Human Raters vs Chat GPT: How Do Their Scorings of EFL Learners' Academic Writing Differ?

Valentina Zapata Villano

Supervisor

Dr. María Eugenia Guapacha Chamorro

A research work submitted in fulfillment of the partial requirements for the degree of Bachelor's in Foreign Languages in English and French
Universidad del Valle
2025

Dedication and Acknowledgment

I want to dedicate this thesis to my mother Martha and to my cat Ramona. They were very important in my life and helped me become who I am, I truly remember them with love and gratitude. I also dedicate this work to my father Diego, who always believed in me and motivated me to keep going when life was hard. To my partner Anderson, thank you for being with me during this time, your support and motivation gave me strength and hope in the most difficult moments.

My deepest appreciation to my classmates and friends from the program. Your ideas, energy, and friendship helped me a lot, working with you in group projects and our discussions in class made this journey special and something I will always remember. I also thank “Los Más Porras Del Valle” scholarship, their financial help gave me the chance to continue my education and achieve this goal.

I want to express my gratitude to Universidad del Valle and the School of Language Sciences for receiving me and helping me grow and improve during these years. Here, I learned many things, evolved as a person and as a future professional; every teacher I had gave me knowledge, support, inspiration and ideas that helped me become the teacher I aspire to be.

Finally, I want to express my heartfelt appreciation to my thesis advisor, Dr. María Eugenia Guapacha. Your patience, empathy, and commitment guided me with generosity and care. Our long tutoring sessions, filled with shared laughter and learning, were essential for the development of this work. Your support and trust made this project possible, and you will always remain in my heart as a source of inspiration in my journey as a teacher.

Table of Contents

Dedication and Acknowledgment **2**
Table of Contents **3**
List of Tables **5**
List of Figures **6**
Abstract **7**
1 Introduction **8**
1.1 Chapter overview 8
1.2 Artificial intelligence in EFL writing assessment 8
1.3 Research question 11
1.4 Description of the context 11
1.5 Significance and contribution of the study 12
1.6 Outline 14
2 Literature Review **15**
2.1 Chapter overview 15
2.2 Empirical research on AI and EFL writing assessment 15
2.2.1 Previous research about the use of AI in EFL writing assessment 15
2.2.2 Previous research about Human Raters' vs AI assessment 16
2.2.3 Previous research on writing assessments in our context 17
2.3 Conceptual Framework 18
2.3.1 Artificial Intelligence 18
2.3.2 EFL Writing 18
2.3.3 EFL Writing Assessment and Fairness 19
2.4 Concluding Remarks 24
3 Research Methodology **25**
3.1 Chapter overview 25
3.2 Research approach and design 25
3.3 Research context 25
3.4 Data collection methods and procedures 27
3.4.1 Writing tasks 27
3.4.2 Writing scorings 28
3.4.3 Analytic scoring rubric 30
3.4.4 Instructions for generating the ChatGPT' scorings 31
3.5 Data analysis 34

3.6 Validity and reliability ... 35
3.7 Ethical considerations ... 36
4 Results ... 37
4.1 Chapter overview ... 37
4.2 Global scorings between ChatGPT and Human Raters ... 37
4.2.1 Descriptive statistics: Identifying trends in global scoring between human raters and ChatGPT ... 37
4.2.2 Inferential statistics: Identifying the reliability of raters across tasks and cohorts ... 40
4.2.3 Inferential statistics: Comparing global ratings across raters, tasks and cohorts ... 40
4.3 Analytic scorings between ChatGPT and Human Raters ... 41
4.3.1 Descriptive statistics: Identifying trends in analytic scoring between human raters and ChatGPT ... 41
4.3.2 Inferential statistics: Identifying the reliability of raters across tasks and cohorts ... 46
4.3.3 Inferential statistics: Comparing analytic ratings across raters, tasks and cohorts ... 48
4.4 Summary of key findings ... 51
5 Discussion ... 53
5.1 Chapter overview ... 53
5.2 Global scoring agreement between Human Raters and ChatGPT ... 53
5.3 Analytic scoring alignment between Human Raters and ChatGPT ... 54
5.4 A balanced integration of AI as a complement in EFL writing assessment ... 55
6 Conclusions ... 57
6.1 Summary of findings and conclusion ... 57
6.2 Broader impact and local contribution ... 57
6.3 Limitations of the study ... 57
6.4 Recommendations for Future Research ... 58
References ... 59
Appendices ... 64
Appendix 1. Informed Consent ... 64
Appendix 2. Analytic Rubric Task 1 ... 66
Appendix 3. Assumption Checks ... 68

List of Tables

Table 3.1 Scoring sample used in this study 29

Table 4.1 Descriptive statistics of human and ChatGPT' global scorings - Task 1 across cohorts 37

Table 4.2 Descriptive statistics of human and ChatGPT' global scorings - Task 2 across cohorts 38

Table 4.3 Inter-rater agreement on global scorings Tasks 1 & 2 40

Table 4.4 Comparison of global scorings across raters, cohorts and tasks 40

Table 4.5 Descriptive statistics of human and ChatGPT' analytic scorings - Task 1 across cohorts . 42

Table 4.6 Descriptive statistics of human and ChatGPT' analytic scorings - Task 2 across cohorts . 44

Table 4.7 Inter-rater agreement on analytic scorings Task 1 46

Table 4.8 Inter-rater agreement on analytic scorings Task 2 47

Table 4.9 Comparison of analytic scorings across raters and cohorts in Task 1 48

Table 4.10 Comparison of analytic scorings across raters and cohorts in Task 2 50

List of Figures

Figure 3.1 Bar chart created by ChatGPT 28

Figure 3.2 A sample of the output produced by ChatGPT 33

Figure 3.3 ChatGPT and human raters' scorings input into Excel 34

Figure 3.4 Data analysis on JASP 34

Figure 4.1 Humans and ChatGPT' global scorings distribution in Task 1 across cohorts 38

Figure 4.2 Humans and ChatGPT' global scorings distribution in Task 2 across cohorts 39

Abstract

Artificial Intelligence (AI) in language assessment has gained attention due to its potential to minimize teachers' complex task of assessing students' writing. Although previous research has explored the use of technological tools to assess EFL learners' writing, there is a need to further investigate how AI, particularly ChatGPT, can be used as an assessment tool in high-stakes writing assessment, and whether the scores provided by the AI are similar to those assigned by human raters. In the context of a high-stakes writing test at Universidad del Valle, this quantitative research investigated how the scorings of EFL university teachers differed from those given by ChatGPT when assessing EFL university learners' written productions (personal opinion essay and data explanatory essay). Two argumentative writing tasks from two cohorts of ninth-semester EFL learners (*N*= 208) were used to compare the global and analytic ratings awarded by a pool of 20 human raters with those of ChatGPT using an analytic scoring rubric. The analytic dimensions included content, coherence and cohesion, sentence structure, grammar, and vocabulary. A total of 7072 scores were analyzed, including 416 global (208 human and 208 ChatGPT) and 6656 analytic scores across two writing tasks. Analytic scoring covered seven criteria for Task 1 and nine for Task 2, with both human raters and ChatGPT providing an equal number of ratings (3328 each). Data were analyzed on JASP, by using correlations and paired-samples *t*-test. According to the results, ChatGPT showed moderate agreement with human raters in surface-level dimensions such as grammar, vocabulary, and sentence structure. However, larger and statistically significant differences were found in discourse-level categories like content development, coherence, and task achievement, especially in the second task. Descriptive statistics and t-tests confirmed that while ChatGPT produced similar scores to human raters in formal language aspects, it differed notably in areas that involved discourse-level skills. It can be concluded that AI can assist in writing assessment, particularly for language accuracy, but cannot fully replace human raters in EFL writing assessments. This study contributes to the understanding of the effectiveness and reliability of AI-based assessment tools in the context of English writing proficiency tests in Colombian higher education. It addresses questions about alignment and differences between human and AI evaluations, potential biases, and areas where AI may have limitations compared to human raters. The findings have implications for educational institutions seeking to improve assessment processes. Additionally, the study contributes to the ongoing discussion about the role of AI in education and writing assessment.

Keywords: *AI- education, AI-based writing assessment, ChatGPT, EFL writing assessment, human raters*

1 Introduction

1.1 Chapter overview

This chapter introduces the growing role of artificial intelligence in educational settings, particularly within the context of EFL writing assessment. It reviews key developments in the use of AI tools for evaluating student writing, identifies both global and local research gaps, and presents the rationale and research questions that guide the present study. By combining research from other countries with the specific case of a high-stakes proficiency test at Universidad del Valle in Colombia, this chapter sets the foundations for examining both the potential and the limitations of ChatGPT-4 as a tool for writing assessment in higher education due to its growing use in academic contexts and its advanced generative capacity.

1.2 Artificial intelligence in EFL writing assessment

In recent years, artificial intelligence (AI) has become a key part of teaching, learning and assessment. AI refers to computer systems that can perform tasks similar to human thinking, such as learning, reasoning, making decisions, and processing language (Russell & Norvig, 2010). These technologies are now used in many fields, including education, and they are changing how teaching is done and how student work is evaluated (Chaudhry & Kazim, 2021). Through tools that rely on machine learning and natural language processing (NLP), AI has become particularly prominent in tasks related to writing assessment.

Automated Writing Evaluation (AWE) systems, such as ETS's e-rater and Criterion, Project Essay Grade (PEG), and SmartMarq, use AI to assess written work by checking grammar, vocabulary, sentence structure, and overall organization. These tools not only generate immediate feedback but also help teachers manage large volumes of student writing by functioning as secondary evaluators (ETS, 2023). Some platforms go beyond simple scoring and offer support for the full writing process, planning, drafting, and revision, by providing real-time suggestions and criterion-based evaluations (Criterion.ets, n.d.; OER Project, 2023).

Although these systems have shown potential to improve the clarity and flow of students' writing studies also highlight significant challenges. For example, AWE tools often have weaknesses when it comes to evaluating the depth of content, creativity, or the development of arguments (Association for Computational Linguistics, 2020). Klebanov and Madnani (2020) note that these systems usually focus on surface-level issues, such as spelling and grammar, while overlooking higher-level aspects like coherence, rhetorical structure, and audience awareness taking into account that the main research gap addressed in this study is not focused on formal writing but rather on comparing the

scoring accuracy an consistency between human raters and ChatGPT. Van Dijk's (1980) work on the small details (microstructures) and overall meaning (macrostructures) of texts helps explain why this happens. Bronckart (1999) also shows that writing involves thinking and social interaction, which is hard for AI to judge

Marzuki et al. (2023) found that these tools can help learning by improving students' fluency and clarity in writing yet there is a risk that students may rely too much on them, which could slow the development of their critical thinking and problem-solving skills. In addition, as Klebanov and Madnani (2020) pointed out, AI tools may encourage only surface-level changes if they are used without proper teaching guidance.

Several studies have examined how closely automated scoring matches human evaluation. For example, Almusharraf and Alotaibi (2023) compared Grammarly's scoring with that of human raters in EFL contexts and found a positve moderate correlation. Grammarly was better at spotting basic errors, but human raters gave higher overall scores. In another study, Mizumoto and Eguchi (2023) carried out a comprehensive study using over 12,000 essays to assess how well GPT-3 can automatically score written productions. They used scoring guidelines from the IELTS Writing Task 2 rubric and compared the AI's scores to established benchmark scores. The results showed that GPT-3 was very consistent and reliable when evaluating linguistic features such as grammar and vocabulary. Also, the model was able to give steady scores when grading the same essay multiple times. Despite this strong agreement with benchmark ratings, the authors highlight the importance of human involvement in the assessment process, particularly for complex writing criteria like argumentation and overall coherence. Their findings suggest that while AI can be a useful tool for scoring essays, it should not completely replace human reviewers

Sasaki, et al. (2024) studied how well ChatGPT-4 can assess grammar in second language writing. They compared the scores given by ChatGPT-4 with those from human raters using L2 writing samples. The agreement between the AI and human scores was moderate with a reliability measure (ICC) around 0.6 to 0.7. ChatGPT-4 was fairly consistent in finding basic grammar mistakes. The research also showed that ChatGPT-4 had more trouble with broader parts of writing samples, like making sure the task was fully completed and organizing ideas clearly. These parts had lower agreement between AI and humans, highlighting challenges for the AI in understanding overall task achievement, meaning and structure.

The authors suggest that although ChatGPT-4 can provide helpful feedback on grammar, it should not replace teachers' judgment, especially in difficult or high-stakes writing tasks that require deep understanding of context and critical thinking. They recommend using AI feedback along with teacher supervision to get better results in writing assessment.

Kolade, et al. (2023) studied the use of ChatGPT in higher education, focusing on its potential and limitations for academic assessment. They identified significant limitations based on their analysis and previous research, including work by Rudolph et al. (2023). These limitations include occasional imprecisions in feedback, difficulties in citing academic literature properly and challenges in evaluating complex reasoning and deeper rhetorical skills. Their findings revealed that although ChatGPT generates well-formed content, it lacks the critical and complex depth required for fair assessment. The authors highlight that using ChatGPT for academic evaluation without human teachers' supervision can lead to unfairness and bias since the tool may not capture the full complexity of student performance and may provide automated responses without educational context. In summary, they emphasize the crucial role of educators in supervising carefully the ethical use of this technology in academic assessment

Although artificial intelligence research is expanding worldwide, it is important to recognize that AI use and development are regulated differently across countries. According to Modulos AI (2025), "the regulation of artificial intelligence varies significantly by region, creating a complex landscape of requirements and limitations that affect the implementation and development of AI technologies worldwide." For instance, while the European Union has very strict regulations through its AI Act, other countries like China and Brazil have different regulatory approaches and interest. This regulation differences forces research centers to adapt and limit the free global operation of certain AI systems and restrict their availability, especially those developed in the United States. (Modulos AI, 2025)

Guapacha and Chaves (2023) conducted a qualitative synthesis of English as a Foreign Language (EFL) writing studies in Colombia between 1990 and 2020, analyzing 63 research reports published in official Colombian journals. This review evidenced that most studies used qualitative methods and focused on writing teaching and pedagogy. They found a significant lack of research on evaluation, especially on automated scoring using artificial intelligence (Guapacha Chamorro & Chaves Varón, 2023). Since 2020, research on artificial intelligence in educational assessment settings has grown in Latin America, highlighting its potential and weaknesses. For example, Chuang (2025) discusses the rapid advances in generative AI technologies and their transformative effects on language assessment while Latin America is still emerging in this field. The Inter-American Development Bank (IDB, 2023) also promotes AI solutions in Latin American countries to improve educational quality but recognizes infrastructural and training weaknesses. These studies indicate that although progress is being made, there is still a wider opportunity for more systematic research on AI-assisted assessment, particularly in writing evaluation in Latin American contexts.

The Foreign Languages program at Universidad del Valle was chosen for this study as it represents a rigorous academic context in Colombia. Students are required to pass a C1-level English proficiency exam in order to graduate. This important requisite includes a writing section scored by two independent human raters using analytic rubrics that assess multiple dimensions of writing (Universidad del Valle, 2025; Departamento de Lenguas y Culturas Extranjeras, Universidad del Valle, 2025). This context aligns with Colombia's national language education policies, such as the Common European Framework of Reference for Languages (CEFR), assuring standardized proficiency levels (Ministerio de Educación Nacional de Colombia [MEN], 2016.

This research compares the global and analytic scores awarded by ChatGPT and EFL teachers to argumentative essays written by senior students in a high-stakes proficiency exam. The study used the data already collected by the academic program regarding texts and scores and generates scorings with AI using the texts and the rubrics used by the program to assess students' writing. By focusing on the local academic setting and using empirical data, this study offers practical insights into how AI can be responsibly integrated into language assessment. It also seeks to evaluate whether ChatGPT can function as a reliable companion in the scoring process and how its outputs compare to teacher judgment across different scoring dimensions.

The study further aims to identify the strengths and limitations of ChatGPT as an assessment tool in terms of consistency, fairness, and accuracy. It explores whether ChatGPT's ratings reflect human evaluators' judgments on elements such as content, coherence, organization, grammar, vocabulary, and sentence structure. The following research question was posed.

1.3 Research question

How do global and analytic scorings awarded by EFL university teachers differ from those awarded by ChatGPT when assessing EFL university learners' writing performance?

1.4 Description of the context

This study was conducted in the Bachelor's Program in Foreign Languages (English and French) at Universidad del Valle, located in Cali, Colombia. As part of the curriculum, ninth-semester students, as advanced learners, are required to take an English language proficiency test in which they must demonstrate C1-level competence (Universidad del Valle, 2021), as defined by the Common European Framework of Reference for Languages (CEFR) (Common European Framework of reference for Languages, 2002). A central component of this exam is the written production section, where students are given 90 minutes to complete two academic writing tasks. Task 1 asks students to write a personal

opinion essay on a current topic, while Task 2 requires them to describe, interpret, and critically discuss data presented in a statistical graph. Each text is evaluated independently by two trained raters using an analytic scoring rubric.

According to the exam protocol and as defined by Shaw and Weir (2007), the construct of writing proficiency encompasses not only language skills but also cognitive processes such as planning, drafting, monitoring, and revising. The exam is designed to assess the students' ability to produce clear, academically appropriate texts, such as reports and argumentative essays, within a limited time frame and for an academic audience composed of professors or advanced readers (Hyland, 2019).

The assessment criteria also emphasize the students' capacity to present a well-supported position on familiar topics, develop coherent arguments, and incorporate relevant examples. In the case of data interpretation, students must be able to identify and explain key trends or ideas reflected in visual input. The written output is expected to demonstrate cohesion, lexical control, syntactic accuracy, and appropriate use of academic register, punctuation, and spelling (Escuela de Ciencias del Lenguaje, Universidad del Valle, 2021).

Importantly, as outlined in *Proyecto Educativo Resoluciones 056 y 062* for cohorts admitted from 2018 onward, academic writing development is addressed across multiple courses within the program. These include *Oral and Written Academic Genres I and II*, *Composition in English and French*, and *English and French Literature*. Additionally, writing instruction is embedded across the curriculum, with purposeful writing tasks integrated into both coursework and extracurricular activities. For instance, in the *Life Narratives* course, students engage in reflective writing throughout the semester by composing a personal autobiography. In other courses, they produce essays, literary analyses, book reviews, and other types of academic writing. This ongoing engagement with writing aims to strengthen students' critical thinking, textual organization, and communicative competence in academic and professional contexts. Students' writing performance is usually assessed by the lecturers of the academic program who are also the raters of students' English writing proficiency test.

Academic writing is an essential part of this program and is directly linked to graduation requirements. For this reason, it is important to examine current practices in writing assessment. This study focuses on how AI tools, especially ChatGPT, can be integrated into the assessment process, using these tools could help make assessment more consistent and objective, while also reducing both rater bias and teacher workload.

