Annual Review of Applied Linguistics (2025), 1–21
doi:10.1017/S0267190525100160

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance

Aysel Saricaoglu[1] and Zeynep Bilki[2]

[1]Department of English Language and Literature, Social Sciences University of Ankara, Ankara, Türkiye and [2]Department of Foreign Language Education, TED University, Ankara, Türkiye
Corresponding author: Zeynep Bilki; Email: zeynep.bilki@tedu.edu.tr

Abstract
This study examined the capacity of ChatGPT-4 to assess L2 writing in an accurate, specific, and relevant way. Based on 35 argumentative essays written by upper-intermediate L2 writers in higher education, we evaluated ChatGPT-4's assessment capacity across four L2 writing dimensions: (1) Task Response, (2) Coherence and Cohesion, (3) Lexical Resource, and (4) Grammatical Range and Accuracy. The main findings were (a) ChatGPT-4 was exceptionally accurate in identifying the issues across the four dimensions; (b) ChatGPT-4 demonstrated more variability in feedback specificity, with more specific feedback in Grammatical Range and Accuracy and Lexical Resource, but more general feedback in Task Response and Coherence and Cohesion; and (c) ChatGPT-4's feedback was highly relevant to the criteria in the Task Response and Coherence and Cohesion dimensions, but it occasionally misclassified errors in the Grammatical Range and Accuracy and Lexical Resource dimensions. Our findings contribute to a better understanding of ChatGPT-4 as an assessment tool, informing future research and practical applications in L2 writing assessment.

Keywords: ChatGPT-4; L2 writing assessment; accuracy; specificity; relevance

Introduction

The emergence of generative artificial intelligence (AI) technologies marks a significant shift in the landscape of automated writing assessment. Built on transformer architecture and attention mechanisms, large language models (LLMs) can perform complex tasks from summarization, content generation, image creation, code generation, machine translation, etc. (Thompson, 2024). One such model is ChatGPT, a chat-specific model developed by OpenAI (https://openai.com/) and trained on an extraordinarily large database (Gans, 2023). The capabilities of ChatGPT to process

textual input and perform text analysis have attracted considerable attention from second language (L2) writing teachers and researchers.

L2 writing researchers have considered three main aspects when evaluating the capacity of ChatGPT for L2 writing assessment: scoring accuracy, scoring consistency, and error detection accuracy. Studies that have compared ChatGPT scores with human scores have found moderate to strong correlations between the two, considering ChatGPT as a potential and useful tool for automated essay scoring (AES) in L2 writing research and practice (e.g., Latif & Zhai, 2024; Mizumoto & Eguchi, 2023; Mizumoto et al., 2024; Shin & Lee, 2024; Yamashita, 2024). A few studies that have compared ChatGPT-detected errors with human-detected errors have generally yielded positive results, displaying exceptionally high precision, which means that ChatGPT correctly identifies errors, but weaker recall, which means that it frequently misses errors identified by humans (e.g., Pfau et al., 2023; Xu et al., 2024). Regarding its consistency, studies have reported mixed findings. Bui and Barrot (2024) revealed that scoring by ChatGPT was inconsistent across different scoring rounds, while Yamashita (2024) observed "extremely" consistent scoring of ChatGPT.

Previous quantitative studies on ChatGPT have provided valuable evidence regarding ChatGPT's scoring accuracy, consistency, and error detection accuracy, which are important aspects of L2 writing assessment. However, the evidence they have offered is limited in that they have mainly focused on scoring agreement between ChatGPT and humans and generally have addressed ChatGPT's linguistic error detection capacity. While examining agreement between automated scores and human-assigned scores is useful, Attali (2007, p. 2) aptly notes:

> Agreement results tell us little about what is measured by automated scores and thus do not contribute to construct validation of AES [Automated Essay Scoring]. Automated scores may have high agreement with human scores and still suffer from construct underrepresentation or construct-irrelevant variance.

Therefore, when evaluating the potential of ChatGPT for L2 writing assessment, we must look for evidence demonstrating that the system covers the critical dimensions of L2 writing proficiency (Ben-Simon & Bennett, 2007). Moreover, while existing studies have mainly addressed ChatGPT's capacity for assessing linguistic (i.e., grammatical and lexical) accuracy, other equally important but more complex dimensions of L2 writing, such as content development and cohesion and coherence, have been overlooked by researchers. For a complete understanding of the potential of ChatGPT as an L2 writing assessment tool, more comprehensive evaluations focusing on what aspects of L2 writing dimensions ChatGPT covers are needed (Shi & Aryadoust, 2024).

The present study explored ChatGPT-4's capacity to assess L2 writing in an accurate, specific, and relevant way. To that end, we prompted ChatGPT-4 to evaluate 35 argumentative essays written by upper-intermediate L2 writers in higher education using the analytic rubric of a high-stakes standardized L2 writing exam. We examined the feedback generated by ChatGPT-4 across four dimensions of the L2 writing construct as Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. Our exploration of the accuracy, specificity, and relevance of ChatGPT's assessment of L2 writing is limited to ChatGPT feedback on the student texts. Thus, our

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

evaluation is based on ChatGPT-4-identified strengths and weaknesses in L2 writing (i.e., precision) and does not include the strengths and weaknesses that ChatGPT-4 fails to identify (i.e., recall), as we are unable to compare its performance to that of human assessors as the gold standard. A recall analysis fell beyond the scope of our investigation, given that the process of manual examining not only the accuracy but also the specificity and relevance of ChatGPT-4's assessment across four dimensions of L2 writing, based on 1,795 pieces of feedback, requires substantial time and effort.

Automated L2 writing assessment

Assessing students' written work manually presents considerable challenges for language educators as it often requires substantial time and effort due to the need to address various constructs of L2 writing, such as meaning, coherence, cohesion, style, format, vocabulary, and grammar (Graham, 2019). The advent of computer-based educational technologies has introduced alternative methods to manual assessment, such as AES or automated writing evaluation, which are used to support both formative and summative assessment practices.

