REPUBLIC OF TURKIYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGES EDUCATION
ENGLISH LANGUAGE EDUCATION PROGRAM

MASTER'S THESIS

THE IMPACT OF CHATGPT USE IN EFL WRITING CLASSES: A CASE STUDY

ERDEM TALHA KOÇ
22708002

SUPERVISOR
ASST. PROF. DR. GÜLÜMSER EFEOĞLU

2025

REPUBLIC OF TURKIYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGES EDUCATION
ENGLISH LANGUAGE EDUCATION PROGRAM

MASTER'S THESIS

THE IMPACT OF CHATGPT USE IN EFL WRITING CLASSES: A CASE STUDY

ERDEM TALHA KOÇ
22708002
ORCID NO: 0000-0003-1078-2846

SUPERVISOR
ASST. PROF. DR. GÜLÜMSER EFEOĞLU

APRIL, 2025

Erdem Talha KOÇ tarafından hazırlanan “The Impact of ChatGPT Use in EFL Writing Classes: A Case Study’’ başlıklı çalışma, **21/04/2025** tarihinde yapılan savunma sınavı sonucunda oybirliği ile başarılı bulunmuş ve jürimiz tarafından Yabancı Diller Eğitimi Ana Bilim Dalı İngiliz Dili Eğitimi Programında **YÜKSEK LİSANS** tezi olarak kabul edilmiştir.

Danışman	**İmza**
Dr. Öğr. Üyesi Gülümser Efeoğlu	……

Jüri Üyeleri	**İmza**
Dr. Öğr. Üyesi Semin Kazazoğlu	……
Dr. Öğr. Üyesi Manolya Sağlam	……

ABSTRACT

THE IMPACT OF CHATGPT USE IN EFL WRITING CLASSES: A CASE STUDY

Technological advancements have affected education throughout the history and ELT methodologies kept evolving by using what technology has offered. For the past few decades, the researchers have been looking for ways to integrate the computers, and AI into L2 writing classrooms. The current study investigated how ChatGPT could be utilized for L2 writing skill development, given the limited research comparing it to the human-rater feedback. The current study is a mixed method design with 51 participants from a state university in Istanbul. The participants were randomly assigned into experimental and control groups and received feedback for the essays they wrote. Experimental group received the feedback provided by ChatGPT, meanwhile the control group received the feedback provided by the researcher. None of the participants were informed about the involvement of the ChatGPT until the focus group interview, which was conducted at the end of the study. Both qualitative and quantitative data were collected using questionnaire, demographic form, focus group interview, English proficiency test and pre-post writing proficiency test. Pre-post writing proficiency test results revealed that, although there was a significant difference for the experimental group in terms of score difference between pretest and post-test scores, there was no significant difference between groups in terms of post-test scores. The focus group interview revealed that the participants could not distinguish the feedback provided by ChatGPT from the feedback provided by the researcher and believed the feedback provided by ChatGPT was decent. However, some participants preferred human-rater feedback over ChatGPT feedback. The study concludes that, although the feedback provided ChatGPT is indistinguishable from the feedback provided by a human rater, due to how students may feel towards the extended use of it, it should be used moderately. Also, it can be employed in L2 writing classrooms as an interactive supporting tool to help the instructors. These findings, contribute to the research on ChatGPT based feedback and its role in L2 writing skill development.

Keywords: ChatGPT, Automated Writing Evaluation, L2 Writing

ÖZET

CHATGPT KULLANIMININ YABANCI DİL OLARAK İNGİLİZCE YAZMA SINIFLARINDA ETKİSİ: BIR DURUM ÇALIŞMASI

Teknolojide yaşanan gelişmeler eğitimi tarih boyunca etkilemiş ve İngiliz dili öğretim metodolojileri, teknolojinin sunduğu yenilikleri kullanarak devamlı gelişmiştir. Son yıllarda araştırmacılar, bilgisayarları vey apay zekayı ikinci dil olarak İngilizce yazma sınıflarına entegre etmenin yollarını aramaktadır. Bu çalışma, insan değerlendiri geri bildirimi ile ChatGPT geri bildirimi doğrudan kıyaslayan çalışmaların sayısının azlığı dikkate alınarak, ChatGPT'nin ikinci dil yazma becerisinin gelişiminde nasıl kullanılabileceğini araştırmıştır. Çalışma, İstanbul'daki bir devlet üniversitesinden 51 katılımcı ile yürütülen karma yöntem araştırma deseni çalışmasıdır. Katılımcılar rastgele deney ve kontrol grouplarına atanmış ve yazdıkları kompozisyonlara geri bildirim almışlardır. Deney grubu, ChatGPT tarafından sağlanan geri bildirimi alırken, kontrol grubu araştırmacı tarafından sağlanan geri bildirimi almıştır. Ancak hiçbir katılmcıyla, çalışmanın sonunda gerçekleştirilen odak grup görüşmesine kadar bu bilgi paylaşılmamıştır. Çalışmada anket, demografik bilgi formu, odak grup görüşmesi, İngilizce yeterlilik testi ve ön test son test yazma yeterlilik testleri uygulanarak hem nicel hem nitel veri toplanmıştır. Ön test son test yazma yeterlilik testi sonuçları, deney grubu içerisinde ön test son test arasında anlamlı bir puan farkı olduğunu ancak son test puanları açısından gruplar arasında anlamlı bir fark bulunmadığını göstermiştir. Odak grup görüşmesi sonuçları, katılımcıların ChatGPT tarafından sağlanan geri bildirimi, araştırmacı tarafından sağlanan geri bildirimden ayırt edemediklerini ve yeterli bulduklarını göstermiştir. Ancak, bazı katılımcılar insan değerlendirici geri bildirimini ChatGPT'ye tercih ettiklerini ifade etmişlerdir. Çalışma, ChatGPT tarafından sağlanan geri bildirimin insan değerlendirici tarafından sağlanan geri bildirimden ayırt edilememesine rağmen, öğrencilerin bu tür teknolojik araçların aşırı kullanımına yönelik tutumları göz önünde bulundurularak, bu tür araçların ölçülü bir şekilde kullanılmasının uygun olacağını ortaya koymaktadır. Ayrıca, ikinci dil yazma sınıflarında öğretmenlere destekleyici ve eklileşimli bir araç olarak kullanılabileceği sonucuna varılmıştır. Bu bulgular, ChatGPT tabanlı geribildirimin ikinci dilde yazma becerilerinin gelişimindeki rolüne ilişkin literatüre katkı sağlamaktadır.

Anahtar Kelimeler: ChatGPT, Otomatik Yazı Değerlendirme, İkinci Dilde Yazma

ACKNOWLEDGEMENTS

Although it took longer than expected due to unforeseen circumstances, here we are.

First of all, I am deeply grateful to my advisor Gülümser EFEOĞLU. Her never-ending support and timely feedback throughout this painstakingly long process made it possible for me to complete it. I do not think I could have finished it without her guidance. I feel privileged to have worked with her. Words cannot express how thankful I am.

I want to thank Meral GÜMÜŞ and Ömer Faruk Coşkun for their academic insights, and unwavering belief in me that I would be able to complete this journey.

Finally, I would like to thank every single participant who took the time to be a part of this study. Without their willingness to share their experiences, this thesis would not have been possible.

Erdem Talha KOÇ

April, 2025; İstanbul

TABLE OF CONTENTS

ABSTRACT iii
ÖZET iv
ACKNOWLEDGEMENTS v
TABLE OF CONTENTS vi
LIST OF TABLES viii
LIST OF FIGURES ix
LIST OF ABBREVIATIONS x

1. INTRODUCTION 1
1.1. Background of the Study 1
1.2. Statement of the Problem 2
1.3. Purpose of the Study 3
1.4. Significance of the Study 3
1.5. Research Questions 3
1.6. Key Terms 4

2. LITERATURE REVIEW 5
2.1. L2 Writing Skill Development 5
2.1.1. Contrastive Rhetoric 5
2.1.2. Cognitive Models of Composing 6
2.1.2.1. The Flower and Hayes Model 6
2.1.2.2. Bereiter and Scardamalia's Model (1987) 6
2.1.3. Genre Theories 7
2.1.3.1. Systemic Functional Linguistics 7
2.1.3.2. English for Specific Purposes 7
2.1.3.3. Rhetorical Genre Theory 7
2.1.4. Sociocultural Theory 8
2.2. L2 Writing Assessment 8
2.2.1. Formative and Summative Assessment 9
2.2.2. Use of Technology 10
2.2.2.1. Automated Essay Scoring / Automated Writing Evaluation 11
2.2.3. ChatGPT in ELT 15
2.3. Feedback 17
2.3.1. Feedback Types 17

3. METHODOLOGY 20
3.1. Research Design 20
3.2. Setting 21
3.3. Participants 21
3.4. Data Collection Instruments 22
3.4.1. Demographic Form 23
3.4.2. Oxford Quick Placement Test 23
3.4.3. Pre-Post Writing Proficiency Test 23

3.4.4. Writing Tasks (Intervention) 24
3.4.5. ChatGPT 4 24
3.4.6. Analytic Rubric 25
3.4.7. Feedback Effectiveness Questionnaire 26
3.4.8. Focus Group Interview 27
3.5. Data Collection Procedure 28
3.6. Data Analysis 30

4. RESULTS 31
4.1. Quantitative Data Analysis Results 31
4.1.1. Results of Demographic Form 31
4.1.2. Results of the Oxford Quick Placement Test 35
4.1.3. Results of the Pre-Post Writing Proficiency Test 36
4.1.4. Results of the Feedback Effectiveness Questionnaire 39
4.2. Qualitative Data Analysis Results 43
4.2.1. Results of the Demographic Form 43
4.2.2. Results of the Focus Group Interview 45
4.2.3. Results of the Feedback Effectiveness Questionnaire 55

5. DISCUSSION AND CONCLUSION 57
5.1. Discussion of the Research Questions 57
5.1.1. The Effect of Human and ChatGPT Feedback on L2 Writing Skill Development 57
5.1.1.1. Pre-Post Writing Proficiency Test 57
5.1.2. EFL Learners' Opinion on Receiving Feedback from ChatGPT 59
5.1.2.1. Feedback Effectiveness Questionnaire 59
5.1.2.2. Focus Group Interview 60
5.2. Conclusion 63
5.2.1. Limitations of the Current Study 64
5.2.2. Pedagogical Implications 65
5.2.3. Recommendations for Further Research 66

REFERENCES 68
APPENDICES 78

LIST OF TABLES

Table 1. Descriptive Statistics of the Participants 22
Table 2. Analytic Rubric 26
Table 3. Data Collection Timeline 29
Table 4. Participant Background 32
Table 5. Participants' Self-Perceived Proficiency in English Writing 34
Table 6. Descriptive Statistics of the Placement Test 35
Table 7. Descriptive Statistics for Pre-Test Scores of Both Groups 36
Table 8. Results of the Wilcoxon Signed Rank Test for the Pre-Test and Post-Test Scores of Experimental Group 37
Table 9. Results of the Wilcoxon Signed Rank Test for the Pre-Test and Post-Test Scores of Control Group 37
Table 10. Descriptive Statistics for the Post-Test Scores of Both Groups 37
Table 11. Mann-Whitney U Test to Compare the Post-Test Scores of the Experimental and the Control Groups 38
Table 12. Descriptive Statistics for Score Increase of Both Groups 38
Table 13. Mann-Whitney U Test to Compare the Achievement of the Experimental and the Control Groups 39
Table 14. Engagement with FB2 40
Table 15. Perception of FB 1's Effectiveness 41
Table 16. Perceived Necessity of FB2 42
Table 17. Understanding and Influence of FB1 42
Table 18. Analysis of Perceived Difficulties During Writing 44
Table 19. Participants' Preference of Additional Sources 45
Table 20. Participants' Preference of Digital Tools 45
Table 21. Thematic Analysis of the Focus Group Interview 46

LIST OF FIGURES

Figure 1. Research Design 21

LIST OF ABBREVIATIONS

AI: Artificial Intelligence
EFL: English as a Foreign Language
ESL: English as a Second Language
AWE: Automated Writing Evaluation
AES: Automated Essay Scoring
CALL: Computer Assisted Language Learning
MALL: Mobile Assisted Language Learning
LLM: Large Language Model
ELT: English Language Teaching
L2: Second Language
ZPD: Zone of Proximal Development

1. INTRODUCTION

1.1. Background of the Study

Writing is a productive skill with numerous theories trying to explain the underlying processes to understand how it can be developed. One thing that is common among these theories is the importance of practice. However, practice alone is not enough for improvement. Feedback has been regarded vital for improving writing skills (Hyland & Hyland, 2006).

The feedback, which learners will receive, can be given in many forms. It can be given orally, it can be written, it can be given by peers, or even by a computer. Depending on the purpose of it, the feedback can be summative or formative.

Regardless of the form, in traditional classroom settings the feedback is mostly provided by the teacher/instructor. However, providing feedback is time-consuming. To tackle the time limitations, technology has been integrated into writing process.

The idea of using technology for feedback to save up classroom time is not new. Starting from the 60s with PEG (Warschauer & Grimes, 2008), educators have been trying to integrate automatized feedback into the writing process. Over the years many companies have come up with numerous tools to provide learners with feedback. These tools can be categorized into two main categories. The first one is AES or automated essay scoring. These tools mainly designed to provide feedback to the final product. The feedback comes in the form of a score. The score these tools give is based on how similar the writing is to the database these tools were trained on (Warschauer & Grimes, 2008; Liu & Kunnan, 2016). E-rater by ETS, Intelligent Essay Assessor by Person, My Access by Vantage Learning are some of the AES tools. The second category is AWE or automated writing evaluation. These tools, similar to AES tools, provide feedback based on writing. However, unlike the feedback provided by the AES tools, AWE tools focus on aspects of writing such as grammar, coherence and sometimes content. These tools help learners to see and correct their mistakes, thus

improving their writing skills. MY Access, Grammarly, IntelliMetric. and Criterion are some of the popular AWE tools (Barrot, 2024).

However, due to rapid technological advancements, the line between the AWE and AES tools are getting blurry as AWE tools are getting scoring engines and AES tools are giving feedback on errors. Thus, the terms AWE and AES will be used interchangeably from now on.

1.2. Statement of the Problem

Although AWE and AES tools were developed to provide learners with adequate feedback, thus decreasing the workload of the instructors/teachers (Grimes &Warschauer, 2010), there are still a few questions about their effectiveness.

First problem is the fact that most of these tools were designed in countries where English is the native language. As these tools are marketed in countries where English is a foreign language, they were not designed with ESL or EFL learners in mind (Warschauer & Ware, 2006). This leads to raise of questions such as whether these tools are really effective in L2 classroom settings as they were designed using data acquired from native speakers.

Second problem is the reliability of the studies. A good portion of the existing studies on AWE/AES is either funded by the developers developing these tools or conducted directly by these developers (Liu & Kunnan, 2016). Even if these tools were to be given the benefit of the doubt and assumed that the studies conducted or funded by the developers are reliable, learners show mixed reactions to use of these tools (Huang & Renandya, 2018).

Third problem is reliability of the feedback. Even though studies revealed that these tools are good at scoring essays and can be used for high stakes writing assessments (Shermis, 2014), when scoring and providing feedback to these essays, these tools compare the aforementioned essays to authentic materials written by native speakers of the language, which leads them to overlook some grammar errors and provide vague, confusing feedback for non-native speakers of English.

Last problem is the actual usefulness of the feedback. Although learners may have direct access to both corrective and metalinguistic feedback provided by these tools,

learners may opt to copying the corrective feedback only, without gaining any long term L2 writing development (Ranalli, 2021).

1.3. Purpose of the Study

Although the researchers have been working on these technological tools since 1970s, with the educational institutions switching to online education due to the global pandemic, the integration and the use of tools such as these into education only accelerated. As stated above, most of the available Automated Writing Evaluation and Automated Essay Scoring tools are developed in countries where English is the native language. As a result of the using data from native speakers, the score and the feedback these tools provide are based on natives of English. Current study aims to test whether an artificial intelligence language model, such as ChatGPT which was trained and still training using the data available on the internet could be utilized as an alternative to existing AWE/AES tools and the feedback they provide.

1.4. Significance of the Study

This study will have two main contributions to the field. First contribution will be investigating whether a Large Language Model such as ChatGPT can be utilized as an AWE/AES tool to provided L2 writing feedback. The second contribution, in most of the studies regarding the use of AWE/AES tools, the learners know the involvement of the said tools. However, in this study, the source of the feedback, which was ChatGPT, was hidden from the participants until the end of the study.

1.5. Research Questions

In the light of the information provided above, the current study aims to answer the following questions;

1- Does the feedback provided by a human rater or an AWE have any effect on L2 writing skill development?
2- How do EFL learners perceive receiving feedback from ChatGPT?

1.6. Key Terms

Large Language Model: A large language model is a type of AI that is trained on vast amount of data to understand and generate human like and contextually appropriate text (Bonner et al., 2023).

Automated Writing Evaluation: Also referred as Automated Essay Scoring (Warschauer & Ware, 2006), were tools initially designed to rate high stakes exams. Nevertheless, they revised to be able to provide feedback in addition to scores (Stevenson & Phakiti, 2014).

Feedback: An important part of the education process which provides learners with comparison of their performance against the expected outcomes (Schartel, 2012)

Formative Assessment: Kind of assessment that occurs during the learning process to support learning (Vonderwell et al., 2007).

Summative Assessment: A process intended to warrant learner achievement, usually through high-stakes judgements like grades or other performance indicators (Knight, 2002).

2. LITERATURE REVIEW

2.1. L2 Writing Skill Development

Humans are social beings. As a result, they communicate in different ways to convey their feelings. Writing is one of the forms that human beings can use. However, writing is not a skill humans naturally acquire; it has to be developed (Grabe & Kaplan, 1996). There are billions of people on earth and many of them have different backgrounds (nationality, literacy rate, motivations, reasons to write). Naturally, no single theory can account for all of these variables. Hence, there are multiple theories trying to explain what underlies L2 writing skill. There are four main theories; contrastive rhetoric, cognitive models of composing, genre theories, and sociocultural theory.

2.1.1. Contrastive Rhetoric

Proposed by Robert Kaplan back in 1966, contrastive rhetoric focuses on the influence L1 and culture has on the L2 writing. After analyzing 700 compositions written by foreign university students from different countries and cultural background, Kaplan (1966) identified three cultural groups: Semitic, Oriental and Romance. After comparing these compositions to the composition written by L1 English users, Kaplan concluded that rhetoric, which is a mode of thinking and organizing information, differs across different cultures, influencing how individuals analyze, interpret, and synthesize information.

In his article, Kaplan (1966) shared examples from the compositions he had analyzed and concluded that these cultural differences in rhetorical structures can lead to misunderstandings in writing when students write in English. For example, instead of using linear constructions like English, Semitic (Arabic and Hebrew) writing tends to use parallel constructions and coordinate structures, while Oriental (Chinese, Japanese, Korean…) writing often uses a circular or inductive approach, instead of direct approach, which is commonly used in English. Understanding these differences is vital for educators, as it allows them to better support students in adapting their writing to the expectations of the target language. This includes not only teaching

grammar and vocabulary but also helping students recognize and practice the logical and organizational patterns that are characteristic of writing in English.

2.1.2. Cognitive Models of Composing

Research into the cognitive processes of composing focuses on the that occurs in the mind and behaviors involved in writing (Cumming, 2016), mainly on the individual's cognition during writing tasks, to aim model specific aspects of L2 composing, such as planning or word choice. There are two notable models proposed:

2.1.2.1. The Flower and Hayes Model

According to traditional linear stage models, writing is a sequential process consisting of pre-writing, writing and re-writing. However, Flower and Hayes (1981) argue that rather than a linear process, writing process consists of different sub-processes which occur at any given time. These sub-processes like planning, translating, and reviewing, are hierarchically organized and can be embedded within one another.

The model also focuses on the role of the task environment, which includes the assignment or the writing task and the text the writer writes, the writer's long-term memory, where the writer's knowledge about the topic and audience is stored and tapped into when needed, task processes including planning (generating and organizing ideas), reviewing(evaluating and refining), translating (writing the ideas), and the monitor, which is an important element of the model (overseeing and coordinating the writing processes). According to Flower and Hayes (1981) as the writing is a goal-directed activity, the writer continuously monitors, revises and reviews as writing.

2.1.2.2. Bereiter and Scardamalia's Model (1987)

Bereiter and Scardamalia (1987) argue that there are two kinds of writing, first one is similar to how humans naturally know how to walk, however they have to practice a lot to get good at walking. Knowledge-telling model, which is the first model, is used by new writers. In knowledge-telling model, the writer makes use of what is already available to him/her about the given topic. The model relies on what is already learned through social experience and only uses that knowledge, hence limited by that. Knowledge-transforming model, the second model, on the other hand, used by more-experienced writers, involves setting a goal and working towards it through a problem-

solving approach, which includes organizing and reorganizing their knowledge about the topic and the composition's content.

2.1.3. Genre Theories

Genre theories aim to promote L2 writers' understanding of the communicative aspects and purposes of a written text through authentic sources (Harman, 2013). Genre theories posit that texts genres are not static textual forms but dynamic, socially embedded tools that reflects values of communities (Tardy, 2012). They guide how writers convey their feelings across specific communities to meet their audience in academic or professional context. According to the genre theories, learning to write requires mastering these genres (Tardy, 2012). There are three sub-categories of genre theories which are the following:

2.1.3.1. Systemic Functional Linguistics

Introduced by Halliday, systemic functional linguistics argues that the way people use language changes depending on the time, and purpose of the speaker or writer (Butt et al., 2000). Language is viewed as a set of systems, which lets users to make choices to create meanings (Hyland, 2004). SFL argues that explicit language awareness is the key for learning to write (Hyland, 2004), which can be increased through discourse and linguistic awareness (Huang et al., 2017).

2.1.3.2. English for Specific Purposes

English for Specific Purposes (or ESP), focuses on helping learners to communicate effectively in specific academic or professional contexts. ESP writing instruction focuses on teaching learners how to participate effectively in these contexts, writing skills that learners learn are directly related to their specific goals and needs (Hyland, 2013).

2.1.3.3. Rhetorical Genre Theory

Rhetorical Genre Theory argues that, rather than being a fixed set of rules and features, genres are ‘’typified rhetorical actions’’ that occur as a result of recurring social actions (Miller, 1984), which means that genres are, in sum, what people use them to accomplish in a community. As for its application in L2 writing, in addition to mastering grammatical and lexical features of the target language, it is also required to

gain formal knowledge (text organization and language use), rhetorical knowledge (awareness of the audience), process knowledge (effective writing strategies) and content knowledge (knowledge about the topic) (Tardy, 2012).

2.1.4. Sociocultural Theory

Sociocultural theory is a framework which explains how human consciousness and abilities develop through culturally organized activities, tools, and concepts (Cumming, 2016). It is based on Vygotsky's (1978) principals of Zone of Proximal Development (ZPD), scaffolding, and mediation. ZPD refers to the difference between the learner's current level and what the learner can achieve with the help of the others (Marginson & Dang, 2016). Scaffolding refers to the support given to the learner by the more knowledgeable one/facilitator during the learning process (Malik, 2017). Finally, mediation refers to the process by which cultural tools, language, and social interactions facilitate and shape cognitive development and learning (Cumming, 2016).

In sum, every single one of these theories are important as they provide a basis for understanding how L2 writing skill develops over time. For example, cognitive models explain the mental processes involved in composing, which allows more effective instruction practices to be developed. In addition to shaping the teaching practices, these theories also influence assessment practices.

2.2. L2 Writing Assessment

Assessment can be defined as interpretation of the evidence of knowledge (Green, 2022). Assessment helps institutions make fair placement or admission decisions, certifies a certain degree of knowledge or language proficiency, and in terms L2, gives insights about the learning progress of the learner (Lee, 2016; Li & Li, 2022). In terms of L2 writing, assessment has been dominated by short, product focused tasks that assess the learner's ability to produce a meaningful text under a limited time (Cumming, 2002). After assessing the text put out by learners based on the pre-set criteria and getting the results, the results can be used for generating scores (summative assessment) or to promote deeper learning by providing individualized feedback for the learner (formative assessment) (Lee, 2017).

2.2.1. Formative and Summative Assessment

Formative assessment and summative assessment are two notable types of assessment and they serve different purposes. Summative assessment, or assessment of learning, aims to learning achievement of learners at a specific time period by comparing their performance against the pre-determined criteria and using the scores obtained to assess their learning outcomes (Ahmad, 2020). It is often employed for administrative purposes such as assigning grades, passing a class etc. (Genesee & Upshur, 1996). Formative assessment or Assessment for learning, described by Black et al. (2004) as ‘’An assessment activity can help learning if it provides information to be used as feedback... Such assessment becomes ‘‘formative assessment’’ when the evidence is actually used to adapt the teaching work to meet learning needs.’’ (p., 3). Summative assessment is better to test whether the learner has achieved the pre-determined learning outcomes meanwhile formative assessment is better to see the learning gaps (Llamas-Nistal et al., 2013).

In terms of L2 writing, formative assessment focuses on helping the learner through constructive feedback which is provided at multiple stages of writing. On the other hand, summative assessment focuses on assigning a final grade to the completed product (Lee, 2011). Formative assessment is more beneficial for the learners as it offers multiple chances to revise the written product and the feedback the learners receive can raise their motivation. Meanwhile, summative assessment offers an easy way to assess the what have the learner has learned based on the preset criteria, which may make learners focus on the grade, rather than improving their writing skill (Lee 2011; Burner, 2015).

As for the reasons they are employed for and when they are employed; summative assessment is employed to see what the learners have learned up until a certain point in time, which may be at the end of a unit, term, or a high-stakes exam, which helps to track the learners’ progress, and offers a formal proof of learner outcomes, which may be used for future instruction planning (Hall & Sheehy, 2018 ; Torres, 2019). As for the formative assessment; it is used for supporting learning as it happens, it gives instructors a chance to adjust their teaching based on what learners need and offers a chance for learners to be more involved in their learning process (Hargreaves et al., 2018).

In sum, in terms of L2 writing, summative assessment is the score the learner receives after submitting an essay, while formative assessment is the feedback the learner receives from the teacher or the peers during different stages of writing such as pre-writing, during-writing and post-writing. If the assessment can be used by the learner to improve her/his performance it is considered formative, if not, it is summative (Huot, 2002).

2.2.2. Use of Technology

Computer Assisted Language Learning, or CALL, aims to facilitate the language teaching process by integrating technology, started in the 1960s and kept evolving with technology (Warschauer & Healey, 1998). The way computers was employed also reflected the pedagogy of the period. For example, during the behaviorist CALL (1960-1970s), the computer programs were essentially drill tutors that focused on grammar and vocab through exercises. Communicative CALL, rose in the 1980s, with the focus on the actual use of grammar and vocab, and lastly with integrative CALL, which started in 1990s, programs started blending four language skills (Warschauer & Healey, 1998). Around the same time, widespread internet and increased power of the computers led to Intelligent CALL, with the help of AI and algorithms, the programs started to adapt materials to the needs of the learners to offer a personalized learning experience (Schulze & Heift, 2013). In addition to these developments, the gap between computers and mobile devices such as mobile phones and tables got closed in terms of possible use cases due to technological improvements. That led to the Mobile Assisted Language Learning, or MALL, which made it possible for learners to learn at their own pace (Kukulska-Hulme, 2012).

As for how these technological developments used for L2 writing, these technological improvements appeared in classrooms through synchronous and a-synchronous writing activities such as writing activities through e-mails, social platforms and wikis which boosted learner motivation (Warschauer & Healey, 1998). In addition to boosting motivation, computers programs also increase fluency and accuracy in writing (Jafarian et al., 2012). In one study, an Automated Writing Evaluation (AWE) tool, Grammarly, was employed in L2 writing classes to see how it would impact the learners in terms of L2 writing proficiency. The results revealed that overall L2 writing

proficiency of the learners had increased, especially for low proficiency learners (Wei et al., 2023).

The idea of using computers for evaluating writing skills, however, started back in 1970s. Automated Essay Evaluation (AEE) can be defined as the use of computers in the evaluation process of written products (Shermis & Burstein, 2003). The first attempt to use computers in the writing evaluation process started with Page's Project Essay Grader in 1973 (Shermis et al., 2013). Page, to save the teachers from the burden of grading papers, which was the biggest obstacle preventing students from writing more, hence improving their writing skills, designed a computer program with help from his colleagues (Shermis et al., 2013). To put simply how it worked, PEG classified the words, calculated the word counts and gave a score. In the end, although it performed better than some raters in terms of inter-rater reliability, due to computers were expensive and not widespread, it did not gain any popularity.

2.2.2.1. Automated Essay Scoring / Automated Writing Evaluation

Automated Essay Scoring (AES) tools are tools which are designed to assign scores to written products (Stevenson, 2016). Automated Writing Evaluation (AWE) tools on the other hand, are the tools designed to give feedback on the quality of the writing (Cotos, 2014). While AWE tools focus on the feedback, AES tools focus on assigning scores consistently based on predetermined criteria, which is why they are used at some of the high-stake exams. Although these tools have different goals, the way they work is similar. They use natural language processing (NLP), artificial intelligence (AI), and machine learning to analyze written texts based on previously set criteria such as grammar, syntax, mechanics etc. (Cotos, 2014). As AWE tools combine scoring and formative feedback in a single system (McNamara et al., 2015; Grimes & Warschauer, 2010), the terms AWE and AES will be used interchangeably from now on.

Although there are many AWE tools; Criterion, developed by ETS, WriteToLearn, developed by Pearson Education, and MyAccess, developed by Vantage Access and Grammarly can be given as some of the popular AWE tools available on the market.

2.2.2.1.1. Advantages of AWE/AES Tools

One of the advantages of the use of AWE tools in classroom is that it saves time. These tools save the overburdened teachers by saving them from the burden of feedback.

Warschauer and Grimes (2008) observed four different schools which were using AWE tools, two of them MyAccess, other two, Criterion. Interviews with the teachers and the principals of schools revealed that, all of them considered these tools to be time-savers for the teacher. One of the teachers said 'it is a second pair of eyes watching over the classroom'. In addition, these tools keep track of the learner progress by storing each draft the learner submits, which, in turn, creates a form of e-portfolio and gives learner a chance to track his/her progress meanwhile gives the teacher a clear picture of learner growth, his/her strong and weak points in terms of writing skill development (Shadiev & Feng, 2023; Ersanlı & Akman Yesilel, 2022).

These tools, apart from increasing the learner motivation to revise by providing immediate feedback (Khasawneh, 2024), they also increase learner autonomy by giving them a chance to practice without the need of a teacher (Steinhart, 2001). In 2010, Grimes and Warschauer observed 2 schools which were AWE tool MyAccess. Observations, interviews with the students and the surveys revealed that, although students did not take the feedback MyAccess provided as serious as the feedback their teacher provides, they were still happy to receive instant feedback, revised their essays more and were motivated to write more. The teachers in this study also reported that these tools also helped with classroom management.

In another study, two AWE tools, namely Grammarly, and Grammark, were used by 24 undergraduate students (Parra & Calero, 2024). The results of the pre-test post-test design revealed that both tools significantly enhanced students' writing performace, especially in grammar, punctuation and style. No significant difference was found between the tools' effectiveness. Students expressed that they liked friendly interface of the tools and the immediate error correction they offered. The researchers concluded that although these tools may improve autonomy, and save time for the teachers, the teacher feedback still required and these tools should be used as supplementary tools.

In addition, AWE tools are reliable. As a criterion for reliability, researchers compared the scores of human raters and these tools, and found them to be reliable (Attali, 2013). Shermis and Hamner (2013), after collecting 22029 student essays and then comparing the nine AES engines based on the scores given to the essays by these AES engines concluded that, they are now on par with human raters in terms of scoring like a human rater. Thanks to the reliability, these tools are even used for high stakes exams such as Graduate Record Exam (GRE), Test of English as a Foreign Language (TOEFL)

(Williamson et al., 2012). The use of these tools for such high-stake exams not only lowers the amount of time required to evaluate by providing rapid and reliable scoring, it also lowers the cost considering how labor intensive the process is (Weigle, 2013; Ersanlı & Akman Yeşilel, 2022).

In their extensive literature, Nunes et al. (2022) review where they reviewed AWE studies conducted between 2000 and 2020 that examined the effectiveness of these systems in schools (Grades 1-12), they found that these tools were especially effective in improving at least one aspect of student's writing when integrated with teacher feedback. In the studies they reviewed, students appreciated immediate and non-judgmental feedback they received and increased their engagement with writing. Teachers, on the other hand, found AWE systems but said that low-proficiency learners still needed help to make use of the feedback. The literature review concludes that, these tools are promising when used in addition to teacher feedback.

In conclusion, AWE tools are useful for supporting the instruction of writing by increasing learner autonomy and motivation through immediate and non-judgmental feedback they provide. Receiving feedback from an AWE tool also decreases learner anxiety. In addition, these systems provide instant and reliable scoring which saves time for the teachers and give them opportunity to focus on other aspects of the language. However, their efficacy is greatest when they are used in combination with teacher feedback and when teachers help learners with understanding and using the feedback they provide.

2.2.2.1.2. Criticism Against AWE/AES Tools

One big criticism about AWE tools is that the considerate number of existing studies on AWE tools were conducted either by researchers affiliated with the developers of these tools or directly conducted by the developers (Stevenson, 2016; Liu & Kunnan, 2016). In his article 'when the state of the art is counting words', Perelman (2014) challenged assertation made by Shermis and Hamner's study (2013) which after nine AES engines scored more than 22000 essays concluded that these tools were on par with human raters. Using data presented in Shermis and Hamner's study and two additional sources, Perelman concluded that AES tools prioritized essay length consistently more than human raters. He argues that although word count is not the only criterion these tools use to score essays, but these tools also assume that good essays should be long, thus give higher scores to longer essays.

These tools, although they were initially developed in countries which English is the native/second language, are being marketed in countries where English is foreign language. As the data these tools were trained on collected from native speakers, these tools may not be able to detect some of the non-native errors or the feedback. For example, most of the errors native speakers commit are related to syntax errors (Gamon et al., 2013) However, in addition to syntax errors, non-native speakers commit many usage errors. As the data these tools were trained on gathered from the native speakers, the feedback these tools give may not be suitable for non-native speakers as they may fail to detect some of the errors (Liu & Kunnan, 2016; Chen & Cheng, 2008). Liu and Kunnan (2016), wanted to test the scoring ability of WriteToLearn on Chinese undergraduate English majors' essays. They compared the scores given by WriteToLearn and those of four human raters. Although it was consistent in terms of scoring, it did not meet the minimum threshold of precision (90%) for it to be considered reliable. It had difficulty detecting the errors made by Chinese undergraduates. In addition, when they contacted the developers about the seven essays WriteToLearn could not score, they learned that it cannot score nor give feedback if the essay is not similar enough to the ones the machine was trained on. In addition, these tools fail to address cultural nuances and focus on superficial aspects of a text such as its length (Khasawneh, 2024).

Although these tools may get high scores in terms of reliability in terms of assigning scores close to human-raters, this does not make them valid (Attali, 2013). As these tools use predetermined parameters to check and give a score or feedback similar to a human rater, it is possible to fool these tools with nonsensical essays. Powers et al. (2002) wanted to test the limits of the e-rater. They gathered 67 essays which were written to ''challenge'' the e-rater. One of the essays were basically a few paragraphs copy-pasted 37 times, and got the highest possible score from e-rater.

A different take on the AWE tools comes from Herrington and Moran as early as 2001. They claim that these AWE tools are marketed as solutions to save teachers from the burden of reading papers. However, 'the burden of reading papers' is not a problem. It is the symptom which leads to the real problem: overcrowded classrooms. Using these tools would only ensure that the problems persist. They also say that putting AWE tools in charge means they (the staff of the institution) would no longer decide what it means to be ''proficient''.