1.5 Significance and contribution of the study

This study offers a contribution to the growing field of artificial intelligence in language assessment, specifically within the context of writing assessment in an EFL setting. By exploring the

use of ChatGPT as a potential writing assessment tool, the research sheds light on how emerging technologies can support and complement human judgment in high-stakes academic evaluations. It adds to ongoing international discussions around the viability, fairness, and limitations of machine-assisted assessment, and introduces empirical findings from a local educational setting where this topic remains largely unexplored.

The study adopts a quantitative, correlational and comparative approach to examine the similarities and differences between the scores assigned by human raters and those generated by ChatGPT. Through statistical analysis, including intra-class correlation coefficients and paired-samples *t*-tests, the research identifies levels of agreement across global and analytic scoring dimensions. This comparison not only contributes to the validation of AI-assisted scoring but also opens avenues for future studies interested in replicating or extending the methodology in other EFL contexts.

From a professional standpoint, the study also represents a learning opportunity for me as a researcher, who, as a future language teacher and academic, gains valuable insights into the pedagogical use of AI for evaluating student writing. It demonstrates how digital tools can be critically analyzed, adapted, and integrated into teaching practice with the aim of enhancing objectivity, reducing workload, and providing timely feedback.

By testing an AI model against existing evaluation practices, the study proposes realistic strategies to improve assessment consistency and to support teacher decision-making. It also offers pedagogical recommendations that can guide rubric calibration, rater training, and the responsible use of AI-generated feedback.

Importantly, the study is grounded in data generated within the local academic community. It relies on authentic student texts and scores produced by qualified university instructors, making the findings directly applicable to the institutional context from which they emerged. This local relevance not only strengthens the study's practical value but also contributes new evidence to a national research landscape where studies on AI in writing assessment remain scarce.

To date, this is the only known study conducted within the Foreign Languages program that directly investigates the use of ChatGPT as a scoring companion in high-stakes academic writing. As such, it provides a pioneering foundation for future inquiry in the region and serves as an example of how local data and global technological advancements can be brought together to inform educational practice and innovation.

1.6 Outline

This thesis is organized into six main chapters that guide the reader through the entire research process. Chapter 1 introduces the study by presenting its background, purpose, the gap, research questions, institutional context, and significance. Chapter 2 reviews relevant literature on artificial intelligence in EFL writing assessment, including comparisons between human and AI scoring, conceptual framework and studies from the local and international contexts. Chapter 3 describes the research methodology, including the study site, data collection and analysis procedures. Chapter 4 presents the results related to both global and analytic scorings, offering descriptive and inferential analyses across tasks and cohorts. Finally, Chapter 5 summarizes and discusses the main findings and their implications. Chapter 6 concludes the study, reflects on ethical and pedagogical considerations, and provides recommendations for future research. The thesis includes references and appendices, which include assumption checks, statistical outputs, visual representations, and the analytic rubrics used that support the study's transparency and methodological rigor.

2 Literature Review

2.1 Chapter overview

This chapter reviews the existing research and theoretical foundations relevant to this study, including artificial intelligence (AI) in the evaluation of EFL writing, with particular focus on how these technologies have been incorporated into language instruction, learning, and assessment. While numerous studies have emerged at the international level, especially regarding the use of Automated Writing Evaluation (AWE) systems, scholarly work in the Colombian context remains limited. This gap underscores the importance of exploring how AI may be applied to language assessment in underrepresented educational settings. Along with the empirical findings, this chapter includes a section that explains the concepts supporting the present research. The conceptual framework is based on writing pedagogy, cognitive and sociocultural perspectives, and the changing role of AI in academic evaluation; providing the lens used to interpret the results of this study.

2.2 Empirical research on AI and EFL writing assessment

2.2.1 Previous research about the use of AI in EFL writing assessment

Initially, Zhang (2021) carried out a theoretical review of Automated Writing Evaluation (AWE) systems, explaining how they operate, how they are designed, and what their implications are for teaching and learning English as a Foreign Language (EFL). The review recognized that these tools can make feedback easier and are helpful in teaching, especially for large-scale tests, Zhang also warned that too much dependence on them could harm the development of advanced writing skills. To understand the long-term effects of AWE in language learning, the author suggested doing long-term studies to see how repeated use of these tools affects students' writing over time.

Rudolph et al. (2023) reviewed how advanced AI tools, especially ChatGPT, may affect assessment in higher education. Their work combined a literature review with experiments to see how the tool influences both teaching and learning, as well as issues linked to academic honesty. The results showed that ChatGPT can improve assessment by giving quick, personalized feedback and helping educators design evaluation tools. Even so, the study also found clear limits: ChatGPT had trouble citing academic sources correctly and could not process world events that happened after 2021. For this reason, the authors suggested that future research should focus on developing AI systems that are more aware of context and better adapted to academic use.

A further study tested how large language models (LLMs), especially ChatGPT, perform in evaluating written work across different types of language generation, such as summarizing, creating

dialogues, writing stories, and paraphrasing. Kolade et al. (2023) set three evaluation methods that did not need reference texts, they found that ChatGPT performed well, particularly in summarizing and paraphrasing. On the other hand, they saw that it often gave lower scores than expected, which they linked to the poor quality of some test data. The authors advised that future research should include a wider variety of writing tasks and compare ChatGPT with other LLMs using more representative samples.

2.2.2 Previous research about Human Raters' vs AI assessment

In a 2022 study in Technology, Knowledge and Learning, Almusharraf and Alotaibi reviewed at how human raters and Automated Essay Scoring (AES) systems, focusing on Grammarly, differ when evaluating the writing of EFL students. Applying Corder's Error Analysis framework, they examined 197 essays to compare how each rater type identified and responded to language errors. Their analysis showed that Grammarly detected a higher number of surface-level errors such as incorrect articles, misspellings, and punctuation mistakes, compared to human raters. Still, the software tended to assign lower overall scores and showed limitations in handling more complex aspects of writing, such as run-on sentences, possessive forms, and cohesive devices. The researchers concluded that although AES tools cannot fully replace human evaluators, particularly in assessing more subjective elements of writing, they can serve as valuable tools for formative feedback and self-revision.

Chen and Pan (2022) carried out a more recent investigation comparing teacher-generated feedback with that provided by an automated writing tool called Aim Writing. The study took place in a small private university located on the eastern coast of China and followed a quasi-experimental, mixed-methods design. Participants were split into two groups: one received traditional feedback from their instructors, and the other used the AWE tool. Afterwards the researchers analyzed the results using statistical tests like paired t-tests and Pearson correlations and also gathered qualitative insights through semi-structured interviews. The findings showed that both types of feedback helped improve students' writing.

In large-scale studies comparing human and AI scoring, researchers have reported promising consistency in automated evaluations. For example, Mizumoto and Eguchi (2023) studied the use of an AI language model to score over 12,000 essays written by non-native English speakers, using the ETS Corpus of Non-Native Written English. They found high inter-rater reliability, with intraclass correlation coefficients (ICCs) above 0.80 in multiple analytic categories. These results suggest that AI models like ChatGPT can match human scoring patterns with considerable accuracy when used with large and varied datasets, the researchers also noted that adding linguistic features to the scoring process can further improve accuracy and consistency.

More evidence of ChatGPT's potential in writing assessment comes from a recent study by Mizumoto et al. (2024), which tested how well the model evaluated grammatical accuracy in L2 writing. The results showed strong agreement with human ratings when identifying surface-level errors, such as grammar and sentence structure issues. However, the study also found that ChatGPT had trouble judging discourse-level aspects, like how ideas were organized and whether the task was fully completed. Also, the authors stressed that although ChatGPT can provide quick and consistent feedback, its use should always be paired with teacher oversight and a clear pedagogical framework to ensure fair and meaningful assessment.

2.2.3 Previous research on writing assessments in our context

Although there has been a growing international interest in writing assessment and the use of automated tools in this area, empirical research on these topics within Colombia, particularly in university-level EFL programs, is still very limited. Most studies conducted nationally have centered on teaching strategies and the improvement of students' writing skills, rather than on the processes and tools used to evaluate writing (Guapacha & Chaves, 2023).

In the context of the Foreign Languages program at Universidad del Valle, Guapacha and Chaves (2023) carried out a qualitative synthesis that examined research on EFL writing instruction published between 1990 and 2020. Their analysis showed that most of the academic projects focused on classroom interventions and teaching writing through genre-based approaches. However, only a small number of these studies explored formal assessment methods, and even fewer considered the role of scoring rubrics. Notably, the review found no empirical research that had investigated or compared human rater evaluations with the performance of automated writing tools.

The lack of studies in this area reveals a clear gap in the research, particularly regarding how writing is evaluated in high-stakes situations within Colombian EFL programs. This study aims to help fill that gap by offering what appears to be the first empirical comparison between human scoring and AI-generated evaluations in the context of a formal English proficiency exam for ninth-semester students. This institutional test, required for graduation, includes the assessment of academic writing using rubrics applied by two trained evaluators. Before this research, no systematic investigation had been conducted to explore whether AI tools, such as ChatGPT, could provide scores comparable to those assigned by experienced human raters.

By situating the investigation within an under-researched local context and focusing on a high-stakes assessment scenario, the present study contributes original evidence to the national literature on writing assessment. Furthermore, it provides data that can inform rubric calibration, rater training, and future adoption of AI tools for writing evaluation.

2.3 Conceptual Framework

The previous section explored how artificial intelligence is investigated in EFL writing assessment. Building on that foundation, the current chapter introduces the main concepts that inform this research, which include Artificial intelligence and EFL writing assessment.

2.3.1 Artificial Intelligence

According to the European Commission's High-Level Expert Group (2019), artificial intelligence refers to systems capable of interpreting their surroundings, processing information, and making independent decisions to accomplish specific objectives. McCarthy (1955), one of the first researchers in the field, described AI as the study of building machines that can show intelligent behavior. More recently, Sheikh et al. (2023) said that advances in machine learning and natural language processing (NLP) have improved AI's ability to handle complex tasks that involve human language.

In a recent study, Mizumoto et al. (2024) tested ChatGPT's ability to assess L2 writing accuracy, finding positive moderate correlations with human ratings: r = .85 for grammar and for coherence r = .79. Likewise, Mizumoto and Eguchi (2023) reported intraclass correlation coefficients (ICCs) higher than 0.80 in several analytic scoring features such as grammar and vocabulary. Also, the model was able to give steady scores when grading the same essay multiple times. showing that AI scoring can be similar to human scoring in some areas.

The rapid growth of AI tools has increased in language learning and assessment settings. ChatGPT is well-known for its ability to interact with users in a conversational and user-friendly manner (Gregersen, 2023). Since its release by OpenAI in November 2022, it has received mixed reactions from teachers, researchers, and media. Its ability to produce human-like texts, which are often hard to tell apart from those written by people, has caused interest and discussion. The Encyclopedia Britannica explains that ChatGPT can understand everyday language prompts and create answers that copy natural conversation and produce many types of content, from emails and articles to poetry, so it is a flexible tool for different needs (Gregersen, 2023). In this case, AI is used to identify whether their interpretations of EFL learners' writing quality are similar to those made by human raters.

2.3.2 EFL Writing

Over the years, EFL writing has been studied from different theoretical perspectives. At first, Hyland (2019) stands that it was perceived mainly as a product, focusing on the final text and the grammar accuracy. Later, this view shifted to a process-based approach, where writing is understood as a set of stages that can be repeated, rather than a single finished piece. This approach gives

importance to steps such as planning, drafting, revising, editing, and publishing, all of which help shape the meaning and structure of the text. (Hyland, 2019)

According to Martín-Martín (2013), the genre-based approach sees writing as a social activity influenced by the rules of specific text types used in academic or professional contexts. In this view, genres are patterns of communication that follow the expectations and style of certain groups or situations (Martín-Martín2013). In addition, Elbow (2015) proposed an expressive approach that values personal voice, creativity, and spontaneity, seeing writing as a way to express ideas and emotions. Another perspective comes from Britton and Bruner (2020), who present a social constructivist view, in which writing is shaped by social interactions, cultural background, and the intended audience. According to this perspective, learning to write involves actively constructing knowledge through experience, reflection, and engagement with others.

Fang (2022) defends that writing skills mean being able to express ideas, information, or arguments clearly in written form. This skill includes putting ideas in a logical order. It also means using correct grammar and punctuation, choosing good vocabulary, and changing tone or style depending on the audience or purpose. Fang also points out that writing should be clear, creative, coherent and well structured.

As Cumming (2016) explains, It is important to highlight that second language writing can be complex and different for each person, L2 writing is not just one simple skill; It changes depending on the learner, the type of text, and the situation. People write for many reasons in different cultural, academic, and work settings. This means that second language writing is shaped by personal goals, school rules, and the wider social context (Cumming, 2016).

In this study, writing is seen as a continuous process, a way of creating meaning. However, for the purposes of this study, it will focus on the product, which is the text. This study wants to see the texts received similar scores by humans and AI.

2.3.3 EFL Writing Assessment and Fairness

According to Cumming et al. (2019), assessing writing means looking at how well someone can produce texts that fit a certain situation, this includes seeing how they use their knowledge about topics they have learned through reading, listening, or classwork. Good writing tests give students clear instructions that match academic needs and are done in a fair time and in order to be fair and correct, the scoring also needs to be consistent and objective (Cumming et al., 2019)

Yi (2023) explains that writing assessment is a common practice in both international standardized tests and in secondary schools where English or languages' teachers play an important role in creating these evaluations, starting with a clear idea of what writing proficiency means.

Depending on the student's language level and the teaching context, teachers may use different assessment models. Yi (2023) highlights that defining writing ability clearly makes evaluation more effective; teachers need structured guidelines and resources that help them design consistent and meaningful assessments.

The need for accurate, fair, and consistent writing assessment creates a space where artificial intelligence can be a useful support. AI tools can speed up the scoring process and also help teachers improve their assessment practices in different types of tests, like achievement or diagnostic exams. Used in this way, AI can work with human raters to improve both efficiency and quality, while also helping to ensure fairness when comparing AI and human scoring.

In this study, the writing proficiency of students in the Bachelor's program in Foreign Languages at Universidad del Valle is measured through an institutional exam that is required for graduation. Most sections of this test use multiple-choice questions, but the writing section asks students to produce an original text. This is graded with an analytic rubric, and raters are trained to use it consistently. More details are given in the Methodology section.

Within this framework, artificial intelligence can function as a useful support tool by offering immediate evaluations and expediting the scoring process. This involves prompting the AI to apply the same criteria outlined in the rubric used by human evaluators. The study then compares the results from both sources to assess the degree of consistency between them. Ultimately, the goal is to determine whether AI-generated scores are comparable to those assigned by human raters, which could support the broader integration of AI in high-stakes writing assessment.

2.3.3.1 Scoring rubrics to assess writing

A rubric is a tool used in assessment to make expectations clear and provide organized, criteria-based feedback. According to Stevens and Levi (2023), rubrics describe specific performance levels for each aspect being evaluated, which helps teachers grade more accurately. They also serve as teaching tools by making the grading process easier for students to understand and more consistent among instructors. Rubrics are especially valuable in performance-based assessments, such as essays, projects, presentations, and oral tasks. They are also common in peer reviews, discussions, and group work, especially when the goal is to assess complex thinking skills (Stevens and Levi, 2023).

In second language teaching, there are different types of rubrics: holistic, analytic, and multi-trait (Brown & Abeywickrama, 2019; Weigle, 2002). A holistic rubric gives one overall score based on the general impression of the work, this is quick and useful for large scale testing but does not give detailed feedback. Analytic rubrics evaluate writing into separate areas like content, organization, grammar, vocabulary, and cohesion, so each gets its own score. This helps students see exactly where they are strong and where they need to improve. Multi-trait rubrics are similar to analytic ones but focus on specific genres or contexts; for example, when grading an argumentative essay, criteria might

include the clarity of the thesis, the quality of the evidence, the strength of the argument and other grammar, cohesion and sentence structure alike criteria.

In this study, an analytic rubric was used; it was designed and validated by lecturers in the program (see the Methodology chapter for more details). To check how well ChatGPT's scores aligned with human raters, both global and analytic scores were compared.

2.3.3.2 AI tools to assess writing

Advances in computer science and natural language processing (NLP) have led to the development of many Automatic Writing Evaluation (AWE) tools used in education today. These systems provide quick feedback on students' writing, helping them improve more efficiently, while also reducing teacher's workload by making the evaluation process faster and more consistent (Ramesh & Sanampudi, 2022).

Among the best-known AWE platforms are Project Essay Grade (PEG) by Measurement Inc., Criterion from Educational Testing Service (ETS), and MyAccess! by Vantage Learning. These platforms allow students to submit their work several times and receive immediate feedback after each submission. The comments usually cover grammar, vocabulary, punctuation, and spelling, and may also address other aspects like the clarity of the thesis statement or the overall organization of the text (Woodworth & Barkaoui, 2020).

In addition to identifying surface level errors, these systems often suggest ways to improve writing style, such as using active voice instead of passive or reducing redundancy. Each submission produces both individual component scores and an overall score for the text, this makes them useful for providing feedback during the learning process as well as for final evaluations.

NLP, as a branch of AI, enables computers to interpret and analyze human language (Verspoor & Cohen, 2013). Through computational methods, NLP identifies meaning and grammatical structure, allowing automated systems to evaluate and respond to writing in ways similar to human assessment.

In recent years, generative AI tools like ChatGPT, developed by OpenAI, have become increasingly relevant in writing assessment. Unlike earlier systems based on fixed rules, ChatGPT uses deep learning models to process, generate, and assess language in a way that closely resembles human expression. Studies show that ChatGPT can effectively detect grammar and sentence level errors while offering constructive grammar related suggestions (Mizumoto & Eguchi, 2023). It can also give feedback on other elements, such as idea organization and sentence structure (Mizumoto et al., 2024).

Nevertheless, researchers such as Sasaki et al. (2024) note that ChatGPT has limitations, especially when feedback requires deep reflection or analysis of discourse level features, which is often the case in complex academic or argumentative writing, when used accompanied by teacher guidance, it can be a valuable resource for formative assessment. As part of the newer generation of AWE tools, ChatGPT can deliver feedback on a large scale while adapting to individual learners. Yet, its use also

brings important discussions about fairness, reliability, and the continuing importance of human judgment in critical evaluation contexts. (Sasaki et al. 2024)

In this study, ChatGPT was selected to assess students' writing productions due to its foundation on the advanced GPT-3 language model, which has gained wide recognition for its ability to interact naturally and in a conversational way with users, this makes ChatGPT an accessible and effective tool familiar to both students and educators (Floridi & Chiriatti, 2020). This study used the GPT-4 version since teachers can provide specific prompts that guide the AI in assessing multiple essays considering the criteria prompted.