The potential of automated systems in writing assessment has captured the attention of researchers seeking to evaluate their effectiveness, emphasizing key areas such as validity, teacher and/or student perceptions, impact, and factors that influence students' and/or instructors' utilization of these systems (Shermis & Wilson, 2024). Previous research frequently investigated the validity of automated systems by analyzing the accuracy and range of feedback types provided as well as their reliability by comparing automated scores to human scores (e.g., Dikli & Bleyle, 2014; Vo et al., 2023). Existing research has proved the effectiveness of automated systems in delivering consistent evaluations using predefined criteria and providing tailored suggestions for improvement, particularly in addressing linguistic dimensions and surface-level aspects such as grammar and mechanics (Fu et al., 2022). However, concerns remain about their effectiveness in evaluating more complex features such as idea development and organizational aspects (Fu et al., 2022; Shi & Aryadoust, 2024).

With the remarkable capabilities of generative AI in understanding and generating meaningful content, LLMs such as ChatGPT are now utilized in language assessment practices (Barrot, 2023; Mizumoto & Eguchi, 2023). The use of ChatGPT as an assessment tool has also become a major research topic in L2 writing.

ChatGPT as an emerging L2 writing assessment tool

Although limited in number, existing studies on the effectiveness of ChatGPT to assess L2 writing have mainly focused on its accuracy and consistency in scoring and error detection, with generally positive findings. In one of the earliest studies, Mizumoto and Eguchi (2023) examined the essay scoring accuracy of OpenAI's text-davinci-003 model, part of the GPT-3.5 series. Their corpus included 12,100 essays written to eight prompts by TOEFL test takers in 2006 and 2007 with different first languages. They classified the scores of the essays into three levels as low, medium, and high and prompted the text-davinci-003 model to score each essay from 0 to 9 based on the IELTS TASK 2 Writing band descriptors. According to their findings, despite some

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

variation in the 1–2 points scoring, the text-davinci-003 model reflected the three writing levels of the written responses. In another study, Bui and Barrot (2024) compared ChatGPT-given scores (GPT-3.5) with those given by an experienced human rater. The researchers also examined the consistency of ChatGPT by analyzing its scores taken at multiple time points. Their learner corpus included 200 argumentative essays from the ICNALE-Written Corpus, which college students from Asian countries at four different proficiency levels wrote (i.e., A2 waystage, B1 threshold lower, B1 threshold upper, and B2 vantage or higher). They used the argumentative version of the Common Core State Standards-aligned writing rubric including five main criteria: claim, development, audience, cohesion, and style and conventions. According to their correlational analysis, ChatGPT's scores were not closely aligned with the scores of the experienced human rater and were not consistent across different rounds of scoring. Similarly, Shin and Lee (2024) compared the scores of My GPT, a customized chatbot based on GPT-4, with human scores for 50 English essays written by Korean EFL students. They prompted My GPT to evaluate the essays for four scoring domains as Task Completion, Content, Organization, and Language Use by using a rubric that they provided and to generate constructive feedback for formative assessment purposes. They also provided My GPT with a sample question and sample scores for each scoring domain as a reference. Shin and Lee (2024) observed high correlations between chatbot scores and human scores across the four domains.

A few other studies of L2 writing assessment have addressed the effectiveness of ChatGPT in accurately detecting linguistic errors. Pfau et al. (2023) evaluated ChatGPT-4's precision (i.e., correctly identified errors) and recall (i.e., missed errors) by comparing its error identification to human error identification. Their corpus included 100 essays written by Greek learners of English for a large-scale test that assessed skills at the Common European Framework of Reference's (CEFR's) B2 level. They also examined ChatGPT-4's consistency in identifying errors by reprocessing 30 randomly selected essays through ChatGPT-4 2 months later. Using a 5-point analytic scale – 1 (*fail*), 2 (*narrow fail*; satisfies some, but not all B2 criteria), 3 (*marginal pass*; B2), 4 (*clear pass*; exhibits B2/C1 features), and 5 (*honors pass*; exhibits some C1/C2 features) – the researchers reorganized the original scores given to the essays into five bands, categorized according to the grammar subscores of 1–1.5, 2, 3, 4, and 4.5–5. Pfau et al. (2023) observed that ChatGPT had a high precision: 99% of the errors were correctly identified as errors by ChatGPT. ChatGPT's recall was lower: 69% of the errors were missed. ChatGPT's error detection capacity was stronger with higher score levels. In another study, using the Cambridge Learner Corpus First Certificate in English dataset, including written responses from 232 Asian test-takers, Mizumoto et al. (2024) compared the accuracy of ChatGPT in detecting grammatical errors to human coders and Grammarly. Their findings revealed that ChatGPT's error detection performance was comparable to human evaluators, with a stronger correlation ($\rho = 0.79$) than Grammarly ($\rho = 0.69$). Addressing the prompt effects, Xu et al. (2024) examined the impact of prompt settings on ChatGPT-4 in assessing the grammatical accuracy of 100 essays written by L1 Greek learners of L2 English. They used two prompts (i.e., more detailed prompts with definition and examples of errors and less detailed prompts without definition and examples of errors) to examine ChatGPT-4's performance in detecting grammatical errors compared to human-detected errors. They found out that

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

ChatGPT-4 generally detected fewer errors than humans. Regarding the prompt settings, they found that prompts with a detailed definition of errors resulted in higher correlations with error counts detected by humans than those without a definition. Xu et al. (2024) concluded that more detailed prompts enhance ChatGPT-4's error detection capacity.

As the review above clearly indicates, the potential of ChatGPT as an L2 writing assessment tool is already recognized by L2 writing researchers. The quantitative studies in the literature provide evidence of its scoring accuracy based on high correlations with human scores (except Bui & Barrot, 2024). However, high correlations between automated scores and human scores do not necessarily mean that they are based on the same constructs. High correlations are an essential condition for the validity of automated scoring but are not considered sufficient (Shermis & Burstein, 2003; Weigle, 2013). Studies in the literature also provide evidence for ChatGPT's error detection accuracy. However, the evidence is limited to linguistic error detection. ChatGPT's capacity for assessing other dimensions of L2 writing, such as content development and cohesion and coherence is less studied.

The present study aimed to explore ChatGPT-4's capacity for L2 writing assessment with a specific focus on accuracy, specificity, and relevance. To that end, we prompted ChatGPT-4 to evaluate 35 argumentative essays written by upper-intermediate L2 writers in higher education using the analytic rubric of a high-stakes standardized L2 writing exam. We asked ChatGPT-4 to evaluate each student essay for four main dimensions in the rubric: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. We then qualitatively examined each piece of ChatGPT-4 feedback on each essay to evaluate whether it was accurate, specific, and relevant. We specifically sought to answer the following research questions:

1. To what extent is ChatGPT-4's assessment of student essays accurate, specific, and relevant across different dimensions of L2 writing?
2. To what extent is the accuracy, specificity, and relevance of ChatGPT-4's assessment associated with specific dimensions of L2 writing?

