

Exploring the Potential of ChatGPT for Evaluating English Essays in a Criterion-Based Assessment

ANDREA GJOREVSKI

*East Texas A&M University
Commerce, Texas, USA; and
Partner Language Training Center Europe
Garmisch-Partenkirchen, Germany*

MIMI LI 

*East Texas A&M University
Commerce, Texas, USA*

TROY L. COX

*Brigham Young University
Provo, Utah, USA*

Abstract

Open access to novel AI tools offers unprecedented opportunities for human–AI collaboration in writing instruction and assessment. While research on using generative AI tools like ChatGPT in these contexts is emerging, more is needed to understand their effectiveness as Automated Writing Evaluation (AWE) tools. This study explores the potential of ChatGPT (GPT-3.5) to assist teachers and learners in the North Atlantic Treaty Organization (NATO) by evaluating English writing based on holistic scoring criteria. Using a mixed-methods approach, the study compared ChatGPT's ratings with human ratings on 100 writing tests to assess inter-rater reliability. It also analyzed the justifications provided by both human raters and ChatGPT to evaluate how well ChatGPT understood the rating criteria at different proficiency levels and whether its rationales could provide effective feedback for learners and support teacher feedback practices. Results showed strong agreement between ChatGPT's and human ratings, with ChatGPT demonstrating a similar understanding of the rating scales and offering justifications with elements of effective feedback. These findings indicate that ChatGPT holds promise as an AWE tool,

providing meaningful feedback and valuable insights into holistic rating scales. This study encourages further exploration of AI in the L2 classroom and suggests leveraging AI to enhance writing pedagogy and classroom-based assessment.

doi: 10.1002/tesq.70011

INTRODUCTION

The emergence of artificial intelligence (AI) tools like ChatGPT, released by OpenAI in 2022 has ushered in a new era of human and computer collaboration, with practitioners and researchers in second language (L2) studies engaging with these tools in ways that have the potential to reshape the landscape of language pedagogy and assessment. Among the myriad uses, ChatGPT has provided great potential for supporting learners with writing development (Barrot, 2023; Kim & Chon, 2025; Su, Lin, & Lai, 2023). L2 composition, however, poses distinct challenges due to its inherent complexities, and these challenges are further compounded with classroom assessment, where evaluation should be learning-oriented, criterion-referenced, and facilitative to the development of students' overall writing proficiency (Ferris & Hedgcock, 2014). Teachers are increasingly held accountable for assessment in the classroom (Hamp-Lyons, 2013) with the expectation that they have the expertise and knowledge to not only develop and assess student progress, but also provide tailored feedback for improvement, interpret criteria, and report scores (Cumming, 2009). However, few teacher training programs include formal instruction on language assessment or even focus on language assessment literacy in teacher training (Fulcher, 2012; Weigle, 2007; Yan, Zhang, & Fan, 2018). This gap has left many language teachers struggling with the substantial challenges of writing evaluation of their learners and feedback practices.

Automated writing evaluation (AWE) systems have proved to help facilitate feedback provision. AWE systems can evaluate and score written texts accurately (Warschauer & Ware, 2006) as well as provide constructive feedback to learners (Grimes & Warschauer, 2010). Relying on advanced computational techniques and algorithms using natural-language processing (NLP) and machine learning (ML), AWE technologies have been in use since the 1960s (Cotos, 2014). Yet significant advances in technology since then have led to more current generative artificial intelligence (GAI) tools such as ChatGPT. As ChatGPT can also be prompted to provide evaluations to essays, L2

researchers (e.g., Barrot, 2023; Han & Li, 2024; Su et al., 2023) have examined the advantages and challenges of using ChatGPT for writing feedback. In addition, language testing researchers have begun to explore the use of ChatGPT in writing assessment contexts (Kim et al., 2024; Mizumoto & Eguchi, 2023; Yancey, Laflair, Verardi, & Burstein, 2023) with results indicating the potential for its use in providing reliable ratings as compared with human scores. Furthermore, studies (e.g., Steiss et al., 2024) have investigated ChatGPT's potential to provide formative feedback comparable to teachers' comments on writing assessments. Naismith, Mulcaire, and Burstein (2023) investigated the potential of using ChatGPT to score discourse coherence on written responses from the Duolingo English Test (DET), finding an exact agreement rate of 56% and 53% between GPT-4 and human raters depending on whether the rating was provided before or after the rationale. Their content analysis of human and GPT-written rationales for their ratings highlighted consistency in feedback content and structure, as well as GPT-4's ability to incorporate grading criteria concepts into its justifications.

Enlightened by Naismith et al.'s (2023) study, we sought to evaluate ChatGPT's ability to assess the writing proficiency of personnel working in various positions within the North Atlantic Treaty Organization (NATO) context.¹ These personnel across nations must develop their proficiency to achieve their required scores on an assessment using the NATO language proficiency scale. Given the high-stakes nature of NATO proficiency assessments, test takers are under significant pressure to perform at the required level, and teachers face the challenge of helping develop their learners' writing proficiency to meet the standards. While the language teachers in most nations are highly competent and effective, teachers may lack comprehensive familiarity with the grading criteria for the NATO tests or may not be calibrated to the standards followed by official testers. Similarly, students often face challenges in identifying actionable strategies to improve their writing and achieve their required scores. ChatGPT 3.5—a widely accessible, free, and user-friendly tool—holds significant potential to address these existing gaps by offering appropriate evaluations and effective feedback that can foster positive washback in the classroom. This study thus compares ratings provided by expert human raters with those generated by ChatGPT after training, specifically assessing whether ChatGPT's interpretation of the scoring criteria aligns with that of human raters. Additionally, the study investigates the utility of

¹ The first author works at the Partner Language Training Center Europe, which has the mission “to provide flexible language training for US personnel while offering English and strategic language instruction and expertise for NATO Allies and worldwide partners”.

ChatGPT’s rationales as feedback mechanisms, which could assist learners in revising and improving their writing. By systematically analyzing ChatGPT’s performance, this study seeks to assess its potential to address challenges faced by both teachers and students while preparing for NATO writing tests.

In the following section, we synthesize relevant literature that informed this study and provide the contextual foundation for the research.

LITERATURE REVIEW

Theoretical Frameworks

Given that this study focuses on assessing ChatGPT’s potential as an effective and appropriate automated scoring and feedback tool, we based our inquiry on the framework proposed by Williamson, Xi, and Breyer (2012) for evaluating and implementing automated scoring systems for constructed-response writing tasks. Drawing on trends in assessing computer-based scoring techniques and grounded in Kane’s argument-based validation framework (Kane, 1992, 2006, as cited in Williamson et al., 2012), their work establishes “guidelines, criteria, and best practices under a unified evaluation framework for operational deployment of automated scoring in an assessment for high-stakes purposes” (p.4). The framework consists of five inference-based areas of emphasis that serve as steps to guide the evidence-gathering process to support assumptions about test scores and use. The present study focused solely on the first two areas, *Construct Relevance and Representation* and *Empirical Performance (association with human scores)*, as these were deemed the most germane to our purposes and provided a solid foundation for investigating ChatGPT’s potential as a classroom-based writing assessment tool. The area of *Construct Relevance and Representation* recommends that one critically evaluate the alignment between the assessment construct and the system’s capability (p.6). By addressing a series of questions in this area aimed at determining whether the scoring system adequately represents and measures the intended construct, stronger evidence will be gathered to support the tool’s use. The present study aims to gather evidence about how well ChatGPT’s design as a large language model (LLM) can in fact understand the construct of the test and assess writing based on each scale level descriptor. Furthermore, the study aims to better understand which features of text ChatGPT utilizes in its evaluation and whether these features align with human raters as well as the scale criteria. We posit that ChatGPT’s transformer neural network

architecture and extensive training, which enable it to predict language patterns and contexts effectively (Sarrion, 2023), make it well-suited for evaluating essays at specific test levels and assessing them according to the criteria outlined in the scale. The second area, *Empirical Performance*, outlines criteria for comparing computer-generated scores with human scores, emphasizing that despite challenges identified in research regarding human raters, human scores remain the “gold standard.” Following some of the criteria, our empirical analysis was based on a study design which included three human scores and one ChatGPT score per essay, allowing for inter-rater reliability analysis to better assess the quality of the human scores as compared with ChatGPT scores. To enhance score reliability, we employed trained human raters and further trained ChatGPT using benchmark samples and expert rationales, aiming to align its scoring capabilities with those of the trained raters.

Alongside the theoretical framework used to guide score comparisons and assess ChatGPT’s rating effectiveness, our investigation into ChatGPT’s ability to provide meaningful feedback was also grounded in Vygotsky’s sociocultural learning theory. Vygotsky’s sociocultural theory introduces the concept of the zone of proximal development (ZPD), the space between what a learner can accomplish independently and what they can achieve with assistance and in collaboration with someone more competent (Vygotsky, 1978). Within this space, feedback functions as essential mediation, facilitating progress within the ZPD. When integrated into instruction, feedback extends beyond mere guidance, transforming into a powerful tool for fostering learning and development. By providing structured, criteria-based feedback on a specific writing task, ChatGPT can facilitate understanding of writing tasks, enabling learners to refine their skills and fill the learning gap.

To guide our application of these theoretical underpinnings as they apply to L2 writing contexts, we analyzed ChatGPT’s rating rationales through a typology of feedback devised by Pearson (2022). Pearson’s typology builds on Ellis (2009), who categorized feedback as direct, indirect, and metalinguistic, by expanding on these approaches and including students’ perceptions of feedback and their ability to apply it during the revision process. It includes 10 research-based characteristics² of feedback. Our analysis focused on four of these ten areas most relevant to our objectives: specifically, *mode and tone* (e.g., praising or critical), which have been shown to influence student perceptions of feedback (Hyland & Hyland, 2001); *text specificity*, where

² Focus on target; mode and tone; syntactic structure; text specificity; location; explicitness; length; mitigation strategies; delivery method; and temporality (pp. 4–5).

specific comments rather than generic ones positively impact revisions (Ferris, 1997); *comment length*, long feedback showing the most positive revision results (Ferris, 1997); and the *presence of mitigation strategies* (e.g., hedging or interrogatives), which may positively influence learners' revision processes (Nurmukhamedov & Kim, 2009). To Pearson's framework, we added two additional strategies: *advice for revision*, shown to support learners effectively during revisions (Ferris & Roberts, 2001) and *criteria-based feedback*, as it has been argued that students need to understand the standards in order to monitor their own development (Sadler, 1989).

Automated Writing Evaluation

Initially developed in the 1960s in order to help cut down on the amount of work faced by English teachers (Page, 1966), AWE systems have since been used to comprehensively evaluate essays in educational contexts and reliably score high-stakes assessments (Chapelle, Cotos, & Lee, 2015). In both testing and teaching, AWE systems have been used to not only assign a numerical score to a piece of writing (Shermis, Burstein, & Bursky, 2013) but also perform a wide range of analyses on written text (Warschauer & Ware, 2006). Comprised of two components for scoring and feedback provisions, AWE tools are commonly used for diagnostic and formative feedback purposes to provide information to students on their writing strengths and weaknesses (Grimes & Warschauer, 2010), and research in low-stakes classroom environments has shown that AWE tools can evaluate texts with scores that correlate relatively well with human scores (Warschauer & Ware, 2006). The feedback produced by AWE systems has been found to be of similar quality to teachers' feedback and, as such, is often utilized in classrooms in combination with teacher feedback (Woodworth & Barkaoui, 2020).

A desirable quality of AWE systems is their ability to provide feedback to students incredibly quickly as compared to the time it would take an instructor to assess the work. Teachers and students alike can effectively utilize the features of AWE systems to engage with one another through the writing process, which can not only improve the writing abilities of students, but also facilitate feedback practices of teachers (Jiang, Yu, & Wang, 2020; Wilson & Roscoe, 2019; Zhang, 2020). The feedback AWE systems provide has been welcomed for its immediacy, accuracy, and ability to "transform and enhance learners' language learning experience" (Xi, 2010, p. 298). AWE also enables instructors to provide more intentional, focused and selective feedback (Jiang et al., 2020). Both students' and teachers' favorable

perceptions of AWE were reported in previous research, such as its potential to enhance learners' motivation, autonomy, and metalinguistic knowledge (e.g., Link, Dursun, Karakaya, & Hegelheimer, 2014; Wang, Shang, & Briody, 2013). Moreover, multiple studies (e.g., Chappelle et al., 2015; Li, Link, Ma, Yang, & Hegelheimer, 2014) also supported the validity of AWE by demonstrating significant correlation between AWE scores and human scores. However, one drawback is AWE's inability to truly comprehend a text the way skilled humans do, with the claim that NLP methods used in automated scoring systems cannot interpret a text using the cognitive processes employed by humans, nor can it understand figurative devices, puns, or other idiomatic nuances of language (Shermis et al., 2013). This drawback makes total reliance on their assessments impractical.

ChatGPT for L2 Writing Feedback and Assessment

While AWE systems have long been used for rating writing and providing feedback, new GAI tools, like ChatGPT, open even more potential. Utilizing comparable technology through NLP techniques, ChatGPT may perform even better than traditional AWE systems in classroom and assessment settings due to its user-friendly interface that allows for conversational interactions and its ability to quickly adapt to virtually any specific writing situation. Su et al. (2023) explore ChatGPT's (GPT 3.5) potential to serve as a peer collaborator in writing classrooms, specifically focusing on the chatbot's ability to assist in argumentative writing tasks. They highlight ChatGPT's similarity to AWE systems, noting its potential to assist learners in argumentative writing by offering tailored and augmented support. They also suggest teachers provide ChatGPT with customized writing rubrics that the chatbot can use to provide more targeted feedback. Barrot (2023) explores ChatGPT's (both the 3.5 and 4.0 models) potential in L2 writing classrooms, highlighting its use as a guide through the writing process and for formative feedback. She emphasizes its ability to score writing based on criteria, justify scores, and provide targeted feedback on higher-order skills. Teng (2024) used a mixed-methods approach analyzing questionnaires and interviews of EFL learners in an English Practical Writing course. The results revealed that students who used AI to edit and proofread their texts showed greater measures of motivation, self-efficacy, and engagement in their writing compared with a control group.

While much investigation has gone into comparing human and AI-produced writing as well as its ability to provide feedback and aid writers in the classroom, there are a few studies addressing effects and

usability of GAI in language assessments. Mizumoto and Eguchi (2023) conducted a study to determine the accuracy and reliability of one of the earliest models in the GPT 3.5 series, GPT 3 text-davinci-003 model, as an automated essay scoring (AES) tool to score essays from the TOEFL11 corpus, a collection of English essays produced by test takers of 11 different language backgrounds who had taken previous TOEFL exams. Their results indicated that GPT follows similar rating patterns for the essays of low, medium, and high quality, that aligned with TOEFL ratings, though exact alignment with human scores were far from perfect (54.33%). Based on these findings, they suggest the potential for GPT's use to support, not replace, human scoring. While innovative and one of the first to investigate GPT as an AES tool, a critical gap in their study was the use of zero-shot prompting, that is, not providing essay samples to train the GAI model, which may have had an impact on the results. Furthermore, as they did not train the model nor did they require the model to provide a rationalization for its scores, they note that the "black box-like nature in which the process by which results are obtained remains obscure" (p.10). As such, they recommend future research address these limitations.

Similarly, a study by Yancey et al. (2023) compared GPT-4 ratings with GPT-3.5 ratings and human ratings using the Duolingo English Test (DET) writing section. Essays were scored on a rubric based on the Common European Framework of Reference (CEFR) scale, and researchers experimented with different training techniques such as including calibration samples or asking ChatGPT to provide a rationale. GPT-4 was reported to outperform GPT-3.5, achieving scores comparable to those generated by AWE systems. However, variations in its ratings compared to human scores were noted across different L1 groups. Despite prompting ChatGPT for rationales, the researchers did not analyze these rating processes to gain deeper insights into its reasoning mechanisms, leaving an important aspect of the study unexplored.

The Present Study

Prior studies highlight the potential of ChatGPT in writing classrooms and assessment contexts, yet further investigation is needed into its role in L2 writing instruction. Specifically, there is limited knowledge about ChatGPT's ability to reliably evaluate writing using a specific set of criteria and provide actionable, criteria-based feedback that can support the development of writing proficiency. Addressing these gaps is crucial to leveraging ChatGPT as a transformative tool for language learning. Motivated by Naismith et al.'s (2023) positive findings, this

research explores the potential of ChatGPT (GPT-3.5) as an innovative AWE tool to assist test-takers of the NATO Standardization Agreement (STANAG) 6001 English proficiency tests.³ These tests serve as gateways for military personnel to access critical NATO roles, requiring extensive preparation through national English Language Training (ELT) programs. While these programs employ highly qualified instructors, many lack training in the STANAG 6001 holistic scoring scale, which demands significant expertise and norming. Similarly, learners often struggle to grasp the scale's application to writing, hindering their progress. Despite the release of GPT-4o (May 2024), this study focused on GPT-3.5 as a free, accessible tool suited for nations with varying resources and aimed to evaluate its potential to provide reliable scoring and meaningful feedback which could help enhance teachers' and learners' understanding of scoring standards and thus support the achievement of proficiency standards in preparation for high-stakes testing. The study was guided by the following research questions:

1. How do GPT-3.5's ratings compare to human ratings on STANAG 6001-based writing proficiency tests?
2. Does GPT-3.5 share a similar understanding of the STANAG 6001 rating scale at each level (Levels 1–3) as human raters? If so, how?
3. How can the justifications provided by ChatGPT serve as feedback to writers?

METHOD

Background to NATO Testing

The NATO STANAG 6001 tests are general proficiency tests within the military context and are used to evaluate the language of military personnel for interoperability purposes within NATO. Nations are responsible for developing and scoring their own tests in accordance with NATO language standards as outlined in the STANAG 6001 document and rating scale⁴ (<https://natobilc.org/wp-content/uploads/>

³ The NATO STANAG 6001 tests are general proficiency tests within the military context and are used to evaluate the language of military and civilian personnel for interoperability purposes within NATO. Most NATO and Partner Nations develop their own tests in each of the four skills (listening, speaking, reading, writing).

⁴ The NATO STANAG 6001 scale, from which the scores are based, consists of six levels with plus levels in between Levels 0 and 4: Level 0 “no proficiency”; Level 1 “survival”; Level 2 “functional”; Level 3 “professional”; Level 4 “expert”; Level 5 “highly articulate native”.

2024/11/ATrainP-5-EDA-V2-E.pdf). The rating scale descriptors at each level define the writing construct through statements related to the content of the message (topics), tasks (can-do statements), and accuracy. In writing tests, single-level prompts are aligned to task statements according to their level. As an example, one of the Level 2 tasks in the STANAG 6001 is to narrate in the past. To elicit this task, a test prompt may have the test taker write an email to a friend describing a past training exercise that was recently conducted. Alternatively, some nations have adopted a practice of “spiraling” test prompts, in which the prompt elicits two base levels in a single prompt; this is done for efficiency and practicality. The prompt presents a context along with two points that the test-taker must respond to. In Levels 1 and 2 spiraled prompt (Figure 1), the first point represents a Level 1 task and the second represents a Level 2 task. In relation to prompt alignment, a significant principle of STANAG 6001 testing is that of being a criterion-referenced test; as such, writing is scored according to the “clearly stated expectations or criteria” (Clifford, 2016, p.225) as found in the STANAG 6001 level descriptors. When operationalized, this means that a prompt written to elicit a Level 2 response can be scored as a maximum a Level 2 if it meets those criteria. Spiraled prompts are rated similarly, though as they assess two levels, the scoring range is wider. Ratings in most testing teams is conducted holistically by comparing the writing sample with the criteria in the STANAG 6001 level descriptions. Most writing tests are bi- or multi-level tests, and examinees are allowed up to 90 minutes to respond to all the prompts. A test is typically scored by two raters, with a third rater serving as arbitrator when the scores do not match.

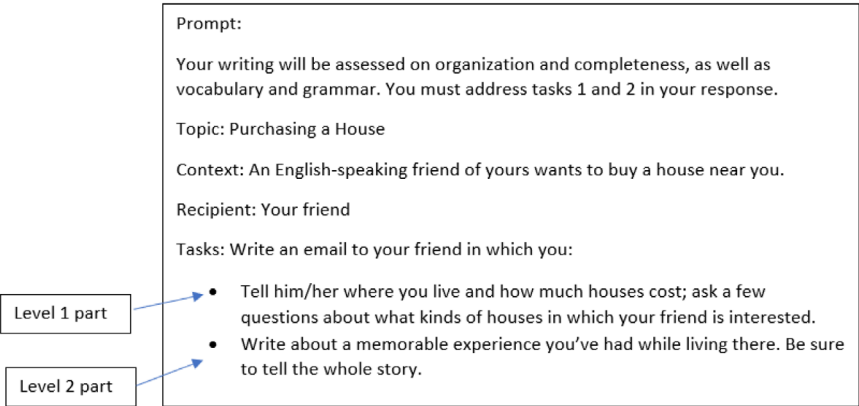


FIGURE 1. Spiraled writing prompt for Levels 1 and 2.

Participants

Participants in the study included 32 human raters representing 12 nations and 1 NATO testing organization.⁵ They are all considered specialists in language testing and have served as official testers and raters in their national testing teams. They all met the study requirements to have at least 3 years' rating experience with STANAG 6001 national tests and to be graduates of the NATO-approved Language Testing Seminar⁶ held at the Partner Language Training Center Europe (PLTCE), which is considered an important first step in the standardization and norming process for NATO STANAG 6001 testers. Most of the participants began their careers in the classroom as language instructors and transitioned to the testing department over time, either on an official, full-time basis or to serve as part-time testers for the productive skills. The L1 of most of the raters was languages other than English.

Dataset

The data used in this study consisted of 100 essays of 100–600 words sourced from a 2019 large-scale study called the Benchmark Advisory Test (BAT) 2 to which the first author had access. Participants in this study signed the consent form indicating that their essays could be used in future studies. The BAT 2 was a multi-national, large-scale investigation of speaking and writing based on the NATO STANAG 6001 language proficiency scale. The BAT 2 writing assessment was a testlet comprised of three different writing prompts, two of which were spiraled at Levels 1/2 and Levels 2/3, while the third prompt was a single-level prompt written at Level 3. The entire testlet was assigned one holistic score based on the evaluation of the three samples together. In order to determine a single score on a single essay for the present study, the first author extracted one hundred essays from the testlets using a stratified random sampling taken from each prompt level (31 Level 1/2 prompt essays, 39 Level 2/3 prompt essays, and 30 Level 3 prompt essays). The first author went through the essays and changed names and specific locations to remove any identifying or

⁵ Participating nations: Bulgaria, Canada, Croatia, the Czech Republic, Estonia, Georgia, Italy, Latvia, Poland, Romania, Spain, the United States, and testers from Supreme Headquarters Allied Powers Europe (SHAPE).

⁶ The Language Testing Seminar is a 2-week course that provides a foundation of language assessment concepts as well as STANAG 6001 norming for language testers in national defense-sector testing organizations.

personal information that the test taker may have added to the writing sample; nothing else in the writing sample was modified.

Data Collection Procedures

The 30 raters were randomly assigned to groups of three to score ten of the same essays, which were labeled according to the prompt level. Each group had a sampling of each of the three prompt types (Level 1/2; Level 2/3; Level 3). To ensure consistency in rating practices and that all raters clearly understood the spiraled-prompt rating process, as this practice may not be used in their nations, raters were sent instructions that they received before explaining the spiraled writing prompts and how to assign the ratings using the STANAG 6001 descriptors. To assist them in rating, they were sent the STANAG 6001 rating scale descriptions for Levels 0+ to Level 3 and two additional rating tools, the Assessment Criteria Key Features and the Writing Factor Grid, both of which are tools used in former trainings to aid raters and help them understand the scale levels for writing. For essays that were at the 1–2 spiral level of prompt difficulty, scores of 1, 1+, and 2 represented the potential range of ratings; for essays at the 2–3 spiral level of prompt difficulty, scores of 1+, 2, 2+, and 3 represented the potential range of ratings; for essays at the 3 level of prompt difficulty, raters could indicate “Non-ratable” (NR), 2+ or 3 as the potential range of ratings. This was due to the fact that if the writing sample did not meet the Level 3 threshold and did not meet the Level 2+ description for inconsistent performance at Level 3, a rating could therefore not be assigned, and it could be deemed “Non-ratable.” In addition to these instructions, the first author selected two essays from each group of 10 (a total of 20 essays) using a stratified random sampling to ensure a variety of levels were represented and asked each rater to provide a written justification for why they assigned that specific rating to each selected essay. Two raters not included in the rating process were asked to provide written rationales on benchmark samples that would be used to train ChatGPT. These raters were selected based on their extensive experience in language testing. They were given identical sets of essays representing each level and asked to provide a written explanation of how the sample met the criteria for a particular level. