FUTURITY
Education

DOI: https://doi.org/10.57125/FED.2026.06.12

How to cite: Wang, Y.- Mei, Harmer, P. A., & Xue, C. (2026). From human raters to AI: A critical review of research on ChatGPT in writing assessment. *Futurity Education, 6*(2), 193–212. https://doi.org/10.57125/FED.2026.06.12

From Human Raters to AI: A Critical Review of Research on ChatGPT in Writing Assessment

Yu-mei Wang

PhD, Department of Curriculum & Instruction, School of Education and Human Sciences, University of Alabama at Birmingham, United States

Peter A. Harmer

PhD, Oregon Research Institute, Eugene, OR, United States

Changsong Xue

PhD, School of Medicine, Jinlin, Tonghua Normal University, China

***Corresponding author:** yuwang@uab.edu.*

Received: March 2, 2026 | **Accepted:** May 26, 2026 | **Available online:** June 18, 2026

Abstract: Writing assessment is a cornerstone of student learning, yet it remains one of the most demanding and resource-intensive tasks in education, owing to the complex, subjective nature of writing and the considerable faculty time it requires. Since the public release of ChatGPT in late 2022, a rapidly growing body of research has explored its potential as an automated writing-assessment tool, but the findings have been mixed and at times conflicting, leaving educators without clear guidance. The purpose of this paper is to systematically synthesize empirical studies that used ChatGPT to assess student writing, guided by three research questions: (1) What is the internal consistency of ChatGPT in writing assessment? (2) How does ChatGPT's assessment performance compare with that of human raters? and (3) What factors influence ChatGPT's performance in assessing student writing? Following a Systematic Literature Review

methodology, four academic databases (AACE Learning and Technology Library, ERIC, JSTOR, and Scopus) were searched, yielding 374 publications, which were screened against defined inclusion and exclusion criteria to retain empirical studies reporting statistical comparisons between ChatGPT and human raters. The review found that ChatGPT's internal scoring consistency was highly variable, with approximately half of the studies reporting unacceptable reliability that tended to decline as essay complexity increased. Inter-rater correlations with human raters were similarly mixed: of the studies examined, ten reported strong, two moderate, and eleven poor agreement. ChatGPT performed more reliably when assessing lower-order elements such as grammar and mechanics, but struggled with higher-order elements such as argumentation, logical development, and contextual accuracy, and its performance varied with the writing elements assessed, the quality of prompt design, the model version, and study methodology. Notably, about half of the studies failed to test ChatGPT's internal reliability before comparing it with human raters, a critical methodological gap. These findings offer educators and institutions evidence-based guidance on the current capabilities and limitations of ChatGPT as a cost-effective writing-assessment tool, and underscore the need for standardised prompt design and research protocols.

Keywords: ChatGPT, assessment, writing, internal consistency, inter-rater correlation.

Introduction

Writing is conventionally how students demonstrate their mastery of curriculum content. However, its role should go far beyond that of presentation and communication " . . . to expand one's own knowledge through reflection rather than simply to communicate information" (Weigle, 2002, p.5). As a cornerstone of student learning (Arnold et al., 2017), it is an "essential tool" for learning subject matter in academic settings (Weiggle, 2002). Furthermore, "of all the skills that support student learning, writing is 'of particular interest', and it is 'significant in all modern occupations ' " (Scardamalia & Bereiter, 1987). Therefore, Graham (2019) concluded, "if students are to be successful in school, at work, and in their personal lives, they must learn to write. This requires that they receive adequate practice and instruction in writing, as this complex skill does not develop naturally" (p.277).

Writing assessment is crucial in developing student writing skills but is widely acknowledged to be daunting (Deane, 2011; Dockrell & Connelly, 2021; Judy, 1973; Haines, 2022; Hessler et al., 2009; Horvath, 1984; Neff-Lippman, 2011; McNamara et al., 2015)), ". . . even for professional, trained raters" (Deane, 2011, p.2) because of: a) the complex nature of writing itself, b) subjectivity in assessing its higher cognitive elements, and c) logistical barriers, primarily faculty time.

Research Problem

Since the release of ChatGPT, research has explored its potential as an automated writing assessment tool. Numerous studies have compared scoring performance between ChatGPT and human raters, but have yielded mixed findings. These inconsistent findings require a systematic review of existing studies to understand better the contexts in which ChatGPT can effectively score student writing.

Research Focus

The purpose of the current paper is to synthesise studies that used ChatGPT to assess student writing. The findings will yield important insights into the accuracy, consistency, and efficacy of ChatGPT as a pedagogically viable and cost-effective method for assessing student writing.

Research Questions

- What is the internal consistency of ChatGPT in writing assessment?
- How does ChatGPT assessment performance compare with that of human raters?
- What factors influence ChatGPT's performance in assessing student writing?

Literature Review

Writing Assessment

The initial obstacle to effective writing assessment is the fact that "writing is not a binary skill in which the answer is right or wrong. An array of combinations of words and sentences may produce effective written communication, making objective evaluation of writing a challenging task" (Hessler et al., 2009, p.71). Unlike grading mathematics for which there is a pre-defined correct answer and optimal process for deriving it, assessing writing is complicated because it can be "creative", "subjective" or "personal" (Kwedor, 2021).

Additionally, writing well requires a high level of intellectual engagement (Palermo & Wilson, 2020; Graham, 2019; Hayes, 2012; Hayes & Flower, 1983; Hermansson & Lindgren, 2019; Valenti et al., 2003; Weigle, 2002), so quality assessment goes beyond grammar and other mechanics to include cognitive elements such as analysis, synthesis, evaluation, and creativity. However, assessment of cognitive elements is essentially subjective and "one of the difficulties of grading essays is the subjectivity, or at least the perceived subjectivity, of the grading process. Many researchers claim that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which students perceive as a great source of unfairness" (Valenti et. al., 2003, p.320).

Finally, to be effective, an assessment must provide rich and descriptive feedback to guide students to fulfil learning expectations, but this can become virtually impossible because of faculty time constraints exacerbated by: a) the need for informative feedback, b) the importance of feedback across the writing process, and c) class size.

For a writing assessment to have meaningful pedagogical value, feedback must be more than just a score or generic/superficial comments such as "Good work!" or "Excellent!", neither of which is beneficial for students striving to perform better. However, providing detailed critiques for each submission to maximise individual improvement is a time-intensive activity (Warschauer & Ware, 2006), subject to the Law of Diminishing Returns, given that faculty time is a finite resource.

Moreover, because writing is a developmental process with multiple phases (prewriting, drafting, revising, editing, finalizing), with each phase playing a significant role in refining/perfecting/enhancing the work, faculty have to closely monitor the process of each student and provide feedback at each stage for best outcomes so, in effect, "assessment of student writing and performance . . . occur[s] at many different stages" (Nebraska Writing Center, 2024), furthering adding to faculty time pressure (Lee, 2020; Smith, 2017; Warschauer & Ware, 2006).