Our evaluation of the accuracy, specificity, and relevance of ChatGPT's assessment of L2 writing is based on ChatGPT-generated feedback on the student texts; in other words, ChatGPT-identified strengths and weaknesses in L2 writing. In our study, accuracy refers to the accuracy of ChatGPT's identification of a strength or weakness in student writing; more specifically, whether ChatGPT-4 correctly or incorrectly identified a weakness or strength in the student text. Specificity refers to ChatGPT's identification of a specific or general strength or weakness in student writing. Relevance refers to ChatGPT's identification of a criterion-related strength or weakness in student writing (i.e., specific criteria within each key dimension).

Methodology

Learner corpus

The learner corpus used in this study was part of a bigger corpus collected for an MA thesis supervised by the second author, and access to the corpus was granted to the

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

authors of this study with the permission of the MA student (Taş, 2024). The corpus consisted of 35 argumentative essays written by undergraduate Turkish students of English enrolled in the English language preparatory program of a private university in Ankara, Türkiye, where English was the medium of instruction. According to the CEFR, students were at the upper-intermediate level of English (B2 CEFR) (Council of Europe, 2001). The corpus included essays from two classes taught by two different instructors. The argumentative essay writing task was completed in class by students responding to the following prompt: "There is no real function of literature or cinema in our society. Do you agree or disagree with this statement? Please explain." They were required to write a minimum of 220 words. The corpus used in this study consisted of 9,495 words, with an average of 271 words per essay ($SD = 48.19$), and a total of 642 sentences, with an average of 18 sentences per essay ($SD = 4.06$).

Type of GPT and prompt

To answer our research questions, we used the generic GPT-4 chatbot, the Generative Pre-trained Transformer 4 developed by OpenAI and released in March 2023, rather than a customized AI model. The publicly accessible form better aligns with our aim of examining ChatGPT's potential for L2 writing assessment, as most teachers, students, test developers, and assessment specialists would use this version in real-world contexts. We used the subscription version (GPT-4) rather than the free version (GPT-3.5), as the free model offers a limited number of chats per session, which would have limited our ability to conduct a consistent evaluation of ChatGPT. The subscription version allowed us to have uninterrupted access to the model, thus using it more systematically.

For the ChatGPT-4 prompt, we used key criteria from the IELTS Writing Task 2 rubric: (1) Task Response, (2) Coherence and Cohesion, (3) Lexical Resource, and (4) Grammatical Range and Accuracy. We prompted ChatGPT-4 to evaluate each essay based on the four key criteria and related descriptors. For example, for Task Response, we asked ChatGPT-4 to evaluate the given essay for Task Response based on the given related sub-criteria and to identify the problems in each task response criterion specifically. We chose the key criteria from the IELTS Writing Task 2 rubric for several reasons. First, the IELTS writing rubric is an analytic rubric that provides criterion-specific assessment addressing different aspects of L2 writing. An analytic rubric better aligns with the aim of this study, as we are interested in exploring the extent to which ChatGPT-4 covers those specific aspects of L2 writing when automatically assessing written responses. Second, the specific descriptors in the four key criteria of the IELTS writing rubric helped with the prompt specificity. Prior research suggests that more specific or detailed prompts increase the performance of ChatGPT (e.g., Xu et al., 2024). Lastly, the dataset used in this study comes from a writing task that is similar to IELTS Academic Writing Task 2, where test takers are presented with a point of view and asked to write about whether they agree or disagree with that view in an essay format. Similarly, the writing task used in this study asked students to write an essay about whether they agree or disagree with an argumentative statement. Such independent argumentative writing tasks are commonly utilized in writing classes of preparatory English language programs in higher education in Türkiye, including at the university where data for this study were collected (for a similar real-world example of an

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

independent argumentative writing task, see the TED University English Proficiency Exam [TEDU EPE]; TED University English Language School, 2020, p. 27). Thus, our decision to use the IELTS Writing Task 2 criteria was deemed appropriate by the parallel writing tasks in this study and in the IELTS writing test.

In our ChatGPT prompt, we included only descriptors from the highest band of the IELTS writing rubric, that is, we did not include the band descriptors from all proficiency levels. The reason for this was because we aimed to explore to what extent ChatGPT addressed the key aspects of L2 writing as defined by the IELTS rubric, but not to evaluate the scoring accuracy of ChatGPT across different proficiency levels. We only included the descriptors from the highest band to indicate what the full expectations were regarding the specific aspects of L2 writing in that band. While our study does not focus on the quality or effectiveness of ChatGPT's feedback on these aspects of writing, one of the best ways to investigate whether ChatGPT assesses L2 writing in a specific, accurate, and relevant way is through a qualitative examination of the feedback it generates on these key aspects. Therefore, we prompted ChatGPT-4 to identify the problems related to each criterion specifically to be able to examine its feedback in detail.

Given that prompt features are critical in successful automated language assessment with AI, we used an iterative process of prompt engineering. Initially, we reviewed the relevant literature to identify sample prompts used in similar studies (e.g., Mizumoto et al., 2024; Pfau et al., 2023; Yoon et al., 2023). We then created and tested our initial prompt to evaluate the specificity of the feedback ChatGPT-4 generated with that prompt. Through repeated trials and qualitative evaluations, we modified our prompt in wording and specificity to achieve the maximum effectiveness from ChatGPT-4 in accuracy, specificity, and relevance.

To facilitate the automated assessment and feedback generation process using ChatGPT-4, we developed a Python script. This script processed essays in batches, submitted them to ChatGPT-4, and retrieved the generated feedback in a structured Excel file. We manually extracted the automated feedback from the Excel file for coding and further analysis. (See the supplementary materials for the Python script, a sample student essay (SA1), ChatGPT-4 feedback on SA1, and the coding of the ChatGPT-4 feedback on SA1.)

Data coding

To examine the accuracy, specificity, and relevance aspects, we manually coded each piece of feedback generated by ChatGPT-4 for each essay. For the coding process, we first carefully read the student-written draft, and then, using an Excel sheet, we coded each specific piece of feedback for that draft in three coding categories with two codes in each category: (1) Accuracy: accurate or inaccurate; (2) Specificity: specific or general; and (3) Relevance: relevant or irrelevant.