2.3.3.3 Reliability

In assessment, validity, reliability, and fairness are considered the three key principles. Reliability refers to how consistent test results are when measured on different occasions or by different evaluators. Sullivan (2011) points out that scoring differences can be caused by variations among raters, the way tasks are designed, or the testing conditions, all of which may affect the dependability of results.

There are several ways to define and measure reliability, Classical Test Theory (CTT) uses reliability coefficients to show how consistently a test measures what it is meant to measure, Generalizability Theory goes further by looking at how scores hold up across various conditions or rater groups, Item Response Theory (IRT) focuses on the accuracy of individual test items and their contribution to overall reliability (Miller, 2019).

Miller (2019) also notes that reliability is especially important in high-stakes tests, such as those that determine graduation from university, in these cases, even small errors in scoring can have serious consequences for both fairness and accuracy. To address this, the study examined the reliability of both global and analytic scores to determine if ChatGPT can be considered a consistent and dependable evaluator.

According to the Center on Standards and Assessment Implementation (2018) Reliability is often measured on a scale from 0 to 1, with values closer to 1 representing stronger consistency in scoring. Objective test formats like multiple-choice questions usually represent higher reliability because they involve minimal interpretation. On the other hand, writing tasks, being performance-based, require subjective judgment and therefore tend to show lower reliability unless raters are properly trained and calibrated (Center on Standards and Assessment Implementation, 2018). In this study, both holistic and analytic scores from ChatGPT and human raters were compared to examine the consistency and reliability of AI as a scoring tool in EFL writing assessment.

2.3.3.4 Rater reliability

Reliability relates to how consistently scores are produced across repeated evaluations or different raters(Shaw & Weir, 2007). As Sullivan (2011) points out, differences in evaluator judgment, task design, or testing conditions can all introduce errors that reduce both the accuracy and reliability of the results.

Researchers have approached the concept through several theoretical models (Miller, 2019). In Classical Test Theory (CTT) reliability is expressed as a statistical coefficient that indicates the degree of consistency in scores. Generalizability Theory goes further by exploring how stable those scores remain when raters, tasks, or testing environments change. Meanwhile, Item Response Theory (IRT) shifts the focus to the precision of individual test items and their contribution to the overall reliability of an assessment. In all cases, Miller (2019) emphasizes that reliability becomes especially important in high-stakes contexts like graduation exams, where even small errors can reduce fairness and credibility.

Reliability values are often placed on a scale from 0 to 1, with higher scores indicating greater consistency (Weir, 2005). Multiple-choice tests tend to reach higher values because they limit subjectivity. In contrast, performance-based tasks such as essay writing usually score lower since they rely on human interpretation. It is important to highlight that careful training and calibration of evaluators can help minimize this variability (Center on Standards and Assessment Implementation, 2018).

There are two main ways to examine rater reliability: inter-rater and intra-rater consistency (Weigle, 2002). Inter-rater reliability looks if different raters give similar scores to the same text, something that suggests the rubric is being applied fairly. Intra-rater reliability, on the other hand, checks if one rater is consistent with themselves over time. Both are important, especially for tasks that require interpretation, as Brown and Abeywickrama (2019) note, rubrics with clear and detailed performance levels can guide raters toward more consistent scoring.

In this study, the focus is on inter-rater reliability to examine how closely ChatGPT's scores align with those of trained human raters when assessing students' writing. To measure this, statistical tools such as intra-class correlation coefficients (ICC) and paired-samples t-tests are used. If the correlation is positive (close to 1.0) it would indicate that ChatGPT's scoring is consistent with human evaluation. However, large differences might suggest that the AI's scoring method or the rater training process needs improvement. In the same way, even if statistical comparisons turn out significant it gives useful clues about the comparability of human and AI ratings.

2.4 Concluding Remarks

The literature reviewed in this chapter shows the increasing interest in using artificial intelligence as a tool for language learning, especially for writing assessment. Tools such as ChatGPT have been shown to provide quick and personalized feedback, support both formative and summative evaluations and help reduce the workload for teachers.

At the same time, research has identified important limitations. Concerns remain about the fairness and accuracy of AI-generated feedback, its limited capacity to adapt to complex academic contexts, and the risk of bias or misinterpretation, especially when evaluating advanced writing skills such as argumentation, organization, and genre-specific criteria (Mizumoto et al., 2024). These challenges suggest that AI should complement rather than replace human judgment in writing evaluation. While global studies have explored comparisons between human and automated raters, few investigations have focused on high-stakes academic settings in Latin America. This gap reinforces the relevance of the current study, which seeks to understand how ChatGPT's scores align with those of trained human raters when evaluating the written performance of university-level EFL learners.

The conceptual framework discussed here integrates writing theory, assessment reliability, and artificial intelligence to support the study's design and interpretation. This framework allows for a comprehensive understanding of the factors that influence how writing is evaluated and how technology can be responsibly integrated into that process. The insights from this review contribute to ongoing conversations about the role of AI in language education.

3 Research Methodology

3.1 Chapter overview

This study aimed to compare the scorings awarded by Artificial Intelligence (ChatGPT) and human raters to texts produced by EFL university students. Such a study allows us to identify the affordances of AI for pedagogical and assessment purposes. This chapter describes the methodological procedures conducted to investigate the research phenomenon.

3.2 Research approach and design

The present study follows a quantitative research approach. According to Creswell and Creswell (2018), quantitative research aims to test objective hypotheses by examining the relationships between measurable variables. This study implemented a comparative and correlational design. The comparative section compares the scorings given by human raters and ChatGPT on the same student essays. This comparison helps examine whether there are differences or similarities in the scores assigned by human raters and ChatGPT (Fraenkel et al., 2019). The correlational section examines the level of agreement in the scores of human raters and AI. Correlational research does not establish cause and effect but shows the strength of the relationship between two variables (Creswell, 2014). By using both methods, the study provides a view of how ChatGPT scores agree with human ratings and whether ChatGPT can be a reliable appliance to support teachers in writing assessment in EFL settings (Mertler, 2016). While the study used pre-existing scores from human raters, it also generated new scores with ChatGPT for the same texts. These two sets of scores were then compared and correlated to examine agreement, variability, and consistency. This approach goes beyond a purely ex post facto design because the research created a secondary dataset for analytical purposes, rather than only analyzing data that already existed.

3.3 Research context

The School of Language Sciences at Universidad del Valle offers a Bachelor's in Foreign Languages with emphasis on English and French. This investigation was conducted in this context because the School of Language Sciences at Universidad del Valle regularly assesses the language proficiency of advanced learners, for whom academic writing is one indicator of communicative competence. Given the central role of writing assessment in the program, exploring how AI-generated scores align with human ratings provides valuable insights into the potential and limitations of AI tools in authentic, high-stakes academic settings. The program is designed to prepare language teachers with a critical social perspective, able to manage different communication situations in English and French.

Throughout the curriculum, students work on intercultural communication, pedagogical foundations, linguistic theory, and reflective teaching practice. They also develop skills in speaking, reading, writing, and listening, together with a solid understanding of the morphological, phonological, lexical, semantic, and syntactic aspects of language. The program follows constructivist, socio-critical and research-based principles and complies with national education standards (Universidad del Valle, Escuela de Ciencias del Lenguaje, 2018).

This is an important context to study, since the writing test is a graduation requirement and the results can have a real impact on the academic future of the students in this program. According to Escuela de Ciencias del Lenguaje, Universidad del Valle (2021) The writing performances also represent the outcome of several years of study in a nationally accredited program, which makes them valuable for analysis.

For this investigation, I used the texts, scorings and rubric already designed, calibrated and validated by the English section. The present study only generated AI scorings. The texts belonged to student writers from cohorts 2022 and 2023, and the scorings belonged to human raters and ChatGPT. The written texts were produced by ninth-semester students enrolled in the Bachelor's in Foreign Languages (English–French) program at the School of Language Sciences, Universidad del Valle. These students took the EFL writing component test, designed to assess their proficiency at an advanced (C1) level according to the CEFR. All students' texts were included in the study to ensure comprehensive coverage of performance variation and to avoid sampling bias in the comparison between human and AI scoring. The human raters consisted of 20 EFL lecturers from the same program, all of whom participated in a rater training process focused on the analytic and holistic criteria used for scoring. The raters were trained prior to this study by the English section for the assessment process. Their ratings provided the human benchmark for comparison with the scores generated by ChatGPT. The scorings underwent a reliability analysis to ensure that students received fair scores.

No sociodemographic variables were required for this study because the focus was not on individual learner characteristics, but on the comparison between human and AI scoring performance. The study aimed to analyze the alignment and consistency of scores, not to explore how variables such as age, gender, or socioeconomic background influence writing ability or assessment outcomes. Similarly, no inclusion or exclusion criteria were applied, since the research involved the complete set of texts produced by the ninth-semester students who took the EFL writing component test during the academic period under analysis. Including all available data ensured representativeness of the test population and avoided sampling bias, which was essential for a valid comparison between human and AI ratings.

The inclusion of two consecutive cohorts was a deliberate methodological choice aimed at examining the consistency of human and AI scoring performance across test administrations. Limiting the analysis to a single cohort would have prevented the identification of variability patterns and contextual influences that are crucial for understanding how AI and human ratings behave under authentic testing conditions. Both cohorts corresponded to advanced EFL students taking the same standardized writing component test, under identical scoring rubrics and rater training procedures.

All students gave their consent to use their texts, as they participated in a survey that collected information about their perceptions of writing proficiency. Their consent can be seen in Appendix 1.

3.4 Data collection methods and procedures

3.4.1 Writing tasks

The writing test tasks were designed and validated by the English Department. They aimed to elicit a range of writing competencies, including textual coherence, grammatical control, lexical accuracy, and the ability to interpret and argue complex ideas. To preserve the confidentiality of the writing test prompts here, I provide an example of similar prompts generated by ChatGPT, carefully modeled on the structure, genre, and cognitive demands of the official tasks. This approach does not compromise internal or external validity, as the linguistic and rhetorical features of the tasks were preserved, ensuring that the writing samples remained comparable in complexity and purpose.

3.4.1.1 Writing task 1: Personal opinion essay (prompt samples)

- **Prompt 1:** In recent years, some educators argue that schools should incorporate emotional intelligence and mental health education as core subjects to support students' well-being. Others believe that schools should prioritize academic subjects and leave emotional education to families and communities. Which perspective do you agree with? Write an essay presenting your viewpoint. Provide clear arguments and specific examples to support your position.
- **Prompt 2:** Environmental activism has gained global attention in the past decade. Some people believe that young climate activists are crucial in raising awareness and influencing political change. Others argue that their actions are disruptive and lack realistic solutions. Do you support youth-led climate activism? Write an essay stating your opinion, using relevant reasons and examples to support your stance
- **Prompt 3:** During public health emergencies, such as epidemics or natural disasters, governments often impose restrictions to control the situation. Some people believe such measures are necessary

to protect the population. Others argue that they violate individual freedoms and should be minimized. What is your opinion on this issue? Write an essay explaining your viewpoint, including specific reasons and real-world examples.

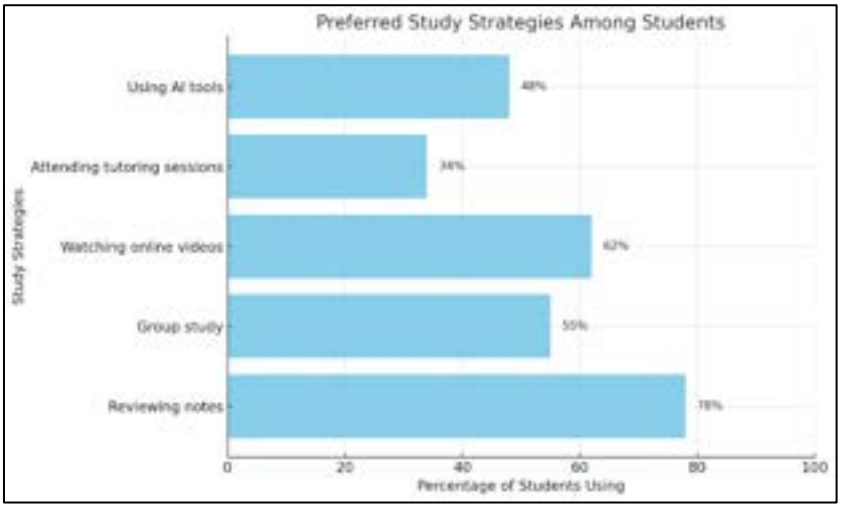
3.4.1.2 Writing Task 2: Data interpretation essay (prompt sample)

The following chart shows the results of a survey conducted among university students about their study habits before final exams. The graph displays the percentage of students who reported using each study method: reviewing notes, group study, watching online videos, attending tutoring sessions, and using AI tools (such as ChatGPT or Grammarly).

Instructions:

- Describe the most relevant features of the data.
- Interpret the results, suggesting possible reasons behind the trends.
- Propose ways to promote study strategies with a population you know.

Figure 3.1 *Bar chart created by ChatGPT*



3.4.2 Writing scorings

The main data analyzed in this study consists of the scores awarded to EFL students' writing proficiency test responses. Access to this data was granted by the English Coordinator of the School of Foreign Languages (See Appendix 1). Although the researcher's TDG tutor also served as the coordinator of the section responsible for managing access to the students' writing tests, potential conflicts of interest were minimized through transparent procedures and formal authorization. The student researcher submitted a formal written request to the coordinator (who was also her supervisor) to obtain access to the anonymized texts used in the study. This request and the subsequent approval

were documented and processed according to institutional research ethics guidelines. The data were handled confidentially, and the identities of both students and raters were anonymized prior to analysis to safeguard objectivity and ethical integrity.

Specifically, the corpus includes a total of 208 texts: 104 from Task 1 (opinion essays) and 104 from Task 2 (data explanation essays). These texts were written by students who took the exam in 2022 (n = 54) and 2023 (n = 50). These writing samples were produced as part of a high-stakes proficiency exam aligned with the C1 level of the CEFR (Council of Europe, 2002). At this level, students are expected to "write clear, well-structured texts on complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion" (CEFR, 2002, p. 6).

Only two cohorts were included in the study because they represented the most recent and complete sets of data available at the time of the research. These cohorts had undergone the same version of the EFL writing component test, under standardized administration and scoring procedures, which ensured consistency and comparability across samples.

The texts were scored by the lecturers of the program (n = 20). Each text was marked independently by two raters using an analytic rubric aligned with CEFR writing descriptors. The final human scores per text were calculated as the average of the two raters' scores. It is noteworthy to mention that, as reported by the English section coordinator to the researcher, the final scorings went through a reliability analyses, indicating that these scorings resulted from a valid and reliable assessment process. These same texts were also assessed by ChatGPT using the same scoring rubric, with prompts carefully designed to replicate the original assessment context. Both global and analytic scores were extracted from ChatGPT's responses and compared to human scores to analyze the degree of agreement and evaluate rater reliability. A total of 7072 scorings were analyzed, as detailed below in Table 3.1.

Table 3.1 Scoring sample used in this study

Scoring Type	Source	Details	Total
Global Scorings	Human Raters	208 texts × 1 score each	208
	ChatGPT	208 texts × 1 score each	208
	Subtotal		**416**
Analytic Scorings	Human Raters	Task 1: 7 criteria × 104 texts = 728 Task 2: 9 × 104 = 936 (728 + 936 = 1664) × 2 tasks	3328
	ChatGPT	Same as above	3328
	Subtotal		**6656**
	Total		**7072**

3.4.3 Analytic scoring rubric

The rubric used in this study is the one designed, validated and used by the English Section (See Appendix 2). The students and raters are informed about the test and the assessment tool. This rubric is composed of seven key criteria: (1) task achievement, which assesses how well the student responds to the prompt and develops relevant ideas with clear arguments and examples; (2) coherence, focusing on logical organization, paragraphing, and effective topic sentences; (3) cohesion, evaluating the use of linking expressions and appropriate referencing; (4) style, considering the use of formal academic tone and hedging expressions; (5) sentence structure, including sentence variety, punctuation, and clarity; (6) grammar, such as verb forms, article use, prepositions, and subject-verb agreement; and (7) vocabulary and spelling, including lexical range, appropriateness, and word form accuracy.

The rubrics are publicly available on Campus Virtual to students and teachers as part of the validity and fairness of the writing test. The writing component of the proficiency test uses two slightly different analytic rubrics depending on the task. Each rubric includes three main facets or categories, evaluated across a scale of five performance levels. The levels range from limited control (score of 1) to consistent mastery (score of 5), allowing for detailed feedback.

For Task 1 (Opinion Essay), the rubric includes the following categories: *Task Achievement and Content Development*: Evaluates whether the candidate addresses the task and fully develops relevant ideas using a clear standpoint, supporting arguments, examples, and a conclusion. *Coherence and Style*: Assesses the logical organization of paragraphs, use of topic sentences, cohesion through linking devices and referencing, and maintenance of an academic register (avoiding informal or oral features). *Form*: Includes sentence structure (variety, complexity, punctuation), grammar (accuracy in verb use, articles, prepositions, agreement), and vocabulary (range, appropriateness, spelling).

For Task 2 (Data Explanation Essay), the same rubric structure is applied, but the first category is replaced by a more detailed breakdown of content criteria based on the specific demands of the task. The other two categories, Coherence and Style and Form, remain consistent across both tasks.

Content:

- **Part 1:** Summary of data – Clarity and accuracy in reporting figures, trends, time/place references, and details.
- **Part 2:** Interpretation of data – Explanation of trends and reasons behind the data.
- **Part 3:** Development of discussion – Expansion of ideas with supported proposals or critical reflection.

3.4.4 Instructions for generating the ChatGPT' scorings

3.4.4.1 Prompt Task 1

The Open AI, Chat GPT-4, was not able to read images or tables since the free version was used during the data input in 2024. For this reason, it was necessary to transform the rubric into a written prompt that the AI could work with. The procedure to obtain the scoring from ChatGPT was to copy and paste every essay into ChatGPT, adding the following prompt: "Could you rate the following essay according to the following criteria:

Task achievement: Task achievement refers to how well the candidate addresses the task and develops the ideas relevant to the task. (Standpoint statement + support with details and illustrative examples)
5) Fully address the task. Presents a clear and fully developed position with relevant, extended and well supported ideas, details and examples.
4.5) In between 4 and 5
4) Sufficiently addresses the task. Presents a clear position, and presents, extends and supports main ideas with some details and examples.
3.5) In between 3 and 4
3) Addresses the task. Although the position is presented, it is not sufficiently supported; there may be irrelevant details and not enough examples.
2.5) In between 2 and 3
2) Address the task only partially/ May lack focus. The position is unclear: supporting ideas, details / examples are unclear / repetitive / irrelevant.
1.5) In between 1 and 2
1) Respond to the task only in a minimal way/the answer is completely unrelated to the task. Does not express a position. Presents a few ideas, which are largely undeveloped or irrelevant.

Coherence: Organisation of whole text; logical paragraphing, topic sentences. Reader engagement
5) Very well organized: logical paragraphs, topic sentences, clear visual introduction and conclusion. Message followed effortlessly.
4.5) In between 4 and 5
4) Well organized: logical paragraphs, topic sentences, relevant introductory statement, may be no concluding statement. Message clear.
3.5) In between 3 and 4
3) Writing organized into paragraphs, but some paragraphs are not well structured. Introduction or conclusion missing. Message generally clear.
2.5) In between 2 and 3
2) Poor organization: illogical/ splintered /misplaced / no paragraphing but ideas can be followed. Hard to follow (needs rereading) at some parts. Introduction and conclusion missing. Shortish script (200 - 240 words)
1.5) In between 1 and 2
1) Lacks organisation: no paragraphing. Incomprehensible / hard to follow throughout. Insufficient length (less than 200 words) and substance.

Cohesion inside text: Use of cohesive devices (linking expressions) and referencing (this, these, those, which, it)
5) Skilful non-formulaic use of a range of cohesion and referencing devices.
4.5) In between 4 and 5
4) Appropriate use but slightly formulaic / lower range of cohesive devices. Good referencing.
3.5) In between 3 and 4
3) Formulaic use /overuse of cohesive devices. Repetitious referencing.
2.5) In between 2 and 3
2) Insufficient/simple or repetitive connectives (and,but,so). Some incorrect cohesive devices and unclear referencing.
1.5) In between 1 and 2
1) Lacks cohesion; many cohesive devices incorrectly used. Disjointed script.

Style: Formal, impersonal, academic tone or register; hedging (e.g., possibly, might, may, could be). Not oral, colloquial or informal. Informal language includes abbreviations, contractions, etc.
5) Clear academic style and good use of hedging
4.5) In between 4 and 5
4) Most aspects of academic style with some hedging. May be some slight informality
3.5) In between 3 and 4
3) Some evidence of academic style, some informal or colloquial language. No / repetitive hedging.
2.5) In between 2 and 3
2) Little evidence of academic style: Frequent informal / colloquial language, wordiness. No hedging.
1.5) In between 1 and 2
1) Overuse of personal and informal / colloquial language.

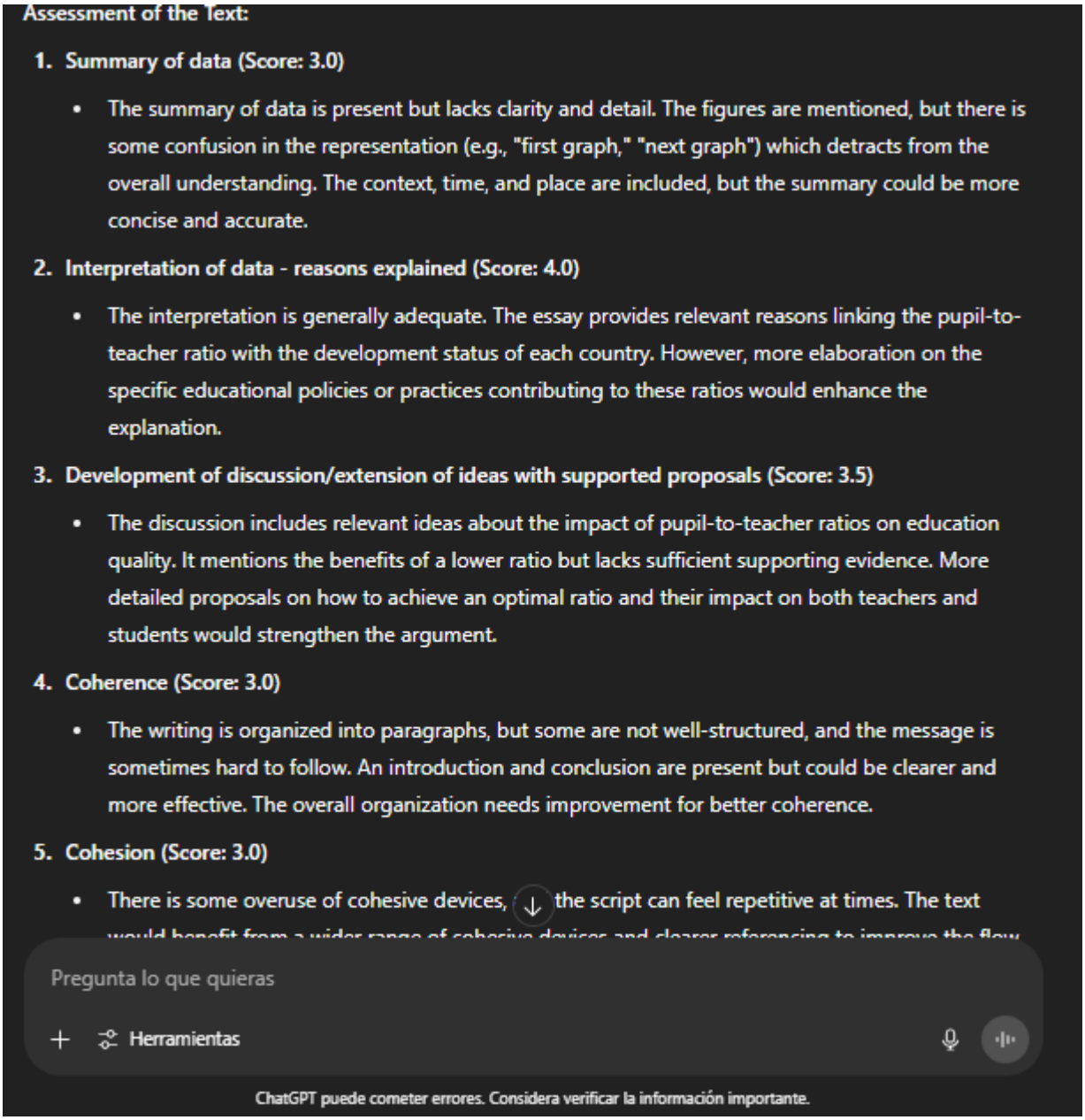
Sentence Structure: accuracy, variety, and complexity of sentences; punctuation
5) Controlled and varied sentence structure.
4.5) In between 4 and 5
4)Adequate variety; occasional errors in complex sentences.
3.5) In between 3 and 4
3) Limited range; errors in complex sentences may be frequent.
2.5) In between 2 and 3
2) Limited control; many incomplete / rambling / convoluted sentences. Lacks punctuation
1.5) In between 1 and 2
1) Lacks control of sentence structure throughout-short script.

Grammar: parts of speech (verbs, articles, prepositions, agreement, plurals)
5) Virtually error free
4.5) In between 4 and 5
4) A few minor errors
3.5) In between 3 and 4
3) Minor/repeated errors
2.5) In between 2 and 3
2) A range of significant errors affect the expression of ideas.
1.5) In between 1 and 2
1) Short script (less than 200 words) with a range of significant errors affecting expression of ideas.

Vocabulary and spelling: Appropriate or accurate use of words; Range (richness or sophistication): Form (spelling and word form
5) Wide range of vocabulary appropriately used. Very few minor misspellings.
4.5) In between 4 and 5
4) Vocabulary generally appropriate and good range. Some minor misspellings.
3.5) In between 3 and 4
3) Vocabulary appropriate but lacks range. Few errors in word use. Some important, intrusive spelling errors.
2.5) In between 2 and 3
2) Vocabulary simple / repetitive / frequently inaccurate or inappropriate. Many intrusive misspellings.
1.5) In between 1 and 2
1) Range and use of vocabulary inadequate. Numerous errors in word use and spelling disproportionate to length and complexity of script.

ChatGPT-4 immediately responded with quantitative and qualitative feedback for every criterion, as seen in Figure 3.2.

Figure 3.2 *A sample of the output produced by ChatGPT*



Subsequently, the numerical data was collected and organized in an Excel spreadsheet together with the data from human raters, as detailed below.

3.5 Data analysis

As a researcher, the information was manually input with the help of the research supervisor who dictated the scorings. The scores were input into an Excel Spreadsheet and then analyzed with the statistical software JASP. Figures 3.3 and 3.4 illustrate how the scorings of human raters and ChatGPT were input into the Excel file and analyzed with the software.

Figure 3.3 *ChatGPT and human raters' scorings input into Excel*

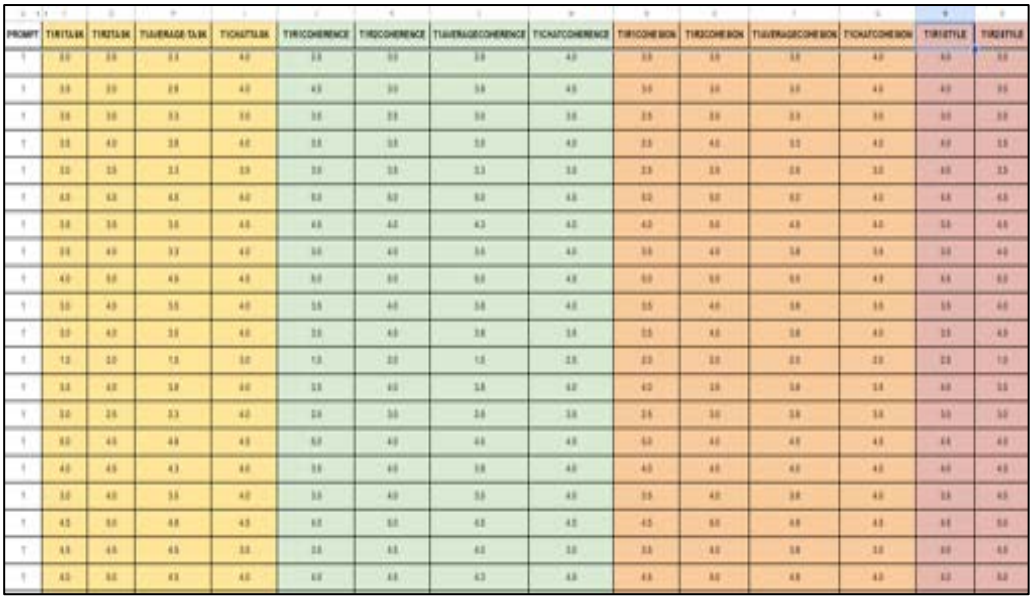
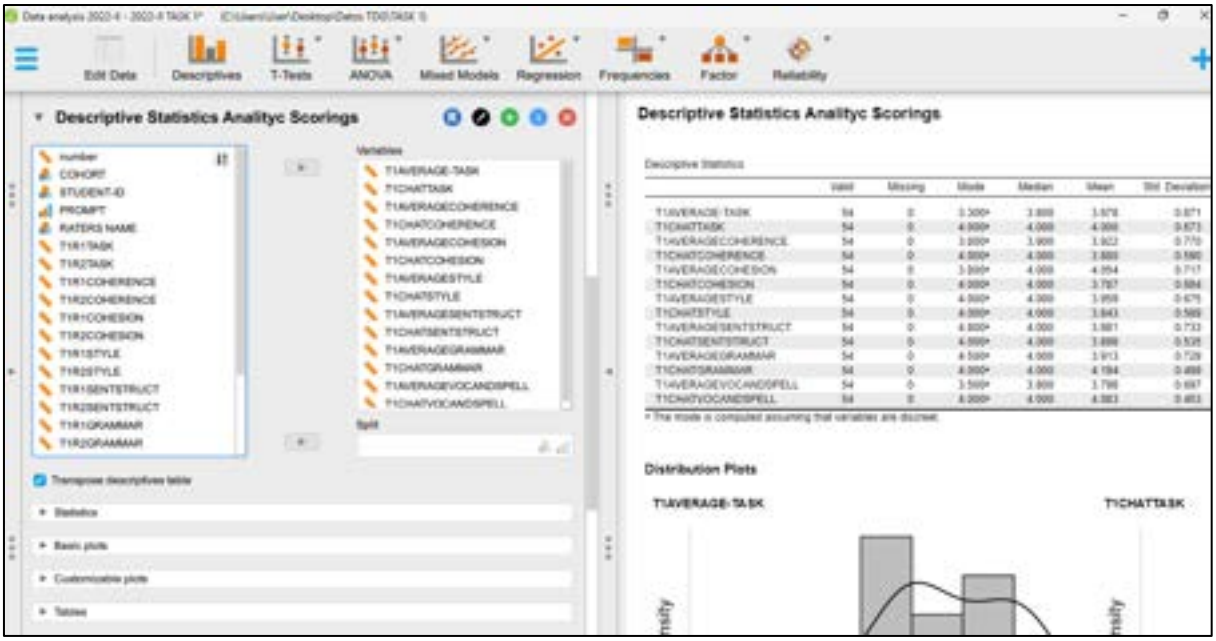


Figure 3.4 *Data analysis on JASP*



To identify how the global and analytic scorings awarded by ChatGPT differed from those of human raters, the average scorings awarded by the human raters and the ones generated by ChatGPT were used. Given that the research objective is to measure the level of agreement between human and machine scoring of EFL academic writing, the study employed inter-rater reliability analysis. This is defined as the extent to which two or more raters provide consistent scores for the same performances (Weigle, 2002). To examine this, Cohen's weighted kappa for ordinal data was used as a statistical measure.

To quantify the agreement for ordinal data, Cohen's weighted kappa was used as the statistical measure because it accounts for the degree of disagreement between raters rather than treating all disagreements equally (Cohen, 1968). The interpretation of Cohen's kappa values follows established guidelines: values between 0.01 and 0.20 indicate slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, and 0.81 to 1.00 almost perfect agreement (Landis & Koch, 1977).

Cohen's kappa interpretation guidelines:

0.01-0.20: None to slight agreement

0.21-0.40: Fair agreement

0.41-0.60: Moderate agreement

0.61-0.80: Substantial agreement

0.81-1.00: Almost perfect agreement

The scores were also compared between ChatGPT and human raters using paired sample t-tests. A paired samples t-test is used to determine if there is a significant difference between the means of two related groups (Pallant, 2016). It is typically used when two sets of measurements from the same individuals are compared. In this case, the sets of data (global and analytic scorings) belong to the same students who were assessed across two different conditions: human and machine raters. By using these statistical methods, the study aims to determine the reliability of ChatGPT as a potential support tool in writing assessment, as well as its degree of alignment with human expert evaluation. Assumptions checks per statistical analysis were performed and are reported in Appendix 3.

3.6 Validity and reliability

The validity of a study is determined by the extent to which its findings accurately reflect the phenomenon being studied, whereas reliability refers to the consistency of a method (Cresswell & Creswell, 2018). In this regard, the constructs, design and methodological decisions were informed by

theory on writing assessment and AI. 208 texts and 7072 scorings constitute a moderate sample of data to be analyzed. Special attention was paid to the data input process to avoid any digitization mistake that can compromise the results. The research supervisor acted as the assistant and supervisor of the data process, management and analysis. Several rounds of analyses were performed to confirm the results.

3.7 Ethical considerations

The study's data was obtained directly from the Foreign Languages program through an Informed consent form and the disclosure of confidentiality and ethical considerations. The names of the students and raters were not disclosed; they were identified by their IDs. **AI use disclosure:** ChatGPT was used for different purposes in this study. ChatGPT was the tool under investigation, from which scorings were obtained and compared with those of human raters. It was also used to create Figure 3.2 of the comparable writing prompt and to validate our interpretations of the results. ChatGPT was also a great tool to polish language use and style, aspects that were also improved by the reseacher's tutor.

4 Results

4.1 Chapter overview

This chapter presents the results of this research that examined how the global and analytic ratings awarded by human raters and ChatGPT differed. The results are organized into two main sections. The first section addresses the comparison of global scorings, whereas the second section focuses on analytic scorings across various dimensions of writing. Descriptive and inferential statistics are presented in each section.

4.2 Global scorings between ChatGPT and Human Raters

To identify similarities and differences in global scorings, I performed a reliability analysis using Cohen's kappa inter-rater agreement and a comparison of means using a paired-samples t-test. The results of the descriptive statistics are presented first to identify trends regarding higher and lower scoring awarded by humans and ChatGPT (See Table 4.1).

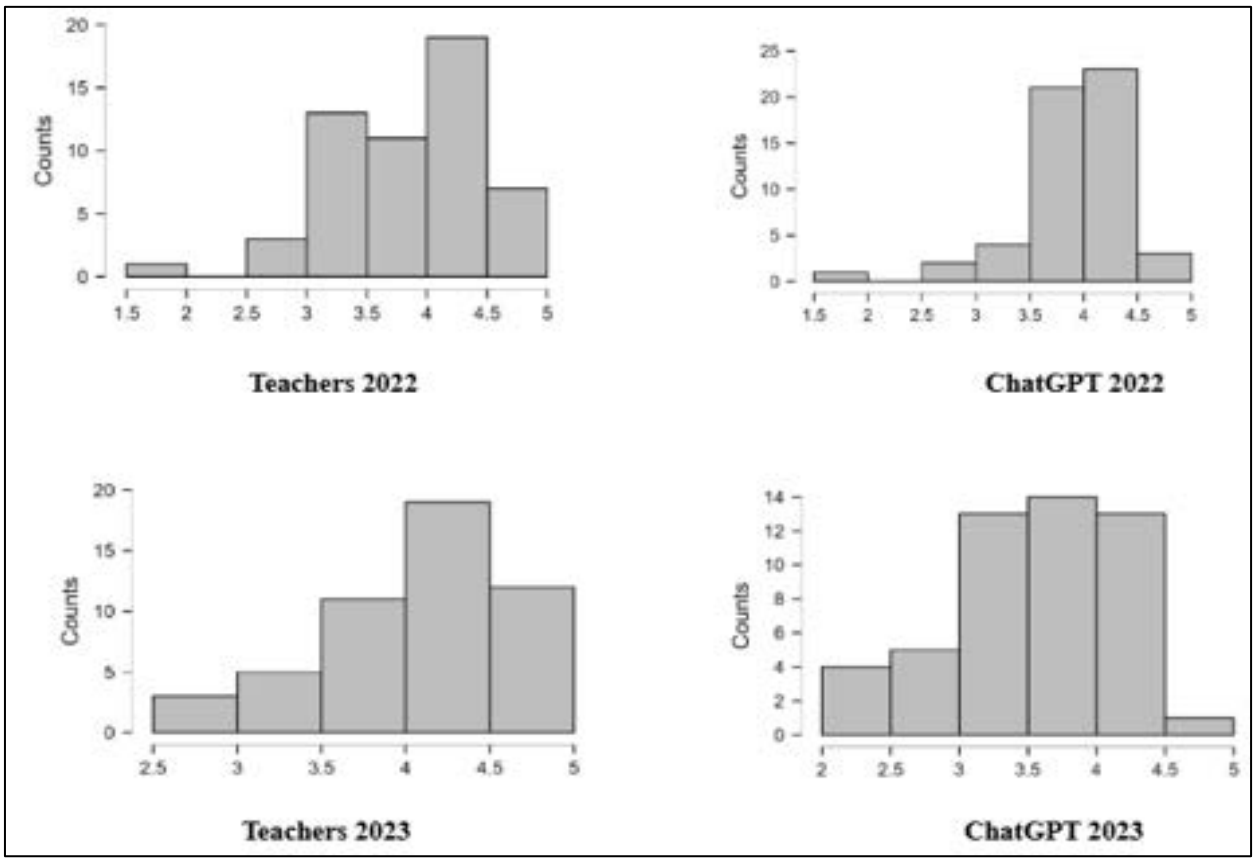
4.2.1 Descriptive statistics: Identifying trends in global scoring between human raters and ChatGPT

Table 4.1 *Descriptive statistics of human and ChatGPT' global scorings - Task 1 across cohorts*

Raters' Global Ratings - Task 1	N	Mode	Mdn	M	SD	Shapiro-Wilk	P-value of Shapiro-Wilk	Min	Max
Teachers 2022	54	4.5	4.0	3.9	0.6	0.962	0.088	1.9	5.0
ChatGPT 2022	54	4.3	4.0	4.0	0.5	0.859	< .001	1.8	4.7
Teachers 2023	50	4.3	4.3	4.1	0.5	0.933	0.007	2.9	4.9
ChatGPT 2023	50	4.0	3.8	3.6	0.6	0.942	0.017	2.0	4.6

As seen in Table 4.1, humans and ChatGPT awarded rather similar ratings across cohorts in Task 1 about personal opinion essays. Most ratings (3.9, 40, 4.1) were classified as C1 level (from 3.7 to 4.6), thus showing students' required performance and indicating a good relationship between human and AI's ratings overall. Even so, there were notable differences in the scores of Cohort 2023, in which ChatGPT awarded lower scores classifying students into a B2 level (3.2 - 3.6), indicating an important difference that might affect students' performance. To have a better view of the scoring distribution in Task 1 across cohorts, histograms were used, as seen in Figure 4.1.

