL learners in Chinese universities, and they receive systematic instruction and practice in English academic writing as part of their curriculum. When selecting EFL learner participants for the study, only those who had completed at least one semester of an English academic writing course by the time of data collection were included. Learners who had not received such instruction were excluded. This criterion was set because the effectiveness of reformulation feedback depends on the students' existing knowledge of writing and their language proficiency.

The final sample included 61 English major students from a university in northern China. Participants were recruited from university academic writing courses and through personal contacts, ranging from first-year undergraduates to first-year graduate students. All the participants were randomly divided into (a) an experimental group (n = 30), who wrote an argumentative essay, compared it against the AI (GPT-4) generated reformulation texts, and composed the rewriting, and (b) a control group (n = 31), who engaged in the same argumentative writing task as the experimental group did, performed self-correction without receiving AI-generated reformulation texts and completed rewriting. One participant in the control group did not finish the rewriting, so their data was excluded. As a result, each group had 30 participants. Standard ethical procedures were followed, approved by CUREC (Research Ethics Reference: EDUC_C1A_24_136, see Appendix C), with participants voluntarily taking part and providing informed consent.

Demographic information was collected from all participants (see Table 1), including gender, age, age of onset of learning English, year of study and language

proficiency. The study chose the CET-4 (College English Test-Grade 4) score as the primary indicator of learners' English language proficiency. The CET-4 (College English Test-Grade 4) is a national standardized English proficiency test administered in Chinese universities. It is designed to assess the English language skills of undergraduate students in listening, reading, writing, and translation. The test is typically taken by students after completing their first two years of university study, and it serves as a benchmark for their English proficiency at an intermediate level. Thirteen of the 60 participants did not provide their CET-4 scores because they had not yet taken the test. Among these, 10 were first-year undergraduate students, 1 was a third-year undergraduate, and 1 was a fourth-year undergraduate.

In the experimental group, 9 participants were male and 21 were female. The average age was 20.40 years (SD = 1.714), with participants beginning to learn English at an average age of 7.17 years (SD = 2.001). The average academic year of study is 2.57 (SD = 1.19) with the distribution of academic years in this group as follows: 20% (n=6) were first-year undergraduate students, 33% (n=10) were second-year, 23% (n=7) were third-year, 17% (n=5) were fourth-year, and 7% (n=2) were first-year postgraduate students. The group's average score on the CET-4 (College English Test-Grade 4) was 581.91 (SD = 49.094). These details suggest that the participants in the experimental group were generally at an intermediate level of English proficiency.

The control group consisted of 4 male and 26 female participants. Their average age was 20.27 years (SD = 1.41), with the average starting age for English learning at 7.00 years (SD = 1.93). The average academic year of study is 2.67 (SD = 1.18) with the distribution of academic years in this group as follows: 13% (n=4) were first-year undergraduate students, 40% (n=12) were second-year, 23% (n=7) were third-year, 13% (n=4) were fourth-year, and 1% (n=3) were first-year graduate students. The control group's average CET-4 score was 580.13 (SD = 42.66). Overall, according to this information, the participants in both groups were generally intermediate EFL learners, which makes the two groups similar and thus comparable.

Table 1. Demographic information of participants

	Gender		Age		Age of onset		Academic year		CET-4 score	
	Male (n)	Female (n)	Mean	SD	Mean	SD	Mean	SD	Mean	SD
AI-group	9	21	20.40	1.71	7.17	2.00	2.57	1.19	581.91	49.09
Control-group	4	26	20.27	1.41	7.00	1.93	2.67	1.18	580.13	42.66

Note. Age of onset refers to the age when learners began learning English.

3.3 Instruments

3.3.1. Writing tasks.

The participants engaged in a three-stage writing task: writing, revision, and rewriting. Their initial writing, subsequent revision, and rewriting after one week were treated as a pretest, post-test, and delayed posttest, respectively. During the comparison stage, only the experimental group received AI-generated reformulation texts. Regarding the writing task, participants were instructed to write whether they agreed or disagreed with banning smoking in public places within 200-300 words. The topic was chosen to replicate the previous study, and because it is a familiar topic related to their daily life. The writing task was pilot-tested with another group of three EFL learners who are first-year graduate students to estimate the time to spend for the task and ensure level appropriateness.

3.3.2. AI-generated reformulation text

OpenAI's GPT-4 was employed to generate reformulation texts as feedback on student writing. It was chosen for this study because it provided the most suitable and accurate feedback among the LLMs tested. Additionally, GPT-4 outperformed other LLMs on its response quality at the time of its release (OpenAI, 2023) and to my knowledge that was still the case when this study was conducted.

The prompt used for ChatGPT to generate reformulation texts consisted of multiple components. First, GPT-4 was given the role of a native English speaker. Second, the prompt of the writing task given to the students was included. Third, GPT-4

was asked to produce a reformulated text of approximately 200-300 words based on the students' writing. An example of the prompt and the resulting feedback see Appendix A.

3.4 Procedure

All participants including the treatment group and the control group completed the three-stage task with the same argumentative writing prompt in the same order. The procedures of the experiment are described below.

During the first treatment session, all participants were given 5 minutes for pre-task planning to generate ideas about the topic. Each participant then completed their initial writing task on Microsoft Word. Participants were informed that there was no strict time limit for this task, although they would receive a reminder at the 30-minute mark. Participants were not allowed to use dictionaries or any other resources. This restriction was implemented to assess whether participants could resolve the issues in their writing process solely by using the provided reformulation texts during the revision stage.

After participants completed their initial writing, they proceeded to the revision stage, where each group employed different revision strategies. For the experimental group, the researcher included participants' submission in the prompt provided to GPT-4 to generate reformulation feedback for each student. Participants were then instructed as follows: "In a moment you will see your feedback on the assignment. We want you to work on the assignment to improve it using this reformulation text. You can freely decide which changes to accept and which to reject." During the revision, the reformulation text was displayed on the left side of the screen and the original writing on the right, both as separate MS Word© files. The editing process was tracked using the 'track changes' function of Word. By contrast, participants in the control group were asked to self-correct their initial writing without additional resources. Participants were instructed as follows: "In this stage, we want you to revise your assignment to a level that you are satisfied with and that reflects the best you can achieve using your own knowledge". The editing

process were also tracked using the 'track changes' function of Word. Still, there was no strict time limit for this session, but participants were reminded at the 30-minute mark.

When participants finished their revision, the researcher conducted the stimulated recall with all the participants from both groups. Markups of editing were used to elicit participants' what they noticed during the revision stage and to explain why they made specific changes. The researcher guided participants to provide recall for every sentence and change they made, asking questions such as: Why did you not make changes here? Why did you not make changes here? What's the difference between the original version and the revised version here? The following examples of stimulated recall episodes were provided to the participants in the experimental group: "I couldn't say X but the reformulation puts Y" and "I have expressed this ideas as X, and the reformulation puts it Y". For the control group, the examples were "I wrote X, but I thought it should be Y" or "I wrote X but I am not sure if it was correct". Each instance of X or Y was counted as a noticed feature in the analysis of noticing during the stimulated recall. Upon completing the above stages, participants' original and revised writings were collected as the pre-test and post-test products respectively.

One week later, the same writing prompt was given to each participant as the delayed post-test. Participants did not have access to their previous drafts while rewriting. Figure 1 shows the procedure of the whole writing task.

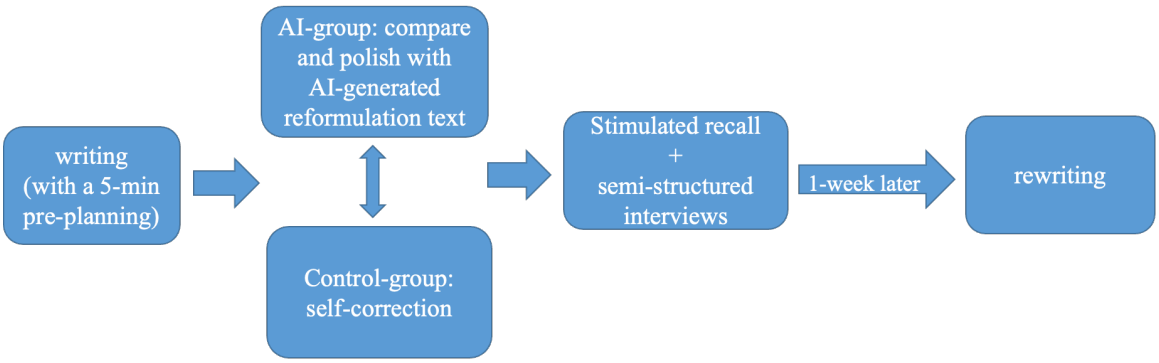

Figure 1. The procedure of the three-stage writing task.

The entire experiment was conducted in an interview room at the participants' university. The duration of task completion in the first week was approximately 90 minutes and 35 minutes in the second week.

3.5 Data Analysis

3.5.1 Scoring rubrics.

To determine whether AI-generated reformulation texts help learners improve their rewriting (Research Q1), participants' initial writing, subsequent rewriting, and delayed rewriting were assessed using an analytical scoring rubric to investigate specific areas of improvement and overall progress. The quality of the participants' writing tasks was evaluated with the same analytic rubric used in Kang's (2024) study, which assesses student writing based on grammar, vocabulary, content, and organization. These four dimensions were specifically chosen due to their frequent assessment in previous studies on the effectiveness of models as a written corrective feedback (WCF) tool (Coyle & de Larios, 2014; Hanaoka, 2007). Grammar and vocabulary were evaluated based on the accuracy of word forms, meanings, and grammatical structures. Content was assessed for the presence of relevant ideas, well-supported topic sentences, and persuasive supporting details. Organization was evaluated in terms of the logical development and cohesive arrangement of ideas. The rubric used a scale of 0 (no score) to 6 (excellent performance) for each dimension, with a total maximum score of 24. The scores from each dimension were examined separately to identify which areas improved significantly and were then combined to produce a total score.

To ensure the reliability of the scores, these essays were coded "blind" i.e. without knowing whether the person was in the experimental or control condition, or which timepoint the essay was from. Additionally, the researcher invited another rater, a qualified teacher for both English and Chinese teaching at a private school in the UK, as the co-rater. A sample of 18 essays was randomly selected, and both raters independently scored each piece of writing to standardize the scoring process. Interrater reliability was

assessed using correlation analysis, and the results showed a high correlation coefficient of 0.68 (p=.002) using Pearson correlation. The researcher then scored the remain papers and the scores were used as the dependent variable in data analyses.

3.5.2 Statistical analysis.

The raw data from the scoring were entered into SPSS for both descriptive and inferential analysis. Initially, descriptive statistics were used to summarize the data and provide an initial understanding of the distributions of scores in both groups through means and standard deviations etc. To address the first research question, a Mixed Repeated Measures ANOVA was conducted, with time as the within-subjects factor and group as the between-subjects factor, to test for differences in overall scores and scores from four specific aspects between the experimental and control groups across different time points (pre-test, post-test, delayed post-test). This analysis specifically aimed to identify any interaction effect of time point and group on writing scores in general and in the four dimensions of the EFL writing.

To answer the second question, multiple regression analysis was used to identify specific predictors of writing gains within the experimental group. To gauge participants' acceptance on AI-generated reformulated text, we calculated the text similarity between the AI-generated text and students' initial writing and their revision in the post-test using Python (programming codes see Appendix B). The increase in similarity from the AI-initial writing to the AI-revised text indicates participants' acceptance on the AI-generated reformulation text in their own writing. The author termed this measure 'AI-acceptance' and used it as an independent variable in our analysis. Hierarchical regression analysis was conducted to determine which factors (i.e., language proficiency and AI-acceptance) predict the participants' writing score gains.

This comprehensive approach allowed for a thorough examination of the effect of the intervention of AI-generated reformulation feedback and an understanding of the factors that contribute to improvements in writing scores.