Lastly, class size complicates the quality of feedback-faculty time availability conundrum in that, as class size increases, the units of time available to provide individual feedback on student writing assignments decrease.

The interaction between the need for detailed feedback across recurring revisions and class size can produce a multiplicative effect that results in physical and/or psychological fatigue for faculty, further jeopardising the quality of writing assessment necessary for students to improve their writing ability.

As a result of the time constraints and effort required, some faculty avoid writing assignments (National Commission on Writing in American Schools and Colleges, 2003) or take a one-shot approach to providing feedback, which does not allow students to revise their writing in response to faculty comments (Uymaz, 2019). Neither of these approaches are conducive to developing competent student writers.

To address the challenges of limited time and resources educators have attempted to tap technology for help in assessing student writing, driving the development and adoption of Automated Writing Evaluation (AWE) systems in educational settings (Chen & Cheng, 2008; Fleckenstein et al., 2023; Hussein et al., 2019; Ramesh & Sanampudi, 2022; Valenti et al., 2003; Wilson & Roscoe, 2019; Lee, 2020).

Automated Writing Evaluation

AWE (also known as Automated Essay Scoring (AES)) is a computer-based assessment system designed to reduce faculty workload in evaluating papers and to improve student writing by motivating students to revise their work through real-time scores and diagnostic feedback.

Despite the recent advances in interactive artificial intelligence (AI) that have brought its potential into the mainstream, the first AWE emerged more than 60 years ago. Historically, the three most prominent AWE programs are Project Essay Grading (PEG), Intelligent Essay Assessor (IEA), and E-rater.

Project Essay Grading (PEG) was developed in 1966 to emulate human raters and score essays for the College Board, which was seeking ways to reduce the burden on human raters of grading large volumes of student essays each year. Although PEG had initial success due to the high correlation between its scores and those of human raters, technology barriers faced by schools, primarily limited access to mainframe computers needed to run it (Page, 2003), basically "put it on the shelf" for 25 years.

As microcomputers became more readily available to educational institutions in the early 1990's, interest in automated writing evaluation re-emerged. Intelligent Essay Assessor–(IEA) was designed to evaluate the conceptual relevance and coherence of content, rather than surface-level features such as grammar, style, or other mechanics, and thus represented an advance over PEG. IEA proved that AWE has the capacity to evaluate the quality of writing (Kukich, 2000; Landauer et al., 2000; Landauer et al., 2003; Lim et al., 2021)

E-rater, released in 1999, combined the strengths of both PEG and IEA. Developed by the Educational Testing Service (ETS) to grade its widely used standard tests, including the Graduate Record Exam (GRE), Test of English as a Foreign Language (TOEFL), and Praxis. Currently, E-rater is the most robust AWE system available (Burstein, 2023; Hussein et al., 2019; Rudner, 2001).

Research on the effectiveness of AWE in classroom settings highlights both its benefits and drawbacks.

One of the most consistent findings of AWE integration into educational settings is that students generally appreciate the immediate feedback it provides (Ariyanto et al., 2021; Correnti et al., 2022; Chen & Cheng, 2008; Gao, 2021), unlike the traditional teacher-rated approach, where students often wait days or even weeks to receive feedback (Shermis et al., 2001). A student in Wang's (2015) study described the experience as ". . . like magic. . . . Since its response was so immediate, I knew where I had made mistakes right away, and I was more willing to revise my writing again and again. It indeed pointed out some errors which I had ignored" (p. 88). As hypothesised early in the development of AWE, immediate feedback

appears to motivate students to write and revise more (Grimes & Warschauer, 2010; Lee, 2004; Lee, 2020; Tang, 2017; Wang, 2015; Wang et al., 2013; Warschauer & Grimes, 2008). Tang (2017) found that over 60% of students reported revising their essays at least once or twice, and about 28% revised three to four times.

Revision is a key to strengthening writing ability, and Links et al. (2022) found that students who interacted with an AWE system had better long-term retention of writing improvements than those without access to AWE. Attali (2004) identified 30 error types between the first and final submissions of writing assignments and found that error rates decreased significantly in the final submission. The study also measured the development of discourse elements such as background, thesis, main points, supporting ideas, and conclusion, and found significantly increased rates across all features except the thesis statement. In a study by Correnti et al. (2022), in which students submitted argumentative papers to AWE and revised them according to its feedback, about 94% of students believed their revised copies were better than their original submissions. A paired-samples t-test comparing students' original copies and revised copies confirmed significant improvements. On average, students added more than three pieces of evidence to support their arguments in the final copies of their essays.

AWE also has broad support among educators for several reasons, particularly its potential to deliver the immediate feedback shown to be important for motivating students to improve their writing, but which is logistically simply not possible for teachers to provide (Hann et al., 2021), especially as class size increases. Additionally, with the time "reclaimed" by using AWE to evaluate student submissions, educators can focus on strengthening other aspects of writing pedagogy, including developing more effective writing instruction strategies, focusing on higher-order writing skills, and working with individual students to accommodate their diverse needs (Correnti et al; 2022; Hann, 2021; Tang & Rich, 2017; Warschauer & Grimes, 2008; Wilson et al., 2021).

Nevertheless, AWE has drawbacks, with the major concerns centred on the quality of feedback provided (Stevenson, 2016). Although AWE can help provide feedback on surface features of writing, such as lexical and grammatical errors (Warschauer & Grimes, 2008), traditional AWE often fails to handle more complicated writing features, such as coherence, organisation, logic, and content development or provide content-specific feedback because each system operates on a set of pre-programmed fixed responses. Therefore, they cannot adjust feedback based on specific contexts because they are not programmed to recognise them. In the study by Chen and Cheng (2008), students perceived AWE feedback as "vague," "abstract," "unspecific," "formulaic," and "repetitive", especially in the areas of essay coherence and content development.

Cheville (2004) argued that AWE promoted formulaic writing at the expense of logic and creativity: ". . . subordinating meaning, rendering it subservient to structure" (p.49). Thus, AWE may be unable to identify essays that are either illogical or, more problematically, creatively unconventional. For example, in the Chen and Cheng (2008) study, a student wrote a story as the introduction of her essay and received a low score because the AWE rated the introduction as illogical and incompatible with the conclusion. The student replaced the story with a conventional introduction and received a much higher score. Wang (2015) found that students quickly exploited AWE's emphasis on formulaic writing. After several submissions, they realised that the AWE being used could only detect whether an essay contained four components (introduction, thesis statement, two supporting paragraphs, conclusion) but not whether the components' content was meaningful to the topic, and that they could receive higher scores if they followed the predefined structure in their essays, even if the information they inserted was irrelevant.

Similarly, AWE appear to favour length, and certain linguistic and grammatical features (Chen & Cheng, 2008; Cheville, 2004; Shermis & Burstein, 2002), which results in students padding their work with

long sentences or additional paragraphs even if they are unrelated to the topic (Grimes & Warschauer, 2006), a move away from competent writing. When Wang (2015) asked students whether they knew how to achieve higher scores on AWE, most responded by adding more words/paragraphs, regardless of content.