In the accuracy coding category, we evaluated whether ChatGPT-4 correctly or incorrectly identified a weakness or strength in the student text. In other words, we evaluated whether the ChatGPT-4 feedback on the student text was accurate or inaccurate. For example, we coded the following feedback on the Lexical Resource criterion as accurate: "The phrase 'on different topics' in 'looking on different topics' should be 'at different topics'" (S19A). On the other hand, we coded the following feedback on the

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

Lexical Resource criterion as inaccurate since the error was grammatical, not lexical: "'affected' should be 'affect' in: 'people can affected by some things'" (S22A).

In the specificity coding category, we evaluated whether ChatGPT-4 identified a specific or general strength or weakness in the student text. We coded feedback as specific when it referred to specific parts of the student text, which often offered actionable suggestions for improvement. We coded feedback as general when it did not point to any specific aspects of the student text, but was vague, providing general statements. For example, we coded the following feedback on the Cohesion and Coherence criterion as specific: "The second paragraph introduces history books and movies but jumps to a quote by Ataturk without adequately connecting the quote to the initial idea" (S3A). On the other hand, we coded the following feedback on the Cohesion and Coherence criterion as general: "Each body paragraph introduces one of the reasons but lacks sufficient elaboration or supporting details" (S4A).

In the relevance coding category, we evaluated whether the feedback was relevant or irrelevant to the target criterion. For example, we coded the following feedback on the Grammatical Range and Accuracy criterion as irrelevant since spelling errors were addressed in the Lexical Resource dimension of the rubric: "'they are colorfull things in a black and white world' the word 'colorfull' should be 'colorful'" (S9A). On the other hand, we coded the following feedback on the Grammatical Range and Accuracy criterion as relevant: "'This essay discuss' should be 'This essay discusses'" (S24A). Table 1 provides coded samples of feedback from our dataset.

In cases where ChatGPT-4 feedback partially belonged to a specific code, which was infrequent in our dataset, we divided the feedback into parts and coded each part separately. For example, we divided the following feedback on the Coherence and Cohesion dimension into two parts and coded the first part of the feedback as general and the second part of the feedback as specific: "The essay could benefit from better-organized paragraphs. The ideas presented are relevant but often feel disjointed and not fully developed within each paragraph." (general) "For example, the second paragraph introduces history books and movies but jumps to a quote by Ataturk without adequately connecting the quote to the initial idea" (specific) (S3A).

The automated feedback that ChatGPT-4 generated on some essays also included formative feedback that provided suggestions for improving the student text based on specific criteria. For example, for Essay 3A and the third criterion of the Task Response (i.e., "Ideas are relevant, well-extended, and supported"), ChatGPT-4 provided the following feedback:

> Extend your arguments with specific examples, statistics, or studies that support your claims. For instance, if stating that history books impact politics, cite particular books and historical events they influenced. If arguing that crime is linked to television programs, provide data or case studies.

For some other essays, ChatGPT-4 provided examples of revised introduction, body, or conclusion paragraphs. Since our focus in this study was not the potential of ChatGPT-4 as a formative assessment tool, we did not code such feedback and removed it from our analysis. In total, we coded 1,795 pieces of feedback, with the majority falling under the Grammatical Range and Accuracy dimension ($N = 514$), followed by an equal number of pieces of feedback in the Lexical Resource dimension ($N = 468$) and the

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

Table 1. Coded samples from the dataset

Essay	Dimension	Criterion	ChatGPT-4 Feedback	Accuracy	Specificity	Relevance
S19A	**Task Response**	1 – The task prompt is appropriately and sufficiently addressed and explored in depth.	While the essay touches on various effects, it lacks depth in its exploration, providing general statements without thoroughly examining how these impacts manifest in specific examples or studies.	Accurate	General	Relevant
S7A	**Task Response**	1 – The task prompt is appropriately and sufficiently addressed and explored in depth.	The essay mentions some notable figures in Turkish literature and cinema but does not delve into how their work contributes to the functions of literature and cinema in a broader societal context.	Accurate	Specific	Relevant
32A	**Task Response**	2 – A clear and well-developed position is presented in response to the task prompt.	Although the essay presents a clear position (disagreement with the statement), it is not well-developed. The position needs to be elaborated more comprehensively to enhance clarity.	Accurate	General	Relevant
S10A	**Task Response**	2 – A clear and well-developed position is presented in response to the task prompt.	The position of the writer is not very clear and slightly contradictory. The writer begins by agreeing with the statement that "there is no real function" but then discusses how people in Turkey give value to cinema, which suggests cinema does have a function.	Accurate	Specific	Relevant
S21A	**Task Response**	3 – Ideas are relevant, well-extended, and supported.	Ideas are generally relevant but are not sufficiently extended or well-supported with concrete examples.	Accurate	General	Relevant
S3A	**Task Response**	3 – Ideas are relevant, well-extended, and supported.	The arguments provided (e.g., the impact of history books and movies, and the influence of sensational TV series on crime rates) are not sufficiently developed with concrete examples or detailed analysis.	Accurate	Specific	Relevant

Table 1. (*Continued.*)

Essay	Dimension	Criterion	ChatGPT-4 Feedback	Accuracy	Specificity	Relevance
S33A	**Coherence and Cohesion**	1 – Information and Ideas Are Logically Organized Within and Across Paragraphs, and There Is a Clear Progression Throughout the Response	The essay follows a logical structure with an introduction, body paragraphs addressing different points, and a conclusion.	Accurate	General	Relevant
S11A	**Coherence and Cohesion**	1 – Information and Ideas Are Logically Organized Within and Across Paragraphs, and There Is a Clear Progression Throughout the Response	The progression of ideas within paragraphs is somewhat linear, but transitions between some ideas are not very smooth. For example, the transition between discussing the societal importance of written works and then jumping to the influence of cinema is abrupt.	Accurate	Specific	Relevant
S35A	**Coherence and Cohesion**	2 – A range of cohesive devices, including reference and substitution, is used flexibly.	Words such as “Furthermore,” “Thanks to,” “As we getting socialize,” and “Additionally” are used but often incorrectly or awkwardly.	Accurate	Specific	Relevant
S14A	**Coherence and Cohesion**	2 – A range of cohesive devices, including reference and substitution, is used flexibly.	Transitions between ideas are present but could be more fluid.	Accurate	General	Relevant
S32A	**Coherence and Cohesion**	3 – Paragraphing is used effectively and appropriately to support overall coherence.	There is a need for better topic sentences. The paragraphs often begin in a way that doesn’t clearly indicate what the paragraph will discuss. For example, “Firstly I will mention how film or literature genres effect our society” could be more specific.	Accurate	Specific	Relevant