3.5.3. Selection of participants for stimulated recall analysis.

To address the third research question, a representative sample of participants was selected for an in-depth qualitative analysis of stimulated recall based on performance and diversity factors identified through quantitative analysis. To improve the validity and representativeness of the sample, the selection criteria include language proficiency, task performance, and intervention type. Task performance was divided into four categories: 1, total score gains in both post-test and delayed post-test; 2, gains only in post-test, but no gains in delayed post-test; 3, No gains in post-test but gains in delayed post-test; 4, No gains in both post-test and delayed post-test.

For each category, we selected four participants: two from the control group and two from the AI group. Within each category, one participant had high language proficiency, and one had low language proficiency. Language proficiency was primarily determined by CET-4 scores, supplemented by year of study. An exception was noted for category 3, where no participants with low proficiency were available. Consequently, we have eight high-proficiency participants equally distributed between the two groups across all four performance categories, and six low-proficiency participants equally distributed between the two groups across three of the four performance categories. Below is the information about the selected participants (Table 2).

Table 2. Demographic information of selected participants

Pseudonym	Gender	Language proficiency	CET-4 score	Year of study	Task performance	Intervention Type
William	male	Intermediate-low	523	Third	1. Total score gains in both post-test and delayed post-test	AI
Simon	male	Intermediate-low	429	Third	4.No gains in both post-test and delayed post-test.	AI
Lucy	female	Intermediate-low	507	Second	4.No gains in both post-test and delayed post-test.	Self-correction
Cindy	female	Intermediate-high	556	Fourth	2. Gains only in post-test, but no gains in delayed	AI

					post-test	
Gillian	female	Intermediate-high	623	Fourth	1. Total score gains in both post-test and delayed post-test	AI
Linda	female	Intermediate-high	605	Second	3. No gains in post-test but gains in delayed post-test.	control
Daisy	female	Intermediate-high	621	Second	4.No gains in both post-test and delayed post-test.	AI
Rose	female	Intermediate-low	NA	First	3. No gains in post-test but gains in delayed post-test.	Self-correction
Sophia	female	Intermediate-high	594	Third	4.No gains in both post-test and delayed post-test.	Self-correction
Fiona	female	Intermediate-low	NA	First	1. Total score gains in both post-test and delayed post-test	Self-correction
Felicity	female	Intermediate-low	NA	First	3. No gains in post-test but gains in delayed post-test.	AI
Kate	female	Intermediate-high	600	Fifth	2. Gains only in post-test, but no gains in delayed post-test	Self-correction
Helen	female	Intermediate-high	625	Third	1. Total score gains in both post-test and delayed post-test	Self-correction
Winter	female	Intermediate-high	582	Fourth	3. No gains in post-test but gains in delayed post-test.	AI

3.5.4. Coding for stimulated recall analysis.

In order to investigate the problems and features they noticed at the comparison stage (Research Q3), the transcripts of the stimulated recall of the selected participants were analyzed. The noticed features were coded according to a coding scheme adapted from Hanaoka (2007), categorizing them into lexis, grammar, content, organization, or other issues (see, Table 3).

Table 3. Coding scheme for noticed features.

Category	Description	Examples (originally written in Chinese)
Content	Generation and development of ideas	-I think the example is more relevant to the argument.
Organization	Structure or organization of ideas	-I don't know how to arrange these two opposite ideas.
Grammar	Particular features of grammar	-The reformulation text uses the past tense here, but I used present tense
Vocabulary	Lexical items, collocations, etc.	-I used 'breathing system' while the reformulation text gave the 'respiratory system'.
Others	Features that did not fit into any specific category	-The reformulation is well-written, but I couldn't learn its language style.

To answer the RQ4, the researcher independently performed a thematic analysis of the data on rejection of AI-generated reformulation feedback from the stimulated recall and then consulted on salient themes found in the data with the supervisor.

4. Results

This chapter examines potential confounding variables and presents descriptive and inferential statistics relevant to the research questions.

4.1 Effects of AI-generated Reformulation Feedback on Improving Writing Scores

4.1.1 Potential covariates

To ensure a robust analysis of the AI intervention's effect on writing scores, it is essential to investigate potential confounding variables. Comparing the distribution of these variables across groups helps determine if they should be included as covariates in the analysis. This step is crucial to control for any baseline differences between groups, preventing them from confounding the results and ensuring more accurate and reliable findings.

As presented in the Table 1. the experimental group and the control group were similar and thus comparable across several demographic and proficiency measures. For the gender distribution, there were 9 male participants and 21 female participants in the experimental group, while there were 4 male and 26 female participants in the control group. The average age in the experimental group was 20.40 years (SD = 1.714), and in the control group, it was 20.27 years (SD = 1.413). The average age of onset for learning English was 7.17 years (SD = 2.001) in the experimental group and 7.00 years (SD = 1.930) in the control group. Notably, second- and third-year students made up the largest proportion of participants, accounting for 56% in the experimental group and 63% in the control group. Additionally, the average CET-4 (College English Test-Grade 4) score was 581.91 (SD = 49.094) in the experimental group and 580.13 (SD = 42.663) in the control group.

It can be seen from Table 1 that the groups are similar on each of these measures. Therefore, there is no need to account for these potential covariates in the subsequent analysis.

4.1.2 Comparison of writing scores

Recall that the study assessed four different aspects of writing: content, organization, grammar, and vocabulary. Scores were recorded for each individual component, as well as for the total score. All participants' writing—including the initial writing, revised writing, and rewriting—was evaluated on a scale from 0 (very poor) to 6 (very good) for each of the four areas. The total score ranged from 0 (very poor) to 24 (very good).

Descriptive statistics for both groups for the pre-test (T1), post-test (T2), and delayed post-test (T3) are reported in Table 4. Figure 2 presents the line plots showing the total score mean and the mean scores for each of the four components of two groups across three time points (T1, T2, T3), with the error bar indicating 95% confidence interval.

Table 4. Descriptive statistics of writing scores.

		Total		Content		Organization		Grammar		Vocabulary	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
T1	AI group	16.50	3.59	4.03	1.0	4.17	1.1	4.17	1.2	4.13	0.9
	Control group	15.63	2.68	3.70	0.7	3.87	0.8	4.17	0.9	3.90	0.9
T2	AI group	17.57	2.65	4.13	0.6	4.37	0.9	4.30	0.8	4.77	0.8
	Control group	16.13	2.92	3.77	0.6	4.0	0.8	4.23	0.9	4.17	1.0
T3	AI group	17.53	2.91	4.20	0.7	4.37	0.7	4.40	1.2	4.57	0.8
	Control group	17.43	3.69	4.27	0.9	4.33	1.1	4.27	1.0	4.57	1.0

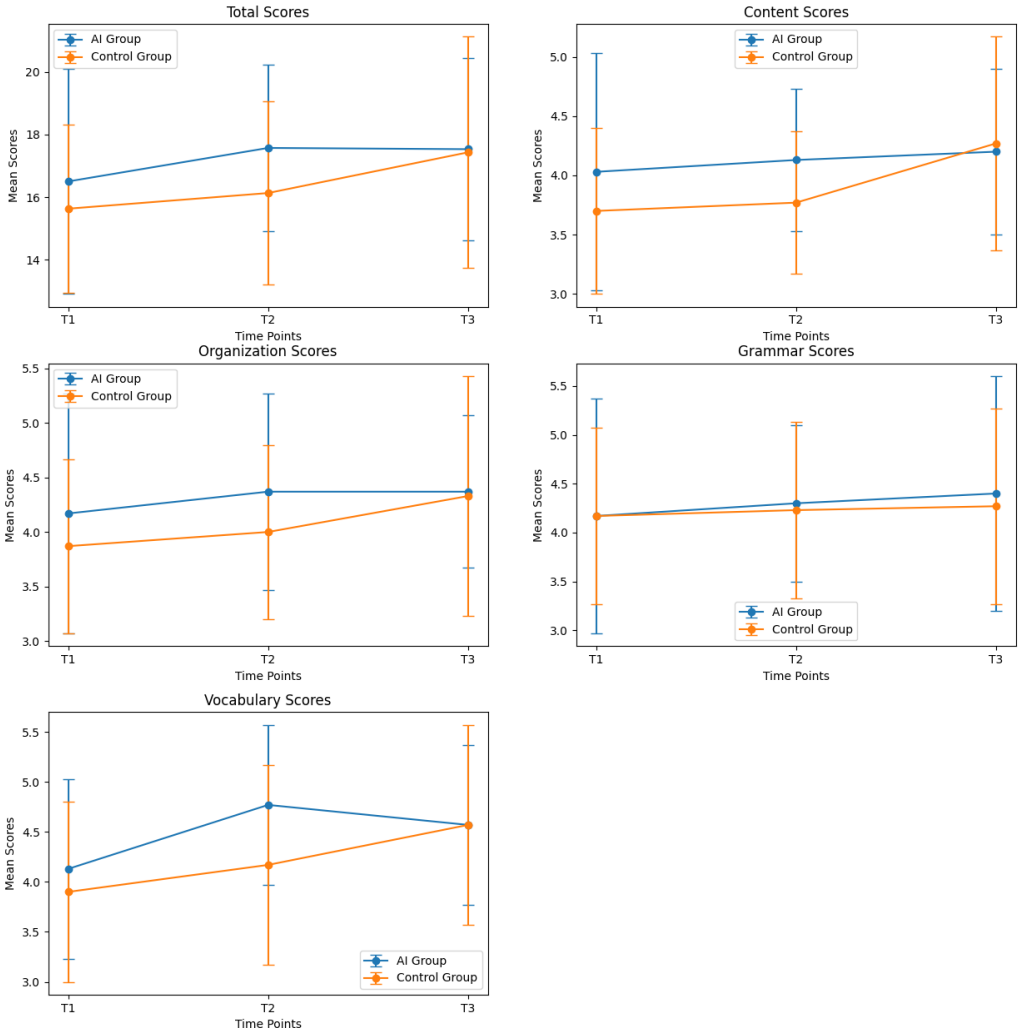

Figure 2. Line plots of mean scores for comparison of the two groups.

The data for the overall writing scores and scores for each of the components were analyzed statistically in five separate Repeated Mixed ANOVA (RM-ANOVA). The dependent variables are participants' writing scores. Time is a within-participant variable with three levels: T1, T2, and T3. The between-group variable is group: experimental group or control group. Before each RM-ANOVA was conducted, the

normality of the measures was checked by values of the skewness and kurtosis and the assumption of normal distribution was fulfilled. After performing each RM-ANOVA, the assumption of homogeneity of variances was checked using Levene's test. This assumption was met for all measures but for organization scores. In the RM-ANOVA for organization scores, the homogeneity of variances assumption was violated, as indicated by significant results from Levene's test at Time 1 ($p = .039$) and Time 3 ($p = .002$). Consequently, the ANOVA with Welch's adjustment was reported for organization scores.

4.1.2.1 Total scores

The first RM-ANOVA analysis was conducted with total writing scores. The results of RM-ANOVA show that there is a significant main effect of time with medium effect size ($F(1.550, 89.882) = 7.498$, $p = .002$, $\eta2=.114$), indicating that participants' total scores of writing differ at the three time points. Breaking this down and looking at pairwise contrasts, there is a significant change in the students' total writing scores both between pre-test and post-test ($F(1, 58) = 8.495$, $p = .005$, $\eta2 =.128$), and pre-test and delayed post-test ($F(1, 58) = 10.155$, $p = .002$, $\eta2 = .149$). The effect size for the change from pre-test to delayed post-test is larger ($\eta2 = .149$) than for the change from pre-test to post-test ($\eta2 = .128$).

Critically, the interaction between group and timepoint was not significant with a small interaction effect ($F(2, 116) = 1.667$, $p = .200$, $\eta2 = .028$). Looking at pairwise contrasts, the interaction between group and the contrast between pre-and the first post test was not significant ($F(1, 58) = 1.111$, $p = .296$, $\eta2=.019$) nor was the interaction between group and the contrast between the pre-test and delayed post-test ($F(1, 58) = .744$, $p = .392$, $\eta2= .013$). Thus although both groups improve their writing from the pre test to each of the post tests, there is no evidence that the AI-reformulation feedback leads to greater improvements than in the control group.