Figure 4.1 *Humans and ChatGPT' global scorings distribution in Task 1 across cohorts*



According to the histograms in Figure 4.1, in Cohort 2022, teachers had a wider range of ratings, whereas ChatGPT provided a limited range. In contrast, in Cohort 2023, there was an opposite trend, with ChatGPT generating a wider variety of scorings. These results indicate that ratings can vary across cohorts, based on students' performances and characteristics. The next section presents the descriptive statistics of scoring across cohorts of Task 2 (Data explanatory essay).

Table 4.2 *Descriptive statistics of human and ChatGPT' global scorings - Task 2 across cohorts*

Raters' Global ratings - Task 2	N	Mode	Mdn	M	SD	Shapiro-Wilk	P-value of Shapiro-Wilk	Min	Max
Teachers 2022	54	4.4	3.5	3.4	0.856	0.951	0.027	1.2	4.9
ChatGPT 2022	54	2.9	3.1	3.1	0.545	0.943	0.012	1.8	3.9
Teachers 2023	50	4.3	3.6	3.7	0.599	0.945	0.022	2.1	4.8
ChatGPT 2023	50	2.9	2.9	2.9	0.47	0.979	0.506	1.9	3.9

As shown in Table 4.2, in Task 2 human raters awarded higher scores than ChatGPT across cohorts. In Cohort 2022, the ratings (3.4, 3.1) classify as B2 level (3.2 -3.6), indicating that this cohort had a lower performance in Task 2. However, there were notable differences in the scorings awarded to Cohort 2023, with ChatGPT awarding lower scores (2.9) compared to human's higher ratings (3.7).

These results indicate that there can be variability in scorings across tasks and cohorts and between human and machine raters.

It can also be seen that the ratings given to Task 2 were lower than those of Task 1 across raters. It seems that Task 2 was more complex than Task 1 for students. It can also be stated that ChatGPT appears to fail interpreting more complex prompts such as the description of a graph. To have a better view of the scoring distribution, I also used histograms, as seen in Figure 4.2.

Figure 4.2 *Humans and ChatGPT' global scorings distribution in Task 2 across cohorts*

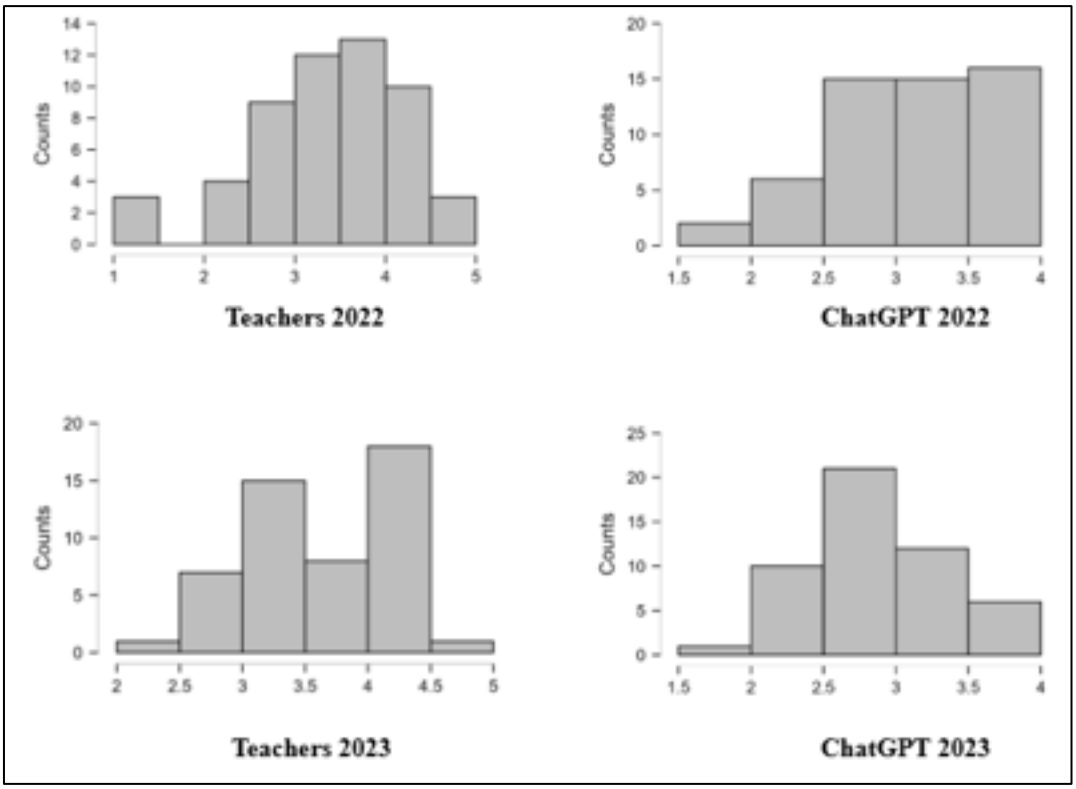


Figure 4.2 illustrates the distribution of global scores awarded by human raters and ChatGPT in Task 2, which involved data explanation essays, across two cohorts. The histograms reveal distinct differences in how each type of rater assessed the students' performance. In the 2022 cohort, both humans and ChatGPT assigned scores that predominantly fell within the B2 range (3.2–3.6). This indicates a relatively closer alignment between the two, though human raters still gave slightly higher scores on average. However, in the 2023 cohort, a more marked divergence appears. Human raters continued to award scores within the C1 range (3.7–4.6), recognizing more advanced levels of performance. In contrast, ChatGPT's scores dropped, falling below the 3.2 threshold, thereby classifying a majority of students at a lower level. The lower scores by ChatGPT also suggest less flexibility in distinguishing between varying levels of performance, particularly in more complex

writing tasks. The next section presents the results of the inferential statics using inter-rater agreement and comparisons of means on global scores.

4.2.2 Inferential statistics: Identifying the reliability of raters across tasks and cohorts

This section presents the results of the reliability levels of humans and ChatGPT as raters, to identify how reliable they were in awarding global scorings and how lenient they were. The ratings awarded by the two raters were analyzed with Cohen's kappa inter-rater agreement.

Table 4.3 *Inter-rater agreement on global scorings Tasks 1 & 2*

				95% CI	
Raters' global scorings	**Agreement**	**Weighted kappa**	**SE**	**Lower**	**Upper**
Task 1					
Teachers vs ChatGPT 2022	Slight	0.045	0.038	-0.028	0.119
Teachers vs ChatGPT 2023	Fair	0.342	0.112	0.112	0.561
Task 2					
Teachers vs ChatGPT 2022	**Moderate**	**0.513**	0.092	0.333	0.692
Teachers vs ChatGPT 2023	Fair	0.248	0.062	0.126	0.370

According to the results in Table 4.3, the inter-rater reliability was predominantly low between humans and ChatGPT across tasks and cohorts. This indicates that there was little similarity in how the essays were scored between these two types of raters. In Task 2 for the 2022 Cohort, there was a moderate level of agreement, meaning the human raters and ChatGPT were more consistent in how they rated the data interpretation essays. As the reliability levels are not consistent across tasks and cohorts, it can be said that ChatGPT might not be a reliable tool yet, more comparisons are needed, and more rating training needs to be provided.

Overall, these results suggest that ChatGPT's reliability as a scoring tool is mixed. While some tasks showed moderate similarity, especially those with a clearer format and more detailed instructions, the general lack of consistency means that AI-based scoring should be used carefully in important evaluations. These findings highlight the ongoing need for human feedback and raise concerns about the role of AI in formal writing assessment.

4.2.3 Inferential statistics: Comparing global ratings across raters, tasks and cohorts

A paired-samples t-test was conducted to compare global scores awarded by human raters and ChatGPT across tasks and cohorts to identify whether the scorings were statistically significantly different or not (Table 4.4).

Table 4.4 *Comparison of global scorings across raters, cohorts and tasks*

Rraters' global scorings	Student test	df	P	Effect Size	SE Effect Size
Task 1					
Teachers vs ChatGPT 2022	-1.236	53	0.222	-0.168	0.113
Teachers vs ChatGPT 2023	6.263	49	**< .001**	0.886	0.166
Task 2					
Teachers vs ChatGPT 2022	3.427	53	**0.001**	0.466	0.126
Teachers vs ChatGPT 2023	10.507	49	**< .001**	1.486	0.194

The results in Table 4.4 revealed that the differences in ratings were statistically significant for three out of four comparisons. For Task 1 - 2022 cohort, there were not statistically significant differences in scorings, indicating reliability in scorings as they were similar between raters. In contrast, the ratings for the other cohorts and tasks yielded statistically significant differences, thus disconfirming the reliability and similarity of ratings across raters. These outcomes confirm that ChatGPT tends to score lower or is harsher than human raters, especially in tasks requiring deeper interpretation and critical thinking.

Taking all these results together, it can be seen that teachers provide a greater variety of scores, distinguishing students' levels of performance in complex writing tasks. This suggests that human raters seem to be still more reliable in writing assessment, particularly in content and in tasks that require higher-order thinking and interpretation.

4.3 Analytic scorings between ChatGPT and Human Raters

The study also examined how closely the analytic scorings awarded by ChatGPT aligned with those provided by human raters across two writing tasks and cohorts. To assess this, Cohen's kappa was used to measure inter-rater agreement for each analytic dimension.

4.3.1 Descriptive statistics: Identifying trends in analytic scoring between human raters and ChatGPT

This section presents the results of the descriptive statistics to identify trends regarding higher and lower ratings awarded by humans and ChatGPT on analytic ratings (See Table 4.5). In this section, histograms are not presented given the large number of graphs. Numerical results will be presented in tables.

Table 4.5 Descriptive statistics of human and ChatGPT' analytic scorings - Task 1 across cohorts

Raters' analytic ratings - Task 1	N	Mode	Mdn	M	SD	Shapiro-Wilk	P-value of Shapiro-Wilk	Min	Max
Task Achievement									
Teachers 2022	54	3.3	3.8	3.7	0.871	0.949	0.024	1.5	5.0
ChatGPT 2022	54	4.0	4.0	4.0	0.673	0.782	< .001	1.0	5.0
Teachers 2023	50	4.0	4.3	4.2	0.690	0.864	< .001	2.0	5.0
ChatGPT 2023	50	4.0	4.0	3.9	0.572	0.897	< .001	2.0	5.0
Coherence									
Teachers 2022	54	3.8	3.9	3.9	0.770	0.944	0.013	1.5	5.0
ChatGPT 2022	54	4.0	4.0	3.8	0.590	0.808	< .001	1.5	4.5
Teachers 2023	50	4.3	4.3	4.2	0.664	0.918	0.002	2.5	5.0
ChatGPT 2023	50	4.0	3.5	3.7	0.550	0.897	< .001	2.0	4.5
Cohesion									
Teachers 2022	54	3.8	4.0	4.0	0.717	0.922	0.002	2.0	5.0
ChatGPT 2022	54	4.0	4.0	3.8	0.684	0.857	< .001	1.5	5.0
Teachers 2023	50	4.5	4.3	4.1	0.584	0.935	0.009	2.5	5.0
ChatGPT 2023	50	4.0	3.5	3.5	0.643	0.911	0.001	2.0	4.5
Style									
Teachers 2022	54	4.0	4.0	3.9	0.675	0.895	< .001	1.8	5.0
ChatGPT 2022	54	4.0	4.0	3.8	0.589	0.823	< .001	1.5	4.5
Teachers 2023	50	4.3	4.3	4.2	0.549	0.899	< .001	2.3	5.0
ChatGPT 2023	50	4.0	3.5	3.5	0.626	0.917	0.002	2.0	4.5
Sentence structure									
Teachers 2022	54	4.8	4.0	4.0	0.733	0.940	0.009	2.3	5.0
ChatGPT 2022	54	4.0	4.0	3.9	0.535	0.769	< .001	2.0	4.5
Teachers 2023	50	4.3	4.3	4.1	0.579	0.949	0.032	2.5	5.0
ChatGPT 2023	50	4.0	4.0	3.5	0.639	0.835	< .001	2.0	4.5
Grammar									
Teachers 2022	54	4.5	4.0	3.9	0.729	0.928	0.003	2.3	5.0
ChatGPT 2022	54	4.0	4.0	4.2	0.499	0.750	< .001	2.0	5.0
Teachers 2023	50	4.5	4.3	4.0	0.637	0.931	0.006	2.3	5.0
ChatGPT 2023	50	4.0	4.0	3.6	0.855	0.835	< .001	2.0	5.0
Vocabulary and spelling									
Teachers 2022	54	3.5	3.8	3.8	0.697	0.961	0.073	2.0	5.0
ChatGPT 2022	54	4.0	4.0	4.1	0.453	0.721	< .001	2.5	4.5
Teachers 2023	50	4.3	4.0	4.0	0.576	0.945	0.021	2.8	5.0
ChatGPT 2023	50	4.0	4.0	3.6	0.749	0.855	< .001	2.0	4.5

Table 4.5 presents the descriptive statistics for the analytic scores given by human raters and ChatGPT in Task 1, which required students to write a personal opinion essay. The results allow us to identify patterns of agreement and divergence between human and AI scorings across seven scoring criteria: Task Achievement, Coherence, Cohesion, Style, Sentence Structure, Grammar, and Vocabulary and Spelling. In general, the results show that both human raters and ChatGPT awarded scores that mostly fall within the C1 range, suggesting an overall adequate level of student writing. However, some trends highlight key differences in scoring behavior.

For Task Achievement, ChatGPT tended to give slightly higher scores than human raters in 2022 (M = 4.00 vs. 3.7), but slightly lower scores in 2023 (M = 3.9 vs. 4.2). This pattern suggests that while ChatGPT may be more lenient with less recent data, it becomes slightly more conservative with newer texts.

In the case of Coherence, the 2022 scores were relatively aligned (M = 3.9 for humans, M = 3.8 for ChatGPT), while in 2023, a larger gap appears (M = 4.2 vs. 3.7). This indicates that ChatGPT might struggle more with identifying logical flow and structure in recent essays. The same trend is evident in the Cohesion criterion, where the difference increases in 2023 (M = 4.1 vs. 3.5). These differences might reflect ChatGPT's harshness in evaluating paragraph transitions, referencing, or the variety of linking devices, whereas teachers might be more moderate.

Regarding Style, human raters consistently gave higher scores than ChatGPT across both cohorts, especially in 2023 (M = 4.3 vs. 3.5). This result suggests that ChatGPT may not fully grasp formal tone, particularly in essays where subtle shifts in voice, hedging, or formality are expected, or it can be that it is harsher than teachers penalizing informal language.

For Sentence Structure, the differences are smaller in both cohorts. In 2022, the mean scores are very similar (M = 4.0 for humans, M = 3.9 for ChatGPT), showing alignment in identifying punctuation, sentence variety, and structure. However, in 2023, the difference is notable (M = 4.1 vs. 3.5), suggesting that while ChatGPT maintains consistency in harshness, detecting more complex syntactic variations.

In Grammar, ChatGPT scored rather similarly to human raters in 2022 (M = 4.2 vs. 4.0), which indicates an alignment in this component. However, in 2023, the scores between raters slightly differed (M = 3.6 vs. 4.0), pointing to misalignments. This might be due to ChatGPT's ability to detect errors of grammatical appropriateness easier than raters.

Finally, Vocabulary and Spelling scores show that ChatGPT awarded slightly higher scores than human raters in 2022 (M = 4.1 vs. 3.8), but lower in 2023 (M = 3.6 vs. 4.05). The results might be due to differences in performance across cohorts.

Overall, these findings show that ChatGPT’s ratings tend to align more closely with human judgments on surface-level features (e.g., sentence structure, grammar, vocabulary) in 2022. However, greater discrepancies appeared in 2023 and in higher-level dimensions like coherence, cohesion, and style. This indicates that ChatGPT's scoring consistency varies across time and dimensions, and it may not yet fully replicate human evaluators’ sensitivity to discourse-level elements.