These drawbacks, primarily due to inherent technological limitations, mean that AWE systems need further refinement to be truly effective for writing assessment.

ChatGPT for Writing Assessment

The release of Chat Generative Pre-Trained Transformer (ChatGPT) in 2022 signified an epoch-making breakthrough in natural language processing models by integrating multiple iterative features into a single application: a) a Chatbot that allows users to interact with it via written or verbal inquiries, b) Generative capacity to create human-like conversations, c) Pretraining by scanning the Web for years and incorporating a vast amount of information from books, articles, and websites, "essentially the entire Internet" (UBS Science, 2024), and d) employing a Transformer architecture to sort through information to generate responses. With its unique blend of attributes, ChatGPT can emulate human conversation, and users generally feel they are communicating with a human being.

Some ChatGPT characteristics are universally acknowledged: (1) it excels in understanding human languages at sophisticated levels; (2) it has remarkable ability in recognizing nuanced contexts, e.g., it can produce written output based on contexts specified and described by users; (3) it is adaptive, e.g., it could produce a poem to celebrate the birthday of a 10-year old or 50-year old, adjusting the content based on age, sentiments, or life-changing events; (4) it is interactive, producing dynamic dialogue and supporting two-way communication with users (users can follow up responses from ChatGPT, make inquiries, ask for clarifications, and refine their inputs); (5) it has a version that is powerful and fully functional that anyone anytime anywhere can access at no cost.

Conventional AWE systems lack these features and primarily focus on superficial text features, rely on predefined rules to generate rigid feedback, lack understanding of the given context, and allow users only one-way interaction. Therefore, ChatGPT holds great promise for assessing student writing.

Research Methodology

Following the Systematic Literature Review (SLR) methodology (Pati & Lorusso, 2018), four academic databases (AACE Learning and Technology Library, ERIC, JSTOR, Scopus) were searched using three targeted keywords: *ChatGPT, writing and assessment*.

A total of 374 publications were retrieved, and inclusion and exclusion criteria were applied (Table 1). Excluded publications included conceptual/theoretical papers, literature review articles, practitioner- or opinion-based papers, and empirical studies that did not include inter-rater statistical analyses comparing ChatGPT and human raters in scoring student writing.

Table 1

Inclusion and Exclusion Criteria

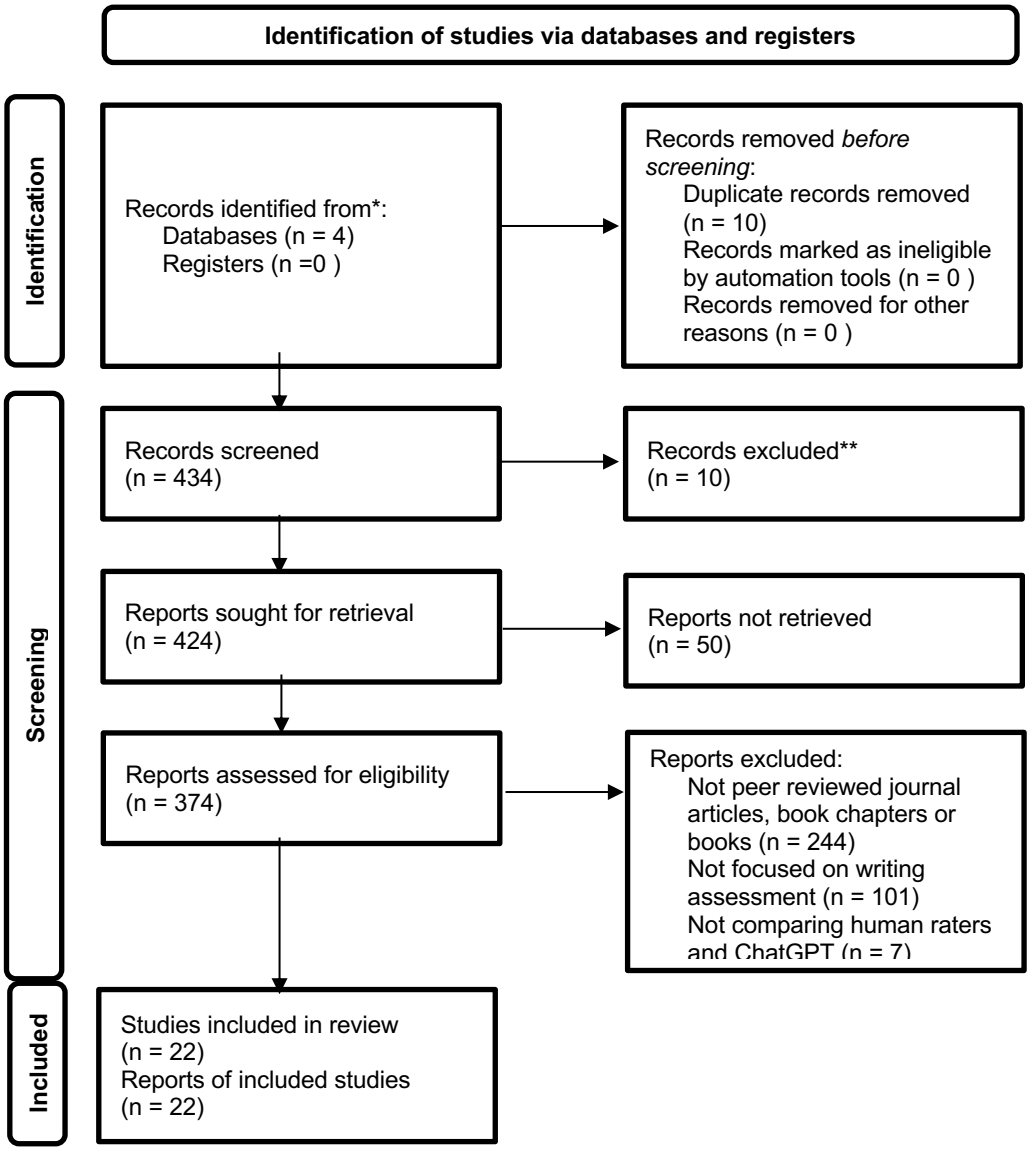
Inclusion Criteria	Exclusion Criteria
• Published during 2023-2025 • Written in English • Peer reviewed (e.g., journal articles, book chapters, or books)	• Publications that • were not peer-reviewed (e.g., SSRN) • did not focus on applying ChatGPT in writing assessment (e.g., ChatGPT in ESL learning)

• Investigated ChatGPT as a writing assessment tool • Studies included statistical analysis to compare ChatGPT's scoring performance with that of human raters in assessing writing	• did not include the comparison between the scoring performance of ChatGPT and human raters in writing assessment

The flowchart (Figure 1) below illustrates the screening process used to select publications for this review.