4.1.2.2 Contents

Separate RM-ANOVA analysis were conducted to examine the effect of AI reformulation feedback over time on students' writing in four aspects: content, organization, grammar and vocabulary. Recall that scores on contents are about assessing relevant ideas, well supported topic sentences, and convincing supporting sentences.

The results of the repeated measures ANOVA (RM-ANOVA) indicate a significant main effect of time with a medium effect size ($F(2, 116) = 5.531$, $p = .005$, $\eta 2 = .087$), suggesting that participants' content scores in writing differed across the three time points. Pairwise contrasts reveal a significant improvement in writing scores between the pre-test and delayed post-test ($F(1, 58) = 8.425$, $p = .005$, $\eta 2 = .127$), while the change between the pre-test and post-test was not significant ($F(1, 58) = .661$, $p = .420$, $\eta 2 = .011$), suggesting that the major improvements occur between pre-test and delayed post-test.

Again, the interaction between group and timepoint was not significant with a small interaction effect ($F(2, 116)=2.176$, $p=.118$, $\eta 2 = .036$). Looking at pairwise contrasts, the interaction between group and the contrast between pre-and the first post test was not significant ($F(1, 58) = .026$, $p = .871$, $\eta 2=.000$) nor was the interaction between group and the contrast between the pre-test and delayed post-test ($F(1, 58) = 2.507$, $p = .119$, $\eta 2 = .041$). Therefore, although both groups improve their writing overtime, particularly from the pretest to the delayed post tests, there is no evidence that the AI-reformulation feedback leads to greater improvements than in the control group.

4.1.2.3 Organization

The RM-ANOVA was conducted to examine the effect of time (pre-test, post-test, and delayed post-test) and group (control vs. AI-intervention) on organization scores.

The results revealed a significant main effect of time on organization scores ($F(1.723, 99.823)=3.293$, $p=.048$, $\eta p2=.054$). Breaking this down and looking at pairwise contrasts, there is a significant change in the students' total writing scores between pre-test and the delayed post test ($F(1,58)=4.940$,$p=.030$,$\eta p2=.078$)with a small effect size

(ηp2=.078), while the change between the pre-test and post-test was not significant (F(1,58)=0.232,p=.632,ηp2=.004), suggesting that the major improvements occur between pre-test and delayed post-test.

The interaction effect between time and group was not significant (F(1.723,99.823)=1.061, p=.342, ηp2=.018). Looking at pairwise contrasts, the interaction between group and the contrast between pre-and the first post test was not significant (F(1,58)=0.232, p=.632, ηp2=.004) nor was the interaction between group and the contrast between the pre-test and delayed post-test (F(1,58)=0.790, p=.378, ηp2 =.013). Thus, there is no evidence that the AI-reformulation feedback leads to greater improvements than in the control group.

4.1.2.4 Grammar

In this section, RM-ANOVA was conducted to examine the effect of time and group on grammar scores.

Different from other aspects, the results revealed that the main effect of time on grammar scores was not significant (F(2,116)=.746, p=.477, ηp2=.013), indicating that grammar scores did not change over time. Breaking this down and looking at pairwise contrasts, the main effect of time between the pre- and the first post-test was not significant (F(1,58)=.678, p=.414, ηp2=.012) nor was the contrast between the pre-test and the delayed post test (F(1,58)=1.150,p=.288,ηp2=.019).

Similarly, the interaction between time and group was not significant (F(2,116)=.118, p=.889, ηp2=.002). Looking at pairwise contrasts, the interaction between group and the contrast between pre-test and the first post test was not significant (F(1,58)=.075,p=.785,ηp2=.001), nor was the interaction between group and the contrast between the pre-test and delayed post-test (F(1,58)=.184, p=.670, ηp2=.003). Therefore, both groups did not improve their writing from the pre test to each of the post tests, and there is no evidence that the AI-reformulation feedback leads to greater improvements than in the control group.

4.1.2.5 Vocabulary

The RM-ANOVA was conducted to examine the effect of time (pre-test, post-test, and delayed post-test) and group (control vs. AI-intervention) on vocabulary scores. The results of RM-ANOVA show that there is a significant main effect of time with medium effect size ($F(1.761,102.162)=11.425$, $p<.001$, $\eta p2=.165$), indicating that participants' total scores of writing differ at the three time points. Breaking this down and looking at pairwise contrasts, there is a significant change in the students' total writing scores both between pre-test and post-test ($F(1,58)=20.231$,$p<.001$,$\eta p2=.259$), and between pre-test and delayed post-test ($F(1,58)=15.473$,$p<.001$,$\eta p2=.211$).

The interaction between group and timepoint was not significant with a small interaction effect ($F(1.761,102.162)=3.044$, $p=.058$, $\eta p2=.050$). Looking at pairwise contrasts, the interaction between group and the contrast between pre-and the first post test was not significant ($F(1,58)=3.358$,$p=.072$, $\eta p2=.055$) nor was the interaction between group and the contrast between the pre-test and delayed post-test ($F(1,58)=.696$,$p=.407$,$\eta p2=.012$). Therefore, although both groups improve their writing from the pre test to each of the post tests, there is no evidence that the AI-reformulation feedback leads to greater improvements than in the control group.

4.1.3 Summary

There is overall evidence that both groups improve - both from the pre to the first post test and from pre to the delayed post test. For specific aspects the writing scores for content, organization and vocabulary of both groups improve from the pre-test to the delayed post-test, while only the vocabulary scores of both groups improve from the pre-test to each of the post-tests. There is no evidence that the grammar scores improved over time in both groups.

There is no evidence that the improvements of both the overall scores and scores of the four components are greater in the experimental group than the control group and thus no evidence that the intervention of AI-generated reformulation feedback was effective in improving writing scores.

4.2 Factors Predicting Writing Score Gains

This section concerns the research question (RQ2) focused on identifying the factors that predict gains in students' writing scores when using AI-generated reformulation feedback. The hypothesis posits that higher AI acceptance and higher language proficiency will predict greater gains in writing scores with the AI intervention. The analysis aims to determine whether the gains from original writing to revised writing are influenced by AI-acceptance. More intriguingly, the study also examines whether AI-acceptance correlates with gains from the pre-test to the delayed post-test. The question here is whether students will perform better or worse on the delayed post-test, where AI feedback is no longer available, depending on how much they relied on the feedback during the intervention.

Multiple regression was conducted within the experimental group with AI reformulation feedback. The independent variables AI-acceptance (i.e. the ratio that participants accepted the changes from AI-generated reformulation) and language proficiency (indicated by CET-4 scores) are continuous and normally distributed and there are no outliers. The dependent variables are writing score gains at different timepoints. The writing score gains from the pre-test to the first post-test were computed by subtracting the pre-test score from the first post-test score. Similarly, the writing score gains from the pre-test to the delayed post-test were computed by subtracting the pre-test score from the delayed post-test score. The dependent variables were found to meet the assumptions of normal distribution. The data are independent from one another — each set of scores, after all, comes from a different student in this group. The data for the overall writing scores and scores for each of the components were analyzed statistically in five separate multiple regression analysis.

4.2.1 Total Scores

We are primarily interested in the effect of AI acceptance on total writing score gains. Some previous literature shows that language proficiency might influence the writing gains by influencing the effectiveness of using the writing feedback, so we want

to see what additional variance in the writing score gains we can explain with the language proficiency, indicated by the CET-4 scores in this case. Therefore, we decide the order in which I entered the two IV: Block 1: AI-acceptance. Block 2: language proficiency (CET-4 scores).

The residuals of these models were approximately normally distributed, satisfying the assumption of normality, and the assumption of homoscedasticity was met, as indicated by a P-P plot showing constant variance of the residuals. Multicollinearity was not a concern, with all VIF values below 10 (AI acceptance: VIF = 1.093; 1.104 CET-4: VIF = 1.093; 1.104).

The hierarchical multiple regression was conducted to first account for AI acceptance and then to evaluate the contribution of CET-4 scores to the total score gains in performance from post-test to pre-test. The initial model, which included AI acceptance as the predictor, explained 18.0% of the variance in total score gains ($R2=0.180$, $F(1,21)=4.594$, $p=0.044$). Adding CET-4 scores to the model resulted in a negligible increase in R2 (0.001), with the total variance explained remaining at 18.0%. This suggests that CET-4 scores do not make a significant contribution to the model beyond what is already explained by AI acceptance. Therefore, AI acceptance is a significant predictor of total writing score gains, while CET-4 scores do not appear to significantly affect these gains.

As for the contribution of AI acceptance and CET-4 scores to the gains in performance from delayed post-test (Time 3) to pre-test (Time 1), AI acceptance as a predictor, explained 0.0% of the variance ($R2=0.000$,$F(1,21)=0.004$,$p=0.953$). Adding CET-4 scores to the model resulted in a minimal increase in R2 (0.007), with the total variance explained being 0.7% ($\Delta R2=0.007$,$\Delta F(1,20)=0.131$,$p=0.721$). These results indicate that neither AI acceptance nor CET-4 scores significantly contribute to explaining the variance in total score gains in delayed post test. Consequently, AI acceptance and CET-4 scores do not appear to be meaningful predictors of performance

gains from Time 3 to Time 1 in this context, despite the contribution of CET-4 scores was marginally larger than AI acceptance over the longer term (one-week later).

4.2.2 Content

For content score gains from Time 2 to Time 1, AI acceptance explained 7.9% of the variance ($R^2 = 0.079$, $F(1,21) = 1.801$, $p = 0.194$). Adding CET-4 scores to the model resulted in a small increase in explained variance to 10.6% ($\Delta R^2 = 0.027$, $\Delta F(1,20) = 0.594$, $p = 0.450$), indicating that neither AI acceptance nor CET-4 scores significantly contribute to the variance in gains in content from post-test to pre-test.

For content score gains from Time 3 to Time 1, AI acceptance explained 1.0% of the variance ($R^2 = 0.010$, $F(1,21) = 0.214$, $p = 0.648$). Adding CET-4 scores increased the explained variance slightly to 2.2% ($\Delta R^2 = 0.011$, $\Delta F(1,20) = 0.233$, $p = 0.634$), with both models remaining non-significant. Consequently, AI acceptance and CET-4 scores do not appear to be meaningful predictors of gains in content performance from Time 3 to Time 1 in this context. The assumptions of normality, homoscedasticity, and multicollinearity were adequately met.

4.2.3 Organization

For organizational gains from Time 2 to Time 1, AI acceptance explained 24.3% of the variance ($R^2 = 0.243$, $F(1,21) = 6.739$, $p = 0.017$), indicating a significant effect. Adding CET-4 scores did not improve the model ($\Delta R^2 = 0.000$, $\Delta F(1,20) = 0.003$, $p = 0.955$). These results indicate that AI acceptance is a significant predictor of gains in organization performance from post-test to pre-test, while CET-4 scores do not contribute significantly.

For organizational gains from Time 3 to Time 1 (T3T1_gains_organization), AI acceptance explained 3.0% of the variance ($R^2 = 0.030$, $F(1,21) = 0.645$, $p = 0.431$). Adding CET-4 scores increased the explained variance to 16.8% ($\Delta R^2 = 0.138$, $\Delta F(1,20) = 3.329$, $p = 0.083$). These results indicate that while AI acceptance alone does not significantly predict gains in organization performance, the addition of CET-4 suggests

potential importance, although the combined model did not reach statistical significance. The assumptions of normality, homoscedasticity, and multicollinearity were adequately met.

4.2.4 Grammar

For grammar score gains from Time 2 to Time 1, AI acceptance explained 2.0% of the variance ($R^2 = 0.020$, $F(1,21) = 0.424$, $p = 0.522$). Adding CET-4 scores resulted in a small increase to 4.2% ($\Delta R^2 = 0.022$, $\Delta F(1,20) = 0.458$, $p = 0.506$), both models being non-significant.