To summarize the advantages and disadvantages of the AWE tools, in their recent literature review Shi and Aryadoust (2023) concluded although these tools can score rapidly, consistently and lessen the workload of the teachers, they can still be tricked into providing high scores by increasing word count, still have problems with deeper aspects of writing such as creativity or cohesion and lastly, how well these tools can score certain subgroups such as L2 learners is still questioned.

2.2.3. ChatGPT in ELT

ChatGPT is a Large Language Model (LLM), designed to generate human-like text, developed by OpenAI and publicly released in November 30, 2022. Ever since its launch it became a hot-topic in the field of ELT as it offered endless possibilities. As it is a rather new topic, there are limited number of studies on it. High amount of the existing studies focuses on how it could be utilized in the ELT field (Kotska & Toncelli, 2023), teachers' perspective of it (Mabuan, 2024), and how it can possibly assist the teachers. A number of studies (Yan, 2023; Zou & Huang, 2024), however, focus on how it can be utilized in terms of L2 writing or students perception of it.

For example, a researcher (Al-khresheh, 2024) investigated how English teachers from different parts of the world perceived the ChatGPT. He reached out to 46 teachers from 39 different countries with various educational qualifications (PhD, Master's, Bachelor's), academic ranks (professors, lecturers, teachers) and with varying lengths of teaching experience. The participants agreed that ChatGPT has a potential to enhance autonomy and language learning by offering personalized support, immediate feedback, and interaction. However, they also expressed concerns about possible overreliance on it may hinder creativity and critical thinking skills of the learners. In addition, its limitations in some skills (speaking and listening), some challenges related to cultural differences were also among the concerns. In the end, they viewed it as a valuable support tool, rather than an alternative to traditional teaching methods.

One study that tried to see the ChatGPT's potential from the L2 perspective focused on its assessment potential. Researchers (Pfau et al., 2023) wanted to see how it would compare to a human-rater, in terms of L2 writing. ChatGPT 4.0 was given 100 essays written by L2 learners from various levels and asked to identify and classify the errors. These essays were also rated by a human-rater. The results revealed that linguistic accuracy wise, ChatGPT was on par with human raters however had problems with

error recall. In addition to that, it was less reliable for rating the essays written by less proficient learners of English. The researchers concluded that despite the limitations, it offered a great potential as a supplementary tool.

Meniado et al. (2024) explored the use of ChatGPT for second language writing among EFL learners in Thailand and Vietnam. 357 students from three different institutions: two universities, one is from Thailand and one is from Vietnam, and a high school from Vietnam joined the study. Throughout a semester, all of these students were shown ways to integrate ChatGPT into writing process by their teachers and used ChatGPT for pre-writing, during-writing and post-writing processes. At the end of the semester, a questionnaire and semi-structured interview were used as means of data collection. The results revealed that the students perceived ChatGPT to be useful for enhancing their writing skills.

A similar study, however, with a much smaller group of participants (Yan & Zhang, 2024), focused on how four Chinese undergraduate learners with varying language proficiency and technological competence would interact with ChatGPT-generated feedback. During their writing course, participants interacted with ChatGPT to receive feedback, and revised drafts for five weeks. Journals, observation of the students in the classroom and interviews were among the data collection tools. The data analysis revealed that students with higher language proficiency and better technological skills used different prompt writing strategies and had better revision quality. Low proficiency learners struggled to integrate ChatGPT feedback. Overall, students liked ChatGPT's non-judgmental tone, felt less embarrassment for receiving negative feedback, at least compared to receiving negative feedback from a teacher, and were willing to use it in the future. The study concluded ChatGPT has potential as an interactive tool however teacher scaffolding is still needed.

In sum, it can be said that although the use of ChatGPT for L2 writing is a recent concept in the field, the number of studies and different approaches to see how it can be best applied already showed that ChatGPT can help learners with different stages of writing, help them self-practice and can help improve their grammar related problems (Su et al., 2023; Yan, 2023).

2.3. Feedback

One of the earliest and the most prominent definition of feedback is ‘’ information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way.’’ (Ramaprasad, 1983, p. 4). In terms of education, although there have been many feedback theories trying to define the word feedback and the underlying processes, what all theories and their various definitions have in common is that apart from them deeming feedback as an essential part of improvement, it is a dynamic process which aims to close the gap between the current level of the learner and the desired level (Lipnevich & Panadero, 2021).

In terms of writing skills, feedback can be defined as the input from the readers that helps the writers to improve their work (Keh, 1990). Feedback has been regarded as vital for writing skills development (Hyland & Hyland, 2006).

2.3.1. Feedback Types

The feedback can be provided by different sources in different forms. It can be provided by the teacher, the peer, or even by a machine. It can be given orally, or written (can be written on a paper or can be sent via an e-mail). It can be analytic, focusing on different aspects of the text such as content, organization or grammar, or holistic, offers overall feedback. It can be direct, offering explicit correction or be indirect, hinting learners the errors without correcting them (Srichanyachon, 2012). It can be provided during a writing conference, giving teacher and the student a chance to discuss it one-on-one (Goshu & Gebremariam, 2024). It can be metalinguistic, instead of correcting the grammar error, teacher could just mark it and write what type of error it is (Nguyen & Chu, 2024). It can be formative or summative. Summative feedback, as the name implies, is given at the end of the writing process. It is given to the final product, and it is usually given as a score to evaluate the quality of the writing, formative feedback, on the other hand, is given throughout the writing process, starting from drafting to final product and offers learners many chances to revise their product (Lee, 2017). It can be focused, providing feedback only on a certain aspect of the text, or it can be unfocused, providing feedback on every aspect of the text (Ellis et al., 2008). It can even be automated and be provided by AWE tools. There are many ways to offer feedback, the type of feedback to be offered should be chosen based on the developmental needs of the learners (Kepner, 1991).

For example, although indirect feedback is considered to be more useful in the long term L2 writing skill development as it fosters critical thinking and language awareness (Miceli, 2006), direct feedback can also be used when providing feedback to a lower-proficiency level learner, or when correcting complex errors or untreatable errors such as idiom errors (Ferris, 2006). For the feedback to be effective, it should aim to close the gap between the current level and the desired level of the learner (Lee, 2017).

However, the learners' current proficiency, or the desired level for the learner is not the only criteria when it comes to choosing appropriate type of feedback for the learner. For example, Saito (1994) investigated the teachers' feedback practices and the preferences of the learners. For the study, the participants included 39 ESL learners in Toronto, enrolled in two intensive English courses and a university-level engineering writing class. Participants were from different cultural backgrounds and the results revealed that participants from teacher-centered/Eastern cultures preferred the feedback provided by their teachers, rather than their peers. In a different study which aimed to see the impact of teacher feedback and peer feedback, Tsui and Ng (2000), tested how secondary L2 learners reacted to the combination of peer and teacher feedback. The results revealed that while most of the students preferred teacher feedback as it was more reliable, and it was coming from an authority some of them preferred peer feedback as it was seen as useful for raising awareness, and it increased collaborative learning.

Zhang and Hyland (2018) conducted a study in which two Chinese students, Flora and David, with different level of English proficiency engaged with teacher feedback and the feedback provided by PigAI, an AWE tool. They analyzed how students engaged with the both feedback for a semester and found that highly motivated, also high proficient student, Flora, preferred AWE feedback as it was immediate, it made it possible to send multiple drafts and receive feedback for each one and it increased learner autonomy. Flora also stated that teacher feedback was quite detailed and it increased her motivation. On the other hand, the other student, David, who was less proficient preferred explicit teacher feedback as he revealed that he struggled with indirect feedback. In addition, he did not like AWE feedback as he received low scores. The study concluded that the student engagement, which is shaped by preference,

motivation and proficiency of the learners, is crucial when it comes to feedback types and how effectively they are used.

Another example would be Liu's (2008) study where 12 university level ESL learners, were provided direct and indirect feedback. The results revealed that although both types of feedback reduced errors in drafts, direct feedback was found to be more effective for revisions indirect feedback led to greater improvement in the long term. When asked, students expressed that they would prefer a combination of the two types of feedback.

Time constraints may also force instructors to rely on certain type of feedback. For example, a case study (Junqueira & Payant, 2015) which explored the feedback beliefs and practices of a novice teacher, Kim, who, then was during her first semester of teaching English composition course to ESL students. Two interviews, her journal and the essays she commented on revealed that although she wanted to focus on global concerns such as content and organization, she had to focus on local concerns such as grammar as providing comprehensive feedback for each essay took almost an hour. As for the reason of her focus on the local concerns, she revealed that practical constraints such as time pressures and workload led her to rely on explicit feedback. Other studies with wider range of participants also confirm the workload, lack of training and time constraints create problems for instructors (Ferris et al., 2011; Hyland, 2013).

In conclusion, there are many theories that try to explain what underlies writing skill development. In addition to explaining the fundamentals for the skill development, these theories also come up with methods/techniques to teach writing better. One thing these theories and their methods/techniques have in common is the value they posit in practice. However, the practice requires feedback for it to be actually useful. There are many ways and forms to offer feedback. However, one of the biggest problems with providing feedback, at least from an instructor's perspective, is that it is quite time consuming. Although we have developed AWE tools to overcome this problem, the quality and effectiveness of the feedback these tools provide is still questionable. The latest developments in technology such as the existence of LLMs led to the rise of the questions about whether these LLM's can be employed as AWE tools in foreign language teaching.

3. METHODOLOGY

In the following section, the design of the study, procedure and methods to collect and analyze data are presented in detail. Initially, research design is presented, following with setting and participants which then followed by the data collection tools.

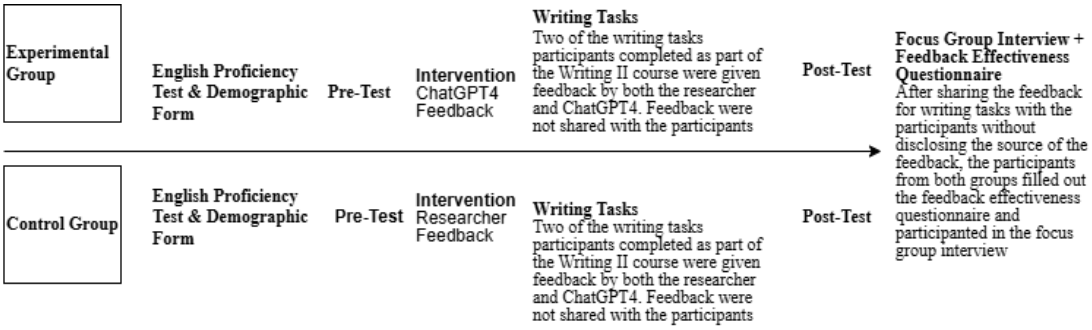
3.1. Research Design

Current study is an embedded mixed design study, which was conducted in natural classroom settings. Mixed method study was chosen as one of the research questions required qualitative data to be collected while the other question required quantitative data to offer a broader and better definition of the problem (Almeida, 2018).

One key procedure of the research design was assigning participants who joined the study were to either control or experiment group and comparing their performance against each other, which makes this study a randomized controlled trail (RCT) (Plano Clark & Creswell, 2015). The experiment group received the feedback provided by ChatGPT4, meanwhile control group received the feedback provided by the researcher. Randomized controlled trail was chosen as it helps to control the confounding variables which may affect the results, which makes it a good option to decide whether a treatment is successful (Plano Clark & Creswell, 2015). In addition to pre-test post-test method, Likert scale questions were also employed to gather quantitative data.

Qualitative part of the study, on the other hand, consisted of collecting qualitative data via open-ended questions and focus group interview. Thematic analysis was employed to detect certain patterns in data (Plano Clark & Creswell, 2015). The research design can be seen in Figure 1.

Figure 1. Research Design



3.2. Setting

The study took place during the spring term of the 2022-2023 academic year and was administrated in the English Language Teaching department of a state university in Istanbul, Turkiye. At the beginning of the semester, freshmen ELT students were informed about the nature of the study, a study related to L2 writing which would require them to write essays and receive feedback from the researcher. The involvement of ChatGPT4 had to be kept from the participants as a result of the nature of the study.

3.3. Participants

51 participants, who were also high proficient, based on the results of the English proficiency test, EFL learners from a state university in Istanbul participated in the study. 35 of them were female and 16 of them were male. Their native language was Turkish. Participants were randomly assigned to experiment or control group. The detailed information about the participants can be seen on the table 1. Participants of the experimental group were coded as GPT, meanwhile participants of the control group were coded as HR (human rater) and they were assigned numbers starting from one.

Table 1. Descriptive Statistics of the Participants

Category	Details	n
Number	Participants	51
Gender	Male	16
	Female	36
Age	18	19
	19	16
	20	12
	21	2
	22	1
	24	1
Native Language	Turkish	50
	Unspecified	1
Years of Studying English	1-3 years	13
	3-5 years	3
	More than 5 years	35

3.4. Data Collection Instruments

Current study adopted a mixed-method design. Thus, a combination of qualitative and quantitative data collection administrated. It is worth nothing that after the earthquake which struck Turkey on 6th February 2023, Higher Education Council declared that the spring semester would continue online across the country. Consequently, the tools administrated online. The data collection tools included; Demographic form, Oxford Quick Placement Test, Feedback Effectiveness Questionnaire, Focus group interview and Pre-Post Writing Proficiency Test.

3.4.1. Demographic Form

A demographic form which consisted of open ended, multiple choice and 13 5-point Likert scale items which aimed to acquire deeper knowledge about participants perception about their own proficiency in writing skills was administrated at the beginning of the study. Open ended questions aimed to gain insight on the English level and background, and the digital tool knowledge of the participants. Likert scale questions, on the other hand; aimed at detecting the participants perceived opinion on their writing skills.

3.4.2. Oxford Quick Placement Test

Prior to the study, to ensure all participants were at the same or close proficiency levels, Oxford Quick Placement Test (*Quick Placement Test*: Version1, 2004) was chosen. Pen and paper version of Oxford Quick Placement Test which assesses reading, vocabulary and grammar includes 60 multiple choice questions (Oxford University Press, 2001), was administrated through Google Forms and participants were given 30 minutes to answer the questions. To ensure they would start at the same time and would not take more than that, the test was shared during a class. According to the test manual, the number of correct answers between 48-54 indicates a proficiency level of C1. The results revealed that the English level of the group was C1 (M=52.73, SD= 5.63018).

3.4.3. Pre-Post Writing Proficiency Test

The Pre-Post Writing Proficiency Test was administrated at the beginning of the study, which was also the 2nd week of the term, and at the end of the study, which was also the 11th week of the term, to compare the overall performance of the experimental and control groups.

The pre-test which was designed by the researcher after investigating IELTS writing tasks from various IELTS preparation books (Cole, 2012; Cambridge University Press & Assessment, 2024), was administrated at the beginning of the term. Similar to IELTS writing tasks, the participants were supposed to write at least 250 words about the given topic in 40 minutes. When it comes to L2 writing, topic familiarity is important. It has been found that learners perform better in terms of lexical complexity when they are required to write about the topics, they are familiar with, rather than

they are not (Yang & Kim, 2020; Yoon, 2017). The familiarity of the topic was also confirmed by the instructor.

The task was the following:

> Some people believe that formal education is the only way to become successful. Others believe that success comes from experience and practical skills.
>
> Discuss both views and give your opinion.

Another argumentative essay prompt was given as the post-test writing task.

> Some people believe that technology has made our lives easier and more convenient, while others think that it has made us more dependent and less social.
>
> Discuss both views and give your opinion.

3.4.4. Writing Tasks (Intervention)

After inspecting the syllabus of Writing 2 course and discussions with the instructor of the course, the researcher decided to use one of the essay topics as the first task and then use the two other essays which students were going to write throughout the semester as the second and the third tasks. Here is the task;

Write a cause-and-effect essay that examines the effects of social media on mental health. Discuss the ways in which social media use can impact mental health and provide examples of how individuals, communities, and governments can address this issue.

3.4.5. ChatGPT 4

ChatGPT 4 is a large language model developed by OpenAI. Its main premise lies in its ability to ‘’*understand’’* the prompts provided by humans and responds accordingly.

The time ChatGPT 4 became available coincides with the onset of the study. Both versions of the ChatGPT was given random essays and the following prompt;

> Provide these essays feedback as if a teacher was writing feedback to his/her student. Make sure you base your feedback on the following criteria; introduction and thesis, organization and structure, evidence and support, grammar, spelling, mechanics and finally conclusion and overall impressions

After reviewing the feedback provided by both versions, the researcher decided to proceed the study with ChatGPT 4 as the answers provided by it deemed more human-like by the researcher.

Before receiving the feedback from ChatGPT it was asked the following;

Let's say I want you to give feedback to essays which will be provided by me. What prompt should I use to get the best results?

ChatGPT provided the following answer;

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

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

By following this prompt, you can ensure that I will provide feedback tailored to the specific essay and student while maintaining a suitable tone for your communication.

Most of the feedback provided by ChatGPT were based on the five criteria that was previously determined. The five criteria were; Introduction, structure and coherence, spelling and grammar, supporting evidence and the conclusion. These criteria were determined based on Jacob's et al.'s (1981) scoring rubric, with the difference of ''content'' being split into introduction, support and conclusion. Yet, it forgot the five pre-determined criteria sometimes and had to be reminded them and rewrite the feedback.

3.4.6. Analytic Rubric

An analytic rubric (Table 2), which consisted of five criteria, which are; introduction and thesis, organization and structure, evidence and support, grammar, spelling, mechanics and, conclusion and overall impressions were designed. The analytic rubric was designed as analytic rubrics are good at separating students across a wide range of criteria compared to holistic rubrics (Wiseman, 2012).

Table 2. Analytic Rubric

Criteria	1	2	3	4	5
Introduction and Thesis					
Organization and Structure					
Evidence and Support					
Grammar, spelling and Mechanics					
Conclusion and Overall Impressions					

To provide feedback to the members of the control group, the researchers individually rated the essays based on the five criteria. Afterwards, average score for each criterion was given as the final score. If the average score contained a decimal, the average score was rounded up to next whole number. Further notes added to address significant issues with writing such as obvious errors when it was deemed necessary by the researchers.