Table 1. (*Continued.*)

Essay	Dimension	Criterion	ChatGPT-4 Feedback	Accuracy	Specificity	Relevance
S4A	**Lexical Resource**	1 – Wide range of vocabulary used accurately and appropriately to convey precise meanings	Some words are incorrectly or awkwardly used, leading to unclear or imprecise meanings. For example, "product art" should be "produce art," and "appearance of real functions" is not idiomatic.	Accurate	Specific	Irrelevant[1]
S8A	**Lexical Resource**	2 – Collocational, idiomatic, and sophisticated lexical features used accurately and appropriately.	The essay lacks idiomatic expressions and more sophisticated lexical features.	Accurate	General	Relevant
29A	**Lexical Resource**	3 – Spelling and word formation are used accurately and appropriately.	There are instances of word form errors. For example, "doing a research" should be "doing research" (uncountable noun).	Accurate	Specific	Irrelevant[2]
S4A	**Grammatical Range and Accuracy**	1 – A wide range of complex structures is used accurately and flexibly.	The essay employs a limited range of complex grammatical structures, and there are several inaccuracies.	Accurate	General	Relevant
S9A	**Grammatical Range and Accuracy**	2 – Punctuation and grammar are used accurately and appropriately.	"Firstly, literature is a magic that makes people think about other people and other people's feelings." The comma is correctly used, but "magic" should be "magical."	Inaccurate	Specific	Irrelevant[3]
S30A	**Grammatical Range and Accuracy**	2 – Punctuation and grammar are used accurately and appropriately.	"To begin antisocial society" should be "to begin with antisocial society."	Accurate	Specific	Relevant

Notes: [1]Relevant to Criterion 2 of Grammatical accuracy and range – Punctuation and grammar are used accurately and appropriately.
[2]Relevant to Criterion 3 of Lexical resource – Spelling and word formation are used accurately and appropriately.
[3]Relevant to Criterion 3 of Lexical resource – Spelling and word formation are used accurately and appropriately.

Coherence and Cohesion dimension ($N = 476$). The Task Response dimension had the lowest number of pieces of feedback ($N = 346$).

After an initial collaborative coding session in which we co-practiced applying the codes to ChatGPT-4 feedback, we cross-checked each other's judgments and resolved ambiguities and disagreements. Subsequently, we divided the dataset equally, each coding half. To assess inter-coder reliability, the first researcher independently coded eight randomly selected feedback files of the set that the second researcher coded ($N = 7$; 20%). The reliability analysis was conducted separately for each assessment feature based on percentage agreement, a common inter-coder reliability metric in applied linguistics (Kim et al., 2024; Rau & Shih, 2021). The results demonstrated high levels of agreement between the two researchers for the three aspects examined: 98.5% for accuracy, 91.2% for specificity, and 92.7% for relevance.

Data analysis

To answer the first research question, we used descriptive statistics (i.e., counts and frequencies) to report the accuracy, specificity, and relevance of ChatGPT-4's assessment of the learner essays for overall and specific criteria within the dimensions. To answer the second research question, following the common statistical practices in L2 writing research, we used a Chi-square test of independence to find out the associations between categorical variables (e.g., Lan et al., 2019; Seo & Oh, 2024). In our study, the categorical variables were assessment features of ChatGPT-4 (i.e., accurate – inaccurate; specific – general; and relevant – irrelevant) and key dimensions of writing (i.e., Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy). To calculate the strength of the associations, we used Cramer's V, a widely accepted effect size measure for Chi-Square tests, particularly dealing with categorical data in contingency tables (Cohen, 1988; Field, 2018). We interpret the Cramer's V results as the following: $\geq$0.1 weak association, $\geq$0.3 moderate association, $\geq$0.5 strong association (Cohen, 1988). Since we analyzed multiple dimensions of L2 writing against multiple assessment features in our study, Cramer V was appropriate as it accounts for varying table sizes, making it suitable for interpreting relationships across multiple categorical variables in L2 writing research (Plonsky & Oswald, 2014). To pinpoint which specific dimensions contributed to the association between writing dimensions and assessment features, we conducted a post hoc test using adjusted residuals to find out which writing dimension largely contributed to the association (Lan et al., 2019, 2022; Seo & Oh, 2024). Adjusted residuals greater than 2.00 are considered significant contributors to the Chi-square value based on the ±2 criteria (Agresti, 2018; Lan et al., 2022). To avoid alpha inflation due to multiple Chi-square tests, we applied Bonferroni adjustment, which resulted in a significance value of $p \leq .002$. Based on the adjusted alpha, we determined the critical value as ±3.09.

Findings

The descriptive results for the accuracy of ChatGPT-4 assessment, as shown in Table 2, demonstrate a high level of accuracy across the four dimensions of L2 writing, with the highest accuracy rate in Coherence and Cohesion (99.4%), followed by an equal

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

Table 2. Descriptive results for accuracy across L2 writing dimensions

	Accurate	Inaccurate	Total
	N (%)	*N* (%)	*N* (%)
Task Response	340 (98.3)	6 (1.7)	346 (100)
Coherence and Cohesion	464 (99.4)	3 (0.6)	467 (100)
Lexical Resource	460 (98.3)	8 (1.7)	468 (100)
Grammatical Range and Accuracy	493 (95.9)	21 (4.1)	514 (100)
Total	1757 (97.9)	38 (2.1)	1795 (100)

accuracy rate in both Task Response and Lexical Resource (98.3%), and with the lowest accuracy rate in Grammatical Range and Accuracy (95.9%).