For grammar score gains from Time 3 to Time 1 (T3T1_gains_grammar), AI acceptance explained 0.9% of the variance ($R^2 = 0.009$, $F(1,21) = 0.188$, $p = 0.669$). Adding CET-4 scores increased the explained variance to 4.6% ($\Delta R^2 = 0.037$, $\Delta F(1,20) = 0.785$, $p = 0.386$), with both models remaining non-significant.

Therefore, AI acceptance and CET-4 scores do not appear to be meaningful predictors of gains in grammar performance at any time points in this context. The assumptions of normality, homoscedasticity, and multicollinearity were adequately met.

4.2.5. Vocabulary

For vocabulary score gains from Time 2 to Time 1, AI acceptance explained 2.6% of the variance ($R^2 = 0.026$, $F(1,21) = 0.562$, $p = 0.462$). Adding CET-4 scores resulted in a negligible increase to 3.1% ($\Delta R^2 = 0.005$, $\Delta F(1,20) = 0.110$, $p = 0.744$).

For vocabulary score gains from Time 3 to Time 1, AI acceptance explained 1.9% of the variance ($R^2 = 0.019$, $F(1,21) = 0.398$, $p = 0.535$). Adding CET-4 scores resulted in a minimal increase to 2.1% ($\Delta R^2 = 0.002$, $\Delta F(1,20) = 0.050$, $p = 0.825$), with both models being non-significant.

Consequently, AI acceptance and CET-4 scores do not appear to be meaningful predictors of gains in vocabulary performance neither from Time 3 to Time 1 nor from Time 2 to Time 1 in this context. The assumptions of normality, homoscedasticity, and multicollinearity were adequately met

4.2.6. Summary

The regression analyses revealed that AI acceptance was a significant predictor of total score gains from the pre-test to the first post-test. Additionally, AI acceptance significantly predicted gains in organization scores from the pre-test to the post-test, but not in other aspects of writing. However, AI acceptance did not significantly predict total score gains nor scores in individual components on the delayed post-test, the rewriting task conducted one week after the pre-test. This contrast suggests that while participants who closely adhered to the AI-reformulated essay initially achieved improved scores, there is no evidence that this strategy had a lasting impact on their ability to rewrite the same essay when the AI reformulation was no longer available.

Language proficiency, indicated by CET-4 scores in this case, did not significantly contribute to any of the models, suggesting that language proficiency as measured by CET-4 does not predict writing score gains in this context.

4.3 Results of Stimulated Recall

The third research question focused on the features EFL learners would notice during the revision stage and the fourth question focused on the subsequent revision behaviors after noticing. To investigate the two questions, a stimulated recall analysis was conducted on a sample of 14 participants, evenly selected from the control group and the experimental group. Following the coding scheme of Kang (2020), the features noticed by participants were categorized into lexis, grammar, content, organization, or other features (see Table 5).

Table 5. Coding categories for noticed features in the stimulated recall.

Categories	Explanations	Examples
vocabulary	features related to the use of words, including collocations, phrases, etc.	-“I think the word ‘zone’ can replace my use of the word ‘place,’ as the word ‘place’ has been used too frequently.”

grammar	features related to any grammatical rules.	- “this verb is missing an ‘s’ to match the third person singular form.”
content	features related to the generation of ideas.	-“I think this topic sentence is very well-written; it concisely conveys the idea I wanted to express.”
organization	features related to the logical arrangement of ideas.	-“the two points have been combined into one paragraph, but I still believe they should be separated into two paragraphs because they address different aspects.”
other	any features not fitting into the categories of lexis, grammar, content, or organization.	-“I can’t mimic the AI’s language style, which is quite different from mine.”

In terms of the quantity of noticing, the results (Table 6) indicated that learners with the AI-generated reformulation noticed significantly more features than control group (p=.000, Cohen’s d=2.618). Most learners with the AI-generated reformulation text as the reference noticed lexical features at the revision stage (65.5%), compared to grammar (8.4%), content (8.4%), organization (18%). In contrast, in the control group, learners noticed most grammatical features (36.8%), rather than vocabulary (33.3%), content (10.5%), organization (10.5%). Comparing the two groups, the experimental group noticed significantly more features in vocabulary (p=.001, Cohen’s d=3.198) and content (p=.007, Cohen’s d=1.739) than the control group, while the control group noticed significantly more features in grammar (p=.041, Cohen’s d=1.289) than the experimental group. There is no significant difference in the noticing in organization features. This suggests that students with AI-generated reformulation feedback outperformed those without the intervention in the quantity of noticing, particularly in features related to vocabulary and content.

Table 6. Quantity of noticing.

		total	lexis	grammar	content	organization	other
AI- Group	William	23	9	1	8	4	1
	Simon	15	8	0	5	1	1
	Cindy	21	12	1	4	3	1
	Gillian	16	4	0	4	5	3
	Daisy	25	15	0	6	2	2
	Felicity	31	15	1	8	3	4
	Winter	29	11	0	6	9	3
	total	160	74	3	41	27	15
Control	Lucy	3	2	1	0	0	0
Group	Linda	8	1	2	4	1	0
	Rose	6	1	1	3	0	1
	Sophia	6	0	3	1	1	1
	Fiona	8	0	0	3	4	1
	Kate	14	2	1	4	5	2
	Helen	16	3	3	5	4	1
	total	61	9	11	20	15	6

The next crucial step is whether students can correctly implement the noticed features into their own writing, which should directly lead to enhancements in their work. For the control group, 86% (n=53) of the noticed features were correctly implemented into their writing, while 13% (n=8) of the noticed features were incorrectly handled, and 1% (n=6) of the noticed features were disregarded due to a lack of resources for solutions.

For the AI group, of all the noticed features, 93% (n=150) of the noticed features (n=160) were correctly corrected. 93% (n=150) can be traced back to/related to the AI-generated reformulation text in the experimental group. From the AI-generated reformulation, 80% (n=120) of them were accepted and changes were made to their own

initial writing, while 20% were unaccepted. For those who accepted the AI-generated feedback, three key conditions emerged: the need for solutions, the availability of better alternatives, and the inability to find solutions, leading them to turn to the AI version. In terms of actions taken upon acceptance, three were identified: copying the entire suggested part, making adjustments to align with the original, and using the feedback flexibly. Of these being accepted, 94% (n=113) were correctly implemented into their own writing. This also suggests the effect of the reformulation feedback, that participants with reformulation at hand can notice more features, and most of these features are related to the reformulation. This can supplement the previous findings that the AI-acceptance can predict their subsequent writing score gains.

For the 20% of them which is not accepted, there are some themes for the reasons why participants did not accept these changes from the AI-generated reformulation from inductive coding (data-driven) (see Table 7). These decisions of unacceptance were based on their evaluation of the AI-generated reformulation texts and a comparison with their own writing. Factors contributing to were identified through thematic analysis were displayed in the table.

Table 7. Code for reasons for rejection.

Code for reasons for rejection	Explanation	Example
Capacity-inappropriate change	Students did not accept the AI-generated reformulation texts because they are too advanced to match students' developmental level.	- "I thought writing in the way of what AI wrote is too difficult for me at the current level, so I decided to keep my own commonly used phrases." (Winter) - "I think if I blindly imitate the advanced vocabulary used by the AI, it might backfire, leading to spelling errors or disrupting the overall flow of my writing, so I kept to use the vocabulary I am more familiar with. This way, I can respond to my writing tasks more confidently." (William)

Differences in original idea	Students did not accept the AI-generated reformulation texts because AI version changed their original idea.	-I wrote that "smoking in some special public zones would even threaten other people's lives," while the AI stated, "Environmentally, smoking poses a severe risk in specific public zones, like mountainous regions, where a discarded cigarette can lead to catastrophic wildfires." The focus differs between the two, and I insisted on my version. (William) - Last sentence, I did not change this sentence, because I also mentioned pregnant women, but it did not specifically say that it is the impact on pregnant women. I think since I mentioned there's a pregnant woman, I should keep this place (Felicity).
Removal of original content	Students did not delete their own texts when they found no corresponding part in the AI-generated reformulations.	-I did not delete this point about air quality, despite that I did find the corresponding part in the AI version (Felicity) - I did not change the last sentence because I previously mentioned pregnant women, I felt it was important to retain this part for consistency, while the AI version does not specifically address the impact on pregnant women here. (Felicity)
Change on original logic	Students did not accept the AI-generated reformulated texts when the AI changed the logic of their original writing.	-It combines my paragraph with the next one, and I don't understand why it merges and changes my structure. However, I believe this paragraph focuses on the individual, while the next one is about the economy. Therefore, I decided to keep my original structure with two separate paragraphs. (Gillian).
Insisting personal writing style	Students did not accept the AI-generated reformulated texts because of their personal writing styles.	- I haven't changed the beginning of this part, which, like the second and third parts, starts with "Firstly," "Secondly," and "Thirdly." I retained my own words because this structure

		aligns with my logical flow and habits. The AI's version does not clearly reflect this sequential logic. (William) - I think my paragraph is quite detailed. The AI's version is also detailed, but I cannot mimic its language style, so I kept my original paragraph unchanged. (Gillian)
Avoiding plagiarism	Students avoid copying large chunks of AI-generated reformulation texts to prevent possible plagiarism.	- I feel that the AI-generated reformulation is much better than mine. I really want to make changes here, but I've already accepted too much changes from the AI version. Since this is just a revision of my own writing, if I change too much, won't it be like copying others? (Daisy)
word count	Students did not accept some changes from the AI-generated reformulation due to concerns about word count.	-I think these three points of view could be written in three independent paragraphs, just as the AI did. However, due to my relatively small word count, I did not separate them into individual paragraphs. (Gillian)
Rigidity of AI texts	Students did not accept the AI-generated reformulations because they thought some of its changes were rigid.	- I felt that the sentence generated by AI lacked coherence, so I didn't delete the phrase "according to the research," to better introduce the supporting evidence below. (Cindy)
Redundance of AI texts	Students did not accept the AI-generated reformulations because they thought some of its changes were redundant.	-This point seems overly complicated because the AI provided an extensive explanation, including precedents and justifications, which made it seem verbose. I originally intended to add a brief supplementary comment to this point, so I kept my sentence concise and did not make other significant changes. (Winter)
Irrelevance of AI texts	Students did not accept the AI-generated	-I think this essay topic should focus on whether smoking should be banned or not.

	reformulations because they thought some of its changes were irrelevant.	However, the AI elaborated extensively on how we should implement the ban, which I found irrelevant. Therefore, I used my own last sentence. (Cindy)

Based on the Grounded Theory (Glaser & Strauss, 1967), we have identified three primary categories under which these factors can be organized: internal factors, external factors, and AI-based factors (see Figure 3).