Figure 1

Screening Process for Selecting Publications for this Review



After systematic screening and review, 22 publications were included in this study (Table 2).

Table 2

Studies on Applying ChatGPT as a Writing Assessment Tool (2023-2025)

Study	Field & Location	Purpose	Areas of Assessment	ChatGPT Version	Writing Samples	Writing Source	Human Raters	Conclusions
Almegren et al. (2024)	English Language Education Saudi Arabia	Compare ChatGPT 3, Bing, Gemini with each other and with 3 human raters	Vocabulary Ideas & Content Organization & Coherence Mechanics (grammar, punctuation)	3	N=30	English as a Foreign Language (EFL) students	3 experienced English language professors	AI programs tested, including ChatGPT, graded lower than the human raters but provided "high quality feedback"
Awidi (2024)	Engineering Australia	Compare writing assessment of ChatGPT and human raters	Depth Analysis Logic Clarity	Not Specified	N=108	Students in 1st year Engineering courses	39 expert tutors	With appropriate prompts ChatGPT can evaluate reflective essays on par with human raters although both had problems with detailed feedback
Bouziane & Bouziane (2024)	Not specified Morocco	Assess the accuracy, efficiency, and cost-effective-ness of ChatGPT essay correction compared to human raters.	Mechanics (grammar, spelling sentence structure, punctuation) Coherence Relevance Structure and clarity	Not Specified	N=100	University students in various disciplines	English Department faculty - Number not specified	ChatGPT performs well in evaluating writing mechanics but not for "thematic consistency"
Bucol & Sangkawong (2024)	English Language Education Thailand	Compare writing scores generated by ChatGPT and human raters	Content Organization Length	Not Specified	N=10	University students in General English class	10 EFL professors	ChatGPT can provide "reliable evaluations across a wide range of users"
Bui & Barrot (2024)	English Language Education 10 Asian countries	Compare writing scores generated by ChatGPT and a human rater	Claim Development Audience Cohesion Style	3.5	N=200	College students studying English	1 experienced EFL professor	Unlike the scoring of a human rater ChatGPT intraclass correlation values indicated that ChatGPT scoring was not stable or reliable.
Fokides & Peristeraki (2024)	Elementary Education Greece	Compare ChatGPT and human raters on accuracy and feedback for elementary school student essays in Greek and English	Mechanics Feedback Expression Structure Support/Guidance	Turbo	N=40 Greek=20 English=20	Elementary school students	20 ESL teachers: 10 primary school teachers; 10 EFL instructors	ChatGPT was better in quantity and quality of assessment and feedback for essays in English but less effective for essays in Greek, even for mechanics
Geçkin et al. (2023)	English Language Education Turkey	Compare writing scores generated by ChatGPT and human raters using rubric	Stages of writing rubric: Emerging, Beginning, Developing, Expanding, Fluent, Proficient	3.5	N=43	Advanced level EFL students	5 experienced college faculty	Overall, the scores assigned by the human raters varied significantly from each other. The correlation between the average of the human raters and ChatGPT scoring was "negligible"
Jackaria et al. (2024)	Education Philippines	Compare writing scores generated by ChatGPT and human raters using rubric	Content Organization Language use Mechanics	3.5	N=20	College students - professional education Course	3 experienced college faculty	Although ChatGPT showed moderate consistency in its ratings, its scoring correlation with the human raters evaluations was poor
Kim et al. (2024)	English Language ability USA	Compare writing assessment of ChatGPT and human raters	Organization Arguments and Details Grammar Conventions	4	N=74	College students whose first language is not English	6 doctoral students – Applied Linguistics & Technology	Compared to human raters, ChatGPT reliability was moderate and its evaluations were "limited in detecting content related issues and integrating source text information"
Kinik & Çetin (2024)	English Language Education Turkey	Compare ChatGPT using rubric with a human rater	7-item rubric including: Ideas Word choice Sentence fluency Organization	3.5	N=20	English Language Teaching student teachers	1 experienced teacher educator	The association between ChatGPT and human rater scoring was inconsistent. Although ChatGPT can give rapid feedback,

								inaccuracies undermine its utility
Ju. Li et al. (2024)	English Language Education China	Compare writing scores generated by ChatGPT and human raters	Language (Grammar, sentence structure, word choice, etc) Content (Depth, clarity, persuasiveness) Organization (Cohesion & flow)	3.5 & 4	N=30	College students who were non-English majors	4 university English faculty	Both ChatGPT 3.5 and 4 provided scores consistent with human raters. However, the quality of feedback differed between the three with ChatGPT 4 quantitatively the most effective.
Jo. Li et al. (2024)	University Education Australia	Compare ChatGPT with 2 rubrics and human raters using iterative prompts	Coding rubric (e.g., Program structure, named constraints, control structures) Reflective rubric (level of self-refection, quality of writing)	4	N=100	First year Computer science Students Postgraduate Public Health students	4 discipline-specific experts	ChatGPT was able to evaluate and assess both coding and reflective assessments and to distinguish between assignments of different quality, being more reliable with higher quality essays. Overall, it was comparable to human raters across all assessments.
Lu et al. (2024)	Education China	Compare ChatGPT with human raters and assess impact of combined use on student writing in Chinese	Elements of an academic abstract writing sample (purpose, method, findings, implications, language conventions)	3.5	N=46	Under-graduate Education majors	2 experienced academics	ChatGPT provided more feedback than human raters but students used rater feedback more readily. ChatGPT assessment reliability varied with quality of its prompts and quality of the writing samples.
Mizumoto & Eguchi (2023)	English Language acquisition Internation-al: 11 different languages	Compare writing scores generated by ChatGPT and human raters for TOEFL 11	TOEFL scores	3 text-davinci-003	N=12,100	ESL students	Trained TOEFL scorers: 2-3/essay	ChatGPT most closely tracks with human rates when ChatGPT scores are used with multiple linguistic criteria (lexical, syntactic, cohesion).
Mizumoto et al. (2024)	English Language Education Internation-al: 3 Asian countries	Comparing writing assessment of ChatGPT, human raters and Grammarly	Grammatical structures	4	N=232	Asian ESL learners (80 Japanese, 86 Korean, 66 Chinese)	Cambridge English Qualification scorers – Number Not Specified	ChatGPT assessment showed a high correlation with human raters and performed better than Grammarly.
Quah et al. (2024)	Dentistry Singapore	Compare writing assessment of ChatGPT and human raters	Rubric related to appropriate clinical care	4	N=69	Final year under-graduate dental students	3 School of Dentistry faculty	ChatGPT performed similarly to human raters, although it was more likely to be more strict and less likely to identify content errors.
Shabara et al. (2024)	English Language Education Egypt	Compare writing scores generated by ChatGPT and human raters	Organization Content Relevance Language Use Communicative quality	3.5	N=100	University students: Arabic-speaking English language learners	11 experienced teachers	ChatGPT reliability was problematic and its assessment performance was not well-aligned with human raters.
Shermis (2024)	AWE research USA	ChatGPT evaluated against human raters	Comparison of 4 prediction models: linear regression, random forest, gradient boost, xgboost)	4o	Essays N=22,029; Short-form constructed responses N=25,683	Students in grades 7, 8, 10	Not Specified	ChatGPT performance was inconsistent and usually lower than human raters. However, it may have "potential for scoring efficiency" if it is properly trained.
Shin & Lee (2024)	English Language Education South Korea	Compare writing assessment of ChatGPT and human raters	Task completion Content Organization Language use	Plus	N=50	South Korean high school English language students	2 experienced in-service English teachers	ChatGPT performance closely aligned with human raters indicating potential to be useful tool for teachers of English as a second language in South Korea.