Pre-post writing proficiency tests were rater by the researcher. However, both researchers rated the writing tasks. To ensure inter rater reliability, before scoring the essays, both researchers scored some essays together and agreed on the details.

For the first task, there were 19 participants in control group. Hence, rather than using Pearson correlation, which requires at least 20 cases (Pallant, 2020) or even more as claimed by others (Bonett & Wright, 2000), Spearman Correlation was used. The result was (p = .834) for the control group. The value indicates a strong positive correlation between the raters (p = .834, $p < .05$).

For the second task, Spearman Correlation was employed once more to calculate the inter-rater reliability, and it was (p = .857), which, again, indicates a strong positive correlation.

For the third task, spearman correlation was used to calculate the inter-rater reliability and it was (p = .893), again indicating a strong positive correlation between the raters.

3.4.7. Feedback Effectiveness Questionnaire

A second questionnaire which offered a combination of open-ended questions and 16 6-point Likert-scale questions about the feedback provided by the researchers and the feedback provided by ChatGPT, without disclosing the source of the feedback, was

administrated before the focus group interview. The questionnaire was inspired from a study which focused on the comparison of the feedback provided by an AWE to feedback provided by humans (Huang & Renandya, 2018). The questionnaire designed for that study offered a combination of items to gain deeper insight on learner perception of the automated feedback and feedback provided by humans, which is also one of goals of the current study. Due to these similarities, items from that study were adopted. For the questionnaire they had used, the researchers grouped the items into four groups and reached the following reliability score for each one was above (α = 0.74). The items were 6-point as unlike the usual 5-point scales, as the extra point increases the discrimination rate, thus discriminates better (Chomeya, 2010). To further confirm the reliability, a reliability test was used on the feedback effectiveness questionnaire and the results indicated a very high reliability (α = 0.988).

3.4.8. Focus Group Interview

Focus group interview, which lasted 90 minutes, was conducted right after the participants completed the feedback effectiveness questionnaire. Every single participant who took a part in the study was invited for the focus group interview and participants joined one of the two focus group interview groups based on their preference. First group had 14 participants, meanwhile the second group had six participants. In total, 20 participants were present for the focus group interview, out of 20 participants, 13 of them were from the control group, meanwhile 7 of them were from the experimental group. The interview began with the revealing of the source of the Written feedback, which was ChatGPT4. Afterwards, 13 questions were posed to get the opinions of participants on the feedback types, human versus AI feedback, feedback form and their emotional reactions to learn that the feedback was provided by ChatGPT4. To analyze the data acquired through focus group interview, thematic analysis was employed. Thematic analysis was chosen to analyze the data as it offers flexibility in identifying, analyzing and reporting patterns within and across the data (Clarke & Braun, 2017).

3.5. Data Collection Procedure

The study was initiated on January 1, 2023, after receiving approval from the Institutional Review Board (see Appendix 1). In addition to the approval from the institution, to ensure privacy, ethical forms were handed out to participants.

After informing the participants about the study, which they were told that they would join a study, write essays and receive feedback. The participants who were willing to join the study filled out the demographic form which was administrated in order to gain a deeper understanding of their self-perceived writing skills, writing habits and knowledge of the available digital tools. Most of them were using translation apps such as Google Translate and DeepL, most of them were aware of and using the Grammarly to correct their Grammar and punctuation errors. When asked about the tools they knew about, only one of them mentioned ChatGPT.

Participants who filled out the demographic forms continued with the Oxford Quick Placement Test which included 60 multiple choice questions focusing on vocabulary, grammar and reading. After the completion the placement test, a Google Classroom was set up and participants were asked to join. On the fourth week of the study, the first writing task was shared with the participants via Google Classroom. The participants received feedback based on their group and the feedback was shared with participants as private comments to ensure the participants would not see the feedback the other participants had received.

In addition to the first task which the participant had completed on the 4th of the study, on 4th and 8th weeks, participants were required to submit essays for their writing class. These essays were assigned by the instructor of the class, and with her permission, used as the second and the third tasks, which they would receive feedback. Both ChatGPT4 feedback and the human-rater feedback was given to these essays. However, the feedback was not shared until the 12th week, right before the onset of the focus group interview.

The post test was conducted on the 11th week, after the completion of the 3rd task. The next week, the researcher, with the permission of the instructor of the class, joined the second half of the class and shared the feedback, provided by both the human-raters and the ChatGPT 4, for the second and third tasks. Participants were given a few minutes to analyze the feedback.

After the participants analyzed both types of feedback, the researcher shared the feedback effectiveness questionnaire which consisted of 16 6- point Likert scale questions about both types of feedback which participants had just investigated and three open ended questions about feedback.

With the completion of the questionnaire, the researcher continued with the focus group interview. Focus group interview was chosen as it capitalizes on the group dynamics to reveal new perspectives on the topic at hand (Guest et al., 2017). The interactions occur during focus group interview helps the researcher to collect additional data that would not be available otherwise (Agar & Macdonald, 1995). The timeline of the study can be seen on the table 3.

Table 3. Data Collection Timeline

Data Collection Tool		**Timeline**
Demographic form		2nd week of the term
English Proficiency Test		2nd week of the term
Pre-writing proficiency test		3rd week of the term
Intervention (first task and two in-class tasks)	a-Cause and effect essay - b- Argumentative essay c- Process essay	4th-11th weeks of the term
Post-writing proficiency test		11th week of the term
Feedback effectiveness questionnaire		12th week of the term
Participants reviewed the feedback for the second and the third tasks and were informed about		12th week of the term

Data Collection Tool	Timeline
the involvement of the ChatGPT	
Focus Group Interview	12th week of the term

3.6. Data Analysis

Due to the nature of the study, both qualitative and quantitative data were collected throughout the study. Quick Placement Test, Pre-Post Writing Proficiency Test were used to quantitative data. Focus group interview and open-ended items inside the feedback effectiveness questionnaire were used to gather qualitative data.

To analyze the quantitative data, SPSS version 27 was used. As for analyzing the qualitative data obtained from focus group interview and the open-ended questions, thematic analysis, following Braun and Clarke's (2006) six phase coding framework, which includes the steps of familiarization with the data, generating the codes, turning codes into themes, revision of the themes, determining the significance of the themes and reporting the findings, was employed.

4. RESULTS

This chapter aims to present the results of the collected qualitative and quantitative data which were collected to answer whether the feedback provided by ChatGPT4 would have any effect on the writing performance of the participants, and the opinions of participants about receiving feedback from it.

Quantitative data was collected through pre and post writing tests, Demographic Form, Feedback Effectiveness Questionnaire. Qualitative data was collected through focus group interview and open-ended questions included in Demographics Form and Feedback Effectiveness Questionnaire.

Here are the research questions which will be answered throughout this chapter:

1- Does the feedback provided by a human rater or an AWE have any effect on L2 writing skill development?
2- How do EFL learners perceive receiving feedback from ChatGPT?

4.1. Quantitative Data Analysis Results

Data acquired from the quantitative part of the demographic form, Oxford Quick Placement Test, 6-point Likert scale items of the feedback effectiveness questionnaire, pre-post writing proficiency test and statistics of the data acquired from focus group interview constitute the quantitative data. For quantitative data analysis, The Statistical Package for Social Sciences (SPSS), Version 27 was employed. The SPSS version 27 was used for descriptive statistics, for reliability tests and to determine the effectiveness of the treatment. To test the effectiveness of the treatment, two-tailed t test was employed.

4.1.1. Results of Demographic Form

To gain deeper insights on the participants' background and their relation with writing skill, the demographic form was employed. The demographic form revealed that the majority of the participants were female with 35 (68%), 47 participants were between

18-20 years old (92%) and 50 of them were native Turkish speakers (98%). 35 participants (68,6%) had spent more than five years studying English, 29 participants had learned English in school (56,9%) followed by 12 self-taught (23,5%).

In terms of involvement with writing; 13 (25,5%) of them wrote frequently 29 (56,9%) of them wrote occasionally. As for the kinds of writing activity they were engaged in, academic writing was the prominent (58,8%). Only 18 participants (35,3%) had taken a course on writing and only nine participants (17,6%) had worked with a tutor or a writing center. As for the use of technology to help them write in English, only one participant (1,9 %) stated that he/she never used technology to help with writing. Detailed information about the participants can be seen on the table 4.

Table 4. Participant Background

Category	**Details**	**n**	**%**
Gender	Male	16	68,6%
	Female	35	31,4%
Age	18-20	47	92,1%
	21-24	4	7,9%
Native Language	Turkish	50	98,04%
	Unspecified	1	1,96%
English study duration (years)	1-3	13	5,9%
	3-5	3	25,5%
	5+	35	68,6%
How did you learn English as a second language	In school	29	56,9%
	Self-Taught	12	23,5%
	Other	10	19,6%

Category	Details	n	%
Current involvement in writing English	Frequently	13	56,9%
	Occasionally	29	25,5%
	Rarely	9	17,6%
Type of writing usually engaged in (participants could choose more than one	Academic	30	52.63%
	Business	2	3.51%
	Creative	25	43.86%
Ever taken a course focused on writing	Yes	18	35,3%
	No	33	64,7%
Ever worked with a tutor/writing center	yes	9	17,6%
	No	42	82,4%

The demographic form also included 13 5-point Likert scale items which designed by the researcher and aimed to see the participants self-perceived proficiency in English Writing. The items had a good reliability with the reliability rate of (α = .727). The items revealed that, the participants were struggling with choosing the right vocabulary, using the right tone and had difficulty in organizing their thoughts and ideas before writing. The rest of the items and their results can be seen in the table 5.

Table 5. Participants' Self-Perceived Proficiency in English Writing

Item	Mean	SD	Min	Max
I feel confident in my ability to write effectively in my L2.	3.45	1.064	1	5
I feel confident using a variety of sentence structures in my writing.	3.52	1.046	1	5
I struggle with choosing the right vocabulary when writing in English.	3.05	1.255	1	5
I find it easy to organize my ideas in written English.	3.21	1.045	1	5
I feel comfortable expressing complex ideas in written English.	3.01	1.048	1	5
I feel confident in my ability to write coherent paragraphs in English.	3.50	1.046	1	5
I struggle with using appropriate tone and style in my written English.	2.58	1.151	1	5
I have difficulty organizing my thoughts and ideas before I begin writing.	2.72	1.372	1	5
I am able to adjust my writing style to suit different purposes and audiences.	3.52	1.007	2	5
I feel confident in my ability to revise and	3.72	0.960	2	5

Item	Mean	SD	Min	Max
edit my own writing in English.				
Feedback is helpful in improving my L2 writing skills.	4.41	0.753	2	5
I am able to incorporate feedback into my writing effectively.	4.00	0.894	2	5
I am comfortable with using technology and digital tools to aid in my writing process in English.	4.25	0.868	2	5

4.1.2. Results of the Oxford Quick Placement Test

The quick placement test required participants to complete 60 multiple-choice questions targeting reading, vocabulary, and grammar under 30 minutes. According to the test manual (Oxford, 2001), a score between 48-54 indicates a C1 level of proficiency. C1 level learners, in terms of writing, are expected to produce well-structured and detailed texts on various complex subjects using a wide range of vocabulary and grammar rules with occasional minor errors that do not affect understanding (Verhelst et al., 2009). Test results revealed that overall group proficiency was C1 with the mean score of (M= 52.7347), detailed statistics of the results can be seen in table 6. In addition, the reliability of test was confirmed with Cronbach's alpha with good internal consistency ($\alpha = .890$).

Table 6. Descriptive Statistics of the Placement Test

	N	Mean	SD	Min	Max
Experimental Group	25	52.2800	6.04511	40	60
Control Group	24	53.2083	5.25009	45	60
Total	49	52.7347	5.63018	40	60

4.1.3. Results of the Pre-Post Writing Proficiency Test

Before analyzing the results, the researcher had set some criteria such as, to a participant's results to be eligible for the analysis, said participant had to complete demographic form, at least two out of three tasks, placement test, and both pre and posttests. Throughout the study some participants had dropped out and the results of some participants could not be included due to previously mentioned criteria. In the end both experiment and control groups had 17 participants whose data could be used.

In the initial assessment of the data, it was observed that scores were ranging between 5-25. The post writing proficiency test scores were evaluated on a scale of 100. Descriptive statistics for Pre-Post test scores of both groups can be seen on the table 7.

Table 7. Descriptive Statistics for Pre-Test Scores of Both Groups

	N	M	SD	Min	Max
Control Group	17	83.7647	10.43749	56.00	96.00
Experimental Group	17	81.4118	10.57745	56.00	96.00

To ensure the homogeneity, the Levene's Test was employed twice, one for the pre-test scores and one for the post-test scores. The Levene's Test result for pre-test scores between groups was $F(1, 32) = 0.007$, $p = .936$., indicating there was no significant difference between the groups. The Levene's Test results for post-test scores resulted in a similar score, $F(1, 32) = 0.96$, $p = .335$, again, indicating there was no significant difference between groups.

To ensure the normal distribution of the groups, Shapiro-Wilks test was used to analyze within-group comparisons (pre-test vs. post-test) as well as between-group comparisons (experimental vs. control). The test revealed that pre-test scores for both groups were not normally distributed which might stem from the low number of participants. Thus, instead of the parametric option, paired-samples t-test, which is used to compare the means of two related groups to determine whether there is a statistically difference (Pallant, 2020), Wilcoxon Signed Rank Test was used as it does not assume normality and more robust to violations.

The Wilcoxon Signed Rank Test was administrated to see within-group differences between pre-test and post-test scores for the control and experimental groups due to the non-parametric nature of the data.

Table 8. Results of the Wilcoxon Signed Rank Test for the Pre-Test and Post-Test Scores of Experimental Group

	Post-Test Scores - Pre-Test Scores
Z	-2.152[b]
Asymp. Sig. (2-tailed)	.031

b. Based on positive ranks.

Any *p* value below *p=0*.05 indicates a significant difference. Based on the pre-post test scores of the experimental group (Table 8), the null hypothesis, which assumes the treatment had no significant effect, is rejected. Rejection of the null hypothesis indicates that the pre-post test score difference for the experimental group is statistically significant (Z = –2.152, *p* = .031) which indicates that the treatment, in this case, the feedback ChatGPT4 had provided to participants, was effective and the post-test scores of the experiment group increased significantly.

Table 9. Results of the Wilcoxon Signed Rank Test for the Pre-Test and Post-Test Scores of Control Group

	Post-Test Scores - Pre-Test Scores
Z	-.762[b]
Asymp. Sig. (2-tailed)	.446

b. Based on negative ranks.

As can be seen on the table 9, (*p=0.446)* indicates that null hypothesis, which assumes that the treatment had no effect, is not rejected. This signifies that the difference between post and pre-test scores for the control group is not significant, which indicates that the feedback provided by human-raters did not lead to any significant difference.

Table 10. Descriptive Statistics for the Post-Test Scores of Both Groups

	N	**M**	**SD**	**Min**	**Max**
Control Group	17	87.0588	7.28415	72.00	100.00
Experimental Group	17	85.1765	9.97644	68.00	100.00

Table 11. Mann-Whitney U Test to Compare the Post-Test Scores of the Experimental and the Control Groups

Total N	**34**
Mann-Whitney U	138.000
Wilcoxon W	291.000
Z	-.227
Asymptotic Sig. (2-sided test)	.821
Exact Sig. (2-sided test)	.838 [b]

b. not corrected for ties

As illustrated in the Table 10, post test scores of the control group (M= 87.0588, SD= 7.28415) is higher than the experimental group (M=85.1765, SD= 9.97644). Manny-Whitney U test was used to see whether there was a significant difference in terms of the post-test scores between the control group and the experiment group. As can be seen on the Table 11, the results revealed that there was no significant (U=138.000, *p*=.838>.05) difference between the control and the experimental groups in terms of post-test scores. Based on the p value, the null hypothesis, which assumes that there is no difference in post-tests scores between experimental and control groups, could not be rejected. As a result, the treatment did not lead to a significant difference.

However, the table 12 indicates that the score increase of experimental group seems to be higher (M=3.7647, SD= 6.85136) than the control group (M=3.2941, SD= 14.08796).

Table 12. Descriptive Statistics for Score Increase of Both Groups

	N	**M**	**SD**
Control Group	17	3.2941	14.08796
Experimental Group	17	3.7647	6.85136

Table 13. Mann-Whitney U Test to Compare the Achievement of the Experimental and the Control Groups

Total N	**34**
Mann-Whitney U	136.500
Wilcoxon W	298.500
Z	-.279
Asymptotic Sig. (2-sided test)	.780
Exact Sig. (2-sided test)	.786 [b]

b. not corrected for ties

The Mann-Whitney U test was used again, to compare the achievement of the control and experimental group and to see whether there is a significant difference between them in terms of score increase. As presented in the table 13, the results revealed that there was no significant (U=136.500, Z= -.279, p=.786>.05) difference between the control and the experimental groups in terms of achievement, which means null hypothesis, which stated that there is no difference between groups in terms of achievement, could not be rejected.

In conclusion, the analysis of the pre-post writing proficiency test both in-group and between groups revealed that, although both groups had increased their post test scores compared to the pre-test, only the experimental group's score increase was significant. However, a comparison of the post test scores between groups revealed that there was no significant difference between groups in terms of achievement. Thus, it can be concluded that although the feedback provided by ChatGPT4 had significant effect on the post-test scores of the experimental group compared to their pre-test scores, the significance was not enough to cause a statistically significant achievement.

4.1.4. Results of the Feedback Effectiveness Questionnaire

Before sharing the feedback effectiveness questionnaire with the participants, participants were given the feedback for tasks 2 and 3, which had been provided separately by both the researcher and the ChatGPT. The participants were not informed about the involvement of the ChatGPT and they reviewed both types of feedback believing that they were provided by the researcher. After viewing the both types of

feedback, the Feedback Effectiveness Questionnaire, which in total had 16 6-point Likert scale items, divided in two sections, and three open ended questions, were shared with the participants. The 16 6-point Likert scale items were divided into two sections. The first section had 8 items about the feedback provided by the human rater, which, in questionnaire was coded as FB1, The second section had the same items about the feedback provided by the ChatGPT4, which was coded as FB2.