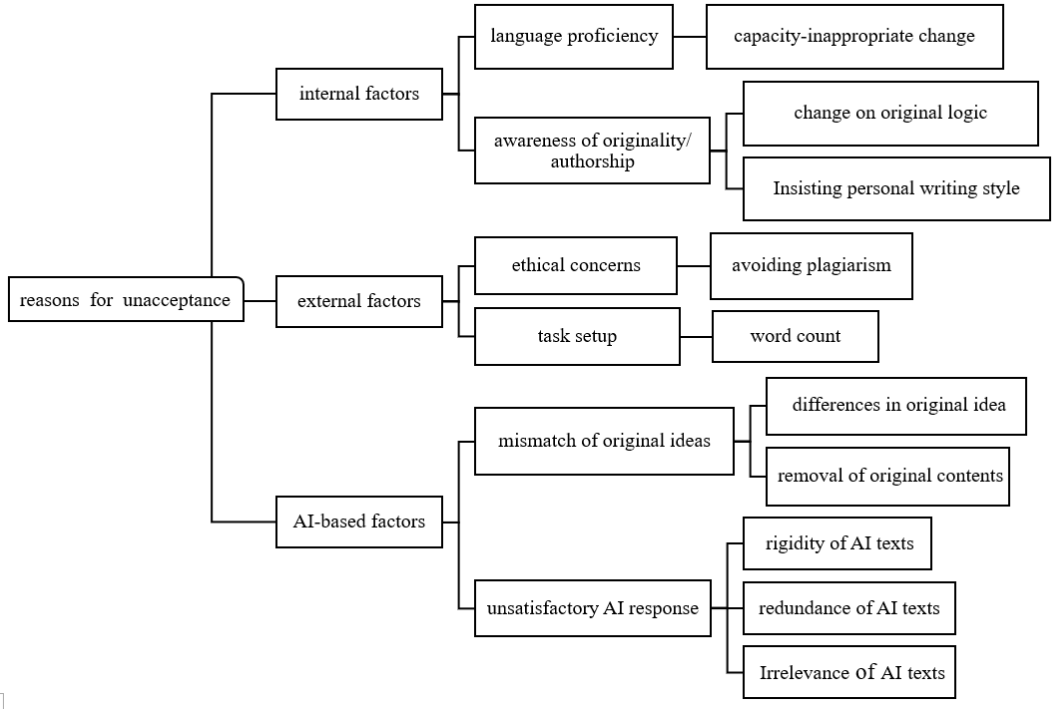

Figure 3. Influencing factors for rejection of feedback

5. Discussion

The previous chapter suggests a discrepancy in the effectiveness of AI-generated reformulation as a form of feedback for improving EFL writing. While the quantitative analysis shows no evidence that such feedback can better improve students' writing scores compared to the control group using self-correction, a significant correlation between the acceptance of AI-generated reformulation revisions and the writing score gains was detected. Additionally, the qualitative analysis of the stimulated recall with a sample of participants revealed evidence that this feedback can enhance noticing. This chapter will explore the disconnection in these findings to draw a cautious conclusion on the effect of AI-generated reformulation feedback on EFL writing. It will first interpret the positive effect of AI-generated reformulation on noticing, situating this within a wider literature. Following this, the chapter will try to explain the lack of a significant effect of the AI-generated reformulation on improving writing scores in this study. Subsequently, implications for writing pedagogy will be summarized. Finally, the chapter will address the study's limitations and implications for future research.

5.1 The Effects of AI-generated Reformulation on Noticing

5.1.1 Improving the quantity of noticing

This section discussed the positive effects of AI-generated reformulation on improving the quantity of noticing. The essay by Qi and Lapkin (2001) differentiated the concept of 'quantity of noticing' and 'quality of noticing', arguing that both play important roles in improving EFL writing. In their study, the quantity of noticing was measured by the number of language-related episodes where participants identified differences or issues in their writing while comparing their text to a reformulation, while the quality of noticing was measured by the amount of noticing where their participants provided reasons for accepting or rejecting the reformulation. In line with this, the current dissertation measured the quantity of noticing by tallying the number of noticed features across four sub-aspects: vocabulary, grammar, content, and organization, reflecting the criteria used in the writing scoring rubric. Besides, the amount of correct

implementation of noticed features into the original writing was measured, allowing for a comparison of the quality of noticing with the control group. The reasons behind participants' acceptance or rejection of the AI-generated reformulations were also examined within the experimental group.

The results for RQ3 found no evidence that the AI-generated reformulation feedback could improve the quality of noticing in this study, compared with the control group using self-correction. The triangulation of results of the stimulated recall and content analysis of participants' original writing and the revised writing revealed that the 14 participants from both groups have reached similarly high rates of successful correction of the issues noticed during the revision stage. This overall success can be attributed to several factors: the participants' familiarity with the writing topic, their language proficiency as intermediate-or-above level English majors, and the fact that they were working from their own production, which means that the issues they identified were likely linked to target language forms that were already partially internalized, enabling them to correct most of the errors they reported (Lázaro-Ibarrola, 2013).

For the quantity of noticing, the dissertation found that participants with the AI-generated reformulation noticed significantly more features than the control group, supporting that AI-generated reformulation feedback helps improve the quantity of noticing. This finding is consistent with previous research on the effect of reformulation on noticing, which has demonstrated that reformulation pushes learners to "notice the gaps" in their L2 production (e.g., Azimian et al., 2023; Coyle et al., 2020; Kim & Bowles, 2019; Lázaro-Ibarrola, 2013, 2023; Milla & Garcia Mayo, 2024; Tsai et al., 2024). In most of the literature, the reformulation feedback was generated by human native or native-like teachers of the target language. To the best of our knowledge, only the study by Tsai et al. (2024) used ChatGPT to generate reformulation texts as a form of feedback to assist EFL English college majors in their writing, although the current dissertation utilized a more advanced version of ChatGPT (GPT-4, released on March 14,

2023) than the one (released on February 13, 2023) used in Tsai et al.'s study. Their findings align with the results in the current dissertation. In both studies, participants were able to notice features through comparison with AI-generated reformulations, despite differences in how ChatGPT was prompted. In the current dissertation, it was the researcher who provided standardized prompts to ChatGPT for generating one-off reformulation text for each participant. In contrast, in the study by Tsai et al. (2024), students were allowed a more flexible, autonomous and interactive use of ChatGPT, with necessary instructions provided beforehand. Despite these differences, a consistency of findings regarding the effect of AI-generated reformulation feedback on enhance noticing has been observed.

The discussion now turns to exploring to what extent the findings regarding noticing in this dissertation may align with the broader body of literature on human-generated reformulation as a form of written feedback in L2 writing. While no research has yet directly compared the AI-generated reformulation feedback with human-generated reformulation, it is still insightful to consider and compare our findings with the extensive body of research on human-generated reformulations. Many studies have provided substantial evidence of the positive impact of reformulations, as defined by Cohen (1983b), generated by native or native-like teachers of the target language, on improving L2/FL learners noticing in the writing processes (e.g., Azimian et al., 2023; Coyle et al., 2020; Qi & Lapkin, 2001; Swain & Lapkin, 2002). For example, Lázaro-Ibarrola (2013) compared the effects of reformulation with self-correction on EFL narrative writing of story retelling. The participants were 16 adolescent students whose L1 was Spanish. Focusing on error detection, this study found that the reformulation group detected more errors (79.23%) than the self-correction group (44.80%). This result suggested that the reformulation group outperformed the self-correction group in the quantity of noticing errors, as error detection can be considered a key purpose and outcome of noticing in language learning. However, it is important to clarify that the focus of this dissertation extends beyond mere error detection. The concept of "noticing"

in the current dissertation encompasses a broader range of features, including not only the identification of problematic issues that need correction but also the recognition of better alternatives, new expressions, and other linguistic improvements. Qi and Lapkin (2001) used the similar concept of noticing to our dissertation. They conducted a within-subjects comparison to demonstrate the effect of reformulation on enhancing the quantity of noticing from the composing stage to the revision stage in EFL writing. The study involved two adult Mandarin-speaking students with advanced English proficiency, studying at a Canadian university. They were required to perform a three-stage narrative writing task based on a given prompt of a picture-story. Researchers used think-aloud protocols to investigate noticing: the participants were asked to verbalize their thoughts while they composed their narratives (Stage 1) and while they compared their original texts to reformulated versions provided by the researchers (Stage 2) four days later. One week after stage 2, participants were asked to revise their draft on the original version as the post-test (Stage 3). In their study, noticing was measured by tallying the language-related episodes (LREs) identified through the analysis of their think-aloud protocols in the first two stages. The study found noticing started during the composing stage, with one participant identifying 25 LREs and the other 16 LREs. However, the quantity of noticing, as indicated by the number of LREs, increased during the revision stage, where one participant identified 29 LREs and the other 31 LREs. Since the study by Qi and Lapkin (2001) was a case study involving only two advanced learners, it is important to approach the findings with caution. Additionally, one concern is the possible distraction from the think-aloud protocols which demand additional cognitive resources and could be very resource-intensive along with the L2 writing processes. Despite these limitations, their study observed the enhanced noticing that occurred when learners reflect on their language output by comparing it with reformulated texts, a finding that aligns with the results of this dissertation's comparison between the reformulation group and the control group.

In this dissertation, given that only 14 participants (7 from the experimental group and 7 from the control group) were involved in the analysis relevant to the RQ3, we consider that this result constitutes 'tentative rather than definitive answers' (Allwright et al., 1998, p. 250). The small sample size also raises the possibility of a false positive in the statistical comparison of the number of noticed features between the two groups. However, the consistency of our findings with positive results in existing literature enhances our confidence that reformulations generated by ChatGPT may have the same positive effect as those provided by human teachers on enhancing the quantity of noticing, which serves as a crucial stepping-stone for writing improvement.

5.1.2 Noticing on specific linguistic aspects

The results for RQ3 also indicated that AI-generated reformulation had a potential effect on improving the quantity of noticing in specific aspects of writing. The 7 students with AI-generated reformulation feedback outperformed the 7 participants without the intervention in the quantity of noticing, particularly in features related to vocabulary and content, but less in grammar and no difference in organization.

Specifically, first, no difference was found in the quantity of noticed features related to organization between the two groups. By contrast, previous literature has found evidence that reformulation can improve noticing of organizational features. For example, Coyle et al. (2020) demonstrated that reformulation can enhance the use of cohesive devices, such as pronominal references and sequence markers when the reformulation texts were targeted as improving organizational features, specifically, the reference cohesion. They suggested that it is necessary to explicitly target specific cohesive devices in the reformulated feedback to increase learners' chances of noticing and incorporating them into their written texts, which echoing the suggestion by Qi and Lapkin (2001) that guiding learners to focus on specific aspects of a reformulation can improve their capacity to derive benefits from it. The dissertation did not include prompts for ChatGPT to generate reformulations targeting specific features, which likely explains why there was no effect on noticing organizational features. However, an interesting

point is that, under the same conditions where no specific features were targeted, although students did not notice more organizational features, they did notice more features in other areas, such as vocabulary and content. This suggests that students allocated their attentional resources differently when noticing various aspects, and there is a prioritization among these aspects. We will continue to explore this in the following discussion.

Second, the reformulation group noticed significantly more features related to vocabulary and content compared to the control group. It is easy to understand that, compared to self-correction without any additional resources, reformulation serves as a valuable tool, offering richer lexical options. Furthermore, since the reformulation is based on the students' own writing, it directly addresses their specific vocabulary needs. This finding is consistent with previous literature (Adams, 2003; Sachs & Polio, 2007). In addition, we found that AI-generated reformulation had a positive effect on noticing features related to content. Previous literature presents mixed findings on the effect of reformulation on noticing content-related features and suggests that learners' language proficiency is a key factor in determining their ability to notice content through reformulation feedback. In the study by Lázaro-Ibarrola (2013), which involved low-level adolescent EFL learners as participants, they found that learners reported noticing only lexical and grammatical changes at the sentence level, completely overlooking other modifications such as punctuation, removal of redundant information, and content reorganization. Lázaro-Ibarrola (2013) therefore concludes that the learners "are missing the forest (content-related errors) for the trees (lexical errors)" (p. 40). This finding contrasts with previous studies involving adult participants who did report text-level or content related features (Cohen, 1989; Qi & Lapkin, 2001) as well as the findings in our dissertation. To explain this contrast, Lee (1997) explained that low level learners tend "to perceive editing primarily in terms of bringing about surface changes" (p. 471). It is possible that the lower proficiency level of the students prevented them from considering

the text as a whole, while the participants in this dissertation, who are college English majors, were capable of noticing content-related features.

Finally, this dissertation found that the reformulation group noticed less grammatical features than the control group. This finding can be explained by the limited attentional capacity theory (Skehan, 1996), which claims that it is challenging for learners to distribute their attention resources effectively to focus on both meaning and form simultaneously due to cognitive overload. Facing such dilemma, empirical evidence supported that learners often prioritize meaning over form (Kang, 2020). In this dissertation, the features on vocabulary and content are meaning-based, while the grammatical features are form-based. For the reformulation group, the cognitive process of comparing their own writing with the reformulated text was complex, involving both perception and production of language, which could be very attention-intensive. Given that our participants were generally intermediate language learners, they may find it difficult to focus on both form and meaning and thus automatically sacrifice attention on noticing grammatical features in favour of prioritizing resources on content and vocabulary features. In contrast, participants in the control group, who engaged in self-correction without additional resources, may have been able to focus more on grammatical features because they had already completed the content generation in the initial stage and had sufficient attentional resources available for grammar during the revision stage.