Steiss et al. (2024)	English Language Education USA	Compare ChatGPT and human raters on writing feedback	Criteria-based feedback: Directions for improvement Accuracy Essential features Supportive tone	3.5	N=200	English language learners: 3 ability levels (6-12 grade)	16 experienced educators	Human raters provided higher-quality feedback overall but ChatGPT quality was close "without requiring any training." However, Chat GPT performance varied with the quality of the writing.
Yang (2024)	English Language Education China	Evaluate ChatGPT reliability and compare with human raters	Language Content Organization	3.5	N=82	EFL students in Chinese universities	3 experienced EFL faculty	ChatGPT is not currently reliable enough to be useful in evaluating EFL compositions.
Yoon et al. (2023)	English Language Education USA	Compare writing feedback by ChatGPT and human raters	Cohesion Coherence	4	N=50	English Language Learner students: grade 12	2 experienced ELL teachers	ChatGPT does not provide effective feedback for ELL essays.

Results

Internal Consistency of ChatGPT in Scoring Essays

Internal consistency is one of the most important measures of ChatGPT's robustness as an assessment tool - a prerequisite for determining its suitability as a replacement for or adjunct to human raters of student writing. Despite this, the extant literature is very problematic, with some studies reporting good values, some reporting poor values, and many not conducting analyses of internal consistency at all.

For example, Awidi (2024) used ChatGPT to assess writing elements such as depth, analysis, logic, and clarity, as well as student feedback. Each sample was evaluated twice by ChatGPT using the same prompt, yielding Cronbach's alpha coefficients of 0.775 and 0.819, respectively, indicating good internal consistency. Bucol and Sangkawong (2024) also reported excellent consistency for ChatGPT, with a higher Cronbach's alpha than human raters (0.980 vs 0.926, respectively). They concluded: "The findings consistently demonstrate . . . its potential as a dependable automated evaluation tool for multiple teachers teaching in the same subject" (p. 8). Mizumoto et al. (2024) also found that ChatGPT showed minimal variation in evaluating linguistic accuracy in second-language writing (Cronbach's alpha = 0.96). Jo. Li et al. (2024) found ChatGPT to be highly consistent in grading student reflective essays as measured by standard error. They concluded that "ChatGPT provided a high degree of reliability in marking the reflective essay in alignment with the provided rubric . . ." (p. 63). The findings of these and other studies (Quah et al., 2024; Shin, 2024; Shin & Lee, 2024) show that ChatGPT appears to follow established criteria well and could accurately apply instructions for assessment tasks, pointing to its potential as a reliable writing assessment tool.

Approximately half of the studies in this review found that ChatGPT demonstrated unacceptable variability in its scoring consistency. Steiss et al. (2024) found that its scoring consistency decreased as essay quality increased, indicating an incompatibility with more complex writing tasks. Although Bui and Barrot (2025) investigated ChatGPT's evaluation of argumentative essays, using a 5-element rubric (claim, development, audience, cohesion, style and convention) with two rounds of scoring to strengthen its reliability, scores in both rounds showed poor reliability across all five criteria, "making it challenging to trust its results" (p. 2052). Similarly, when Shabara et al. (2024) used ChatGPT to score the same essay twice, they found statistically significant differences between the initial scores and regenerated scores, demonstrating that the internal reliability of ChatGPT was "questionable." The researchers attributed this to inherent flaws in ChatGPT's algorithm to interpret complex arguments. However, ChatGPT's reliability can also be influenced by the version. For example, ChatGPT 4 is more reliable than ChatGPT 3.5 in writing assessments (Ju. Li et al., 2024).

Correlation between ChatGPT and Human Raters in Evaluating Writing

The ultimate purpose of this review was to compare ChatGPT's scoring with human raters' scoring in evaluating student writing, given that human raters are the historical gold standard. Examination of the correlation between ChatGPT and human raters in the included studies yielded mixed results: 10 reported a strong correlation, 2 a moderate relationship and 11 a poor association (Table 3).

Table 3

Inter-rater Correlation Between ChatGPT and Human Raters

Author(s)	High	Moderate	Low
Almegren et al. (2024)	X		
Awidi (2024)		Average Measures X	Single Measures X
Bouziane & Bouziane (2024)	X		
Bucol & Sangkawong (2024)		X	
Bui and Barrot (2024)			X
Fokides & Peristeraki (2024)	X		
Geçkin et al. (2023)			X
Jackaria et al. (2024)			X
Kim et al. (2024)			X
Kınık & Çetin (2024)			X
Ju. Li et al. (2024)	X		
Jo. Li et al. (2024)	X		
Lu et al. (2024)			X
Mizumoto & Eguchi (2023)	X		
Mizumoto et al. (2024)	X		
Quah et al. (2024)	X		
Shabara et al. (2024)			X
Shermis (2024)			X
Shin & Lee (2024)	X		
Steiss et al. (2024)	X		

Studies reporting positive correlations suggest that ChatGPT is a reliable writing assessment tool. Its performance was "comparable to human marking, showing a particularly higher degree of confidence in evaluating the reflective essay…." (Jo. Li et al., 2024, p. 66). It was also found that comments provided by ChatGPT aligned well with the scores it assigned based on essay quality (high, average, poor). In evaluating ChatGPT's assessment of ESL writings, Mizumoto et al. (2024) stated their "finding lends support to the potential of using ChatGPT as an automated tool for evaluating linguistic accuracy in L2 writing, as it demonstrates that its evaluations are as reliable as those of human raters in terms of their association with overall writing quality" (p. 10). ChatGPT has also been found to detect more errors than human raters and provide more comprehensive feedback (Fokides, 2024). It correlates well with human raters when used to evaluate aspects of writing, such as content and organisation (Shin et al., 2024). Encouraged by the findings

of their study, Mizumoto & Eguchi (2023) stated: "AI language models, such as ChatGPT, can be effectively utilised as AES tools, potentially revolutionising methods of writing evaluation and feedback in both research and practice (p. 9).