The researcher employed Exploratory Factor Analysis using Principal Components and Varimax rotation. SPSS version 27 was used for the analysis. Factor analysis is used as a data reduction technique, enabling the simplification of a large set of variables into a smaller, more manageable number of factors by identifying clusters among the intercorrelations of the variables (Pallant, 2020).

Factory analysis revealed four factors responsible for 80.194% of variance. These factors and the items will be discussed in detail.

Table 14. Engagement with FB2

Items	N	Min	Max	Mean	SD
I can understand the feedback provided in FB2	20	4.00	6.00	5.5500	.60481
FB2 can improve the quality of this composition	20	2.00	6.00	5.2500	1.11803
The feedback in FB2 is clear	20	4.00	6.00	5.500	.68825
FB2 can help me get a higher score in future assignments	20	1.00	6.00	5.2500	1.11803
FB2 can enhance my writing performance	20	1.00	6.00	5.1000	1.25237

I feel motivated to improve my writing after receiving FB2	20	1.00	6.00	4.8000	1.36111
Knowing that FB2 is provided by instructor influences my initial effort to writing the composition	20	3.00	6.00	5.0500	.94451

The first factor is the engagement with FB2. As it can be seen on the Table 14, participants agreed that the feedback provided by the FB2 to be easy to understand (M=5.5000), clear (M=5.500), can help them improve the quality of their work (M=5.2500), can help them get higher scores in the future (M=5.2500), can enhance their writing performance (M=5.1000), can motivate them to improve their writing skills (M=4.8000), and finally, they believed that receiving it is a good source of motivation (M=5.0500).

Table 15. Perception of FB 1's Effectiveness

Items	**N**	**Min**	**Max**	**Mean**	**SD**
I can understand the feedback provided in FB1	20	3.00	6.00	5.3500	.81273
FB1 can improve the quality of this composition	20	2.00	6.00	4.9000	.96791
The feedback in FB1 is clear	20	3.00	6.00	5.1000	.91191
FB1 can help me get a higher score in	20	2.00	6.00	4.6000	.94032

future assignments					
FB1 can enhance my writing performance	20	2.00	6.00	4.5500	.99868
I feel motivated to improve my writing after receiving FB1	20	1.00	6.00	4.3000	1.30182

Second factor is the perception of the FB1's effectiveness (Table 15). Participants found the feedback provided by the FB1 to be easy to understand (M=5.3550), clear (M=5.1000), can help them improve the quality of their work (M=4.9000), can help them get higher scores in the future (M=4.6000), can enhance their writing performance (M=4.5500), and finally they believed it would motivate them to improve their writing (M=4.3000).

Table 16. Perceived Necessity of FB2

Items	N	Min	Max	Mean	SD
It is necessary to receive FB2	20	1.00	6.00	5.1000	1.44732

The third factor is the perceived necessity of Feedback 2 (Table 16). Based on the table 16, it is clear that the participants believed that receiving the ChatGPT4 feedback was more necessary (M=5.1000), compared to FB1 (Table 17).

Table 17. Understanding and Influence of FB1

Items	N	Min	Max	Mean	SD
It is necessary to receive FB1	20	2.00	6.00	4.6500	1.22582
Knowing that FB1 is provided by instructor influences my initial effort to	20	3.00	6.00	4.9000	1.02084

writing the composition

The last factor is understanding and influence of FB1 (Table 17). Even though the items under this factor did not receive scores as high as their FB2 counterparts, the participants perceived FB1 to be necessary (M=4.6500) and believed that receiving it is a good source of motivation(M=4.9000).

In summary, there were 16 6-point Likert questions, 8 for FB1 and 8 for FB2. Each item about FB2, were rated higher than its FB1 counterpart. Drawing on these results; it can be concluded that participants perceived the feedback provided by FB2, which was unbeknownst to them at the time, ChatGPT4, to be more clear, more useful and deemed it as a better motivator than the feedback the FB1, the feedback provided by human raters. One of the two human raters was an assistant professor at the same state university where the current study was conducted and the second rater was a master's student at the same university.

4.2. Qualitative Data Analysis Results

Qualitative data collected to answer the second research questions which aimed to gain deeper insights about the participants opinion on receiving feedback from ChatGPT. The methods used for collecting the qualitative data were; focus group interviews, demographic forms, and the three open-ended questions embedded in the Feedback Effectiveness Questionnaire. The data obtained from demographic form and the focus group interview were analyzed using thematic analysis and the results will be discussed in detail.

4.2.1. Results of the Demographic Form

51 participants completed the demographic form. In addition to general questions to gain knowledge about the backgrounds of the participants, the demographic form included some open-ended questions to learn about the participants' experiences with writing, and their perceptions and beliefs about writing skill.

One of the open-ended questions aimed find out what writing task was the hardest for the participants. When the participants were asked about what kind of writing task was the hardest for them, out of 51 participants, 46 of them said essays (90.196%), which followed by reports (5.88%), e mail (1.96%) and literature reviews (1.96).

Another open-ended question in the demographic form intended to explore the participants' perceptions about difficulties they had faced while writing in more detail. The answers were read multiple times and categorized for ease of analysis. The results revealed that problems related to vocabulary was the most common issue for the participants with 25 (38.45% of 51) answer mentioning that, which followed by problems related to general writing process such as the writing styles, required number of words etc. The detailed analysis of the answers can be seen on the table 18 below.

Table 18. Analysis of Perceived Difficulties During Writing

Code	**f**	**%**
Vocab	25	38.45%
Cohesion &Expression	15	23.08%
Grammar	6	9.23%
Creativity	6	9.23%
Idea Generation	4	6.15%
No problems	3	4.62%
Length &Word count	3	4.62%
Lack of Writing Experience	2	3.08%
Feedback	1	1.54%
Total	65	100%

The next open-ended question inquired the participants' preference of additional resources to improve their L2 writing skills. Results revealed that most participants used online sources, followed with textbooks. The detailed results can be seen on the table 19.

Table 19. Participants' Preference of Additional Sources

Code	f	%
Online	21	41.17
Textbook	14	27.45
Teachers at school	8	15.68
Pleasure Reading	8	15.68
Total	51	100

The last open-ended question aimed to understand the participants' preference of digital tools to help them improve L2 writing skills. The results revealed that most frequently used category of digital tool was grammar checkers. Quillbot and Grammarly were the most used grammar checkers, after grammar checkers online dictionaries and translators such as Thesaurus and Google Translate were popular. Rest of the tools employed by participants can be seen on the table 20.

Table 20. Participants' Preference of Digital Tools

Category	f	%
Grammar Checkers	23	42,59
Online Dictionaries/Translators	20	37,03
Office Tools	5	9,25
Search Engines	5	9,25
Other	1	1,85
Total	54	100

4.2.2. Results of the Focus Group Interview

Once participants completed the feedback effectiveness questionnaire, the focus group interview, which aimed to develop a deeper understanding about the participants' opinion on the use of AI feedback in writing process, started by first informing the participants about the involvement of the ChatGPT in the feedback process. After the

revelation, the researcher started the focus group interview with 20 participants who were present. The interview, in total, lasted 90 minutes. Focus group interview was chosen instead of one-on-one interviews as focus group interview as it makes it possible to collect data from multiple participants in a short amount of time, makes it possible to see whether participants share similar or different opinions on a certain topic and participant interaction leads to more detailed data (Patton, 2002).

In order to analyze the data collected via the focus group thematic analysis was employed. Thematic data analysis includes the following steps; recording, transcription of the records, reading and taking notes based on the transcription, generating codes, collecting data related to codes, exploring themes, collecting codes under themes, defining and naming themes, and keep repeating the process as new themes emerge (Braun & Clarke, 2006; Patton, 2002).

90-minute long focus group interview was transcribed verbatim, which resulted in a transcription around 2900 words. After the completion of the transcription, the transcription was read multiple times to generate codes. Intra-coder reliability was tested by the revisit of the raw data after an interval to ensure the same codes would be given. After the code generation, themes were created and some recurring themes appeared. The codes can be seen in detail in table 21.

Table 21. Thematic Analysis of the Focus Group Interview

	Codes	f	%
Emotional Reactions to feedback	Human emotions are essential	3	3.16%
	AI has no emotion	1	1.05%
	Distrust to AI feedback after learning the source	2	2.11%
	Humans can understand what is written	2	2.11%
AI vs Human Feedback	AI feedback gives accurate feedback	3	3.16%
	AI feedback always focuses on the good parts	2	2.11%
	In the end feedback is feedback	4	4.21%
	Not as knowledgeable as a teacher	13	13.64%

Detection	AI feedback is not detectable	8	8.42%
Opinions on various types of feedback	Teacher is more knowledgeable	8	8.42%
	Peers cannot be objective/knowledgeable	4	4.21%
	Peer feedback is less stressful	1	1.05%
	Ai+ instructor feedback is the best combination	3	3.16%
	Teacher feedback for emotional reasons	2	2.11%
	Ai and human feedback are equal	8	8.42%
Reactions to Excessive Use of AI Feedback	Request human feedback	2	2.11%
	neutral	5	5.26%
	If instructor uses it, I can use it too	3	3.16%
Opinions on Assessment by AI	AI assessment criteria are not clear	2	2.11%
	Instructors are biased; AI may be better	2	2.11%
	Ai is neutral	2	2.11%
	Instructor must give feedback if he/she is giving the assessment	6	6.32%
	Acceptable when teacher is overloaded with work	2	2.11%
Opinions on form of feedback	Written feedback is permanent	3	3.16%
	Oral feedback offers interaction	1	1.05%
	Oral feedback is stressful	1	1.05%
	Oral feedback is more genuine	2	2.11%
Total		97	100%

For an easier presentation of the answers, to report what each participant said, an excel file containing participant and group info was used. For example, HR6 code was

assigned to the sixth person whose name was listed under the Control group. During the transcription, in total, the participants were asked 13 questions throughout the interview (questions that were asked during the focus group interview can be seen on the appendix 2). The first question aimed to see the participants reaction toward the AI generated feedback and how they were feeling about it. In total, eight participants expressed that they had not realized the ChatGPT's involvement.

> '*'I think these feedback it provided are very good, I thought you were the one providing them (GPT 4*).''

> *''...I agree with my friend, it provides detailed feedback which I can fully understand (HR 6).''*

Although some participants were surprised and found the feedback the ChatGPT had provided to be good, not every reaction towards the involvement of the ChatGPT was positive. For the negative perception, reasons such as the ChatGPT's bias towards giving good feedback, the fact that it was developed by a private company, and the wish to be evaluated humanely were given.

> *''I don't think that it is fair to learn that the feedback was provided by ChatGPT. ChatGPT, or other AI systems are developed by companies and they change based on the changes done by those companies. I prefer instructor feedback (HR 15)''*

> *''I feel less confident towards it. One other thing I have realized is that it always made good comments (HR 11).''*

> *''I believe feedback provided by an instructor, a human is better as both parties are human, they can see the errors and give better feedback, they can evaluate humanely. (GPT 10).''*

Now that the participants knew the involvement of the ChatGPT in the feedback process and had seen the feedback it had provided, the second question tapped into the perceived value of the feedback it had provided. Participants expressed that they did not view the feedback ChatGPT had provided as less valuable than the feedback provided by the researcher.

> *'' I believe both feedback are equally valuable, there is no such thing as feedback by AI worths this much and feedback by instructor worths that much. I check both*

feedbacks, evaluate them and use the better one. However, I believe both feedbacks are pretty constructive (GPT 4).''

Next, the participants were asked what would they do if they were given feedback provided by AI for a whole semester without their prior knowledge and how it would affect their motivation towards writing. A good portion of the participants viewed such a situation from a pragmatic angle. However, some participants still expressed that they would prefer feedback to be provided by their instructor.

''If the feedback, has no problems, I am fine with it. It would not lower my willingness to write. However, if I realize the feedback AI provides is sloppy, I would ask my instructor or another human to provide feedback (GPT 4).''

'' ...Be it AI or instructor feedback, at the end of the day both of them aim to show us our mistakes (HR 13).''

'' The feedback I received was good however, receiving ChatGPT feedback constantly would make me feel bad. I am viewing this more emotionally. I would prefer feedback from both my instructor and the ChatGPT (HR 11).''

In addition to the comments about the participants perception of the feedback, there were also some comments about when the participants deemed it was acceptable for instructors to resort to ChatGPT for providing feedback.

''As students, we are already employing tools such as ChatGPT and Midjourney for our presentations and such. If the teacher is using these tools due to heavy workload, it is fine. However, if an instructor is using it for everything, he/she is not qualified to be a teacher, I would want him/her to quit (GPT 4).''

Next question tapped into the participants perceived value of the ChatGPT feedback. The participants were asked whether how seriously they took the FB2 (the feedback provided by the ChatGPT for the writing tasks) had changed now that they had learned that it came from ChatGPT.

''Never. I view this from a pragmatic angle. If the feedback is good and offers me a chance to improve, there is no problem (GPT 4).''

Although the participants, at least the ones described that themselves as pragmatics expressed that they would not mind receiving feedback from ChatGPT. However, some participants still expressed that they would prefer an instructor over ChatGPT.

''I think the feedback provided by the instructor is the one we take more seriously as we get scores based on that. The AI feedback on the other hand, was normal and showing us our mistakes, the instructor feedback, however, is to make sure we correct our mistakes and never repeat them (HR 16).''

'' I can also copy paste the text and ask ChatGPT to give feedback. Hence, any feedback my instructor provides in addition would make me happier (GPT 10).''

In the next question, the participants were asked how would they feel if the essays they wrote would not be read by the instructor and whether it would affect their motivation to write. The general consensus around this question can be summarized with the following quote from one of the participants;

''I would lose motivation. the instructor who gave the homework should also check it (HR 11).''

Five more participants agreed with the aforementioned statement and made similar comments. In addition to these, some participants took the discussion even further and discussed when the AI use by both parties, which are learners and the instructors, was justified;

''If there is a valid reason for the instructor using AI to give feedback, I do not have any objection as we are already using it… Let's say the instructor gave a homework to 70 students and each student wrote 2-3 pages answers to each question in that homework, a huge burden, let alone the instructor, even his/her assistants cannot read it all. In this case, using AI can be considered normal. However, if the instructor has a lot of free time and yet uses AI to give feedback, it is disrespectful (GPT 4).''

'' Both teachers and students have responsibilities. Assignments and grades should be given accordingly. If a teacher let the AI deal with assessment all the time, I see no problem with using AI to complete the said assignments (GPT 10).''

There were other comments supporting that the assignments should be given accordingly, leaving no space for the instructors to resort to using AI to give feedback. In addition, there were some voices raising their concern about the quality of the feedback AI provides;

''In the feedback you gave me (FB1), I got score of 20 out of 25, the feedback AI gave me for the same assignment said everything was fine. For this reason, I do not think its feedback is reliable. In addition, in case of AI assessing our assignments, I do not think we will receive the grades we deserve as we do not know its criteria (HR 3).''

This comment led to a new perspective for the assessment by AI where a participant raised concern about the reliability of the human raters, and asked if people should be prejudiced against it.

'' Sometimes, there can be problems. Let's say, I wrote an essay, one teacher gives it 20, and another one gives it 80 as a result of every teacher has their way, their philosophy of assessment… However, the system we call AI, is a robot, hence it uses the sources it already has, without adding any philosophy of its own (GPT 4).''

When this participant was asked whether he thought AI to be more reliable than human raters, the participant claimed that both can be unreliable at times. No other comment was made by participants about this topic.

Later, the students were presented with a hypothetical scenario where they wrote essays weekly throughout the term and received feedback which was provided by AI, however they were never informed about it. Later, they were asked whether they would have preferred to know the AI's involvement from the beginning and why or why not. Participants, in unison, expressed that they would prefer to be informed about the situation from the beginning.

''If the reason for it kept a secret is an experiment, it is fine. Otherwise, I would like to know the source from the beginning (GPT 10).''

Next question tapped into whether the revelation that the FB2 was written by ChatGPT had changed the participants motivation to use it. All participants, with one exception, agreed that as long as the feedback was sufficient, would use it. However, not everyone was on board with the use of ChatGPT feedback. Here is the comment of the participant who preferred instructor feedback.

''I would prefer instructors' because the teacher knows me more and knows like where my shortcomings are and such and she's going to continue with me throughout the semester. But the AI is just a one-time thing (HR 17).''

Considering how the feedback AWE tools are generic and these tools do not ‘’know’’ the writer, this one can be considered a valid criticism in favor of the human feedback over AI.

Next, the participants were asked about what they felt when they had received the feedback written by ChatGPT before knowing it was from ChatGPT and whether their feelings had changed now that they knew it came from ChatGPT.

> *‘’If I were to compare two feedbacks, I think ChatGPT’s feedback was better. I do not think analytic feedback is constructive, I mean, there should be extra explanations… I was surprised to see how far the technology has come and it could give feedback just like a teacher (GPT 4).’’*

Later, the participants were asked how would their motivation would change if their instructor were to give feedback to their essays only using ChatGPT and they knew about it. This question led to some complaints among the participant about the constant use of AI for feedback.

> *‘’I would question the reason for you to being the teacher, after all, if you are giving the assignment, you should be the one who provides the feedback. If you are not going to provide the feedback, then why are you giving the assignment in the first place? (HR 5).’’*

> *‘’It is fine by me, I will just use AI to complete the assignments. As long as I am getting high grades, it is fine (GPT 10).’’*

Although the participants expressed that they would use AI tools as they see no point in writing essays that would not be read by their instructors, there was one participant who unknowingly pointed out one of the benefits of the use of AWE/AES tools;

> *‘’AI should not be fully responsible for the process, at least if the instructor is not providing the feedbacks, he/she should make up for it by focusing on the students in classroom or other shared platforms (GPT 4).’’*

With the next question, the participants were asked to express their opinion on instructor, AI and peer feedbacks. The participants expressed their opinions and rated them based on their perceptions.

''If I want to get professional feedback, I would prefer my teacher. However, if what I want less stressful and clearer feedback, I choose peer feedback as I can understand my peers better. We kind of speak the same language (GPT 10).''