5.2 Acceptance and Rejection of AI-generated Reformulation

The previous section provides substantial evidence of the impact of AI-generated reformulations on increasing the quantity of noticing through a between-group comparison. In this section, the focus shifts to a within-group analysis, exploring the role of AI-generated reformulation feedback specifically within the experimental group and aiming to build the connection between the noticing and the behaviors of students' revisions. This discussion will cover the results related to both RQ2 and RQ4, interpreting students' acceptance and rejection of AI-generated reformulations.

5.2.1 Acceptance

Noticing is the first step toward improvement, yet effectively implementing what is noticed is crucial for actual progress to happen. The behaviors of students' revisions can be categorized as acceptance and rejection of the noticed features derived from the AI-generated reformulation feedback.

The acceptance of the AI-generated reformulations reflected the good quality of the texts generated by ChatGPT. The results of RQ2 found that the ratio of AI-acceptance significantly predicted writing score gains in the first revision, particularly in organization, but not in other areas or in the delayed post-test rewriting. This means that the more students adhered to the AI-reformulated texts, the better improvements in overall writing score gains they made, supporting the effectiveness and usefulness of ChatGPT (Hsu, 2023; Thi et al, 2023). To supplementary this finding, in the within-group analysis of the 7 participants in RQ3, the overwhelming 80% of acceptance also supported the high quality of the AI-generated reformulation text, suggesting that students can find solutions they expected in the AI-generated reformulation texts.

Additionally, the acceptance of the AI-generated reformulations also unveiled students' critical and effective use of ChatGPT. Learners in this dissertation demonstrated three types of acceptance: they either copied the entire AI-generated text, made adjustments to the AI-provided content, or used the feedback as a flexible resource. For example, some participants adopted expressions from the AI-generated version and incorporated the vocabulary wherever needed in their writing. The various, flexible and autonomous use of AI-generated texts reflect on students' critical use of GenAI and the evaluation of the contents generated by ChatGPT. Of these being accepted, 94% (n=113) were correctly implemented into their own writing. This also suggests students' engagement with this feedback is overall effective and successful.

However, the impact did not last for the delayed post-test, indicating that students may have difficulties to transfer the noticed features into the intake in their interlanguage system (i.e., Schmidt, 1990). This lack of long-term effect could be attributed to various factors, such as learners' language proficiency and motivation. Given that the

participants in this dissertation were intermediate English learners, it is possible that they were not fully capable of transferring the input (i.e., the noticed features) into intake within the short duration of the experiment. Additionally, the outcome of noticing is vulnerable to time-based decay, meaning that without reactivation the improvements made during revisions may easily fade (Godfroid et al., 2013). Furthermore, in the laboratory setting of this research, students may have lacked the motivation to convert noticing into intake, especially since they were unaware that they would be required to rewrite the essay as a post-test.

5.2.2 Rejection

One innovative finding of this dissertation on learners' rejection of AI-generated reformulation texts. To our knowledge, there is no other studies exploring the rejection of AI-generated reformulations. We believe the behaviour of rejection demonstrated learners' critical thinking and evaluation of the content produced by ChatGPT. The factors contributing to this rejection can be categorized into internal factors, external factors, and AI-based factors (see Figure 3).

The first category of factors influencing rejection is at the individual level. A key reason for rejecting AI-generated reformulations was a lack of language proficiency. Three of the seven participants felt that the AI-generated revisions were too advanced for their current developmental level, making it difficult and stressful to incorporate these features into their writing. Even though they recognized the potential improvements, they believed that accepting such changes would not benefit their future writing because they couldn't produce such sophisticated text independently. This finding suggested that EFL learners may need additional resources to assist their engagement with the reformulation feedback, which will be addressed in the following section. Besides, awareness of originality and authorship played a role in the behavior of rejection. As generative AI becomes more prevalent in academic writing, the issue of authorship in human-AI collaboration has received a lot of attention (e.g., DuBose & Marshall, 2023; Morrison, 2023). Participants often chose to retain their own way of organizing ideas when

ChatGPT altered the original logic of their writing. They also rejected changes to preserve their personal writing style, such as their habitual use of specific cohesive devices, to maintain their authorship and originality.

The second category is external factors including ethical concern and task setup. Some participants rejected a large portion of AI-generated changes to avoid potential plagiarism, even though they recognized that the AI-generated sentences were more concise and refined. The ethical implications of using generative AI have been widely debated, and there is no consensus or absolute definition of plagiarism when it comes to AI-generated content (e.g., DuBose & Marshall). However, one consensus in academia is to thwart "AIgiarism'—AI-assisted plagiarism" (Morrison, 2023). However, there is a growing consensus in academia to prevent "AIgiarism"—AI-assisted plagiarism (Morrison, 2023). In this dissertation, it was exciting to find that participants were mindful of these ethical concerns and sought to avoid cheating by rejecting AI-generated reformulations, demonstrating an awareness sense academic integrity. Another external factor is the task setup, which is context-dependent. The word count for the writing task was approximately 200-300 words. Some participants found that incorporating AI-generated text as supplementary material would cause their essays to exceed the word limit, leading them to discard these changes.

The third category involves AI-based factors. The first factor in this category is the mismatch between the original ideas and the AI-generated reformulations, which led some students to reject these changes. According to the definition, the reformulation texts should preserve as much of the student's original ideas as possible. However, in this dissertation, some participants noted that when they compared their own writing with the AI-generated reformulation, the AI occasionally altered their original ideas by either changing them or omitting some of the key points. When the AI-generated text deviates from the intended reformulation by altering original ideas, participants are more likely to reject these changes and retain their original content. This issue may have arisen because the prompts used to generate the reformulations did not clearly define the concept of

reformulation, as previous research has shown that different prompts can produce AI-generated texts of varying quality and relevance (e.g., Cai et al., 2023; Escalante et al., 2023; Giray, 2023). Furthermore, some participants rejected the AI-generated reformulations due to unsatisfactory responses from the ChatGPT. They pointed out that the AI-generated text could be rigid, redundant, or irrelevant to the writing task, and therefore chose to reject these unsatisfactory changes. This demonstrates their critical thinking skills during the comparison stage with AI-generated reformulations—an ability that is increasingly important to the use of GenAI in academia.

5.3 The Null Effect of AI-generated Reformulation on Improving Writing Scores

Previous sections discussed the evidence found in this dissertation regarding the positive effects of AI-generated reformulation and how this feedback influences EFL learners' revision behaviours. Scores assessing participants' writing at before and after the intervention in our study reveals revealed that both the experimental and control groups showed improvement from the pretest to both the first and delayed posttests. However, there was no statistical evidence to suggest that the experimental group experienced greater improvement, indicating that the intervention may not have produced significantly better improvements. This section will discuss the disconnection in our findings and explore the impact of AI-generated reformulation on improving writing scores.

5.3.1 Type II errors

As with other studies that report non-significant results, one possibility is a Type II error, meaning the study may have been underpowered. Given that there was evidence of learners benefiting from AI-generated reformulation in terms of noticing, and a correlation was found between the acceptance of noticed features from AI-generated reformulations and writing score gains, the type II error seems plausible. Future research could replicate this study with a larger sample size to address these limitations and improve the test's statistical power.

5.3.2 Language proficiency

In addition to the possibility of a Type II error, the null effect observed in this study may be due to the participants' insufficient language proficiency to fully engage with the feedback. Learners' proficiency levels are crucial for understanding and effectively utilizing the feedback provided by AI-generated reformulation. In the qualitative analysis of the rejection of AI-generated reformulations, as discussed above, some participants reported that certain advanced expressions were beyond their current level, which could have made it difficult for them to incorporate these features into their writing. It is likely that this issue existed among other participants who were not selected for qualitative analysis, which leads to the overall null-effect of this feedback. Previous research also emphasized the role of learner's language proficiency in the engagement with the reformulation feedback. Recall that the study by Qi & Lapkin (2001), which involved two advanced EFL learners as participants, provides evidence of the effect of the reformulations on noticing on an individual basis. They found that during the comparison stage, learners with higher L2 proficiency tend to accept more reformulated items or structures and articulate their reasons for doing so, compared to learners with lower L2 proficiency. To explain this difference, the researchers argued that learners with lower proficiency may struggle more to identify the gap between their interlanguage and the target language, even when a "target language model"(Qi & Lapkin, 2001, p. 295) (i.e., reformulations) is provided. One concern of this study is that participants' language proficiency was based on self-reports. However, this may not pose a significant issue. The researchers aimed to examine the influence of language proficiency, and the differences between the two participants' English proficiency levels were evident. The participant with higher language proficiency had studied English as a foreign language for over 10 years and had been living in Canada for three years at the time of the study. In contrast, the participant with lower proficiency had been in Canada for only five months and had never lived in any other English-speaking country before. These distinctions make it relatively easy to differentiate between higher and lower language proficiency levels, ensuring that language proficiency remained a relevant variable in the

study on noticing. Although their study included only two participants, making the results tentative and difficult to generalize, the finding that higher L2 proficiency leads to better noticing aligns with Cohen's (1983) assertion that reformulation is likely to benefit "learners at intermediate levels and above" and may have the greatest impact on advanced students (p. 5). The participants in the current dissertation were mostly second- and third-year college majors who primarily used their L1 in class and had limited exposure to English in their daily lives. Their language proficiency can be considered intermediate, as indicated by their average CET-4 scores and their language exposure. Compared with the participants in study of Qi and Lapkin (2001), those in the current dissertation had lower language proficiency. This may explain why no evidence was found for the effect of AI-generated reformulation feedback on writing score improvements.

To achieve significant improvements with AI-generated reformulation, especially for students with lower proficiency, it may be necessary to provide them with additional resources to help engage with the feedback. For example, students may benefit from tools such as dictionaries and strategies such as peer discussions and consultations with teachers (Lázaro-Ibarrola, 2013; Yang & Zhang, 2010). The study by Escalante et al. (2023) provides valuable insights that interpersonal interactions can enhance engagement with feedback. The research investigates the preferences of 43 ENL students who received feedback from both AI and human tutors over a six-week longitudinal quasi-experimental study. The writing task required participants to write a 300-word paragraph on diverse academic topics related to the material covered in class each week, and they received feedback from both AI and human tutors weekly. It is important to note that the feedback in this study was not in the form of reformulations but rather comments on various aspects of their writing. The results from the questionnaires and surveys showed an almost equal preference between AI-generated and human-generated feedback. Some students valued the clarity, consistency, and detail of AI feedback, while others preferred the personal interaction and immediacy offered by human tutors. Participants noted that

interpersonal interactions during the revision process provided affective benefits, such as increased engagement, which were not as pronounced when merely reading AI-generated feedback. Based on these findings, it is reasonable to consider that incorporating interpersonal interactions during the noticing stage with AI-generated reformulation feedback could be beneficial in enhancing the use of feedback.

Moreover, it could be effective for provide students with interactive use of ChatGPT to assist them in their engagement with the AI-generated reformulations (e.g., Escalante et al., 2023; Su et al., 2023; Tsai et al., 2024). In the current dissertation, it was the researcher who provided prompts to ChatGPT on behalf of the students, meaning that the students did not directly use the AI. This operation was due to time constraints, which did not allow for training students in the effective use of ChatGPT, and to ensure the secure use of the tool, avoiding potential misuses such as leaking personal data. However, the interactive use of ChatGPT by learners has two main benefits that may lead to significant writing improvements. First, providing students with opportunities to ask follow-up questions to the AI could help better incorporate AI-generated reformulation feedback into their practice, giving them another chance to interact with ChatGPT and clarify their needs. For example, in the study by Tsai et al. (2024), college English majors with similar proficiency levels to our participants were given more autonomy in using ChatGPT after receiving professional training and instruction, which led to significant improvements in their writing scores. Second, the interactive use of ChatGPT can help improve the quality of the texts it generates. The quality of ChatGPT's responses is closely tied to the quality of the prompts provided (Giray, 2023; Escalante et al., 2023). In this dissertation, while standardized prompts were used to generate reformulations, the students' writing also served as part of the prompts. Since ChatGPT's responses are highly dependent on the input, poor initial writing from students may result in low-quality AI-generated reformulations. To minimize this risk, it is crucial to provide students with opportunities to interact with ChatGPT, allowing them to clarify their needs in their writing and find better solutions from the AI-generated feedback.