Below are some representative examples of accurate and inaccurate feedback by ChatGPT-4 for each L2 writing dimension:

- The writer mentions that literature and cinema have educational and cultural impacts but does not explore these impacts deeply. (S1A, Task Response, Accurate)
- The essay begins with a rhetorical question and a comparison between the popularity of literature and cinema, but it doesn't clearly state the writer's stance on the given topic. (S18A, Task Response, Inaccurate as the student wrote "There is real function of cinema in our society")
- While each paragraph starts with a topic sentence, the sentences could more clearly relate to the thesis and outline the specific focus of the paragraph. (S5A, Coherence and Cohesion, Accurate)
- "Give precious to" is an incorrect collocation. The writer should use "value" or "appreciate" instead. (S10A, Lexical Resource, Accurate)
- Incorrect word formation: "an artworks" should be "a work of art" or "artworks are." (S15A, Lexical Resource, Inaccurate)
- There is limited evidence of advanced grammatical structures such as conditional sentences, relative clauses, or complex noun phrases. (S13A, Grammatical Range and Accuracy, Accurate)
- Incorrect use of singular and plural forms: "Every country has different thoughts and cinema plots." (S1A, Grammatical Range and Accuracy, Inaccurate)

The descriptive results for the specificity of ChatGPT feedback (see Table 3) showed high variability across the four dimensions of L2 writing. Most of the feedback in the Grammatical Range and Accuracy and Lexical Resource dimensions were specific (79.7% and 78.6%, respectively). The specificity rates in the Coherence and Cohesion dimension were moderate, with around half of the feedback being specific (46%) and the other half general (54%). The feedback in the Task Response dimension had the lowest specificity rate, with the most feedback being general (63.3%).

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

Table 3. Descriptive results for specificity across L2 writing dimensions

	Specific	General	Total
	N (%)	*N* (%)	*N* (%)
Task Response	127 (36.7)	219 (63.3)	346 (100)
Coherence and Cohesion	215 (46.0)	252 (54.0)	467 (100)
Lexical Resource	373 (79.7)	95 (20.3)	468 (100)
Grammatical Range and Accuracy	404 (78.6)	110 (21.4)	514 (100)
Total	1119 (62.3)	676 (37.7)	1795 (100)

Below are some representative examples of specific and general feedback by ChatGPT-4 for each L2 writing dimension:

- Financial problems are mentioned but not explained in terms of how they specifically affect access to literature and cinema. (S4A, Task Response, Specific)
- Some sentences do not clearly follow from the preceding ones, leading to a somewhat disjointed feel. (S20A, Task Response, General)
- The mention of Ataturk's quote is a strong point but is not sufficiently linked to the essay's main argument about the function of literature and cinema. (S3A, Coherence and Cohesion, Specific)
- The paragraphs vary in length and detail, which may affect the readability and coherence of the essay. (S19A, Coherence and Cohesion, General)
- The phrase "gain cultural accumulation" is awkward and unnatural. A more appropriate phrase could be "gain cultural insights" or "enhance cultural understanding." (S5A, Lexical Resource, Specific)
- The essay lacks appropriate collocational use and does not incorporate idiomatic or more complex vocabulary forms. (S22A, Lexical Resource, General)
- Basic grammatical structures are mostly correct, showing a solid foundational understanding. (S17A, Grammatical Range and Accuracy, General)
- "sixty percent of the people doesn't agree" should be "sixty percent of the people don't agree." (S24A, Grammatical Range and Accuracy, Specific)

The descriptive results for the relevance of ChatGPT-4 feedback, as shown in Table 4, indicate its strongest performance in the Task Response and Coherence and Cohesion dimensions. Almost all feedback generated in the Task Response dimension was relevant to the target L2 writing criterion assessed (98.8%). Similarly, most of the feedback in the Coherence and Cohesion dimension was related to the target criterion within this dimension (94.2%). The relevance rates for feedback in the Grammatical Range and Accuracy were slightly lower, with 85.8% being relevant to the target criterion assessed. The weakest performance was observed in the Lexical Resource dimension, with 73.5% of the feedback being relevant to the target criterion within that dimension.

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

Table 4. Descriptive results for relevance across L2 writing dimensions

	Relevant	Irrelevant	Total
	N (%)	*N* (%)	*N* (%)
Task Response	342 (98.8)	4 (1.2)	346 (100)
Coherence and Cohesion	440 (94.2)	27 (5.8)	467 (100)
Lexical Resource	344 (73.5)	124 (26.5)	468 (100)
Grammatical Range and Accuracy	441 (85.8)	73 (14.2)	514 (100)
Total	1567 (87.3)	228 (12.7)	1795 (100)

Below are some representative examples of relevant and irrelevant feedback by ChatGPT-4 for each L2 writing dimension:

- The writer does present a position, which is that literature and cinema indeed have a real function in society. However, this position is not clearly or systematically developed. (S16A, Task Response, Relevant)
- There is some repetition in stating how literature and cinema affect society, which could be streamlined for clarity and conciseness. (S1A, Task Response, Irrelevant)
- There is also a lack of variety in transitions, with over-reliance on simple connectors like "first of all" and "secondly." (S23A, Coherence and Cohesion, Relevant)
- Phrases like "authors was always" and "Authors always have been used their pencils" show issues with subject–verb agreement and proper use of pronouns. (S13A, Coherence and Cohesion, Irrelevant)
- The essay attempts to use a variety of vocabulary to convey its arguments, but it falls short in accuracy and appropriateness. (S20A, Lexical Resource, Relevant)
- "says" should be "say" (S10A, Lexical Resource, Irrelevant)
- There are issues with the usage of articles (the, a, an). (S19A, Grammatical Range and Accuracy, Relevant)
- "charactes" should be "characters" (S15A, Grammatical Range and Accuracy, Irrelevant)

The results of the Chi-square test (see Table 5) revealed significant associations between the type of dimension and assessment aspects (see Table 5): χ^2 (3, $N = 1795$) = 15.133, $p = .002$ for accuracy; χ^2 (3, $N = 1795$) = 267.666, $p < .001$ for specificity; and χ^2 (3, $N = 1795$) = 143.114, $p < .001$ for relevance. The association between dimension type and accuracy and relevance had small effect sizes (Cramer's $V = .092$ for accuracy and Cramer's $V = .282$ for relevance), while the association between dimension type specificity had a moderate effect size: Cramer's $V = .386$.

Table 6 shows the adjusted standardized residuals of the four L2 writing dimensions for accuracy, specificity, and relevance. Based on the ±3.09 critical value, no dimensions except Grammatical Range and Accuracy contributed to the association between dimension type and accuracy. Grammatical Range and Accuracy also contributed to the association between dimension type and specificity, but not relevance.