Several studies in this review reported moderate correlation between ChatGPT and human raters in writing assessment (Awidi, 2024; Bucol & Sangkawong, 2024; Steiss et al., 2024). Although the correlation was moderate, these studies were positive about ChatGPT's capabilities in writing assessment. Bucol and Sangkawong (2024) found that "ChatGPT is capable of providing reliable evaluations across a wide range of users, making it a highly promising option for automated assessment in collaborative education settings." Ju. Li et al. (2024) compared ChatGPT 3.5 and ChatGPT 4.0 and found that ChatGPT 3.5 had lower reliability than human raters. However, ChatGPT 4.0 was an effective tool for assessing EFL writings "that could replace teacher raters in scoring EFL compositions in the classroom settings" (p. 5).

In contrast to the studies that expressed support for ChatGPT in writing assessment, studies that reported poor correlation between ChatGPT and human raters raised serious questions about its role in writing assessment. Hence, ChatGPT "should be used cautiously in rating students' written works" (Jackaria et al., 2024, p. 485). Compared with human raters, "ChatGPT was limited in detecting content-related issues and integrating source text information" (Kim et al., 2024). This finding was supported by Yang (2024), who conducted a study in which ChatGPT was employed to assess the coherence and cohesion of student essays. Yang's study (2024) found that ChatGPT primarily focused on surface-level cohesive features, such as transition words, but failed to distinguish inefficient and inaccurate usage from efficient and accurate usage. This poor performance was observed in several studies during comprehensive assessments of essential writing elements by ChatGPT. Bui and Barrot (2025) evaluated argumentative essays focusing on five essential writing dimensions: claim, development, audience, cohesion, and style. ChatGPT correlated poorly with a human rater in scoring performance in each area. ChatGPT prioritised error detection and penalised errors more heavily than the human rater and could not capture the nuances of the argumentative essays. The study concluded: "Overall findings indicate that this AI tool's capabilities in automated scoring are highly constrained" (p. 2054).

Discussion

Compared with conventional automated writing evaluation systems, ChatGPT represents a significant technological advance, particularly in its capacity to grade surface-level writing features such as grammar, spelling, word choice and sentence structure. However, the findings from the studies reviewed here indicate that ChatGPT does not perform at a level comparable to that of human raters in holistic writing assessment.

Notably, the findings show considerable variability in assessment outcomes. According to our analysis, this variability appears to originate from differences in three areas: (a) types of writing elements assessed, (b) prompts design, and (c) research methodologies. Each factor plays a critical role in shaping the reported findings.

Types of Writing Elements

When measured by cognitive engagement in the writing process, writing elements can be categorised as lower-order and higher-order. Low-order writing elements refer to linguistic accuracy/mechanics, such as grammar, spelling, word choice, sentence structure, and consistency in style and tone. Higher-order writing elements involve the engagement of analysis, synthesis, and evaluation. Examples include logic, a thesis statement, arguments, supporting evidence, organisation, and a summary.

Multiple studies in this review relied heavily on ChatGPT to assess lower-order writing elements (Fokides & Peristeraki, 2024; Jo. (Li et al., 2024; Mizumoto et al., 2024; Mizumoto & Eguchi, 2024; Pfau, 2023), Moreover, the results consistently demonstrated strong alignment between ChatGPT and human raters. In this regard, ChatGPT does not appear to differ from conventional AWE systems in its evaluation of writing mechanics.

In contrast, studies that investigated ChatGPT's ability to evaluate higher order writing elements (e.g., Bui & Barrot (2025) "claim", "development" "audience", "cohesion", "style"; Awidi (2024) "depth", "analysis", "logic", "clarity"; Jackaria et al. (2024) "ideas/relevance", "organization"; Yang (2024) "content") generally found poor correlation between ChatGPT and human rater. These findings show that ChatGPT is not sophisticated enough to deal with more complex writing assessment tasks. ChatGPT cannot detect contextual errors and often produces a flawed evaluation. For example, Steiss et al. (2024) required students to write on the topic: "How did the Montgomery Bus Boycott succeed?" One student confused Rosa Parks with Jo Ann Robinson and developed the essay on this mistaken assumption. In contrast to the human rater, who readily identified the error and provided appropriate feedback, ChatGPT failed to recognise the mistake. However, the correct information was included in its training/source materials. As the researchers noted: "This reminds us that ChatGPT does not actually 'understand' the text it is given but instead relies on a predictive algorithm to generate feedback" (p. 9). Kim et al. (2024) reported a similar case in which ESL students were asked to write a 300-350-word essay on "robots" in a 30-minute writing test. ChatGPT was found to grade essays based on text length. It penalised students who wrote one-sentence conclusions and flagged them as "cut off," even when the sentence effectively summarised the essay's main points. On the contrary, human raters considered these conclusions satisfactory because they "prioritise pragmatism and content relevance in a time-constrained testing environment" (p. 82).

Prompt Design

Crafting a prompt is both an art and a science. Designing prompts for writing assessment is not merely a technical task but an iterative process that requires testing and refinement (Awidi, 2024; De Winter et al., 2023; Jo Li et al., 2024; Steiss et al., 2024). Prompts are the "instructions" that guide ChatGPT on what to assess and how to assess a writing task. ChatGPT is very sensitive to prompt instructions, so even the slightest changes in wording can alter the assessment output. For example, human raters interpret "assess the organisation of the essay" and "evaluate the coherence of the writing" as equivalent instructions. However, ChatGPT may treat the two as distinct instructions, resulting in inconsistent assessment scores. Despite its importance, approximately 50% of studies in this review did not include a prompt protocol, i.e., details of the process outlining how prompts were designed, tested, and refined.

It is not sufficient for a study to present the prompt(s) applied in the writing assessment. A well-documented prompt protocol should provide answers to critical questions such as: (a) How was the prompt aligned with the established rubrics? (b) How many writing samples were used to test the prompt? (c) How many runs were the prompt tested? (d) How did the change of particular words and sentence structures in the prompt impact the assessment scores? (e) What were the prompt design revised and refined?

High-quality prompts can only be achieved through iterative cycles of design, testing, evaluation, and refinement. Research without such documentation limits its replicability and slows progress in the field of prompt engineering.

Research Methodology Issues

Research on the use of ChatGPT for writing assessment is a recent phenomenon, as ChatGPT was only released in late 2022. The studies in this review provided important information and insights into the use

of AI in educational settings as an automated writing assessment tool. Nevertheless, given the emerging nature of this new area of inquiry, there is significant room for improvement, particularly in methodological design to strengthen the validity and reliability of research findings. This study has identified several key methodological gaps in studies investigating ChatGPT as an automated writing assessment tool.

This review has found that approximately one-half of the studies failed to test the internal reliability of ChatGPT before comparing its scoring performance with that of human raters. This omission is problematic because ChatGPT is not inherently consistent in applying scoring criteria across all student writings. Without first testing the internal reliability of the assessment tool, assessment results may be biased. Otherwise, even when inter-rater reliability between ChatGPT and human raters is high, this alignment may be due to chance.