In conclusion, the lack of significant improvement in writing scores may be due to students not having sufficient proficiency to effectively engage with AI-generated feedback. To address this, students may require additional resources, whether external aids like dictionaries and textbooks, collaborative engagement, or the promising approach—interactive use of ChatGPT.

5.2.3 The genre of writing

The null effect observed in this study may be due to the inherent limitations of reformulation, which might have limited its effectiveness in the context of argumentative writing. As Allwright et al. (1988) noted, a 'good' reformulation may not necessarily be a 'good' example of native writing, as it is constrained by its fidelity to the original writer's intentions. Much of the previous literature that has found evidence of reformulation positively affecting writing improvement has focused on narrative writing tasks, such as story retelling and picture-prompted narrative tasks (e.g., Coyle et al., 2020; Lázaro-Ibarrola, 2013; Milla & García Mayo, 2024; Yang & Zhang, 2010). In narrative writing, content requirements are often less demanding since the elements are typically fixed by the task prompts. In contrast, argumentative writing heavily relies on content, particularly the strength and clarity of the argument, which plays a crucial role in determining the overall quality of the writing. Reformulations, by definition, are constrained in their ability to enhance content significantly, as they must remain faithful to the original ideas of the writer. This limitation raises the question of whether reformulation can be as effective in improving argumentative writing as it has been shown to be in narrative writing.

To address the importance of content in argumentative writing, it is useful to compare the effects of reformulation feedback with model essay feedback. The primary difference between the two lies in the fact that reformulations maintain the learner's original ideas, while model essays are composed according to the task prompts or writing topic. Previous research has provided evidence that model essays can significantly improve argumentative writing scores (Kang, 2020, 2024). While reformulation is also limited in its ability to modify content compared to model feedback, suggesting that

reformulation may be more suitable for narrative writing, whereas model essays could be more effective in improving argumentative writing.

To further explore the use of GenAI on assisting argumentative writing, Su et al. (2023) explored a more varied approach than reformulation when using ChatGPT to assist in argumentative writing. highlights how ChatGPT can enhance argumentative writing by supporting students throughout various stages of the writing process. They displayed that ChatGPT aids in idea generation and provides feedback on outlines during the preparation stage, offers content and language feedback during the editing stage, assists with grammar and language correction during proofreading, and helps students reflect on their writing through chat history. By integrating these functions, ChatGPT can help students develop more structured, well-supported arguments and improve their overall writing skills on the individual basis. This approach may offer valuable insights into how GenAI can be effectively utilized to enhance argumentative writing, which is particularly important in academic contexts.

5.3.4 Multiple “dosage”

One possibility worth exploring is that the noticing and cognitive processes observed in participants might translate into significant improvements if the intervention were extended over a longer period, with students repeatedly engaging in the reformulation process across different essays. The lack of detectable improvement in writing skills may simply be due to an insufficient “dosage” of the intervention in this study. In the current study, the intervention may not have provided enough opportunities for these improvements to fully manifest in measurable writing improvements. A more extended intervention period, allowing for continuous practice and reinforcement, for example, repeated using AI throughout the semester, could potentially yield more substantial and lasting improvements in students’ writing abilities.

5.4 Pedagogical implications

This study highlights the importance of a critical approach to utilizing AI-generated reformulation feedback in EFL instruction.

First, Teachers should consider integrating AI-generated reformulation feedback in EFL writing classes. ChatGPT, in particular, can serve as an effective reformulator, offering rapid feedback that rivals the quality of human reformulation but without the delays typically associated with human intervention. This immediate feedback could enhance students' engagement and noticing, leading to improved writing skills over time.

Second, educators should be reminded that while GenAI offers timely feedback, its effectiveness may not always meet expectations. Data from this study revealed that students sometimes found AI-generated content redundant, unnecessary, or irrelevant. However, the critical engagement shown by students, who did not fully accept AI modifications without careful consideration, is a positive sign. This suggests that students are capable of exercising critical thinking when interacting with AI-generated feedback, which is a crucial skill in the digital age. Therefore, teachers should provide students with explicit instructions on how to use GenAI tools effectively. This training is essential not only for ensuring safe usage but also for promoting interactive and meaningful engagement with AI. Additionally, fostering critical evaluation skills will help students assess the quality of AI-generated content, encouraging them to think deeply about the feedback they receive.

Third, while GenAI has the potential to be a valuable tool in EFL instruction, it should be used in conjunction with teacher involvement and supplementary resources. Especially for learners with lower language proficiency, teachers should provide additional scaffolding to help them understand the features they notice in AI-generated feedback. This might involve breaking down complex feedback into more manageable parts, offering supplementary explanations, or facilitating cooperative writing for peer discussions to ensure that all students can fully benefit from the reformulation process.

5.5 Limitations

This dissertation has several limitations. First, one primary limitation is the reliability of ratings and coding. Due to the time constraint, the rating for most writing was conducted solely by the researcher, with only 10% independently rated by a co-rater. Although there was good interrater reliability for this subset, the acceptance on a single

rater for most of the rating may have overlooked subtle improvements in students' writing. The coding of the stimulated recall data was also performed by the researcher alone, though it was double-checked. This operation could introduce bias and limit the reliability of the findings.

Second, the prompts for ChatGPT used in the study could be optimized, which is essential to elicit the best possible performance of ChatGPT. According to feedback from the qualitative study, some students reported that AI-generated reformulations sometimes altered their original ideas, conflicting with the traditional definition of reformulation feedback. Adjusting prompts to include the definition of reformulation and specific requirements, such as adhering to students' language proficiency, could enhance the accuracy and relevance of the feedback.

Third, the cross-sectional design, which involved a single writing task, also limits our understanding of how AI-generated reformulation feedback might affect different writing tasks and whether its impact endures over time. The lack of reinforcement and multiple interventions in our study may have constrained long-term writing improvements, potentially due to the limited transfer of noticing into meaningful writing outcomes.

Another limitation is that our study focused only on the noticed features, without accounting for unnoticed features, which are also critical for understanding the whole picture of learners' noticing and learning from this process.

Given these limitations, future research should enhance interrater reliability by ensuring a larger proportion of ratings are independently verified by multiple raters, possibly through a double-blind rating process. Additionally, optimizing ChatGPT prompts is crucial for eliciting more accurate and relevant feedback. Furthermore, longitudinal studies with multiple interventions would provide a more comprehensive understanding of the long-term effects of AI-generated reformulation feedback across different writing tasks. A thorough analysis of both noticed and unnoticed features would

also help offer a fuller picture of the learning process. By addressing these limitations and future research directions, educators and researchers can better understand the potential and constraints of AI-generated feedback in EFL writing, ultimately enhancing its effectiveness in EFL academic writing.

6. Conclusion

The dissertation aimed to discover the effect of AI-generated reformulation feedback on EFL academic writing. Statistical analyses of the scores did not find evidence that AI formulated feedback was effective in improving EFL students' argumentative writing scores compared to the control group using self-correction. However, the extent to which students accepted the AI-generated reformulated texts was a significant predictor of writing score gains, particularly in the organizational aspect, suggesting that following the AI-generated reformulation feedback led to better improvements in scores, at least on the first revised writing. Furthermore, analyses of responses from a subset of participants taking part in simulated recall found that those in the experimental group noticed more language features than the control group, indicating that AI-generated reformulation feedback improved EFL learners' noticing.

Further work is needed to understand the discrepancies. As covered in the discussion, the discrepancies may be caused by a false positive on the noticing (type I error), or, more likely, because the effect of AI-generated reformulation may be very small, requiring a larger sample size and a higher dosage (i.e., multiple interventions) to be detectable. It is also possible that the participants in the current study did not have sufficient language proficiency to effectively "learn" from the reformulation after "noticing" the gaps between their interlanguage and the target language. Additionally, the discrepancies might stem from inherent limitations of the reformulation feedback itself in improving the content or argumentation, which are crucial aspects for argumentative writing.

The dissertation also fills the gap in the existing literature by exploring learners' critical use of ChatGPT, specifically investigating their behaviours of rejection to the changes from AI-generated reformulations. The factors contributing to this rejection can be categorized into internal factors, external factors, and AI-based factors, suggesting learner's awareness of critical thinking when using ChatGPT.

This dissertation provides valuable insights for EFL pedagogy and future research. The study encourages both teachers and students to critically engage with AI-generated reformulation feedback. Future research could explore interactive features of ChatGPT and combine additional resources such as teacher guidance and cooperative writing combined with the reformulation feedback. Additionally, longitudinal studies incorporating multiple training sessions with AI-generated reformulation feedback could provide insights into its long-term effects. Overall, this dissertation contributes to a deeper understanding of the potentials and constraints of AI-generated feedback in EFL writing, paving the way for the application of GenAI in educational contexts.

Appendices

Appendix A. The Example Prompt and Resultant Feedback

Sample prompt sent to GPT-4 to generate model-based feedback

You will be a native speaker of English. Here is the writing prompt that students are given: "Do you agree or disagree that smoking should be banned in public places?". Below I will share with you the student's writing. Based on their writing, please reformulate a piece of model writing around 200-300 words. Here is the student's writing.

[The student's writing was inserted here.]

Example AI-generated feedback

Original student writing

For non-smoking people, the smoke is really threatening. Indirect smoking is more harmful for health. By the law, some public places prohibit smoking in there. But smoking people often reject this. I think it should be punished. Smoking in public places is an act of not considering other people. Many people irritate from the smoke and smell of that. Moreover, children or pregnant woman can be susceptible.

AI-generated reformulation text

I strongly advocate for the ban on smoking in public places, aligning with the concerns for public health and societal well-being. Secondhand smoke, a byproduct of smoking,

poses a significant threat to bystanders, proving even more detrimental than direct smoking in some instances. This exposure can have severe health implications, especially for vulnerable groups such as children and pregnant women, who may suffer from increased risks of respiratory diseases and other health issues.

Legal measures have been implemented in certain areas to curb smoking in public spaces, yet compliance remains a challenge. Smokers often overlook these regulations, jeopardizing the health of non-smokers and disrespecting their right to a clean and safe environment. Such disregard highlights a lack of consideration for the well-being of others, exacerbating the discomfort and health risks posed by the smoke and odor of cigarettes.

It is imperative that stringent enforcement of smoking bans in public places be prioritized, accompanied by penalties for violations. This approach not only respects the health and comfort of the majority but also contributes to a more conscious society where public health and mutual respect are valued. By implementing and strictly enforcing these bans, we can protect vulnerable populations and ensure a healthier, more considerate community for all.

Appendix B. Programming Codes for Calculating Text Similarity

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import docx2txt


def read_docx(file_path):
    """
    Read the content of a DOCX file and return it as a string.

    Parameters:
    file_path (str): The path to the DOCX file.

    Returns:
    str: The content of the file.
    """
    return docx2txt.process(file_path)


def calculate_text_similarity(text1, text2):
    """
    Calculate the cosine similarity between two texts using TF-IDF vectorization.

    Parameters:
    text1 (str): The first text to compare.
    text2 (str): The second text to compare.