Table 4.6 Descriptive statistics of human and ChatGPT' analytic scorings - Task 2 across cohorts

Raters' analytic ratings - Task 2	N	Mode	Mdn	M	SD	Shapiro-Wilk	P-value of Shapiro-Wilk	Min	Max
Summary of data									
Teachers 2022	54	4.0	3.8	3.6	0.794	0.932	0.004	1.0	5.0
ChatGPT 2022	54	3.5	3.5	3.4	0.662	0.867	< .001	2.0	4.5
Teachers 2023	50	4.3	3.9	3.8	0.705	0.903	< .001	1.0	4.8
ChatGPT 2023	50	2.0	3.0	2.8	0.664	0.893	< .001	2.0	4.0
Interpretation									
Teachers 2022	54	3.0	3.5	3.5	0.932	0.933	0.005	1.0	5.0
ChatGPT 2022	54	3.5	3.5	3.2	0.711	0.873	< .001	1.0	4.0
Teachers 2023	50	3.8	3.8	3.7	0.755	0.953	0.045	1.5	5.0
ChatGPT 2023	50	3.0	3.0	3.0	0.661	0.928	0.005	1.5	4.0
Discussion									
Teachers 2022	54	3.0	3.1	3.2	1.034	0.957	0.051	1.0	4.8
ChatGPT 2022	54	3.0	3.0	2.9	0.633	0.917	0.001	1.0	4.0
Teachers 2023	50	4.0	3.5	3.6	0.724	0.961	0.095	1.8	5.0
ChatGPT 2023	50	3.0	3.0	2.8	0.673	0.900	< .001	1.0	4.0
Coherence									
Teachers 2022	54	2.0	3.5	3.3	1.012	0.949	0.022	1.0	5.0
ChatGPT 2022	54	3.0	3.0	3.1	0.617	0.879	< .001	1.0	4.0
Teachers 2023	50	4.0	3.6	3.6	0.678	0.942	0.016	2.3	4.8
ChatGPT 2023	50	2.5	3.0	2.8	0.487	0.903	< .001	2.0	4.0
Cohesion									
Teachers 2022	54	3.5	3.6	3.5	0.938	0.946	0.017	1.0	5.0
ChatGPT 2022	54	3.0	3.0	2.8	0.572	0.911	< .001	1.5	4.0
Teachers 2023	50	4.3	3.8	3.7	0.603	0.941	0.014	2.5	5.0
ChatGPT 2023	50	2.5	2.5	2.6	0.567	0.909	< .001	1.5	4.0
Style									
Teachers 2022	54	4.0	3.6	3.5	0.920	0.935	0.006	1.0	4.8
ChatGPT 2022	54	3.0	3.0	2.9	0.498	0.906	< .001	2.0	4.0
Teachers 2023	50	4.3	3.8	3.8	0.645	0.960	0.085	2.5	5.0
ChatGPT 2023	50	3.0	3.0	2.7	0.419	0.856	< .001	2.0	4.0
Sentence structure									
Teachers 2022	54	3.3	3.3	3.3	0.945	0.968	0.165	1.0	5.0
ChatGPT 2022	54	3.0	3.0	3.2	0.562	0.871	< .001	2.0	4.0
Teachers 2023	50	4.3	3.8	3.6	0.588	0.953	0.044	2.5	4.8
ChatGPT 2023	50	3.0	3.0	3.1	0.474	0.878	< .001	2.0	4.0
Grammar									
Teachers 2022	54	3.5	3.5	3.5	1.028	0.945	0.015	1.0	5.0
ChatGPT 2022	54	3.0	3.5	3.4	0.509	0.868	< .001	2.5	4.0
Teachers 2023	50	4.0	3.8	3.6	0.608	0.943	0.017	2.0	4.8
ChatGPT 2023	50	3.0	3.5	3.4	0.458	0.855	< .001	2.0	4.0
Vocabulary and spelling									
Teachers 2022	54	3.3	3.3	3.4	0.901	0.954	0.036	1.0	5.0
ChatGPT 2022	54	3.0	3.0	3.1	0.511	0.895	< .001	2.0	4.0
Teachers 2023	50	4.3	3.8	3.7	0.618	0.934	0.008	2.5	4.8
ChatGPT 2023	50	3.0	3.0	3.0	0.404	0.867	< .001	2.0	4.0

Table 4.6 presents the descriptive statistics for the analytic scores awarded by human raters and ChatGPT for Task 2, which required students to interpret and discuss data from a chart. The scoring categories were slightly different from Task 1, including Summary of Data, Interpretation of Data, Development of Discussion, Coherence, Cohesion, Style, Sentence Structure, Grammar, and Vocabulary and Spelling.

In general, the results show that ChatGPT awarded consistently lower scores than human raters across most categories and cohorts. In the Summary of Data criterion, human raters gave higher average scores in both 2022 (M = 3.6) and 2023 (M = 3.8) compared to ChatGPT (M = 3.4 in 2022, M = 2.8 in 2023). Similarly, in Interpretation of Data, the average scores given by ChatGPT were also lower than those of human raters in both years (2022: M = 3.2 vs. 3.5; 2023: M = 3.0 vs. 3.7). The Development of Discussion scores also show a consistent trend: human raters awarded higher scores than ChatGPT in both years. In 2023, the difference became more pronounced (M = 3.58 vs. 2.84). For Coherence, Cohesion, and Style, the pattern remains consistent, human raters gave higher scores in both cohorts. The Sentence Structure category also shows higher scores from human raters, particularly in 2023 (M = 3.64 vs. 3.07). In Grammar and Vocabulary and Spelling, the differences are less pronounced but still present. In 2023, for Grammar, human raters awarded M = 3.65 while ChatGPT gave M = 3.38. For Vocabulary and Spelling, the scores were M = 3.71 for humans and M = 3.02 for ChatGPT.

These results suggest that ChatGPT is harsher than human raters when marking EFL learners' writing. It could be that teachers are more aware of the context of the learners, that is why they are more gentle marking their writing. It could also be that ChatGPT may not accurately recognize these patterns in these students' writing. The results also suggest differences in views and assessment regarding how well the essay is organized and whether it maintains an appropriate academic tone. Human raters are more flexible in recognizing grammatical variation and complexity, confirming that ChatGPT aligns more closely with human raters on surface-level aspects but differently in other macro skills.

In summary, ChatGPT's scores in Task 2 tend to be consistently lower than those of human raters, especially in categories that require higher-level thinking and discourse organization. The discrepancies are more visible in the 2023 cohort, possibly due to changes in prompt complexity, student writing, or ChatGPT's scoring behavior. These findings suggest that while ChatGPT may offer partial support in evaluating basic language features, it might not yet capable of fully replacing human judgment in content.

4.3.2 Inferential statistics: Identifying the reliability of raters across tasks and cohorts

This section presents the results of the reliability levels of humans and ChatGPT as raters, to identify how reliable they were in awarding analytic scorings, and how harsh or lenient they were. The ratings awarded by the two raters were analyzed with Cohen's kappa inter-rater agreement (Table 4.7).

Table 4.7 Inter-rater agreement on analytic scorings Task 1

				95% CI	
Raters' analytic scorings Task 1	**Agreement**	**Weighted kappa**	**SE**	**Lower**	**Upper**
Task Achievement					
Teachers vs ChatGPT 2022	Fair	0.334	0.097	0.144	0.524
Teachers vs ChatGPT 2023	**Moderate**	**0.489**	0.124	0.246	0.732
Coherence					
Teachers vs ChatGPT 2022	**Moderate**	**0.552**	0.079	0.397	0.707
Teachers vs ChatGPT 2023	**Moderate**	**0.434**	0.101	0.237	0.632
Cohesion					
Teachers vs ChatGPT 2022	**Moderate**	**0.517**	0.096	0.329	0.704
Teachers vs ChatGPT 2023	Fair	0.363	0.103	0.160	0.565
Style					
Teachers vs ChatGPT 2022	**Substantial**	**0.623**	0.079	0.467	0.779
Teachers vs ChatGPT 2023	Fair	0.239	0.074	0.095	0.384
Sentence Structure					
Teachers vs ChatGPT 2022	**Moderate**	**0.525**	0.090	0.348	0.702
Teachers vs ChatGPT 2023	Fair	0.294	0.111	0.078	0.511
Grammar					
Teachers vs ChatGPT 2022	**Moderate**	**0.445**	0.100	0.249	0.640
Teachers vs ChatGPT 2023	Fair	0.226	0.135	-0.038	0.491
Vocabulary and Spelling					
Teachers vs ChatGPT 2022	**Moderate**	**0.505**	0.093	0.323	0.687
Teachers vs ChatGPT 2023	Fair	0.270	0.116	0.043	0.498

For Task 1, which involved writing an opinion essay, the results of the inter-rater agreement between ChatGPT and human raters ranged from fair to moderate revealing noticeable differences in agreement levels. Of 14 reliability analyses, seven were **moderate**, indicating that there was more agreement internally per dimension, compared to the reliability analysis of global ratings. It can also be seen that there was more agreement in the ratings of Cohort 2022 than those of Cohort 2023. It seems that the performance between cohorts varied, which might have influenced the scores.

In 2022, most criteria showed moderate agreement, with the highest being *Style* ($\kappa = 0.623$, substantial agreement), followed by *Coherence, Cohesion, Sentence Structure, Grammar, and Vocabulary and Spelling*, all falling within the moderate range. However, in 2023, agreement levels declined across all dimensions. Overall, while ChatGPT's performance in 2022 aligned more consistently with human raters, the 2023 scores suggest a decline in agreement and greater variability. This trend may reflect changes in ChatGPT's behavior, variations in input data, or differences in human

rater calibration over time. The findings highlight that certain analytic dimensions, such as style, grammar, and cohesion, may be more challenging for AI to assess reliably compared to human raters.

Table 4.8 Inter-rater agreement on analytic scorings Task 2

				95% CI	
Raters' analytic scorings Task 2	**Agreement**	**Weighted kappa**	**SE**	**Lower**	**Upper**
Summary of data					
Teachers vs ChatGPT 2022	**Moderate**	**0.412**	0.117	0.182	0.642
Teachers vs ChatGPT 2023	Slight	0.140	0.066	0.010	0.269
Interpretation of data					
Teachers vs ChatGPT 2022	**Moderate**	**0.460**	0.108	0.248	0.672
Teachers vs ChatGPT 2023	Fair	0.290	0.087	0.120	0.460
Development of discussion					
Teachers vs ChatGPT 2022	**Moderate**	**0.468**	0.090	0.291	0.644
Teachers vs ChatGPT 2023	**Moderate**	**0.410**	0.064	0.285	0.535
Coherence					
Teachers vs ChatGPT 2022	**Moderate**	**0.421**	0.100	0.225	0.618
Teachers vs ChatGPT 2023	Slight	0.197	0.066	0.068	0.326
Cohesion					
Teachers vs ChatGPT 2022	Fair	0.363	0.074	0.217	0.508
Teachers vs ChatGPT 2023	Slight	0.147	0.045	0.059	0.236
Style					
Teachers vs ChatGPT 2022	Fair	0.308	0.073	0.165	0.451
Teachers vs ChatGPT 2023	Slight	0.149	0.035	0.080	0.217
Sentence Structure					
Teachers vs ChatGPT 2022	**Moderate**	**0.439**	0.103	0.238	0.641
Teachers vs ChatGPT 2023	Fair	0.260	0.077	0.109	0.412
Grammar					
Teachers vs ChatGPT 2022	Fair	0.339	0.105	0.133	0.544
Teachers vs ChatGPT 2023	Fair	0.396	0.084	0.232	0.560
Vocabulary and Spelling					
Teachers vs ChatGPT 2022	**Moderate**	**0.480**	0.084	0.315	0.644
Teachers vs ChatGPT 2023	Fair	0.287	0.067	0.154	0.419

The weighted kappa scores comparing human and ChatGPT analytic scorings for Task 2 reveal a general decline in agreement from 2022 to 2023, with several dimensions showing particularly notable drops. In 2022, most criteria demonstrated **moderate agreement,** including *Summary of Data* ($\kappa = 0.412$), *Interpretation of Data* ($\kappa = 0.460$), *Development of Discussion* ($\kappa = 0.468$), *Coherence* ($\kappa = 0.421$), *Sentence Structure* ($\kappa = 0.439$), and *Vocabulary and Spelling* ($\kappa = 0.480$). A few features such as *Cohesion, Style,* and *Grammar* were rated with fair agreement ($\kappa =$; $\kappa =$, correspondingly). However, by 2023, many of these agreement levels weakened. The most pronounced declines occurred in *Summary of Data*, *Coherence*, *Cohesion* and *Style*. Although *Development of Discussion* maintained moderate agreement in both years, the general trend points to a decrease in consistency between human and ChatGPT ratings in 2023. These results suggest that ChatGPT's analytic assessments became less aligned with human judgments over time, particularly in aspects that require interpretive and integrative reasoning, such as summarization, coherence, and stylistic appropriateness. The decline

could be attributed to students' performance, variations in prompts, updates to ChatGPT, or evolving expectations from human raters.

4.3.3 Inferential statistics: Comparing analytic ratings across raters, tasks and cohorts

A paired-samples t-test was conducted to compare analytic scores awarded by human raters and ChatGPT across tasks and cohorts to identify whether the scorings were statistically significantly different or not (Table 4.9).

Table 4.9 *Comparison of analytic scorings across raters and cohorts in Task 1*

Analytic ratings Task 1	Student test	df	P	Effect Size	SE Effect Size
Task Achievement					
Teachers vs ChatGPT 2022	-2.713	53	0.009	-0.369	0.156
Teachers vs ChatGPT 2023	3.948	49	< .001	0.558	0.142
Coherence					
Teachers vs ChatGPT 2022	0.485	53	0.630	0.066	0.125
Teachers vs ChatGPT 2023	6.151	49	< .001	0.870	0.152
Cohesion					
Teachers vs ChatGPT 2022	2.907	53	0.005	0.396	0.136
Teachers vs ChatGPT 2023	7.424	49	< .001	1.050	0.168
Style					
Teachers vs ChatGPT 2022	1.653	53	0.104	0.225	0.112
Teachers vs ChatGPT 2023	8.664	49	< .001	1.225	0.198
Sentence structure					
Teachers vs ChatGPT 2022	0.983	53	0.330	0.134	0.129
Teachers vs ChatGPT 2023	6.074	49	< .001	0.859	0.179
Grammar					
Teachers vs ChatGPT 2022	-3.228	53	0.002	-0.439	0.141
Teachers vs ChatGPT 2023	3.397	49	0.001	0.480	0.183
Vocabulary and spelling					
Teachers vs ChatGPT 2022	-3.864	53	< .001	-0.526	0.125
Teachers vs ChatGPT 2023	4.389	49	< .001	0.621	0.176

Table 4.9 presents the results of paired-samples t-tests conducted to compare the analytic scores awarded by human raters and ChatGPT for Task 1. This task required students to write a personal opinion essay. The analysis helps identify whether the differences in ratings between the two types of raters are statistically significant and in which categories these differences are more noticeable.

The results indicate that significant differences were found in most analytic categories, especially in the 2023 cohort. For example, in the category of Coherence, the difference between human and ChatGPT ratings was not significant in 2022 ($p = 0.630$) but became highly significant in 2023 ($p < .001$). This suggests that ChatGPT and human raters disagreed more in evaluating the coherence of student essays in 2023.

A similar trend can be observed in Cohesion. In 2022, there was a significant difference ($p = 0.005$), and in 2023, the difference became even more pronounced ($p < .001$). Style was another category with notable variations. While no significant difference was observed in 2022 ($p = 0.104$), in 2023, the scores differed greatly ($p < .001$) with the largest effect size among all categories ($d = 1.225$). Task Achievement also showed statistically significant differences in both years, though the effect was stronger in 2023 ($p < .001$, $d = 0.558$) compared to 2022 ($p = 0.009$, $d = -0.369$). In Sentence Structure, Grammar, and Vocabulary and Spelling, the pattern of significant differences continued, particularly in the 2023 cohort. For example, Sentence Structure had no significant difference in 2022 ($p = 0.330$), but a significant difference emerged in 2023 ($p < .001$, $d = 0.859$). These findings reflect differences in assessing EFL learners' writing regarding the use of cohesive devices and referencing, formal tone and academic register, and the completeness and support of arguments.

However, Grammar and Vocabulary showed the opposite pattern across years: Grammar scores were significantly higher for ChatGPT in 2022 ($p = 0.002$, $d = -0.439$) but lower in 2023 ($p = 0.001$, $d = 0.480$), suggesting inconsistency in how the AI interprets grammatical accuracy. Vocabulary and Spelling also followed this pattern, with ChatGPT scoring significantly higher in 2022 ($p < .001$, $d = -0.526$), but lower in 2023 ($p < .001$, $d = 0.621$). This shift could indicate that ChatGPT's vocabulary evaluation either evolved or responded differently to newer student writing, possibly influenced by changes in prompt complexity or AI behavior.

To sum up, the analysis of Task 1 shows that in 2022, the scores from ChatGPT and human raters were more similar in several categories. However, in 2023, the differences between them became larger in all areas. This means that ChatGPT's scorings and criteria might differ from human raters across times, students, prompts and AI updates. Because of this, even though ChatGPT can help in some parts of writing assessment, it should still be used carefully in important exams and always with support from human evaluators.

Table 4.10 *Comparison of analytic scorings across raters and cohorts in Task 2*

Analytic ratings Task 2	Statistic	df	P	Effect Size	SE Effect Size
Summary of data					
Teachers vs ChatGPT 2022	2.711	53	0.009	0.369	0.147
Teachers vs ChatGPT 2023	8.408	49	< .001	1.189	0.220
Interpretation of data					
Teachers vs ChatGPT 2022	2.531	53	0.014	0.344	0.139
Teachers vs ChatGPT 2023	7.093	49	< .001	1.003	0.181
Development of discussion					
Teachers vs ChatGPT 2022	2.631	53	0.011	0.358	0.132
Teachers vs ChatGPT 2023	8.905	49	< .001	1.259	0.159
Coherence					
Teachers vs ChatGPT 2022	1.609	53	0.114	0.219	0.139
Teachers vs ChatGPT 2023	8.424	49	< .001	1.191	0.204
Cohesion					
Teachers vs ChatGPT 2022	6.365	53	< .001	0.866	0.150
Teachers vs ChatGPT 2023	12.863	49	< .001	1.819	0.240
Style					
Teachers vs ChatGPT 2022	5.223	53	< .001	0.711	0.157
Teachers vs ChatGPT 2023	12.695	49	< .001	1.795	0.233
Sentence structure					
Teachers vs ChatGPT 2022	1.146	53	0.257	0.156	0.139
Teachers vs ChatGPT 2023	7.008	49	< .001	0.991	0.185
Grammar					
Teachers vs ChatGPT 2022	1.254	53	0.215	0.171	0.149
Teachers vs ChatGPT 2023	3.378	49	0.001	0.478	0.155
Vocabulary and spelling					
Teachers vs ChatGPT 2022	2.977	53	0.004	0.405	0.129
Teachers vs ChatGPT 2023	9.705	49	< .001	1.373	0.181

Table 4.10 shows the results of paired-samples t-tests comparing the analytic scores given by human raters and ChatGPT in Task 2, which required students to describe and interpret data from a graph. This analysis helps to understand whether the differences between the two raters were statistically significant and how large those differences were.

The results show that most of the differences between human and ChatGPT scores were statistically significant, especially in the 2023 cohort. In fact, all nine analytic categories in 2023 had p-values below .001, meaning the scores were clearly different. For example, the Style category had a

large effect size (d = 1.795), showing that ChatGPT gave much lower scores than human raters. A similar result was seen in Cohesion, indicating strong disagreement in how the structure and flow of ideas were judged.