''... the feedback instructor gives also includes the essays written by the others as he/she also provides feedback to other students. However, when a peer gives feedback, he/she only reads my essays, which may lead to different comments than an instructor's (HR 6).''

Even though some participants preferred peer feedback and justified their preference by stating that peers view essays from different angles and the feedback they provide is clearer and less stressful, some participants preferred instructor feedback, believing peers may have some bias.

''When I give feedback to someone, I cannot stop thinking how my comments may be mean, how I may hurt them. For that reason, I do not think the feedback I give is very constructive (HR 5).''

''I am in no position to provide peer feedback to anyone. Most of my comments are probably wrong in terms of feedback (HR 12).''

''I do not think peers can be objective (HR 15).''

''I think peer feedback goes to the third place as peers cannot be knowledgeable not as AI nor instructors. However, between the AI and instructor. I do not think there is a difference. Even the way we ask questions changes the answer AI gives (GPT 4).''

Due to the reasons stated by the participants in the previous questions, first and third places were not constant. However, AI feedback got the second place in every comment.

''If I have to rank, I would say; instructor, AI and peer (HR 15).''

When asked about their preferred form of feedback some preferred written and justified their preference by saying they were forgetful, or receiving oral feedback in front of the whole classroom was embarrassing for them. However, some participants still preferred oral feedback.

''I prefer written feedback as I am a forgetful person. I may forget what was said to me, but writing is permanent (GPT 10).''

''I also prefer written feedback, as right now the education is online, I would not like to receive oral feedback in front of the whole class. In addition, written feedback is permanent (HR 6).''

''My friend said that written feedback is permanent, however, there are benefits to oral feedback. For example, during oral feedback, as there is a dialogue, both you and the instructor can add points and explain the reasoning behind them even further (GPT 4).''

''I think oral feedback is more genuine… written feedback seems more formal, which makes it hard for me to detect the intention of the instructor. That's why, I prefer oral feedback (GPT 8).''

Lastly, in the light of the qualitative data, the participants were asked about their opinions on AI, its place in the classroom and whether they would use it when they become teachers.

''I believe ChatGPT is good for getting into subjects, and it offers pretty good feedback. However, if you are an expert, ChatGPT cannot offer the same expertise as you. … I believe teachers are better at teaching if you are at a higher level (GPT 14).''

Considering all the findings, the focus group interview revealed a few focal points. First of all, none of the participants could realize its involvement in the feedback process, which can be attributed to its ability to generate human-like texts. As for how they felt it, although most of the participants were pragmatic with its existence as it offered, at least according to them, fine feedback, however some participants expressed they still preferred feedback from their instructor as they believed an instructor would know them better than AI and they valued the human element. As for its use in the classroom by instructors, the participants expressed that they would be informed about its use beforehand and they would not be bothered with its use as long as it was not using extensively and it was used for a good reason.

The participants were also asked to compare different feedback sources. The participants revealed that although they believed the feedback provided by ChatGPT to be adequate, they would still prefer their instructors when they want to get professional feedback, and they would prefer their peers when they want to receive stress free

feedback and feedback that focuses on different aspects of an essay. Yet, some participants said that they did not think they were qualified to provide feedback.

Lastly, when they were asked about their preferred channel for feedback, oral or written, their opinions were mixed. Some participants preferred oral as they believed they could understand more when they receive feedback by talking to the instructor meanwhile some preferred written as they were not comfortable with receiving oral feedback in front of the whole classroom.

4.2.3. Results of the Feedback Effectiveness Questionnaire

In addition to 16 6-Points Likert scale questions, Feedback Effectiveness Questionnaire had 3 open ended questions, which were which were introduced to gain additional insight and understanding of the rationale of the scores they gave to six-point Likert scale questions. The open-ended questions were the following;

1- Which type of feedback (FB1 or FB2) do you find easier to understand? Please explain why.
2- Which type of feedback (FB1 or FB2) do you think helps you improve your writing performance more? Please give examples.
3- Do you think it is necessary to receive feedback?

In this case, FB1 refers to the analytic rubric, meanwhile FB2 refers to the feedback provided by ChatGPT4.

Out of 20 participants 10 (50%) of them considered the feedback provided by ChatGPT4 easier to understand. The amount of detail offered by FB2, the clarity of the feedback it provides, and their perception about FB2 being more specific are some of the reasons given by the participants as for why they held that view. One of the participants (GPT4) answered the first questions by writing ‘*’FB2. is easier to understand since it explains what points are great or bad.’’*. Six (30%) participants on the other hand, considered FB1 easier to understand. Compared to FB2, FB1 deemed to be more basic, clearer, and more comprehensive by six (30%) participants. One of the participants (GPT 11) wrote *‘’ FB1 because "the table view" is simple to understand’’*. Four (20%) participants considered both types of feedback to be quite easy to understand.

Out of 20 participants 14 (70%) of them believed that FB2 would help them improve their writing skills more. As for why would it help more, the participants wrote that it had more ‘’detail’’. The fact that it was detailly written was well received by all participants. Three of the participants believed that any kind of feedback was welcomed, regardless of the source. Finally, three participants said that they would prefer FB1, justifying their decision with statements about FB2’s length, and how FB1 directly states what is lacking. One participant who preferred FB1 said *’’ FB1 because it better highlights the areas where I can improve my writing.’’*

All participants believed that receiving feedback was necessary. Among the reasons for the necessity of the feedback were; seeing one’s mistakes, learning about his/her incompetencies. One of the participants (HR 14) wrote;

> *’’ …because it is crucial to see mistakes in writing. There would be no use in writing if we do not take any feedback…’’*

In light of the all the data analyzed, it can be said that all of the participants believe receiving feedback is necessary to as it shows them their errors and offers possible ways to improve themselves in terms of L2 writing skill. As for the strong points of each feedback; the participants who preferred FB1 based their preference on FB1 being more concise, meanwhile the participants who preferred FB2 expressed that the feedback was more detailed. Hence, it can be concluded that both types of feedback was effective. Participants preference of one over another is based on their preference of what kind of feedback they want to receive.

5. DISCUSSION AND CONCLUSION

In this section, the analyzed qualitative and quantitative data will be investigated in terms of the research questions. In addition to that, the overall effectiveness of the study, limitations of the study and the recommendations for further research will be discussed.

5.1. Discussion of the Research Questions

In this section, the research questions will be answered using the collected qualitative and quantitative data.

5.1.1. The Effect of Human and ChatGPT Feedback on L2 Writing Skill Development

The current study aimed to answer two questions, the first question was *''Does the feedback provided by a human rater or an AWE have any effect on L2 writing skill development?''* A pre-post writing proficiency test was employed to answer this question. Pre-post writing proficiency tests were rated using the aforementioned analytic scoring rubric. The scores of the pre-test and post test were compared both between groups and in-group to see whether the treatment had caused a meaningful difference.

5.1.1.1. Pre-Post Writing Proficiency Test

Although number of participants took the pre-test and the posttest were higher, due to the dropouts and the eligibility criteria, only the data acquired from participants who filled out the demographic form, completed the English proficiency test, and completed at least the second or the third task in addition to the first task, were included. In the end only the data acquired from 34 participants could be used. 17 of these participants were in the control group meanwhile the other 17 were in the experimental group. The data analysis revealed insights which will be discussed in detail.

The first insight from the results was the in-group score increase in terms of the pre and post- test scores for the experimental group was higher. There are many studies that confirm that AWE tools lead to higher score increase in short term (Wei et al., 2023; Li et al., 2017). In addition to existing studies, Mahapatra (2024) compared ChatGPT feedback to instructor feedback using pre-test post-test design with undergraduate ESL students to see its effect on academic writing skills and found that score increase for the experimental group was statistically significant.

The second insight is that, although both groups had increased their post-test scores, there was no significant difference between groups. This also coincides with a similar study which compared teacher feedback to AWE tool feedback. In their study, similar to current study, Ajabshir and Ebadi (2023) confirmed learners' proficiency level to be intermediate using Oxford Quick Placement Test. The experimental group received immediate feedback from Grammarly, meanwhile control group received feedback from the teacher. Post test results showed that although both groups had increased their scores, there was no significant difference. In their study, the authors concluded that this lack of difference could be explained both types of feedback, although they were focusing on different aspects of the language, were equally effective. In another study (Bodaubekov et al., 2025) 134 participants were divided into four control and four experimental groups. Control groups received feedback from their instructor meanwhile the experimental groups received feedback via Write & Improve. The researchers could not find consistent improvement in any groups. As a result concluded that both types of feedback were equally important. One last example for this is a study which directly compared teacher feedback to ChatGPT feedback. Alsofyani and Barzanji (2025) used pre-test post-test design with 102 university students where experimental group received feedback from ChatGPT and the control group received teacher feedback. The post-test scores revealed that there was no significant difference between groups.

In sum, although the experimental group had increased their scores more, there was no statistically significant difference compared to control group. A few reasons could be given as for why that was not the case for the currents study as follows.

Studies found that AWE feedback is most effective on learners with lower proficiency levels (Wang, 2015;Wei et al., 2023). In the current study, however, the overall group proficiency was, confirmed using Oxford Quick Placement Test, C1, which could

explain the lack of statistically significant difference between groups in terms of post-test scores.

Second possible explanation for the lack of difference could be the duration of the study. In the current study participants had received feedback only once. However, it is concluded that medium to longer use of AWE can lead to higher effect (Ngo et al., 2022). So, longer exposure to ChatGPT feedback could lead to a statistically significant difference.

Considering the data analysis and the other studies mentioned in this section to discuss the results deeper, the first research question which aimed to observe instructor and AWE feedback's effect on L2 writing skill development, can be answered, at least in scope of the current study, as both AWE and instructor feedback have similar effect on L2 writing skill development.

5.1.2. EFL Learners' Opinion on Receiving Feedback from ChatGPT

The second research question the current study aimed to answer was *''How do EFL learners perceive receiving feedback from ChatGPT?*''

Two separate data collection tools were employed to answer this question. The first one is the three open-ended questionnaire in the second part of the feedback effectiveness questionnaire, which focused on the comparison of the two types by the participants. Second tool, the focus group interview, addressed the reactions to the use of ChatGPT and receiving feedback from AI tools in general. The results gained from both of these tools will be discussed separately.

5.1.2.1. Feedback Effectiveness Questionnaire

Before starting the questionnaire, the feedback of the task 2 and task 3 were uploaded to a google drive folder. Participants, using the link provided by the researcher, accessed the drive file and checked the feedback provided for them for the second and the third tasks. In addition, the participants were also shown random feedbacks using screen share function to ensure the participants would see both types of feedback. Afterwards, the participants completed the feedback effectiveness questionnaire.

In general, it is possible to deduce that the feedback effectiveness questionnaire has revealed that the participants considered feedback provided by ChatGPT, which was unbeknownst to them at the time, in the context of the current study in which the

participants compared the feedback provided by human-raters to the feedback provided by the ChatGPT to be better in some aspects. This, result that ChatGPT feedback deemed to be better than human-rater feedback in some aspect such as amount of detail, also coincides with similar studies in the field (Teng, 2024; Guo & Wang, 2024).

Amount of detailed the feedback written by ChatGPT could be the reason. Although many participants found FB2 to be more detailed, it would not be realistic for an instructor with more than 50 students to read every single one participants' essay and write a lengthy text to give feedback. Hence, instructors rely on summative methods such as rubrics. In the context of the current study where the participants compared human-rater feedback to the ChatGPT feedback and perceived the one provided by the ChatGPT to be better in most aspects, their preference of FB2, or ChatGPT feedback can be considered as a testament to their preference of a more detailed feedback rather than a general one provided via a rubric.

5.1.2.2. Focus Group Interview

Focus group interview, which was aimed to gain the participants' reaction to use of ChatGPT in the feedback process, revealed valuable information regarding the participants' reaction toward receiving feedback from ChatGPT, which will be discussed in this section.

Although all of the participants had agreed that the feedback provided by ChatGPT was adequate, some of the participants did not like the idea of receiving ChatGPT feedback extensively with or without their knowledge. For the reasons of not wanting to receive feedback from ChatGPT extensively, there were two categories. The first category was the perceived problems related to ChatGPT and the second one was the problems related to the instructor. For the first one, participants mentioned how it cannot ''read'', and how the feedback it gives focuses on the good parts. For the latter, some participants claimed that if an instructor uses AI tools to give feedback even though he/she has the time and does it without any tools, he/she should not teach any longer. In addition, according to the participants, if the instructor resorts to AI tools for feedback due to the number of assignments given, the instructor should not assign any homework if he/she will not be able to provide feedback. So, it is possible to say that the participants, viewed feedback as one of the mandatory roles of the instructor.

However, not every participant was harsh against the idea of utilizing AI for feedback. Some participants admitted that they were using AI tools as well and if the instructor has a good reason to resort to these tools for feedback once in a while, it was acceptable for them. However, if an instructor uses these tools for feedback all the time, he/she makes up for it by dedicating students more time in the classroom. This view is one of the biggest selling points of AWE tools we have. There are many studies in which teachers have confirmed that the use of AWE tools for feedback helped them to focus on other aspects of writing (Li et al., 2015).

> *''I believe feedback provided by an instructor, a human is better as both parties are human, they can see the errors and give better feedback, they can evaluate humanely. (GPT 10)''*

> *''The feedback I received was good however, receiving ChatGPT feedback constantly would make me feel bad. I am viewing this more emotionally. I would prefer feedback from both my instructor and the ChatGPT (HR 11)''*

Harrington and Moran (2001), have a similar opinion about the AWE tools. They mentioned the use of AWE tools would desensitize us as we would no longer write to a human, but to a machine. In addition, during the interviews of a study about utilizing ChatGPT in classroom settings under the instructor's guidance (Teng, 2024) the participants made comments very similar to the GPT 10's comment about ''humane evaluation''. Yancey et al. (2004) also reached a similar conclusion and stated that writing is a social act, hence, it should be assessed and read by humans. In addition to these comments related to how the use of AWE tools impact social aspect of writing, Vojak et al. (2011) argued that these tools were in favor of a certain way of writing (linguistically generic and lengthy texts) and argued that this tendency would change the students' writing habits and force them toward writing in a standard way that would meet the AWE tools' expectations and Link et al. (2020) reached the same conclusion in their study where they compared a combination of teacher feedback with Criterion feedback to only teacher feedback. They observed that some of the students who received a combination of AWE and teacher feedback did not correct their sentences, instead they simply deleted their sentences when Criterion flagged their sentences as incorrect.

One of the participants made a different comment and said;

''I would prefer instructors' feedback because the teacher knows me more and knows like where my shortcomings are and such and she's going to continue with me throughout the semester. But the AI is just a one-time thing (HR 17)''

Even though this may be true for the existing AWE tools on the market, and HR 17's comment can be supported by a previous study (Grimes & Warschauer, 2010), which revealed that students, took the feedback provided by AWE tools more lightly as they were graded based on the teacher feedback, ChatGPT is an LLM model, its main premise is to react to your prompts. In addition, it can ''remember'' the previous dialogue. Therefore, if someone were to provide ChatGPT with essays every week and ask it to evaluate your overall progress at the end, it could, at least in theory, track the progress and tailor the feedback based on the shortcomings. Regardless, in the scope of this study, this comment reveals that if instructors are to use ChatGPT as an additional feedback source, they should inform the students about its capabilities and how to use it efficiently. It is also worth noting that there are other studies that focused on how ChatGPT could be utilized for writing skills which reached similar conclusions (Guo & Wang, 2024; Teng, 2024).

The comment from HR 17 can also lead to whole another discussion. One of the key aspects of the Sociocultural theory, which was discussed earlier in the literature review section, is that humans learn with interaction and learning is a social process (Kozulin, 2002). Viewing the data acquired during the focus group interview including the HR 17's comment through the lenses of Sociocultural theory implies that the negative perceptions against the use of the ChatGPT or any other AI tool can be linked to the loss of the human element in the feedback process. For example, the HR 17's comment can be summarized that HR 17 has a negative perception about the use of ChatGPT to provide feedback because language learning is a social process which cannot be separated from the learner's social and cultural identity (Kozulin, 2002) and an AI tool cannot deduce the writer's social and cultural background just from the essay it is supposed to provide feedback. In addition, learning is optimized when the learner is provided with just enough scaffolding so that the learner can achieve a little more than what he/she could achieve alone. These AWE tools however, are designed to appeal to wider audiences, hence the feedback they provide is usually generic, which may be considered insufficient by high level learners (Ranalli, 2018).

In the light of the information above, the RQ2 can be answered as; the participants are willing to receive feedback from ChatGPT as long as they are informed about its involvement in the feedback process. They are fine with instructors ‘’resorting’’ to ChatGPT when necessary, however they do not want instructors to use ChatGPT to give all the feedback or for scoring as they are not sure about the criteria it uses, they do not believe that ChatGPT can ‘’read’’ and they believe that their instructors know them better. Lastly, they still believe feedback is the responsibility of the instructor as the instructors are the ones who assign the writing tasks in the first place.

5.2. Conclusion

Current study was a different take on the studies related to AWE tools in two ways. First, instead of using an actual AWE tool, ChatGPT was utilized as an AWE tool. Second, the feedback provided by ChatGPT was provided to the participants as if it was coming from an instructor, which makes the current study a different approach on AWE tools and their effect on L2 writing skill development.