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

Table 5. Chi-square test results for the association between dimension type and feedback features

Association	$\chi 2$ ($df = 3$)	Sig. (2-sided)	Cramer's *V*
Dimension * Accuracy	15.133	.002	.092
Dimension * Specificity	267.666	<.001	.386
Dimension * Relevance	143.114	<.001	.282

Note: Bonferroni adjusted *p*-value ≤ .002.

Table 6. Adjusted standardized residuals for the four dimensions across feedback features

	Accuracy		Specificity		Relevance	
Dimension	Accurate	Inaccurate	Specific	General	Relevant	Irrelevant
Task Response	.55	−.55	−10.95	10.95	7.18	−7.18
Coherence and Cohesion	2.57	−2.57	−8.45	8.45	5.22	−5.22
Lexical Resource	.71	−.71	9.02	−9.02	−10.42	10.42
Grammatical Range and Accuracy	−3.67	3.67	9.01	−9.01	−1.21	1.21

The dimensions of Task Response, Coherence and Cohesion, and Lexical Resource all made large contributions to the associations between dimension type and both specificity and relevance.

Discussion

In this study, we explored to what extent ChatGPT-4's assessment of L2 writing was accurate, specific, and relevant across four dimensions of L2 writing: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. We also examined if the accuracy, specificity, and relevance of ChatGPT-4's assessment were associated with specific dimensions. According to our results, ChatGPT-4 performed best in the accuracy aspect. Accuracy rates were exceptionally high (above 95%) across all four dimensions. This result corroborates earlier results by Pfau et al. (2023), who also reported that 99% of the errors that ChatGPT identified were correctly identified as errors. This suggests that ChatGPT-4 accurately identifies the strengths and weaknesses in student writing across all dimensions of L2 writing. However, it is important to reemphasize that our results indicate precision accuracy, pointing to the accuracy in ChatGPT-4-identified strengths and weaknesses across four dimensions of L2 writing. We did not examine ChatGPT's recall accuracy, the strengths and weaknesses ChatGPT-4 failed to identify in student writing. It is possible that ChatGPT might have lower recall accuracy rates as the literature on the evaluation of automated systems, including ChatGPT, generally reports higher precision than recall (e.g., Pfau et al., 2023). Therefore, our findings should be interpreted with caution. Future research should also investigate recall accuracy to provide a more comprehensive evaluation of ChatGPT-4's capabilities as an L2 writing assessment tool.

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

Regarding specificity of ChatGPT-4's assessment across different dimensions of L2 writing, we found that ChatGPT-4's performance in this aspect was weaker, with notable variability in specificity across dimensions. While specificity in the Grammatical Range and Accuracy and Lexical Resource dimensions were relatively high, we observed moderate specificity in the Coherence and Cohesion dimension and low specificity in the Task Response dimension. Low specificity on Task Response is not surprising as generating content-related feedback remains challenging for automated systems (J. Li et al., 2015; Zhang, 2020). Generating specific feedback on task response goes beyond the structural features or linguistic patterns, as in grammar or vocabulary, and it requires an understanding of the specific topic, context, meaning, logic, depth of ideas, audience, and purpose, which are all hard to operationalize in automated systems (Fu et al., 2022; J. Li et al., 2015; Shi & Aryadoust, 2024). Regarding specificity, ChatGPT-4's weaker performance in specificity in Coherence and Cohesion is also not surprising. While cohesion is comparatively easier to assess automatically, as it depends on explicit linguistic devices (e.g., Crossley et al., 2013); coherence poses a challenge. Assessing coherence requires comprehension beyond surface-level text features and involves evaluating the logical and meaningful connection of ideas, which remains a complex task for automated systems (Yoon et al., 2023).

Regarding to what extent ChatGPT-4's assessment was relevant to specific criteria across different dimensions of L2 writing, we found that ChatGPT-4's performance in this aspect was strongest in the dimensions of Task Response and Coherence and Cohesion. This means that ChatGPT-4's assessment was closely aligned with the criteria in these dimensions. However, we observed more irrelevance in the dimensions of Grammatical Accuracy and Range and Lexical Resource, with the highest irrelevance rates in the Lexical Resource dimension. During our coding process, we observed that ChatGPT-4 sometimes failed to distinguish between grammatical and lexical errors. For example, it often identified subject-verb agreement errors (commonly categorized as a grammatical error type in the L2 writing pedagogy and literature) within the Lexical Resource dimension, under the third criterion "Spelling and word formation are used accurately and appropriately." For example, ChatGPT-4 identified the error "who does great jobs" in "We have great directors who does great jobs for the world" as a lexical error and corrected it grammatically as "who do great jobs." It seems that distinguishing between lexical errors and grammatical errors might be a challenge for LLMs like ChatGPT. This might be due to the lexico-grammatical phenomenon that some lexical and grammatical features overlap. This might also be due to the limitations in the training data since ChatGPT is trained on an enormous amount of text, which is not explicitly and linguistically fine-tuned for L2 writing assessment, thus might not align with established linguistic perspectives or rules.

Our results indicated a significant association between ChatGPT's assessment aspects of accuracy, specificity, and relevance and four dimensions of L2 writing. Compared to the small effect sizes for accuracy and relevance, the moderate effect size for specificity suggests that the level of specificity in ChatGPT's assessment is more dependent on the type of dimension than does the level of its accuracy and relevance. This finding supports additional support that certain dimensions of L2 writing might be more challenging for automated systems including LLMs to assess.

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

Regarding the pedagogical implications of our results, the low level of ChatGPT-4's specificity in assessing Task Response and Coherence and Cohesion might negatively impact students' learning processes when it is used as a formative assessment tool. General feedback is often found undesirable by L2 learners since it cannot help them identify and address specific weaknesses in their writing. Lack of specificity in feedback leaves students uncertain about how to respond to feedback or improve their writing and reduces their engagement with automated systems (Z. Li et al., 2014; Liu et al., 2016). Content-specific feedback with clear and precise language and revision suggestions has been found to be more effective in enabling L2 writers, particularly those with low writing self-efficacy, to achieve comparable improvements with fewer rounds of revisions (Wilson & Czik, 2016; Zhu et al., 2020). This highlights the importance of specificity in automated systems including ChatGPT-4 for fostering meaningful engagement with L2 learners. On the other hand, specific feedback is not preferred by L2 writing teachers, as it requires less cognitive effort, thus leading to less cognitive engagement of the learner (Zhang & Hyland, 2023). From that perspective, ChatGPT-4's low specificity in the dimensions of Task Response and Coherence and Cohesion may be cognitively more beneficial for L2 writers, as they will spend efforts on trying to figure out what specific issues the general feedback points to. In light of these considerations, it appears that ChatGPT-4's feedback specificity is double-edged. Future studies should investigate how different levels of specificity in feedback from ChatGPT impact L2 writers' learning outcomes.