Other categories like Development of Discussion, Coherence, and Vocabulary and Spelling also had large differences in 2023. These results show differences in evaluating how students organized their ideas, supported their points, and used vocabulary effectively in more complex writing tasks. In contrast, in 2022, the differences were smaller. For instance, in Grammar (p = 0.215) and Sentence Structure (p = 0.257), there were no significant differences, suggesting that ChatGPT and human raters had more similar opinions in those areas during that year.

Overall, the inferential analysis of Task 2 highlights that ChatGPT differs more in evaluating writing that involves explanation, interpretation, and discussion, especially in newer responses. These findings could suggest that while ChatGPT may be useful for checking grammar or basic writing features, it should not be relied on to assess more complex aspects of academic writing without human input.

4.4 Summary of key findings

This chapter outlined the main results of the study, focusing on how the scores from human raters compared to those generated by ChatGPT in both global and analytic evaluations of student writing. The analysis covered two writing tasks across two different cohorts, using descriptive statistics alongside inferential tests to examine the patterns.

Looking at the global scores, the 2022 data showed a reasonable level of similarity between ChatGPT and human ratings. However, the 2023 results turn other way: the gaps became more visible, particularly in complex tasks like data interpretation essays. In these cases, ChatGPT often assigned lower overall scores, suggesting it may handle straightforward evaluations more consistently than those requiring discourse-level skills

For the analytic scores, ChatGPT's evaluations were closer to human raters in aspects such as grammar, sentence structure, and vocabulary. Even so, the alignment weakened when judging discourse-level elements such as task achievement, coherence, cohesion, and style, especially in the 2023 cohort. One possible reason is that these features demand a deeper understanding of the text's purpose and audience which is something AI still struggles to replicate. Statistical tests supported these observations, indicating that the differences were significant rather than casual. In most cases, ChatGPT gave lower scores than human raters, especially for the more recent cohort and more cognitively demanding criteria. These findings may suggest that while AI tools like ChatGPT can be

useful in certain areas, they are not yet capable of replacing human judgment in high-stakes or complex writing assessments.

In summary, this chapter has shown that ChatGPT may complement but not fully substitute human raters. These results set the foundation for the next chapter, which will interpret these findings in light of the literature and discuss their implications for educational assessment and the responsible use of AI in academic contexts.

5 Discussion

5.1 Chapter overview

This chapter discusses the findings of the present study in relation to previous research and the conceptual framework used to investigate the topic. It starts with the discussion and implications of differences in global and analytic scorings between human raters and ChatGPT. The discussion highlights the strengths and limitations of ChatGPT as an assessment tool and offers suggestions for future practice and research in EFL writing assessment.

5.2 Global scoring agreement between Human Raters and ChatGPT

This study first examined whether the global scores assigned by ChatGPT matched those awarded by human raters. In general, both human raters and ChatGPT placed students at similar overall proficiency levels, suggesting that the AI can capture broad indicators of language ability. On the other hand, the statistical analyses, particularly inter-rater agreement measures and comparisons of mean scores showed inconsistencies. Reliability between raters was often low, and many of the differences across tasks and cohorts were statistically significant.

These results raise questions about the suitability of ChatGPT for formal writing assessment and even more in high-stakes contexts in which inconsistent scoring could lead to misclassification, with potentially unfair academic results. As noted by Almusharraf and Alotaibi (2023) their research suggests that variability among human raters can affect scoring reliability due to factors such as background and experience. Similar concerns appear in earlier studies. For instance, Kolade et al. (2023) reported that ChatGPT's feedback and scoring patterns shifted depending on the task, raising issues of fairness and bias. Rudolph et al. (2023) also stressed that while AI offers speed, it often lacks the judgment needed for decisions with serious consequences.

The present findings are consistent with those of Mizumoto and Eguchi (2023), who observed that AI models tend to underperform with complex, context-dependent writing. Likewise, Mizumoto et al. (2024) found stronger agreement with human raters in grammar scoring, but weaker alignment in discourse-level analysis and task completion. Other studies also point to similar trends. Chen and Pan (2022) showed that although AWE tools can supplement teacher feedback, human evaluators offer more pedagogically useful comments, particularly on coherence and content organization. Similarly, Almusharraf and Alotaibi (2023) reported that while tools like Grammarly are proficient in identifying surface-level features, human raters still stand out when assessing overall writing quality.

Taken together, these studies suggest that although ChatGPT can offer a general estimate of student performance and assist in formative feedback, its use as a primary or sole scoring tool in high-

stakes academic contexts is problematic. Human judgment remains essential to ensure scoring validity, especially in tasks that involve higher-order thinking, argument development, and interpretation.

5.3 Analytic scoring alignment between Human Raters and ChatGPT

The study also explored the consistency of raters on analytic scores across different writing dimensions, including task achievement, coherence, cohesion, style, sentence structure, grammar, and vocabulary and spelling. Results from both the descriptive statistics and t-tests showed key trends.

The highest inter-rater agreement was observed in surface-level aspects such as style, sentence structure, and vocabulary and spelling. These results were also supported by small or non-significant differences in t-tests in 2022 for sentence structure and grammar, indicating closer alignment in those areas. ChatGPT appears to be more reliable in assessing formal language elements, where accuracy and consistency are easier to detect automatically.

In contrast, lower levels of agreement were found in categories requiring higher-order thinking, such as coherence and content development. For example, coherence showed fair agreement and statistically significant differences in 2023, reflecting differences in how ChatGPT and human raters judged the logical flow of ideas. Similarly, cohesion had large differences in 2023, suggesting that AI tools differ in evaluating connections between sentences and paragraphs compared to teachers. This trend was even more visible in Task 2, which involved interpreting visual data. Dimensions such as Summary of Data and Interpretation of Data showed big differences.

These findings are consistent with research by Klebanov and Madnani (2020), who pointed out that AWE tools perform well at surface-level detection but often lack discourse-level insight. Mizumoto et al. (2024) also observed that while AI performs well in detecting grammar and vocabulary issues, it struggles in more subjective or abstract scoring dimensions such as task fulfillment and organization. Similarly, Zhang (2021) emphasized the risks of over-reliance on AI scoring, particularly when it overlooks content quality and rhetorical structure.

Taken together, the descriptive and inferential results confirm that ChatGPT is more consistent in surface-level aspects, but the agreement with human raters drops significantly for higher-level elements of writing. These patterns reinforce the idea that AI can support teachers in assessing language form but should not be relied on or only used for evaluating content quality or textual organization, particularly in high-stakes writing tasks.

5.4 A balanced integration of AI as a complement in EFL writing assessment

Given the lack of consistency and variability in scorings between teachers and AI, it can be suggested that ChatGPT can be used as a complement in writing assessment but cannot substitute completely the assessment and feedback of teachers. This lack of consistency could be triggered by differences in how each rater type interprets discourse-level elements, the subjective nature of some rubric categories, and ChatGPT's limited ability to evaluate context and intention in EFL writing.

This study confirms that ChatGPT aligned more with human raters on surface-level dimensions, such as grammar and sentence structure, supporting findings from Mizumoto and Eguchi (2023), Zhang (2021), and Chen and Pan (2022). However, the significant differences found in discourse-related areas, like coherence, task development, and content interpretation, echo concerns raised by Klebanov and Madnani (2020) and Mizumoto et al. (2024) about AI's struggles with deeper levels of writing analysis.

One key implication for the Foreign Languages program is the need to promote teacher training on how AI systems evaluate writing. Teachers might approach ChatGPT's outputs with a critical view and adapt their feedback as needed. When AI is used particularly in assessment, it is important to explain to students what the feedback means and how it works with other assessment elements and the teachers feedback. In high-stakes tests, such as the C1 graduation requirement, human scoring is still essential to preserve fairness and ensure that decisions reflect the wider context of each student's work. The analytic rubric used by the program could also be calibrated to better support AI scorings. For example, clearer descriptors for abstract categories like coherence or argumentation might help improve the consistency of both human and AI ratings and at the same time, carefully adjusting ChatGPT prompts to match the rubric's language could increase alignment.

From an ethical viewpoint, integrating AI into writing assessment brings opportunities for more efficient feedback but also raises challenges that need careful attention. A major concern is the potential gap between AI-generated scores and human judgment, while human raters consider context, argumentative structure, creativity, and rhetorical impact, systems like ChatGPT often rely more heavily on surface-level language features. This difference can lead to errors in the scores, especially when students present strong ideas but make minor grammar or vocabulary mistakes. In high-stakes contexts, such as graduation exams, this could unfairly disadvantage students and reduce trust in the evaluation process.

There is also a pedagogical risk that learners might shift their focus away from critical thinking, originality, and argumentation, and instead prioritize grammar and vocabulary to fit what they think the AI expectations. This could be especially problematic for students with well-developed analytical

skills who still struggle with linguistic accuracy. If AI and human scores conflict, students may become unsure about what counts as effective writing or how to interpret feedback.

Another risk is that excessive dependence on AI for grading could make evaluation more impersonal and rigid: without enough human supervision AI feedback may lack the deep and contextualized instruction that teachers usually provide, reducing its motivational impact and this can be discouraging for learners, especially those at intermediate or advanced levels who benefit from recognition of creative and critical contextualized choices in their writing.

To address these issues, AI should be used as a complement rather than a replacement for teacher evaluation. One practical approach is to let AI scoring serve as an initial point of view, highlighting common patterns or recurring errors in the surface levels so that human raters can focus on higher-order skills. However, this approach requires that teachers be properly trained to interpret and critique AI-generated scorings.

Pedagogically, AI tools like ChatGPT can still play a valuable role in helping students identify and correct grammar and vocabulary issues. This aligns with findings from Chen and Pan (2022), who found that combining AI and teacher feedback improved writing outcomes. Therefore, in low-stakes situations such as practice essays, class activities or peer review writing, ChatGPT can provide quick feedback that encourages students to work more independently and emphasizes their own particular errors or disadvantages in their writing process in order to better perform in more serious context. As Kolade et al. (2023) enunciated, students need to be aware of AI's limitations and be trained to improve and reflect on their work without depending entirely on this tool.

For both high-stakes and low-stakes contexts, a combined approach works best since ChatGPT could be able assist with reviewing large amounts of writing, identifying recurring language errors and improvements, or providing initial insights. Nevertheless, the final scoring decisions, especially when they affect grades or graduation, should remain in charge of teachers.

To sum up, while ChatGPT can be a valuable tool in certain aspects of writing assessment, its use should follow adequate calibration, clear institutional policies, teacher preparation, and a solid pedagogical framework. Used in this way, it can contribute to a more efficient, fair, and personalized evaluation process in EFL classrooms.

6 Conclusions

6.1 Summary of findings and conclusion

This study examined how global and analytic scores from ChatGPT differed to those given by human raters in assessing EFL writing. The results showed moderate agreement in surface-level components such as grammar, vocabulary, and sentence structure, but significant differences in more complex aspects like coherence, content development, and interpretation of information. These results were most evident in Task 2 (data interpretation essays), where the need to analyze visual information and build arguments exposed the tool's current limitations. Based on these findings, the study concludes that ChatGPT can complement formative writing assessments, particularly in language-based criteria but cannot replace human supervision and feedback in high-stakes and more complex evaluations.

6.2 Broader impact and local contribution

This research expands the understanding of AI supported writing assessment in less studied settings such as Colombian higher education and it provides concrete evidence for the Foreign Languages program at Universidad del Valle, where passing a C1-level writing test is a graduation requirement. The findings are opportune, given the increasing interest in academic contexts on using AI to reduce teacher workload and offer general feedback.

The quantitative, comparative approach reinforced the validity of the results and allowed precise statistical comparisons between AI and human scoring in the presented sample. The methodology could be adapted for use in other programs to assess AI integration or in other studies. Moreover, using authentic student work and a rubric validated by the institution strengthens the value of the findings.

For the Licenciatura en Lenguas Extranjeras program at Universidad del Valle, these results can strengthen curriculum updates, rubric improvement, and teacher training. Identifying where ChatGPT's scores align or fail to align with teacher judgments is essential for maintaining fair, valid, and context align evaluation practices. Beyond this institution, the findings may help other public universities considering AI assisted assessment, particularly in contexts with limited resources where students struggle the most with their autonomous learning.

6.3 Limitations of the study

While the study offers useful insights, several limitations should be acknowledged. First, the analysis focused only on two similar writing genres: opinion essays and data explanatory/ interpretation essay within a single academic program, which limits how much the findings can be applied to other contexts. Second, the study used the free version of ChatGPT available in 2024, which

did not have image-processing capabilities and may perform differently from newer models that continue to update every day. Third, even though human raters received training, some variation in how they applied the rubric likely remained, particularly in subjective areas like style or coherence. Finally, the study compared scores but did not examine the quality or pedagogical value of the feedback generated by ChatGPT, which could be an important direction for future research. Based on these limitations, some recommendations for further research are provided below.

6.4 Recommendations for Future Research

Future studies should expand the analysis to include a wider variety of genres, such as argumentative essays, reflective writing, or research summaries. Comparing ChatGPT to other AI tools like GPT-4, Gemini, or Grammarly could also offer a broader perspective on automated assessment performance. In addition, it would be worthwhile to explore hybrid scoring models where human and AI assessments are combined, highlighting their own particular strengths and achieving a more elaborated and personalized assessment. Future research could also focus on how students and teachers interpret and use qualitative AI generated feedback, as well as its impact on learner motivation, independence, and writing performance over a longer period of time.

Another important area for study is the development of institutional policies and ethical guidelines to ensure that AI use in assessment is fair, transparent, and aligned with educational goals. By addressing these aspects, researchers and educators can help guarantee that tools like ChatGPT contribute meaningfully to writing evaluations and supporting learning processes while protecting fairness and quality in diverse education contexts.

References

Almusharraf, N., & Alotaibi, H. (2023). *An Error-Analysis Study From an EFL Writing Wontext: Human and Automated Essay Scoring Approaches.* Technology Knowledge and Learning, *28*(3), 1015–1031. https://doi.org/10.1007/s10758-022-09592-z

Al-Rashidi, A. H., Asif, M., Vanani, M. G., & Aberash, A. (2022). *Learner-oriented assessment (LOA) practice: the comparative study of self-assessment, peer assessment, and teacher assessment on EFL learners' writing complicity, accuracy, and fluency (CAF), speaking CAF, and attitude*. Language Testing in Asia, 12(1). https://doi.org/10.1186/s40468-022-00209-x

Bronckart, J.-P. (1999). *The sociocognitive foundations of language and thought*. John Benjamins Publishing Company

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (2ª ed.). Pearson Education ESL.

ChatGPT "Definition of ChatGPT." (n.d.). Openai.com. Retrieved December 19, 2023, from https://chat.openai.com/

Chaudhry, M. A., & Kazim, E. (2022). *Artificial Intelligence in Education (AIEd): a high-level academic and industry note 2021*. AI and Ethics, 2(1), 157–165. https://doi.org/10.1007/s43681-021-00074-z

Chen, H., & Pan, J. (2022). *Computer or human: a comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing*. Asian-Pacific Journal of Second and Foreign Language Education, *7*(1). https://doi.org/10.1186/s40862-022-00171-4

Chen, Y., Wang, R., Jiang, H., Shi, S., & Xu, R. (2023). *Exploring the use of large language models for reference-free text quality evaluation: An empirical study*. https://doi.org/10.48550/ARXIV.2304.00723

Chuang, P.-L. (2025*). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions*. Language Assessment Quarterly, 22(1), 45-63. https://doi.org/10.1016/j.laq.2025.01.003

Coe, K., & Scacco, J. M. (2017*). Content Analysis, Quantitative*. In The International Encyclopedia of Communication Research Methods (pp. 1–11). Wiley. https://doi.org/10.1002/9781118901731.iecrm0045

Cohen, J. (1968). *Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit.* Psychological Bulletin, 70(4), 213–220. https://doi.org/10.1037/h0026256

Common European Framework of reference for Languages: Learning, teaching, assessment (CEFR) - Common European Framework of reference for Languages (CEFR) - www.coe.int. (n.d.). Common European Framework of Reference for Languages (CEFR). Retrieved December 19, 2023, from https://www.coe.int/en/web/common-european-framework-reference-languages

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Creswell, J. W., & Miller, D. L. (2000). *Determining validity in qualitative inquiry.* Theory into Practice, *39*(3), 124–130. https://doi.org/10.1207/s15430421tip3903_2

Cumming, A. (2016). 3. *Theoretical orientations to L2 writing.* In R. Manchón & P. Matsuda (Ed.), Handbook of Second and Foreign Language Writing (pp. 65-88). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9781614511335-006

Departamento de Lenguas y Culturas Extranjeras, Universidad del Valle. (2025). *Program assessment policies*. https://lenguaje.univalle.edu.co/departamentos-3

Elbow, P. (2015). " *Personal Writing" and" Expressivism". Critical expressivism: Theory and practice in the composition classroom*, 15.

Emig, J. (1971). *The composing processes of twelfth graders*. https://eric.ed.gov/?id=ED058205

Escuela de Ciencias del Lenguaje, Universidad del Valle. (2021). *Examen de competencias licencitura en lenguas extranjeras [Reglamento].* Universidad del Valle. https://lenguaje.univalle.edu.co/examenes-lenguas-extranjeras/examen-de-competencias-licenciatura-en-lenguas-extranjeras

ETS Criterion writing evaluation service. (n.d.). Ets.org. Retrieved December 19, 2023, from https://criterion.ets.org/default.aspx

European Commission's High-Level Expert Group on Artificial Intelligence. (2019). *A definition of AI: Main capabilities and scientific disciplines* (p. 8). European Commission.

Floridi, L., & Chiriatti, M. (2020). *GPT-3: Its nature, scope, limits, and consequences.* Minds and Machines, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

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

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to design and evaluate research in education* (10th ed.). McGraw-Hill Education

Gregersen, E. (2023). *ChatGPT*. In Encyclopedia Britannica.

Guapacha Chamorro, M. E., & Chaves Varón, O. (2023). *EFL writing studies in Colombia between 1990 and 2020: A qualitative research synthesis*. PROFILE Issues in Teachers Professional Development, 25(1), 247–267. https://doi.org/10.15446/profile.v25n1.94798

Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press.