Traditionally, AWE tools assess written texts by analyzing their features and comparing them to the written samples on which they were trained and subsequently assigning a score, correcting mistakes, or providing generic feedback. These systems offer no additional interaction. Hence, the main reason the researcher had decided to see if ChatGPT could be utilized as an AWE tool was interaction ChatGPT had offered. Any AWE may provide feedback on grammar and other mechanic parts of the text better than ChatGPT, as they are specifically designed for it. However, these tools are not accurate as much when it comes to content and organization of the text (Li et al., 2015) and they offer no interaction, or give any chance to tailor the feedback to individual needs. ChatGPT, on the other hand, can be asked to focus on certain features of the text or provide the feedback in a certain way, just like the current study. In addition, traditional AWE tools were designed using the data from native speakers and they were not designed for non-native speakers although they are marketed in countries where English is the foreign language (Warschauer & Ware, 2006). Hence, the feedback they provide, the essays they can score, cannot be too different than what they were initially trained on (Liu & Kunnan, 2016), which brings into mind questions such as how accurate these tools can be at detecting the errors of non-native speakers (Weigle, 2013), or whether the feedback these tools provide satisfies the non-native

speakers (Lai, 2010). ChatGPT on the other hand, is an LLM trained using the data available on the internet, including the data created by English speakers from all over the world with different cultural backgrounds. Accordingly, there should be no writing too unusual for it to understand or comment on. In addition, it always keeps learning as it interacts with users and the data from all over the internet and it keeps getting updates. Thus, the researcher thought the ChatGPT could be utilized as an AWE tool to provide feedback to learners. Most of the commercially available tools provide feedback. However, the students do not frequently revise their essays and even when they do, the revisions they make are usually limited to grammar (Attali, 2004; Warschauer & Grimes, 2008). For this reason, instead of making it provide feedback on the mechanics, grammar and other sentence level aspects of language, the researcher wanted to test whether ChatGPT could provide feedback as a teacher would and what would the students think about the feedback they receive without knowing that the it is provided by ChatGPT.

The results showed that although the feedback it provided was adequate and indistinguishable from feedback provided by human-raters, due to the participants' emotional reactions toward the excessive use of it, it should be used moderately. Even when it is to be used, the learners should be informed as providing ChatGPT feedback without informing students would not be ethical and they may feel betrayed or get angry and question why would their instructor not provide the feedback and rely on ChatGPT. In terms of the current study, it can be concluded that despite the hesitations of the participants related to impact ChatGPT may have on their writing motivation when used excessively, it might be a handful tool helping teachers decrease their workload and focus on other aspects of writing which they could not before due to time limitations. This result also coincides with the results of some other studies (Chen & Cheng, 2008; Bai & Hu, 2016; Han & Sari, 2022), which concluded that the best approach to use ChatGPT or any other AWE tools would be using them as an additional feedback source, as an addition to instructor feedback, thus saving time for the instructor to focus on the other aspects of writing.

5.2.1. Limitations of the Current Study

Current study has few limitations. First one is the number of the participants. In the beginning of the study, there were 51 participants who had agreed to join the study.

However, after the dropouts, there were 46 participants left. Out of 46 participants, only 34 participants' data could be used for RQ1 as rest of the participants did not meet the criteria, which was to complete the placement test, demographic form, pre-post writing proficiency test and complete at least two tasks out of three (one of the tasks which they were required to complete was the task 1), set by the researcher, and only 20 participants' data could be used for RQ2 as rest of the participants did not join the interview.

Second limitation was the online education in Turkey. As the result of the online classes, some participants were using mobile devices to join classes and complete the tasks. Some of them had technological problems and could not upload the essays, and then had to contact the researcher. This problem resurfaced during focus group interview and feedback effectiveness questionnaire when some participants could not check the feedback for the second and the third task before the interview started.

The last limitation for the current study, or reality of the current classroom environments all around the world, was the form of the feedback that were provided to the participants. The instructor feedback had to be provided via a rubric. This was caused due to the high number of the participants (46) and the number of essays had to be read. If this study were to be conducted with, for example six participants, providing well written feedback which addresses every criterion would not be a problem. However, due to the aforementioned limitations, the study had to be conducted this way. This limitation also applies to real life as one of reasons for the instructors resort the summative feedback and summative assessment is the overcrowded classrooms (Yüksel & Gündüz, 2017).

5.2.2. Pedagogical Implications

Drawing on the collected data, there are a few recommendations which could be made. The first recommendation would be teaching the learners how to use ChatGPT effectively and build up a system to use it in class with the learners. The feedback provided by AWE tools are little to no use without context-sensitive scaffolding and the teachers' help to students to help them understand it (Jiang et al., 2020). It has been confirmed by the studies that the feedback provided by AWE/AES tools is useful for correcting grammar errors, thus freeing up time for the teachers to focus on other aspects of the language such as content (Li et al., 2015; Schroeder et al., 2008; Wilson

& Czik, 2016). In this case, ChatGPT could be used by the learners to get the surface level feedback for their in-class writing tasks. Students could receive feedback on the grammar and the instructor can handle deeper level feedback such as content, organization and so on (Teng, 2024).

Writing is a productive skill which requires constant practice and feedback to improve (Keh, 1990). Hence, the second recommendation would be encouraging learners to interact with ChatGPT in their free times so they get to practice, receive feedback instantly increase learner autonomy. In addition, due to its ability to remember the previous conversations, the students can use it as a writing coach that adapts to their writing style and provides feedback based on individual needs of the learners. The learners can use it to track their overall development and see where their shortcomings are.

5.2.3. Recommendations for Further Research

Based on this study and its results, there are a few recommendations that could be made for further studies.

This study involved 51 participants who were freshmen EFL learners. The results received from 51 participants are hard to generalize due to the small sample size, and the participants dropouts. Hence, the first recommendation to researchers would be implementing the same study with more participants in order to reach more generalizable results. Preferably in face-to-face settings. Second recommendation is conducting same study with participants from various proficiency levels. In this study participants were from similar English level. As mentioned earlier, learners with low proficiency levels get the most from these AWE/AES tools. The lack of significant difference in this study could be explained with the lack of proficiency level difference of the participants. For this reason, a study with participants from different proficiency levels may yield results worth looking into.

The third recommendation is due to the nature of the tool, namely ChatGPT. Data collection process began in February 2023, which coincides with the time ChatGPT 4 was introduced. From the onset of the data collection till the end of it, a few problems occurred. The first problem was the updates. ChatGPT received four updates. Each update changed the way ChatGPT reacted to the same prompt, which was also verified by other researchers (Chen et al., 2024). Sometimes, ChatGPT forgot the previously

set criteria and provided feedback on the criteria itself created. ChatGPT had to be reminded of each criterion and provided feedback samples it had previously provided as an example of what it should provide, which did not work every time and new chats had to be created. These updates also affected the length of the feedback. In order to give participants a sense of consistency, ChatGPT had also been warned regarding the length of the feedback. Sometimes, it provided too short feedbacks. The last problem, collecting the writings of the students, sharing them with ChatGPT, and sending back feedback to students is quite time-consuming. ChatGPT can be integrated into web pages and can be instructed to work in a certain way. Thus, creating a webpage that students can use to send their essays to ChatGPT and receive feedback based on the pre-set criteria would make it possible to work with numerous participants would save the researchers from the burden of being the middleman. It would also increase autonomy by giving participants a chance to receive feedback which they can interact and shape depending on their needs as much as they want. A study letting learners interact with ChatGPT that way would make it a direct competitor of the commercially available AWE/AES tools on the market and would definitively yield interesting results.

Another recommendation would be a direct comparison between ChatGPT and an actual AWE tool. One of the biggest downsides of AWE tool is, as they have to be suitable for wider audiences, the generic feedback it provides may not meet the needs of the more proficient learners (Xu & Zhang, 2022). ChatGPT, on the other hand may provide more tailored feedback for learners. A study which will involve participants using both AWE tools and ChatGPT for a prolonged time and compare the feedback they had received may be worth looking into.

The last recommendation is due to the nature of ChatGPT. On mobile applications students chat with ChatGPT thanks to its speech-to-text feature. As some of the students commented how they preferred oral feedback due to its interactive nature, letting students take audio feedback from ChatGPT and then discuss the feedback with it instantly may yield interesting results in terms of students' perception towards the use of AI feedback for L2 writing skill development. A study which focuses on whether receiving the ChatGPT feedback orally or written would lead to any motivational or skill development differences is worth investigating.

REFERENCES

Agar, M., & MacDonald, J. (1995). Focus groups and ethnography. *Human Organization*, *54*(1), 78–86.

Ahmad, Z. (2020). Summative assessment, test scores and text quality: A study of cohesion as an unspecified descriptor in the assessment scale. *European Journal of Educational Research, 9*(2), 523–535. https://doi.org/10.12973/eu-jer.9.2.523

Ajabshir, Z. F., & Ebadi, S. (2023). The effects of automatic writing evaluation and teacher-focused feedback on CALF measures and overall quality of L2 writing across different genres. *Asian Pacific Journal of Second and Foreign Language Education, 8*, 26. https://doi.org/10.1186/s40862-023-00201-9

Al-khresheh, M. H. (2024). Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers and Education: Artificial Intelligence, 6*, Article 100218. https://doi.org/10.1016/j.caeai.2024.100218

Almeida, F. (2018). Strategies to perform a mixed methods study. *European Journal of Education Studies*, *5*(1), 137–151. https://doi.org/10.5281/zenodo.1406214

Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: a mixed-methods study. *Journal of Educational Computing Research, 63*(2), 431–463. https://doi.org/10.1177/07356331241307297

Attali, Y. (2004). Exploring the Feedback and Revision Features of Criterion.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). Routledge.

Bai, L., & Hu, G. (2016). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Barrot, J. S. (2024). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies*, *29*(6), 7155–7179. https://doi.org/10.1007/s10639-023-12083-y

Bereiter, C., & Scardamalia, M. (Eds.). (1987). The psychology of written composition (1st ed.). Routledge. https://doi.org/10.4324/9780203812310

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8–21. https://doi.org/10.1177/003172170408600105

Bodaubekov, A., Agaidarova, S., Zhussipbek, T., Gaipov, D., & Balta, N. (2025). Leveraging AI to enhance writing skills of senior TFL students in Kazakhstan: A case study using "Write & Improve". *Contemporary Educational Technology, 17*(1), ep548. https://doi.org/10.30935/cedtech/15687

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika, 65*(1), 23–28. https://doi.org/10.1007/BF02294183

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology, 23*(1), 23–41. https://doi.org/10.56297/BKAM1691/WIEO1749

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burner, T. (2015). Formative assessment of writing in English as a foreign language. *Scandinavian Journal of Educational Research*, *60*(6), 626–648. https://doi.org/10.1080/00313831.2015.1066430

Butt, D., Fahey, R., Feez, S., Spinks, S., & Yallop, C. (2000). *Using functional grammar*. National Centre for English Language Teaching and Research.

Cambridge University Press & Assessment. (2024). *IELTS 19 academic: Student's book with answers with audio with resource bank.* Cambridge University Press.

Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94–112. https://doi.org/10125/44145

Chen, L., Zaharia, M., & Zou, J. (2024). How Is ChatGPT's Behavior Changing Over Time?. *Harvard Data Science Review*, *6*(2) 1–44. https://doi.org/10.1162/99608f92.5317da47

Chomeya, R. (2010). Quality of psychology test between Likert scale 5 and 6 points. *Journal of Social Sciences*, *6*(3), 399-403. https://doi.org/10.3844/jssp.2010.399.403

Clarke, V. & Braun, V. (2017) Thematic analysis, *The Journal of Positive Psychology*, *12*(3), 297-298, https://doi.org/10.1080/17439760.2016.1262613

Cole, V. (2012). *IELTS introduction study skills: A self-study course for all academic modules*. Macmillan Education.

Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement*. Palgrave Macmillan.

Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing, 8*(2), 73–83. https://doi.org/10.1016/s1075-2935(02)00047-8

Cumming, A. (2016). Theoretical orientations to L2 writing. In R. Manchón & P. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 65–88). De Gruyter Mouton. https://doi.org/10.1515/9781614511335-006

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

Ersanlı, C. Y., & Akman Yeşilel, D. B. (2022). The use of automated writing evaluation tools to foster ESL writing instruction. In G. Yangın-Ekşi et al. (Eds.), New directions in technology for writing instruction (English Language Education, Vol. 30, pp. 194–207). Springer. https://doi.org/10.1007/978-3-031-13540-8_10

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge University Press. https://doi.org/10.1017/CBO9781139524742

Ferris, D., Brown, J., Liu, H. S., & Stine, M. E. A. (2011). Responding to L2 students in college writing classes: Teacher perspectives. *TESOL Quarterly, 45*(2), 207–234. https://doi.org/10.5054/tq.2011.247706

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600

Gamon, M., Chodorow, M., Leacock, C., & Tetreault, J. (2013). Grammatical error detection in automatic essay scoring and feedback. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 251–266). Routledge/Taylor & Francis Group.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge University Press.

Goshu, K. C., & Gebremariam, H. T. (2024). Revisiting writing feedback: Using teacher-student writing conferences to enhance learners' L2 writing skills. Ampersand, 13, 100195. https://doi.org/10.1016/j.amper.2024.100195

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. Longman.

Green, A. (2022). *L2 writing assessment: An evolutionary perspective*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-15011-1

Grimes, D. & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment, 8*(6).

Guest, G., Namey, E., Taylor, J., Eley, N., & McKenna, K. (2017). Comparing focus groups and individual interviews: findings from a randomized study. *International Journal of Social Research Methodology, 20*(6), 693–708. https://doi.org/10.1080/13645579.2017.1281601

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*, 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hall, K., & Sheehy, K. (2018). Assessment and learning: Summative approaches. In Cremin, T. & Burnett, C. (eds.) *Learning to Teach in the Primary School*, Fourth

Edition. London: Routledge, (pp. 288–303). https://doi.org/10.4324/9781315453736-27

Han, T., & Sari, E. (2022). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning*, *37*(4), 961–985. https://doi.org/10.1080/09588221.2022.2067179

Hargreaves, E., & Gipps, C., & Pickering, A. (2018). Assessment and learning: Formative approaches. In Cremin, T. & Burnett, C. (eds.) *Learning to Teach in the Primary School*, Fourth edition. London: Routledge, (pp. 275–287). https://doi.org/10.4324/9781315453736

Harman, R. (2013). Literary intertextuality in genre-based pedagogies: Building lexical cohesion in fifth-grade L2 writing. *Journal of Second Language Writing, 22*(2), *125–140.* https://doi.org/10.1016/j.jslw.2013.03.006

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, *63*(4), 480–499. https://doi.org/10.2307/378891

Huang, J., Berg, M., Siegrist, M., & Damsri, C. (2017). Impact of a functional linguistic approach to teacher development on content area student writing. *International Journal of Applied Linguistics, 27*(2), 331–362. https://doi.org/10.1111/ijal.12133

Huang, S., & Renandya, W. A. (2018). Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in Language Learning and Teaching*, *14*(1), 15–26. https://doi.org/10.1080/17501229.2018.1471083

Huot, B. (2002). *Rearticulating writing assessment for teaching and learning*. Utah State University Press.

Hyland, K. (2004). Genre and Second Language Writing. University of Michigan Press.

Hyland, K. (2013). Faculty feedback: Perceptions and practices in L2 disciplinary writing. *Journal of Second Language Writing, 22*(3), 240–253. https://doi.org/10.1016/j.jslw.2013.03.003

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

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House Publishers.

Jafarian, K., Soori, A., & Kafipour, R. (2012). The effect of computer assisted language learning (CALL) on EFL high school students' writing achievement. *European Journal of Social Sciences*, *27*(1), 138–148.

Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. System, 93, 102302. https://doi.org/10.1016/j.system.2020.102302

Junqueira, L., & Payant, C. (2015). "I just want to do it right, but it's so hard": A novice teacher's written feedback beliefs and practices. *Journal of Second Language Writing, 27*, 19–36. https://doi.org/10.1016/j.jslw.2014.11.001

Kaplan, R. B. (1966). Cultural thought patterns in inter-cultural education. *Language Learning, 16*(1–2), 1–20. https://doi.org/10.1111/j.1467-1770.1966.tb00804.x

Keh, C. L. (1990). Feedback in the writing process: a model and methods for implementation. *ELT Journal, 44*(4), 294–304. https://doi.org/10.1093/elt/44.4.294

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *The Modern Language Journal, 75*(3), 305–313. https://doi.org/10.2307/328724

Khasawneh, M. A. S. (2024). Investigating the impact of automated instruments used for assessing the writing skill: Perspectives of language e-learners. *Research Journal in Advanced Humanities, 5*(2), 41–54. https://doi.org/10.58256/4fd2qt78

Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education, 27*(3), 275–286. https://doi.org/10.1080/03075070220000662

Kostka, I., & Toncelli, R. (2023). Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations. *The Electronic Journal for English as a Second Language, 27*(3). https://doi.org/10.55593/ej.27107int

Kozulin, A. (2002). Sociocultural Theory and the Mediated Learning Experience. *School Psychology International, 23*(1), 7–35. https://doi.org/10.1177/0143034302023001729

Kukulska-Hulme, A. (2012). Mobile-assisted language learning. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0768

Lai, Y. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology*, 41(3), 432–454. https://doi.org/10.1111/j.1467-8535.2009.00959.x

Lee, I. (2011). Formative assessment in EFL writing: An exploratory case study. *Changing English, 18*(1), 99–111. https://doi.org/10.1080/1358684x.2011.543516

Lee, I. (2016). Putting students at the centre of classroom L2 writing assessment. *Canadian Modern Language Review. 72*, 1–23.

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer. https://doi.org/10.1007/978-981-10-3924-9

Li, J., & L,, M. (2022) Assessing L2 writing in the digital age: Opportunities and challenges. *Journal of Second Language Writing. 57*, 100913 https://doi.org/10.1016/j.jslw.2022.100913

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. Journal of Second Language Writing, 27, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

Li, Z., Feng, H., & Saricaoglu, A. (2017). The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *CALICO Journal*, *34*(3), 355–375. https://doi.org/10.1558/cj.26382

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Lipnevich, A. A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education*, 6, 720195. https://doi.org/10.3389/feduc.2021.720195

Liu, S., & Kunnan, A. J. (2016). Investigating the Application of Automated Writing Evaluation to Chinese Undergraduate English Majors: A Case Study of WriteToLearn. *CALICO Journal*, *33*(1), 71–91.

Liu, Y. (2008). The effects of error feedback in second language writing. *Journal of Second Language Acquisition and Teaching*, *15*, 65–79.