Conclusions

This study contributes to the literature on automated writing assessment in general and the use of ChatGPT for L2 writing assessment in particular by extending the focus of investigations from accuracy to specificity and relevance as well as from grammatical error detection to all dimensions of L2 writing. The high precision accuracy of ChatGPT-4's assessment across all L2 writing dimensions suggests that ChatGPT-4 can accurately identify linguistic and organizational weaknesses and strengths in L2 writing. However, the observed variability in specificity, particularly in dimensions that require higher-order thinking skills (Task Response and Coherence and Cohesion), underscores the importance of enhancing ChatGPT-4's ability to conduct specific assessment.

In this study, we only focused on ChatGPT-identified strengths and weaknesses of L2 writing but did not look into ChatGPT-missed strengths and weaknesses. As we manually and qualitatively examined a high number of feedback cases, the coding process was already demanding and time-consuming; thus, we could not go beyond the precision analysis. Unfortunately, we do not know whether or to what extent there were more issues in the student texts. Without a recall analysis, our understanding of ChatGPT-4 as a valid assessment tool is incomplete. Future studies should evaluate both precision and recall for a more holistic view of ChatGPT-4's assessment potential. Another limitation of our study may pertain to our use of the IELTS Writing Task 2 rubric. Although widely recognized as a valid and reliable tool for large-scale writing assessments, the IELTS Writing Task 2 rubric may not be suitable or applicable for use within specific instructional contexts where students perform various types of writing

https://doi.org/10.1017/S0267190525100160 Published online by Cambridge University Press

tasks. Future researchers may consider using their in-house rubrics that are tailored to their specific learning objectives, instructional contexts, and learner populations. A closer alignment between the assessment criteria and educational contexts might allow researchers to evaluate ChatGPT's capacity for L2 writing assessment in a more contextually relevant way.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0267190525100160

References

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL® essays. *ETS Research Report Series*. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment*, *6*(1), 1–46. https://ejournals.bc.edu/index.php/jtla/article/view/1631

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Crossley, A. S., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education. AIED 2013. Lecture Notes in Computer Science* (Vol. 7926, pp. 134–143). Springer. https://doi.org/10.1007/978-3-642-39112-5_28

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Sage Publishing.

Fu, Q.-K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, *37*(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

Gans, M. (2023). What is ChatGPT? Featuring ChatGPT. Medium. https://medium.com/byte-sized-insights/what-is-chatgpt-featuring-chatgpt-68f52718ffdc

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, *43*(1), 277–303. https://doi.org/10.3102/0091732X18821125

Kim, M., Qiu, X., & Wang, Y. (2024). Interrater agreement in genre analysis: A methodological review and a comparison of three measures. *Research Methods in Applied Linguistics*, *3*(1), Article 100097. https://doi.org/10.1016/j.rmal.2024.100097

Lan, G., Lucas, K., & Sun, Y. (2019). Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. *System*, *85*, Article 102116. https://doi.org/10.1016/j.system.2019.102116

Lan, G., Zhang, Q., Lucas, K., Sun, Y., & Gao, Y. (2022). A corpus-based investigation on noun phrase complexity in L1 and L2 English writing. *English for Specific Purposes*, *67*, 4–17. https://doi.org/10.1016/j.esp.2022.02.002

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, Article 100210. https://doi.org/10.1016/j.caeai.2024.100210

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

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *44*, 66–78. https://doi.org/10.1016/j.system.2014.02.007

Liu, M., Li, Y., Xu, W., & Liu, L. (2016). Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, *10*(4), 502–513. https://doi.org/10.1109/TLT.2016.2612659

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

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

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

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Rau, G., & Shih, Y. S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes*, *53*, Article 101026. https://doi.org/10.1016/j.jeap.2021.101026

Seo, N., & Oh, S.-Y. (2024). Development of clausal and phrasal complexity in L2 writing: A case of argumentative essays of Korean college students. *English for Specific Purposes*, *73*, 46–60. https://doi.org/10.1016/j.esp.2023.09.003

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross disciplinary perspective* (1st ed ed.). Routledge. https://doi.org/10.4324/9781410606860

Shermis, M. D., & Wilson, J. (Eds.). (2024). *The Routledge international handbook of automated essay evaluation* (1st ed.). Routledge. https://doi.org/10.4324/9781003397618

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies*, *29*, 24735–24757. https://doi.org/10.1007/s10639-024-12817-6

Taş, A. T. (2024). *The impact of computer-assisted language learning (CALL) applications on foreign language learners' vocabulary use and their learner autonomy levels and attitudes* (Publication No. 862869) [Master's thesis, TED University]. Council of Higher Education Thesis Center.

Thompson, A. (2024). *ChatGPT for conversational AI and chatbots: Learn how to automate conversations with the latest large language model technologies*. Packt Publishing.

Vo, Y., Rickels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for English learners in a statewide evidence-based writing assessment. *Assessing Writing*, *56*(1), Article 100719. https://doi.org/10.1016/j.asw.2023.100719

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

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

Xu, Y., Polio, C., & Pfau, A. (2024). Optimizing AI for assessing L2 writing accuracy: An exploration of temperatures and prompts. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 151–174). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.10

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics*, *3*(3), Article 100133. https://doi.org/10.1016/j.rmal.2024.100133

Yoon, S.-Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. arXiv. https://doi.org/10.48550/arXiv.2310.06505

Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, Article 100439. https://doi.org/10.1016/j.asw.2019.100439

Zhang, Z., & Hyland, K. (2023). Student engagement with peer feedback in L2 writing: Insights from reflective journaling and revising practices. *Assessing Writing*, *58*, Article 100784. https://doi.org/10.1016/j.asw.2023.100784

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers and Education*, *143*, Article 103668. https://doi.org/10.1016/j.compedu.2019.103668

Cite this article: Saricaoglu, A., & Bilki, Z. (2025). The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. *Annual Review of Applied Linguistics*, 1–21. https://doi.org/10.1017/S0267190525100160