Huyler, D., & McGill, C. M. (2019). *Research design: Qualitative, quantitative, and mixed methods approaches, by John Creswell and J. david Creswell*. Thousand Oaks, CA: Sage publication, inc. New Horizons in Adult Education and Human Resource Development, *31*(3), 75–77. https://doi.org/10.1002/nha3.20258

Inter-American Development Bank (IDB). (2023). *AI Here! Regional call for innovative AI solutions in education.* https://www.iadb.org/en/news/idb-launches-call-artificial-intelligence-solutions-transform-education

Introduction to rubrics. (n.d.). Unl.edu. Retrieved December 19, 2023, from https://teaching.unl.edu/resources/grading-feedback/intro-to-rubrics/

Klebanov, B. B., & Madnani, N. (2020, July). *Automated evaluation of writing–50 years and counting*. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 7796-7810). https://aclanthology.org/2020.acl-main.697/

Kolade, O., Owoseni, A., & Egbetokun, A. (2023). *Assessment on trial? ChatGPT and the new frontiers of learning and assessment in higher education*. Unpublished. https://doi.org/10.13140/RG.2.2.32665.49768

Landis, J. R., & Koch, G. G. (1977). *The measurement of observer agreement for categorical data*. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Licenciatura en Lenguas Extranjeras con énfasis en Inglés y Francés, E. de C. del L. (n.d.). *Protocolo Examen de competencias inglés*. https://docs.google.com/document/d/1a3HqsJFoxapE6LB0PSUzXRpCElMAnBTa/edit

Licenciatura en Lenguas Extranjeras con énfasis en Inglés y Francés, E. de C. del L. (septiembre de 2021). *Reglamento interno del Examen de competencias en inglés y en francés*. https://drive.google.com/file/d/1pkfm1rETAFcBYIX2QQVNoCiSII9iQKJa/view?usp=sharing

McCarthy, J. (1955). *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* Dartmouth College. https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Manning, C. (2020, September). *Artificial Intelligence Definitions*. Stanford.edu. https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf

Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). *The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective*. Cogent Education, 10(2). https://doi.org/10.1080/2331186x.2023.2236469

Martín-Martín, P. (2013). *The teaching of academic writing to English as a second language students: A functional genre-based approach.* Revista de Lenguas para Fines Específicos, 19, 329-351. https://ojsspdc.ulpgc.es/ojs/index.php/LFE/article/view/23

Mertler, C. A. (2016*). Introduction to educational research* (2nd ed.). SAGE Publications

Miller, M. D. (2019). *Reliability in educational assessments.* In Education. Oxford University Press.

Ministerio de Educación Nacional de Colombia (MEN). (2016). Marco común europeo de referencia para lenguas. https://www.mineducacion.gov.co/

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

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). *Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment.* Research Methods in Applied Linguistics, 3(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Modulos AI. (2025). *Global AI Compliance Guide: Regulations & governance landscape*. https://www.modulos.ai/global-ai-compliance-guide/

Moqbel, M. S. S., & Al-Kadi, A. M. T. (2023). *Foreign language learning assessment in the age of ChatGPT: A theoretical account.* Journal of English Studies in Arabia Felix, 2(1), 71–84. https://doi.org/10.56540/jesaf.v2i1.62

Pallant, J. (2016). *SPSS survival manual* (6th ed.). Allen & Unwin.

Pho, P. D. (2013). *Approaches to Genre Analysis. In Authorial Stance in Research Articles* (pp. 10–20). Palgrave Macmillan UK.

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. (2017). Association for Computational Linguistics.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. (2020). Association for Computational Linguistics.

Project score. (n.d.). Https://www.oerproject.com/. Retrieved December 19, 2023, from https://www.oerproject.com/Project-Score

Ramesh, D., & Sanampudi, S. K. (2022). *An automated essay scoring systems: a systematic literature review*. Artificial Intelligence Review, 55(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Ramesh, P., & Sanampudi, S. (2022). *Automated essay scoring systems: Tools and challenges*. International Journal of Computer Applications, 184(20), 1–6.

Rannikmäe, M., Holbrook, J., & Soobard, R. (2020). *Social Constructivism, Jerome Bruner*. In Springer Texts in Education (pp. 259–275). Springer International Publishing.

Roeder, T., & Gatto, R. (Eds.). (2014). *Critical expressivism: Theory and practice in the composition classroom*. The WAC Clearinghouse; Parlor Press.

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

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Prentice Hall.

Sasaki, M., Mizumoto, A., & Matsuda, P. K. (2024). *Machine translation as a form of feedback on L2 writing*. International Review of Applied Linguistics in Language Teaching. https://doi.org/10.1515/iral-2023-0223

Shaw, S., & Weir, C. (2007). *Examining writing: Research and practice in assessing second language writing.* Cambridge University Press.

Sheikh, H., Prins, C., & Schrijvers, E. (2023*). Artificial intelligence: Definition and background*. In Research for Policy (pp. 15–41). Springer International Publishing.

SmartMarq: Essay marking with rubrics and AI. (2019, May 21). Assessment Systems. https://assess.com/smartmarq-ai-essay-scoring/

Sullivan, G. M. (2011). *A primer on the validity of assessment instruments.* Journal of Graduate Medical Education, *3*(2), 119–120. https://doi.org/10.4300/jgme-d-11-00075.1

Types of rubrics. (n.d.). Depaul.edu. Retrieved December 19, 2023, from https://resources.depaul.edu/teaching-commons/teaching-guides/feedback-grading/rubrics/Pages/types-of-rubrics.aspx

Universidad del Valle (2018) *Proyecto Educativo Resoluciones 056 y 062 para las cohortes admitidas a partir del año 2018*.

Universidad del Valle (2021). *Resolución del examen de competencias de la Licenciatura en Lenguas Extranjeras.*

Valid and reliable assessments. CSAI update. (2018). Center on Standards and Assessments Implementation. http://files.eric.ed.gov/fulltext/ED588476.pdf

Van Dijk, T. A. (1980). *Macrostructures: An introduction to the analysis of discourse. Academic Press.*

Verspoor, K., & Cohen, K. B. (2013). *Natural Language Processing*. In Encyclopedia of Systems Biology (pp. 1495–1498). Springer New York.

Walvoord, B. E., Stevens, D. D., & Levi, A. J. (2023). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Routledge.

Wang, M. (2022). Zhuihui Fang: *Demystifying academic writing: Genres, moves, skills, and strategies*. Journal of World Languages, *7*(3), 575–580. https://doi.org/10.1515/jwl-2021-0018

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weir, C. (2005). *Language testing and validation*. Palgrave.

Woodworth, J., & Barkaoui, K. (2020). *Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom*. TESL Canada Journal, 37(2), 234–247. https://doi.org/10.18806/tesl.v37i2.1340

Woodworth, K., & Barkaoui, K. (2020). *Automated writing evaluation tools in language assessment: A critical review of current applications and future directions*. Language Assessment Quarterly, 17(4), 370–390. https://doi.org/10.1080/15434303.2020.1813143

Yi, J.-Y. (n.d.). *Defining writing ability for classroom writing assessment in high schools*. Eric.ed.gov. Retrieved December 19, 2023, from https://files.eric.ed.gov/fulltext/EJ921024.pdf

Zhang, S. (2021). *Review of automated writing evaluation systems.* Journal of China Computer-Assisted Language Learning, *1*(1), 170–176. https://doi.org/10.1515/jccall-2021-2007

Appendices

Appendix 1. Informed Consent

Percepciones de proficiencia en inglés

Por favor responda esta breve encuesta sobre sus percepciones de proficiencia en el idioma inglés. Se le compartirán los resultados comparativos entre sus percepciones y desempeño real.

* Indicates required question

1. Email *

2. Autorizo que la información brindada en este cuestionario y los textos del examen se utilicen anónimamente y para propósitos académicos. *

   Mark only one oval.

   - Sí
   - No

Skip to question 3

Información personal

3. Nombres y apellidos completos *

4. Código *

Consentimiento Informado
Coordinadora de la Sección de inglés

Santiago de Cali, 24 de febrero de 2024

Profesora
María Eugenia Guapacha
Coordinadora Sección de Inglés
Licenciatura en Lenguas Extranjeras Inglés-Francés

Informed Consent

You are invited to participate in the research study entitled "*Human Raters vs Chat GPT: How Do Their Scorings of EFL Learners' Academic Writing Differ*?". This research project is conducted by Valentina Zapata as a requirement for the degree of Bachelor in Foreign Languages (English - French) from the School of Language Sciences at Universidad del Valle.

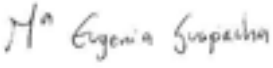
This study aims to compare the ratings awarded to academic essays between human raters/teachers and Chat GPT, an AI-based language model. I am requesting access to the essays written by 9th-semester students during the writing section of the Examen de Competencias to compare the ratings produced by the raters of the English Department and those generated by AI (Chat GPT) . The access will be limited to the essays only and will not include any personal identifying information about the students. The essays will be used solely for research purposes and will be anonymized to protect the privacy of the students. Granting me access to the essays is entirely voluntary, and all information provided will be kept confidential.

There are no potential risks associated with granting access to the essays. The benefits of participating in this study include contributing to the understanding of the use of AI in EFL writing and assessment, which may inform future educational practices.

If you have any questions or concerns about the study, you may contact Valentina Zapata, at valentina.villano@correounivall.edu.co Additionally, if you have any questions regarding your rights as a research participant, you may contact the supervisor Maria Eugenia Guapacha at maria.guapacha@correounivalle.edu.co

Consent:
I have read and understood the information provided in this form, and I voluntarily agree to grant access to the essays written during the Examen de Competencias for the purpose of the study described above.

English Coordination's Name: María Eugenia Guapacha

English Coordination's Signature: 

Appendix 2. Analytic Rubric Task 1

TASK ACHIEVEMENT & CONTENT DEVELOPMENT	5	4.5	4	3.5	3.0	2.5	2.0	1.5	1.0	0
Task achievement refers to how well the candidate addresses the task and develops the ideas relevant to the task.(Standpoint statement + support with details and illustrative examples)	Fully addresses the task. Presents a clear and fully developed position with relevant, extended and well supported ideas, details and examples.		Sufficiently addresses the task. Presents a clear position, and presents, extends and supports main ideas with some details and examples.		Addresses the task. Although the position is presented, it is not sufficiently supported; there may be irrelevant detail and not enough examples.		Addresses the task only partially/ May lack focus. The position is unclear: supporting ideas, details / examples are unclear / repetitive / irrelevant.		Responds to the task only in a minimal way/the answer is completely unrelated to the task. Does not express a position. Presents few ideas, which are largely undeveloped or irrelevant.	Did not take the test
COHERENCE AND STYLE	**5**	**4.5**	**4**	**3.5**	**3.0**	**2.5**	**2.0**	**1.5**	**1.0**	**0**
Coherence: Organisation of whole text; logical paragraphing, topic sentences. Reader engagement	Very well organised: logical paragraphs, topic sentences, clear visual introduction and conclusion. Message followed effortlessly.		Well organised: logical paragraphs, topic sentences, relevant introductory statement, may be no concluding statement. Message clear.		Writing organised into paragraphs but some paragraphs are not well structured. Introduction or conclusion missing. Message generally clear.		Poor organisation: illogical/ splintered /misplaced / no paragraphing but ideas can be followed. Hard to follow (needs rereading) at some parts. Introduction and conclusion missing. Shortish script (200 - 240 words)		Lacks organisation: no paragraphing. Incomprehensible / hard to follow throughout. Insufficient length (less than 200 words) and substance.	Did not take the test
Cohesion inside text: Use of cohesive devices (linking expressions) and referencing (this, these, those, which, it)	Skilful non-formulaic use of a range of cohesion and referencing devices.		Appropriate use but slightly formulaic / lower range of cohesive devices. Good use of referencing.		Formulaic use /overuse of cohesive devices. Repetitious referencing.		Insufficient/simple or repetitive connectives (and,but,so). Some incorrect cohesive devices and unclear referencing.		Lacks cohesion; many cohesive devices incorrectly used. Disjointed script.	Did not take the test
Style: Formal,	Clear academic		Most aspects of		Some evidence of		Little evidence of		Overuse of personal	Did not

impersonal, academic tone or register; hedging (e.g., possibly, might, may, could be). Not oral, colloquial or informal. Informal language includes abbreviations, contractions, etc.	**style and good use of hedging**		**academic style with some hedging. May be some slight informality**		**academic style, some informal or colloquial language. No / repetitive hedging.**		**academic style: Frequent informal / colloquial language, wordiness. No hedging.**		**and informal / colloquial language.**	**take the test**
FORM	**5**	**4.5**	**4**	**3.5**	**3.0**	**2.5**	**2.0**	**1.5**	**1.0**	**0**
Sentence structure: accuracy, variety, and complexity of sentences; punctuation	**Controlled and varied sentence structure.**		**Adequate variety; occasional errors in complex sentences.**		**Limited range; errors in complex sentences may be frequent.**		**Limited control; many incomplete / rambling / convoluted sentences. Lacks punct.**		**Lacks control of sentence structure throughout-short script.**	**Did not take the test**
Grammar: parts of speech (verbs, articles, prepositions, agreement, plurals)	**Virtually error free**		**A few minor errors**		**Minor/repeated errors**		**A range of significant errors affect expression of ideas.**		**Short script (less than 200 words) with a range of significant errors affecting expression of ideas.**	**Did not take the test**
Vocabulary and spelling: Appropriate or accurate use of words; Range (richness or sophistication): Form (spelling and word form)	**Wide range of vocabulary, appropriately used. Very few minor misspellings.**		**Vocabulary generally appropriate and good range. Some minor misspellings.**		**Vocabulary appropriate but lacks range. Few errors in word use. Some important, intrusive spelling errors.**		**Vocabulary simple / repetitive / frequently inaccurate or inappropriate. Many intrusive misspellings.**		**Range and use of vocabulary inadequate. Numerous errors in word use and spelling disproportionate to length and complexity of script.**	**Did not take the test**

Appendix 3. Assumption Checks

Test of Normality (Shapiro-Wilk) Task 1- 2022

		W	P
Task 1 teachers 2022	Task 1 ChatGPT 2022	0.980	0.520

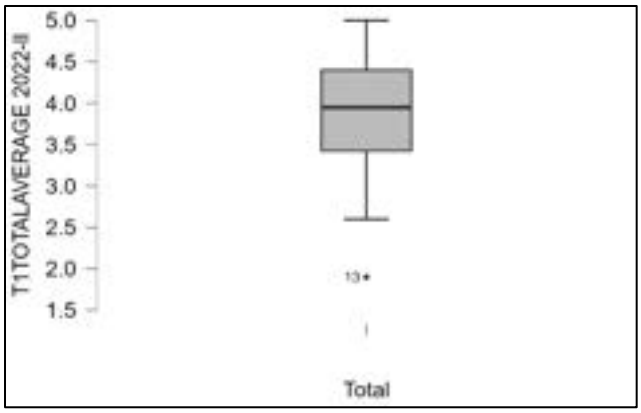
Note. Significant results suggest a deviation from normality.

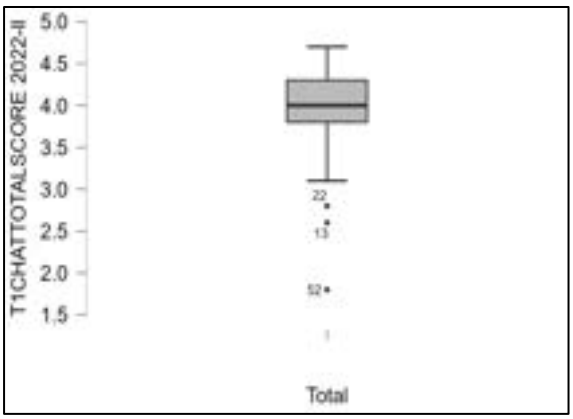
Test of Normality (Shapiro-Wilk) Task 2- 2023

		W	P
Task 2 teachers 2023	Task 2 ChatGPT 2023	0.969	0.206

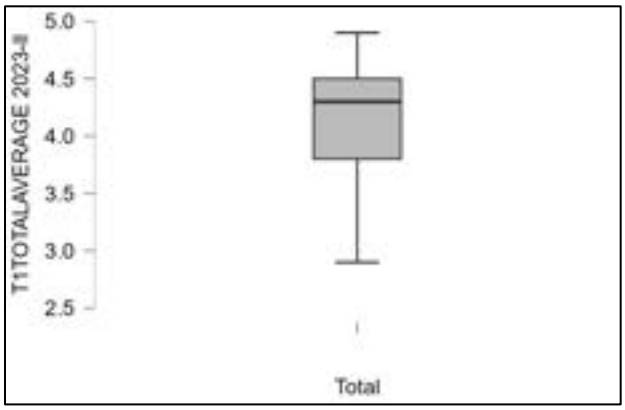
Note. Significant results suggest a deviation from normality.

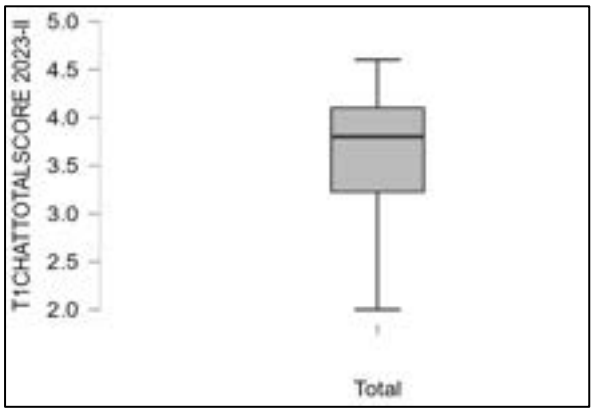
Task 1- 2022 Human Raters vs ChatGPT



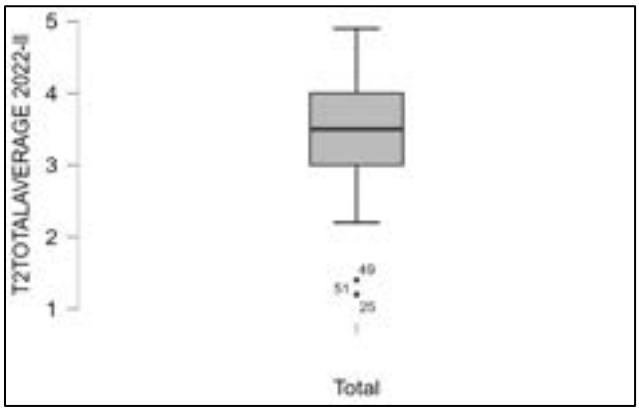


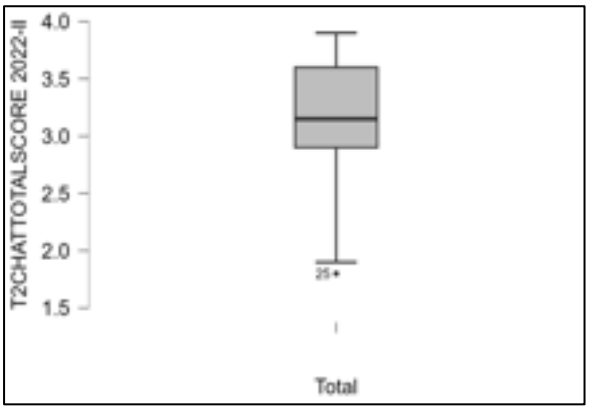
Task 1-2023 Human Raters vs ChatGPT





Task 2-2022 Human Raters vs ChatGPT





Task 2-2023 Human Raters vs ChatGPT