These previously mentioned studies may assume that inter-rater consistency between ChatGPT and human raters can substantiate the internal consistency of ChatGPT and the reliability of human raters. However, intra-class reliability and inter-rater reliability serve different purposes. Intra-class reliability assesses whether a rater -ChatGPT or a human- consistently applies the same standards in evaluating student writings across various criteria (e.g. word choice, style, and content). In contrast, inter-rater reliability measures the agreement between ChatGPT and human raters in scoring student writings. This assumption overlooks the fact that each rating agent (ChatGPT or human) must first demonstrate strong internal reliability before their alignment with another rater can be considered meaningful.

Methodologically sound research should include three indispensable statistical analyses: (a) scoring consistency of ChatGPT; (b) scoring consistency of human raters; and (c) inter-rater consistency. Each one of them is indispensable and serves a different purpose. Scoring consistency of ChatGPT measures whether ChatGPT can apply scoring criteria reliably across all student writing. This verification is fundamental to establishing ChatGPT's capacity as a reliable assessment instrument. The analysis of internal consistency also applies to human raters. Human raters can vary in their evaluation of student writing for many reasons (e.g., misunderstanding of evaluation standards, individual preference for certain writing styles, and fatigue). It is equally important for human raters to demonstrate that they understand and can systematically follow agreed-upon assessment criteria when grading student writing. The inter-rater analysis compares the scoring performance of ChatGPT and human raters. This comparison verifies the degree of alignment in scoring performance between the two rating systems -how closely their gradings match.

To determine whether ChatGPT is ready to be adopted as a writing assessment tool, we must rigorously evaluate its scoring consistency through multiple trials and analysis. Without this basic level of internal consistency, ChatGPT cannot be considered a dependable tool for writing assessment. We must first establish empirically that ChatGPT possesses satisfactory reliability to serve as a writing assessment instrument. The process of testing should be transparent and well-documented. It is a methodological flaw to skip testing ChatGPT's internal reliability. If we never check whether ChatGPT is internally reliable, it becomes impossible to interpret and verify any agreement or disagreement with human raters.

It is worth noting that ChatGPT's scoring consistency is not fixed. Its consistency depends heavily on the quality of the prompts and instructions it receives. ChatGPT does not have intuitive judgment. Its performance is controlled by the parameters and guidance embedded in prompts. Human users: design, test, revise, and refine these prompts and instructions through trial and feedback. This human-guided characteristic means that ChatGPT's scoring consistency can be systematically adjusted and improved through multiple rounds of strategic human intervention.

Conclusion

The findings of studies to date indicate that ChatGPT is not yet consistently on par with human raters in writing assessment. This conclusion is problematic, as it raises questions about the ultimate value of applying ChatGPT as a writing assessment tool in educational settings in terms of cost-effectiveness, given the time and effort needed to develop appropriate prompts for the program for each assessment assignment and the fact that faculty/teachers may still need to review the work for final grades to ensure accurate and fair evaluation of student work.

Based on this review, the following concerns are raised:

a) The heterogeneity of research protocols in the literature examining ChatGPT for writing assessment makes it extremely difficult to provide a consensus on whether it is actually useful.

b) From the disparate results in the research, it appears that ChatGPT may help assess low-order writing elements such as mechanics (spelling, grammar). However, it is not sufficiently reliable for assessing higher-order writing elements (context, logical development of arguments/theses).

c) Even if Chat GPT may be useful for assessing lower-order writing elements, the fact that faculty have to develop the appropriate prompts (usually an iterative process) for it to operate and the fact that faculty still have to read writing assignments to provide the best assessment of higher-order elements indicate that use of ChatGPT may not be particularly cost-effective in terms of faculty time saved.

d) for students, it potentially has value, principally in the speed of feedback (i.e., virtually immediate). However, this advantage is compromised by the uncertainty of the accuracy or reliability of the feedback, especially for higher-order writing elements.

e) Researchers need to develop a standard core protocol for investigating the utility of ChatGPT in writing assessment. While adhering to a core protocol would not preclude incorporating additional unique elements for specific projects, it would help establish a comprehensible internal validity from which to make more rational decisions about the utility of ChatGPT in writing assessment. As this review shows, despite the number of studies published, the current idiosyncratic methodological approaches (e.g., source material, evaluation criteria, outcome variables, statistical analyses, ChatGPT version, approaches to training ChatGPT, etc.) offer little in the way of a coherent answer to the question.

Suggestions for Future Research

1. Future research should use meta-analysis to synthesise findings across studies on ChatGPT in writing assessment. Meta-analysis is a research methodology for identifying overall patterns by statistically combining and analysing quantitative results from multiple studies. Since the release of ChatGPT, the number of studies on its potential as a writing assessment tool has been growing. However, the findings from these studies are inconsistent and sometimes even conflicting due to variations in contexts, prompt designs, and methodological approaches. For example, this review has identified studies reporting positive, negative, and moderate results. By pooling data from multiple studies and conducting a single large-scale analysis, meta-analysis can provide a more comprehensive understanding of ChatGPT's performance in writing assessments and help explain inconsistencies across studies.
2. Future studies should develop and adhere to prompt design protocols when using ChatGPT as a writing assessment tool. As this review shows, ChatGPT's performance is highly sensitive to prompt instructions; even small variations in wording or phrasing may alter its output and produce undesired responses. However, this review shows that many studies did not report on a prompt design process. As a result, information remains unknown on how prompts were designed, tested, and refined. Prompt

design is inherently an iterative and experimental process and should be documented in detail. Standardised prompt protocols are essential in enhancing the reliability and replicability of ChatGPT as a writing assessment tool. A robust prompt design process should include systematically conducting multiple rounds of testing, controlling prompt variables, and documenting changes in assessment outcomes.

3. Future studies should adopt rigorous and standardised research methodologies to strengthen the validity and reliability of findings. This review found that many studies did not assess the reliability of ChatGPT's internal scoring before comparing it with that of human raters. An inter-rater test cannot replace intra-rater assessment of each tool involved, whether it is ChatGPT or a group of human raters. Even when an inter-rater test yields highly consistent results, ChatGPT's reliability may still be low.

It is equally important to assess the scoring consistency of human raters. Many factors can affect human raters' scores, such as individual preferences, fatigue, and human error. Therefore, research on using ChatGPT for writing assessment must incorporate three essential analyses: (a) internal reliability of ChatGPT, (b) internal reliability of human raters, and (c) inter-rater consistency between ChatGPT and human evaluators.

Acknowledgements

None.

Conflict of Interest

None.

Funding

This research received no funding.

References

Ariyanto, M. S. A., Mukminatien, N., & Tresnadewi, S. (2021). College students' perceptions of an automated writing evaluation as a supplementary feedback tool in a writing class. *Jurnal Ilmu Pendidikan*, *27*(1), 41.