    Returns:
    float: The cosine similarity score between text1 and text2.
    """
    # Step 1: Vectorization - Convert texts to TF-IDF vectors
    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform([text1, text2])

    # Step 2: Calculate Cosine Similarity
    similarity_matrix = cosine_similarity(tfidf_matrix[0:1], tfidf_matrix[1:2])

    # Extract the similarity score from the matrix and return it
    similarity_score = similarity_matrix[0][0]
    return similarity_score
```

```
# Paths to the DOCX files
file_path1 = ‘‘
file_path2 = ‘‘

# Read the texts from the DOCX files
text1 = read_docx(file_path1)
text2 = read_docx(file_path2)

# Calculate similarity
similarity = calculate_text_similarity(text1, text2)
print(f"Similarity rate: {similarity}")
```

Appendix C. Curec Approval

Research Ethics Approval

Research Title: Effects of AI-generated Reformulation Text as a Form of Feedback in EFL Writing

Research Ethics Reference: EDUC_C1A_24_136

The above application has been considered on behalf of the Education Departmental Research Ethics Committee (DREC) in accordance with the University's procedures for ethical approval of all research involving human participants.

I am pleased to confirm that, on the basis of the information provided to the DREC, ethics approval has now been granted for this study.

Please note the following:

- **Personal data**: It is the responsibility of the PI to ensure that all personal data collected during the project is managed in accordance with the University's guidance and legal requirements.
- **In-person activities**: Any data collection involving in-person interactions with participants must have an up-to-date fieldwork risk assessment in place; further guidance is available from the Safety Office's website.
- **Amendments**: Please notify the committee if you intend to make any amendments to the information in your ethics application as submitted at date of this approval, as all changes must receive ethical approval prior to implementation. The amendment form is available on the SSH IDREC webpage.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments you might have to staff.curec@education.ox.ac.uk / student.curec@education.ox.ac.uk or ethics@socsci.ox.ac.uk.

Yours sincerely,

Dr Faidra Faitaki, AFHEA
Departmental Lecturer in Applied Linguistics
DREC Member

References

Adams, R. (2003). L2 output, reformulation and noticing: Implications for IL development. *Language Teaching Research, 7*, 347–76.

Allwright, R. I., Woodley, M. P. and Allwright, J. M. (1988). "Investigating reformulationas a practical strategy for the teaching of academic writing". *Applied Linguistics 9*.236-256.

Azimian, E., Rouhi, A., & Jafarigohar, M. (2023). Written languaging, reformulation and EFL learners' writing accuracy. *Porta Linguarum Revista Interuniversitaria de Didáctica de Las Lenguas Extranjeras*, *40*, Article 40. https://doi.org/10.30827/portalin.vi40.23981

Cai, Z. G., Duan, X., Haslett, D. A., Wang, S., & Pickering, M. J. (2024). Do large language models resemble humans in language use? *arXiv.Org*. https://doi.org/10.48550/arxiv.2303.08014

Chen, Y.-S., & Wu, H.-J. (2022). Developing Sustainable Email Pragmatic Competence for EFL Learners through Reformulation. *Sustainability*, *14*, 16868. https://doi.org/10.3390/su142416868

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). Opinion | Noam Chomsky: The False Promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Cohen, A. D. (1983). Reformulating second-language compositions: A potential source of input for the learner. Paper presented at the annual convention of Teachers of English to Speakers of Other Languages, Toronto, Canada. March 15-20. (ERIC ED 228 866)

Cohen, A. D. (1989). Reformulation: A technique for providing advanced feedback in writing. *Guidelines: A Periodical for Classroom Language Teachers, 11* (2), 1 – 9.

Coyle, Y., & de Larios, J. R. (2014). Exploring the role played by error correction and models on children's reported noticing and output production in a L2 writing task. *Studies in Second Language Acquisition, 36*(3), 451-485.

Coyle, Y., Guirao, J. C., & de Larios, J. R. (2018). Identifying the trajectories of young EFL learners across multi-stage writing and feedback processing tasks with model texts. *Journal of Second Language Writing, 42*, 25-43.

Coyle, Y., Férez Mora, P. A., & Solís Becerra, J. (2020). Improving reference cohesion in young EFL learners' collaboratively written narratives: Is there a role for reformulation? *System*, *94*, 102333. https://doi.org/10.1016/j.system.2020.102333

DuBose, J., & Marshall, D. (2023). AI in academic writing: Tool or invader. *Public Services Quarterly*, *19*(2), Article 2. https://doi.org/10.1080/15228959.2023.2185338

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), Article 1. https://doi.org/10.1186/s41239-023-00425-2

Fan, J., Fang, L., Wu, J., Guo, Y., & Dai, Q. (2020). From Brain Science to Artificial Intelligence. *Engineering*, *6*(3), Article 3. https://doi.org/10.1016/j.eng.2019.11.012

Florio, S., & Clark, C. M. (1982). The functions of writing in an elementary classroom. Research in the Teaching of English, 16(2), 115–130.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digital Medicine, 6(1), 75.

Gebhard, J. G. (2002). *Teaching English as a foreign or second language.* Michigan: University of Michigan Press

Giray, Louie. (2023). "Prompt Engineering with ChatGPT: A Guide for Academic Writers." *Annals of Biomedical Engineering 51*, No. 12 (December 2023): 2629–33. https://doi.org/10.1007/s10439-023-03272-4.

Godfroid, A., Boers, F., & Housen, A. (2013). AN EYE FOR WORDS: Gauging the Role of Attention in Incidental L2 Vocabulary Acquisition by Means of Eye-Tracking. *Studies in Second Language Acquisition*, *35*(3), 483–517. https://doi.org/10.1017/S0272263113000119

Hanaoka, O. (2006). Noticing from models and reformulations: A case study of two Japanese EFL learners. *Sophia Linguistica: Working Papers in Linguistics, 54*, 167–192.

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research, 11*, 459–79. https://doi.org/10.1177/1362168810375369.

Hanauer, D. I., Sheridan, C. L., & Englander, K. (2019). Linguistic Injustice in the Writing of Research Articles in English as a Second Language: Data From Taiwanese and Mexican Researchers. *Written Communication*, *36*(1), 136–154. https://doi.org/10.1177/0741088318804821

Hsu, Liwei (2023). "EFL Learners' Self-Determination and Acceptance of LMOOCs: The UTAUT Model." *Computer Assisted Language Learning* 36, no. 7 (September 3, 2023): 1177–1205. https://doi.org/10.1080/09588221.2021.1976210.

Hyland, F., & Hyland, K. (Eds.). (2006). Contexts and issues in feedback on L2 writing: An introduction. In *Feedback in Second Language Writing: Contexts and Issues* (pp. 1–20). Cambridge University Press. https://doi.org/10.1017/CBO9781139524742.003

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, *15*(4), Article 4. https://doi.org/10.30935/cedtech/13605

Kang, E. Y. (2020). Using model texts as a form of feedback in L2 writing. *System*, *89*, 102196. https://doi.org/10.1016/j.system.2019.102196

Kang, E. Y. (2024). Model-based feedback for L2 writing revision: The role of vocabulary size and language aptitude. *International Journal of Applied Linguistics*, *34*(1), Article 1. https://doi.org/10.1111/ijal.12480

Kim, H. R., & Bowles, M. (2019). How Deeply Do Second Language Learners Process Written Corrective Feedback? Insights Gained From Think-Alouds. *TESOL Quarterly*, *53*(4), 913–938. https://doi.org/10.1002/tesq.522

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Pergamon Press Inc.

Lapkin, S., Swain, M., & Smith, M. (2002). Reformulation and the learning of French pronominal verbs in a Canadian French immersion context. *Mod. Lang. J. 86*, 485–507.

Lázaro-Ibarrola, A. (2009). Reformulation and self-correction: Testing the validity of correction strategies in the classroom. *Revista Española de Lingüística Aplicada*, *22*, 189–216.

Lázaro-Ibarrola, A. (2013). Reformulation and Self-correction: Insights into correction strategies for EFL writing in a school context. *Revista Española de Lingüística Aplicada, ISSN 0213-2028, Vol. 22, 2009, Pags. 189-216*, *22*.

Lázaro-Ibarrola, A. (2023). Model texts in collaborative and individual writing among EFL children: Noticing, incorporations, and draft quality. *International Review of Applied Linguistics in Language Teaching*, *61*(2), Article 2. https://doi.org/10.1515/iral-2020-0160

Lee, I. (1997). "ESL learners' performance in error correction in writing: Some implications for teaching". *System, 25*: 465-477.

Levenston, E.A. (1978). Error analysis of free composition: the Theory and the practice. *Indian Journal of Applied Linguistics, 4*, 1-11.

Lim, F.V. & Phua, J., (2019). Teaching writing with language feedback technology. *Comput. Compos. 54*, 102518.

Milla, R., & García Mayo, M. del P. (2024). Collaborative writing, written corrective feedback and motivation among child EFL learners. *Porta Linguarum Revista Interuniversitaria de Didáctica de Las Lenguas Extranjeras*, 27–43. https://doi.org/10.30827/portalin.vi41.23971

Morrison, R. (2023). *A metadata 'Watermark' could be the solution to ChatGPT plagiarism fears.* TechMonitor.

Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of Artificial Intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, *7*(5), Article 5. https://doi.org/10.1016/j.heliyon.2021.e07014

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

Qi, D. S., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing*, *10*(4), 277–303. https://doi.org/10.1016/S1060-3743(01)00046-7

Rahimi, M., & Zhang, L.J., (2018). Writing task complexity, students' motivational beliefs, anxiety and their writing production in English as a second language. *Read. Writ. 32* (3), 761–786.

Radanliev, P. (2024). Artificial intelligence: Reflecting on the past and looking towards the next paradigm shift. *Journal of Experimental & Theoretical Artificial Intelligence*, 1–18. https://doi.org/10.1080/0952813X.2024.2323042

Ranalli, J. & Yamashita, T. (2022) Automated written corrective feedback: Error-correction performance and timing of delivery. *Language Learning & Technology, 26*(1): 1–25. http://hdl.handle.net/10125/73465

Sachs, R., & Polio, C. (2007). LEARNERS' USES OF TWO TYPES OF WRITTEN FEEDBACK ON A L2 WRITING REVISION TASK. *Studies in Second Language Acquisition*, *29*(1), 67–100. https://doi.org/10.1017/S0272263107070039

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

Shankland, S. (2023). Why the ChatGPT AI Chatbot is blowing everybody's mind. *CNET.* https://www.cnet.com/tech/computing/why-the-chatgpt-ai-chatbot-is-blowing-everybodys-mind/.

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, 1–23. https://doi.org/10.1017/S0958344023000265

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*, 38–62.

Storey, V. A. (2023). AI Technology and Academic Writing: Knowing and Mastering the "Craft Skills". *International Journal of Adult Education and Technology*, *14*(1), Article 1. https://doi.org/10.4018/IJAET.325795

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

Sulistyo, T., & Heriyawati, D. F. (2017). Reformulation, text modeling, and the development of EFL academic writing. *Journal on English as a Foreign Language*, *7*(1), 1. https://doi.org/10.23971/jefl.v7i1.457

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

Swain, M., & Lapkin, S. (2002). 'Talking- it- through': Two French immersion learners' responses to reformulation. *International Journal of Educational Research, 37*, 285-304.

Thi, N. K., Nikolov, M., & Simon, K. (2023). Higher-proficiency students' engagement with and uptake of teacher and Grammarly feedback in an EFL writing course. *Innovation in Language Learning and Teaching*, *17*(3), 690–705. https://doi.org/10.1080/17501229.2022.2122476

Tocalli-Beller, A., & Swain, M. (2005). Reformulation: The cognitive conflict and L2 learning it generates. *International Journal of Applied Linguistics*, *15*(1), 5–28. https://doi.org/10.1111/j.1473-4192.2005.00078.x

Tsai, C.-Y., Lin, Y.-T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12722-y

UNESCO (2024). *Guidance for generative AI in education and research—UNESCO Digital Library*. (n.d.). Retrieved 17 June 2024, from https://unesdoc.unesco.org/ark:/48223/pf0000386693?posInSet=2&queryId=6dbb80ed-eab7-4a48-8a73-4398a7631f17

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge. MA: Harvard University Press.

Yang, L. & Zhang, L. (2010). Exploring the role of reformulations and a model text in EFL students' writing performance. *Language Teaching Research*, *14*(4), 464–484. https://doi.org/10.1177/1362168810375369