Llamas-Nistal, M., Fernández-Iglesias, M. J., González-Tato, J., & Mikic-Fonte, F. A. (2013). Blended e-assessment: Migrating classical exams to the digital world. *Computers & Education*, *62*, 72–87. https://doi.org/10.1016/j.compedu.2012.10.021

Mabuan, R. A. (2024). ChatGPT and ELT: Exploring teachers' voices. *International Journal of Technology in Education*, *7*(1), 128–153. https://doi.org/10.46328/ijte.523

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environment*, *11*, 9. https://doi.org/10.1186/s40561-024-00295-9

Malik, S. A. (2017). Revisiting and re-representing scaffolding: The two gradient model. *Cogent Education*, *4*(1), 1331533. https://doi.org/10.1080/2331186X.2017.1331533

Marginson, S., & Dang, T. K. A. (2016). Vygotsky's sociocultural theory in the context of globalization. *Asia Pacific Journal of Education*, *37*(1), 116–129. https://doi.org/10.1080/02188791.2016.1216827

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

Meniado, J. C., Huyen, D. T. T., Panyadilokpong, N., & Lertkomolwit, P. (2024). Using ChatGPT for second language writing: Experiences and perceptions of EFL learners in Thailand and Vietnam. *Computers and Education: Artificial Intelligence*, *7*, 100313. https://doi.org/10.1016/j.caeai.2024.100313

Miceli, T. (2006). Foreign language students' perceptions of a reflective approach to text correction. *Flinders University Languages Group Online Review*, *3*(1), 25–36.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, *70*(2), 151–167. https://doi.org/10.1080/00335638409383686

Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2022). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments*, *32*(2), 727–744. https://doi.org/10.1080/10494820.2022.2096642

Nguyen, H. M., & Chu, T. A. (2024). Written corrective feedback in second language writing: A review of research. *rEFLections*, *31*(2), 858–870. https://doi.org/10.61508/refl.v31i2.275269

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, 38(2), 599–620. https://doi.org/10.1111/jcal.12635

Oxford University Press. (2001). *Quick Placement Test: Paper and pen test, user manual.*

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7th ed.). Routledge. https://doi.org/10.4324/9781003117452

Parra G., L., & Calero S., X. (2019). Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction*, *12*(2), 209–226. https://doi.org/10.29333/iji.2019.12214a

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Sage.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21,* 104–111. https://doi.org/10.1016/j.asw.2014.05.001

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, *2*(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083

Plano Clark, V. L., & Creswell, J. W. (2015). *Understanding research: A consumer's guide* (2nd ed.). Pearson.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*(2), 103–134. https://doi.org/10.1016/s0747-5632(01)00052-8

Quick Placement Test (Version 1). (2004). Oxford University Press.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13. https://doi.org/10.1002/bs.3830280103

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52,* 100816. https://doi.org/10.1016/j.jslw.2021.100816

Saito, H. (1994). Teachers' practices and students' preferences for feedback on second language writing: A case study of adult ESL learners. *TESL Canada Journal*, *11*(2), 46–70. https://doi.org/10.18806/tesl.v11i2.633

Schartel, S. A. (2012). Giving feedback: An integral part of education. *Best Practice & Research Clinical Anaesthesiology*, *26*(1), 77–87. https://doi.org/10.1016/j.bpa.2012.02.003

Schroeder, J., Grohe, B., & Pogue, R. (2008). The impact of Criterion writing evaluation technology on criminal justice student writing skills. *Journal of Criminal Justice Education*, *19*(3), 432–445. https://doi.org/10.1080/10511250802476269

Schulze, M., & Heift, T. (2013). Intelligent CALL. In M. Thomas, H. Reinders, & M. Warschauer (Eds.), *Contemporary Computer-Assisted Language Learning* (pp. 249–265). Continuum.

Shadiev, R., & Feng, Y. (2023). Using automated corrective feedback tools in language learning: A review study. *Interactive Learning Environments*, 32(6), 2538–2566. https://doi.org/10.1080/10494820.2022.2153145

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective.* Lawrence Erlbaum Associates Publishers.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313–346). Routledge.

Shermis, M. D., Burstein, J. C., & Bursky, A. (2013). Introduction to automated essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1–15). Routledge.

Shi, H., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, *28*, 771–795. https://doi.org/10.1007/s10639-022-11200-7

Srichanyachon, N. (2012). Teacher written feedback for L2 learners' writing development. *Humanities, Arts and Social Sciences Studies, 12*(1), 7–17.

Steinhart, D. J. (2001). *Summary Street: An intelligent tutoring system for improving student writing through the use of latent semantic analysis* [Doctoral dissertation, University of Colorado, Boulder].

Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, *42*, 1–16. https://doi.org/10.1016/j.compcom.2016.05.001

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51-65. https://doi.org/10.1016/j.asw.2013.11.007

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

Tardy, C. (2012). A rhetorical genre theory perspective on L2 writing development. In R. Manchón (Ed.), *L2 writing development: Multiple perspectives* (pp. 165–190). De Gruyter Mouton. https://doi.org/10.1515/9781934078303.165

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*. *7*. 100270. https://doi.org/10.1016/j.caeai.2024.100270

Torres, J. O. (2019). Positive impact of utilizing more formative assessment over summative assessment in the EFL/ESL classroom. *Open Journal of Modern Linguistics, 9*(1), 1–11. https://doi.org/10.4236/ojml.2019.91001

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*(2), 147–170. https://doi.org/10.1016/S1060-3743(00)00022-9

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition, 28*(2), 97–111. https://doi.org/10.1016/j.compcom.2011.04.004

Vonderwell, S., Liang, X., & Alderman, K. (2007). Asynchronous discussions and assessment in online learning. *Journal of Research on Technology in Education, 39*(3), 309–328. https://doi.org/10.1080/15391523.2007.10782485

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

Wang, P.-L. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching, 12*(1), 79–100.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22–36. https://doi.org/10.1080/15544800701771580

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching, 31*(2), 57–71. https://doi.org/10.1017/S0261444800012970

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157–180. https://doi.org/10.1191/1362168806lr190oa

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology, 14*, 1249991. https://doi.org/10.3389/fpsyg.2023.1249991

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*(1), 85–99. https://doi.org/10.1016/j.asw.2012.10.006

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31(1), 2–13.* https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100,* 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Wiseman, C. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *International Journal of Language Testing, 2*(1), 59-92.

Xu, J., & Zhang, S. (2022). Understanding AWE feedback and English writing of learners with different proficiency levels in an EFL classroom: A sociocultural perspective. *The Asia-Pacific Education Researcher, 31,* 357–367. https://doi.org/10.1007/s40299-021-00577-7

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

Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities and Social Sciences Communications, 11*(1), 1-14. https://doi.org/10.1057/s41599-024-03543-y

Yancey, K., Lunsford, A., McDonald, J., Moran, C., Neal, M., Pryor, C., Roen, D., & Selfe, C. (2004). CCCC position statement on teaching, learning, and assessing writing in digital environments. *College Composition and Communication, 55*(4), 785 –790. https://doi.org/10.58680/ccc20042788

Yang, W., & Kim, Y. (2020). The effect of topic familiarity on the complexity, accuracy, and fluency of second language writing. *Applied Linguistics Review, 11*(1), 79–108. https://doi.org/10.1515/applirev-2017-0017

Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System, 66,* 130–141. https://doi.org/10.1016/j.system.2017.03.007

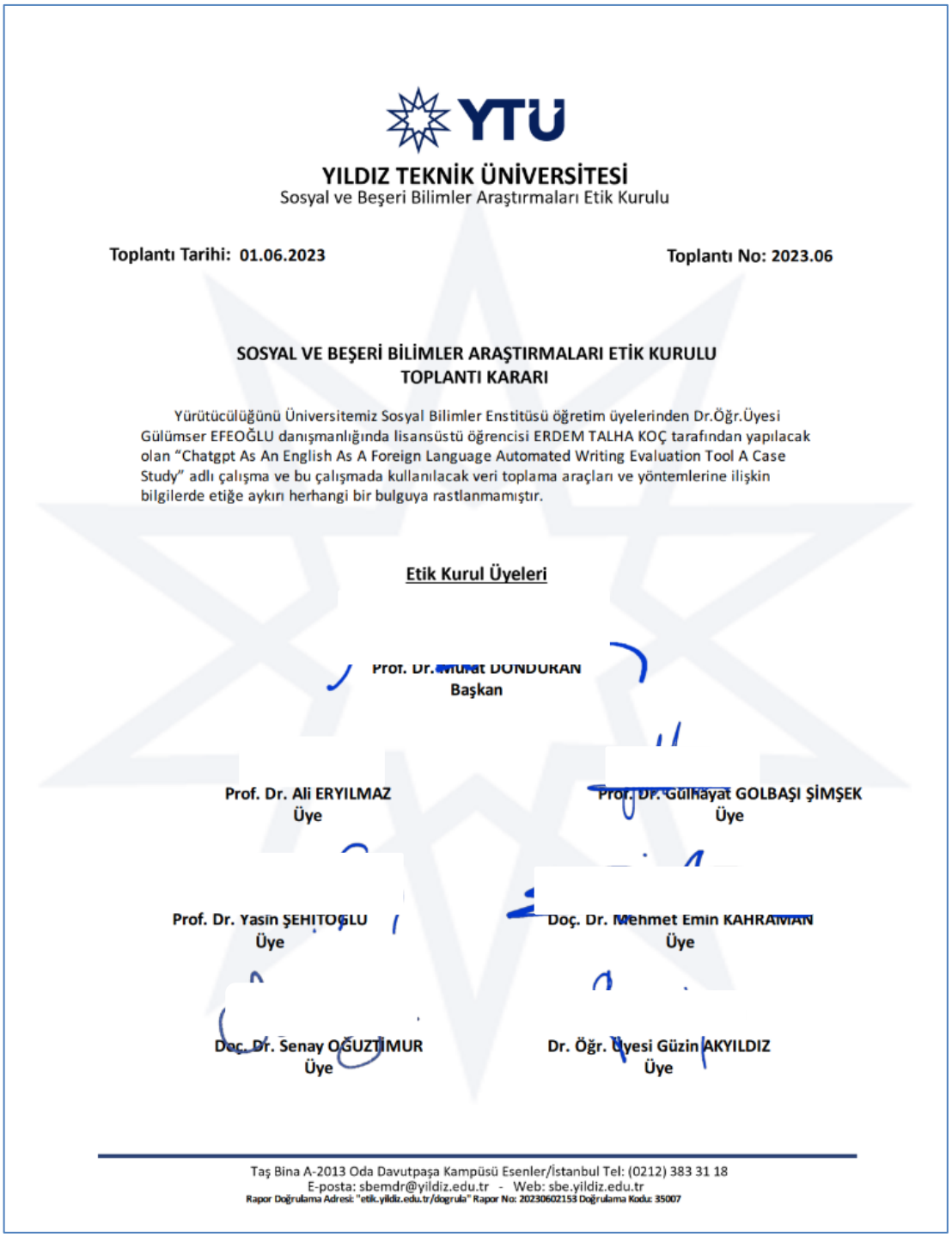
Yüksel, H., & Gündüz, N. (2017). Formative and summative assessment in higher education: Opinions and practices of instructors. *European Journal of Education Studies, 3*(8), 336–356. https://doi.org/10.5281/zenodo.832999

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36,* 90–102. https://doi.org/10.1016/j.asw.2018.02.004

Zou, M., & Huang, L. (2024). The impact of ChatGPT on L2 writing and expected responses: Voice from doctoral students. *Education and Information Technologies, 29*(11), 13201–13219. https://doi.org/10.1007/s10639-023-12397-x

APPENDICES

Appendix 1. Ethics Committee Approval



YILDIZ TEKNİK ÜNİVERSİTESİ
Sosyal ve Beşeri Bilimler Araştırmaları Etik Kurulu

Toplantı Tarihi: 01.06.2023 **Toplantı No: 2023.06**

SOSYAL VE BEŞERİ BİLİMLER ARAŞTIRMALARI ETİK KURULU TOPLANTI KARARI

Yürütücülüğünü Üniversitemiz Sosyal Bilimler Enstitüsü öğretim üyelerinden Dr.Öğr.Üyesi Gülümser EFEOĞLU danışmanlığında lisansüstü öğrencisi ERDEM TALHA KOÇ tarafından yapılacak olan "Chatgpt As An English As A Foreign Language Automated Writing Evaluation Tool A Case Study" adlı çalışma ve bu çalışmada kullanılacak veri toplama araçları ve yöntemlerine ilişkin bilgilerde etiğe aykırı herhangi bir bulguya rastlanmamıştır.

Etik Kurul Üyeleri

Prof. Dr. Murat DÜNDAR
Başkan

Prof. Dr. Ali ERYILMAZ
Üye

Prof. Dr. Gülhayat GOLBAŞI ŞİMŞEK
Üye

Prof. Dr. Yasin ŞEHİTOĞLU
Üye

Doç. Dr. Mehmet Emin KAHRAMAN
Üye

Doç. Dr. Senay OĞUZTİMUR
Üye

Dr. Öğr. Üyesi Güzin AKYILDIZ
Üye

Taş Bina A-2013 Oda Davutpaşa Kampüsü Esenler/İstanbul Tel: (0212) 383 31 18
E-posta: sbemdr@yildiz.edu.tr - Web: sbe.yildiz.edu.tr
Rapor Doğrulama Adresi: "etik.yildiz.edu.tr/dogrula" Rapor No: 20230602153 Doğrulama Kodu: 35007

Appendix 2. Focus Group Interview Questions

1- How does knowing that FB2 was ai-generated, rather than instructor-generated, impact your feelings towards the feedback you received? Why?

2- Do you feel the ai-feedback was less, equally, or more valuable than the feedback from your instructor? Why?

3- How does this revelation influence your trust in the feedback process? Do you feel more or less motivated to write in the future?

4- Given that FB2 was not written by your instructor, would this change how seriously you take the feedback or how you would use it to revise your composition?

5- How do you feel about your instructor not reading your essays? Does this affect your motivation or engagement with the class?

6- Would you have preferred to know that the feedback was coming from an ai from the beginning? Why or why not?

7- Does knowing that the detailed feedback came from chatgpt rather than your instructor change how motivated you are to use this feedback?

8- How did you feel when you received the detailed feedback? How do you feel now knowing that it was actually written by chatgpt?

9- Would you feel more or less motivated to write and improve your writing skills if you knew from the start that you wouldn't be receiving any feedback from your instructor? Why?

10- Would you prefer to receive instructor feedback or peer feedback? Why?

11-Would you prefer to receive instructor feedback or AI feedback? Why?

12- Would you prefer to receive oral or written feedback? Why?

13- What do you think about AI, its place in classroom, and its relation to teachers? When you become teachers, would you choose it to give feedback?

Appendix 3. Informed Consent Form

BİLGİLENDİRİLMİŞ OLUR FORMU

1. Aşağıda imzası olan ben ‘Chatgpt as an English as a Second Langauge Writing Evaluation Tool? A Case Study’ başlıklı çalışmaya katılmayı kabul ediyorum.

2. Bu çalışmayı yürüten Erdem Talha Koç çalışmanın yapısı, amacı ve muhtemel süresi, hakkında ayrıntılı sözlü ve yazılı bilgi verdi.

3. Araştırmacı Erdem Talha Koç'a çalışmasıyla ilgili her soruyu sorma fırsatını buldum. Cevapları ve bana verilen bilgiyi anladım.

4. Araştırmacı Erdem Talha Koç'a bilgilerin ayrıntılarını açıklamama ve benimle ilgili sırları koruması şartıyla benimle bu çalışmayı yapmasına izin veriyorum.

5. Çalışma boyunca tüm kurallara uyacağıma, araştırmacı Erdem Talha Koç ile tam bir uyum içinde çalışacağıma ve konuyla ilgili herhangi bir sorun çıktığında hemen onu arayacağımı kabul ediyorum.

6. Bu çalışma sonuçlarının kullanılmasını kısıtlamayacağımı ve yayın, rapor ve benzeri bilimsel dokümanlarda kullanabileceğini kabul ediyorum.

7. Bu çalışmadan istediğim zaman çıkabileceğimi anladım.

OKUDUM VE ONAYLADIM.

Katılımcının Adı ve Soyadı, Adresi: Tarih, İmza:

Araştırmacının Adı ve Soyadı, Adresi: Tarih, İmza:

Appendix 4. Informed Consent Form – 2

DENEKLERİN GÖNÜLLÜLÜĞÜ

VE

AYDINLATILMIŞ ONAM FORMU

Erdem Talha Koç'un Yıldız Teknik Üniversitesi İngiliz Dili Eğitimi Yüksek Lisans programını tamamlamak için Yıldız Teknik Üniversitesi Yabancı Diller Eğitimi Bölümü'nde yürüteceği 'Chatgpt as an English as a Foreign Language Automated Writing Evaluation Tool? A Case Study' isimli, 1 yıl sürecek olan çalışma için katılımcılardan yazılı ve sözlü veri toplanacağı, toplanacak olan yazılı ve sözlü verilerin yalnızca araştırma amacı ile kullanılacak olup bu verilerin çalışmanın hiçbir aşamasında üçüncü taraflar ile paylaşılmayacağı konusunda bilgilendirildim. Katılımcılar üzerinde herhangi bir olumsuz etkisi bulunmayan, istediğim zaman ayrılabileceğim bildirilen bu çalışmaya katılmayı kabul ediyorum.

OKUDUM VE ONAYLADIM.

Katılımcının Adı ve Soyadı, Adresi: Tarih, İmza:

Araştırmacının Adı ve Soyadı, Adresi: Tarih, İmza:

Appendix 5. Demographic Form Question Samples

1- How long have you been studying English as a second language?

a- 1-3 years

b- 3-5 years

c- 5+ years

2- Have you ever taken any standardized English tests, such as TOEFL or IELTS? If so, what were your scores?

…….

3- Have you ever taken any courses specifically focused on writing in English as a second language?

a- Yes

b- No

4- What do you think are the most challenging aspects of writing in English as a second language?

……

5- How often do you use technology (e.g., Word processing software, grammar checkers) to help you write in English?

a- Always

b- Often

c- Sometimes

d- Rarely

e-Never