Arnold, K. M., Umanath, S., Thio, K., Reilly, W. B., McDaniel, M. A., & Marsh, E. J. (2017). Understanding the cognitive processes involved in writing to learn. *Journal of Experimental Psychology: Applied*, *23*(2), 115.

Attali, Y. (2004). Exploring the feedback and revision features of Criterion. *Journal of Second Language Writing*, *14*(3), 1-20.

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence*, *6*, 100226.

Bouziane, K., & Bouziane, A. (2024). Exploring the role of AI in essay evaluation: A comparative analysis of ChatGPT and human corrections. Research Square. https://doi.org/10.21203/rs.3.rs-4139088/v1

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, 1-16.

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*(2),2041-2058.

Burstein, J. (2003). The E-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 113– 121). Routledge. https://doi.org/10.4324/9781410606860

Chen, Chi-Fen, & Cheng, Wei-Yuan. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning and Technology, 12*(2), 94–112.

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal*, *93*(4), 47-52.

Correnti, R., Matsumura, L. C., Wang, E. L., Litman, D., & Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Computers and Education Open*, *3*, 100084.

Cotos, E. (2011). Potential of automated writing evaluation feedback. *Calico Journal*, *28*(2), 420-459.

Deane, P. (2022). The importance of assessing student writing and improving writing instruction. Research notes. *Educational Testing Service*.

Dockrell, J. E., & Connelly, V. (2021). Capturing the challenges in assessing writing. *Executive Functions and Writing*, 103.

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, *6*, 1162454.

Fokides, E., & Peristeraki, E. (2025). Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek. *Education and Information Technologies*, *30*(2), 2577-2621.

Gao, J. (2021). Exploring the feedback quality of an automated writing evaluation system pigai. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(11), 322-330.

Geçkin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. Human raters. *Journal of Educational Technology and Online Learning*, *6*(4), 1096-1108.

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, *43*(1), 277-303.

Grimes, D., &Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 4–44

Hahn, M. G., Navarro, S. M. B., Valentín, L. D. L. F., & Burgos, D. (2021). A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access*, *9*, 108190-108198.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369-388.

Hayes, J. R., & Flower, L. S. (1983). Uncovering cognitive processes in writing: An introduction to protocol analysis. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), Research on writing: Principles and methods (pp. 207-220). New York: Longman.

Hessler, T., Konrad, M., & Alber-Morgan, S. (2009). Assess student writing. *Intervention in School and Clinic*, *45*(1), 68-71.

Horvath, B. K. (1984). The components of written response: A practical synthesis of current views. *Rhetoric Review*, *2*(2), 136-156.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208.

Jackaria, P. M., Hajan, B. H., & Mastul, A. R. H. (2024). A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *International Journal of Learning, Teaching and Educational Research*, *23*(2), 478-492.

Judy, S. N. (1973). Writing for the here and now: An approach to assessing student writing. English Journal, *62*(1), 69-79.

Kim, H., Baghestani, Sh., Yin, Sh., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), Exploring artificial intelligence in applied linguistics (pp. 73–95). Iowa State University Digital Press. https://doi. org/10.31274/isudp.2024.154.06

Kukich, K. (2000). Beyond automated essay scoring, the debate on automated essay grading. *IEEE intelligent systems*, *15*(5), 22-27.

Landauer, T. K., Laham, D., & Foltz, P. D. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Shermis, M.D. and Burstein, J.C. (eds), *Automated essay scoring: A cross-disciplinary perspective*, (pp. 87-112). Lawrence Erlbaum Associates.

Lee, Y. J. (2020). The long-term effect of automated writing evaluation feedback on writing development. *English Teaching*, *75*(1), 67-92.

Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications*, *11*(1), 1-9.

Li, J., Jangamreddy, N. K., Hisamoto, R., Bhansali, R., Dyda, A., Zaphir, L., & Glencross, M. (2024). AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology*, *40*(4), 56-72.

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A comprehensive review of automated essay scoring (AES) research and development. *Pertanika Journal of Science & Technology*, *29*(3), 1875-1899.

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605-634.

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education*, *49*(5), 616–633. doi.org/10.1080/02602938.2024.2301722

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35-59.

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.

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.

Neff Lippman, J. (2003). Assessing writing. In I. L. Clark (Ed.), *Concepts in composition: Theory and practice in the teaching of writing* (pp. 199-240). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Page, E.B. (2003). Project essay grade: PEG. In Shermis, M.D. and Burstein, J.C. (eds), *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum, 43–54.

Parra G, L., & Calero S, X. (2019). Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction*, *12*(2), 209-226.

Pati, D., & Lorusso, L. N. (2018). How to write a systematic review of the literature. *HERD: Health Environments Research & Design Journal*, *11*(1), 15-30.

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, *2*(3), 100083.

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education*, *24*(1), 962. *doi.org/10.1186/s12909-024-05881-6*

Palermo, C., & Wilson, J. (2020). Implementing automated writing evaluation in different instructional contexts: A mixed-methods study. *Journal of Writing Research*, *12*(1), 63-108.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495-2527.

Rudner, L. M., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research, and Evaluation*, *7*(26),1-6.

Scardamalia, M., & Bereiter, C. (1991). Literate expertise. In K. A. Ericsson & J. Smith (Eds.), Toward a general theory of expertise (pp. 172–194). Cambridge, UK: Cambridge University Press.

Shabara, R., ElEbyary, K., & Boraie, D. (2024). Teachers or ChatGPT: The issue of accuracy and consistency in L2 assessment. *Teaching English with Technology*, *24*(2), 71-92.

Shermis, M. (2024). Using ChatGPT to score essays and short-form constructed responses. doi.org/10.48550/arXiv.2408.09540.

Shermis, M. D., & Burstein, J. C (Eds.). (2003). Automatic Essay Scoring: A Cross Disciplinary Perspective. Mahwah, NJ: Erlbaum.

Shermis, M. D., Rasmussen, J. L., Rajecki, D. W., Olsen, J., & Marsiglio, C. (2001). All prompts are created equal, but some prompts are more equal than other prompts. *Journal of Applied Measurement, 2*(2), 154-170.

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies*, 1-23.

Smith, D. A. (2017). Collaborative peer feedback. *International Association for Development of the Information Society*.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894.

Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, *42*, 1-16.

Tang, J., & Rich, C. S. (2017). Automated writing evaluation in an EFL setting: Lessons from China. *The JALT CALL Journal*, *13*(2), 117-146.

Uymaz, E. (2019). The effects of peer feedback on the essay writing performances of EFL students. *International Journal of Curriculum and Instruction*, *11*(2), 20-37.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, *2*(1), 319-330.

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

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, *26*(3), 234-257.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies*, 3, 22–36.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research 10*(2), 1–24.

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

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208.

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87-125.

Yang, Y. (2024). The reliability of using ChatGPT in rating EFL writings. *Shanlax International Journal of Education*, *12*(4), 49-59.

Yoon, S. Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion. *arXiv preprint arXiv:2310.06505*.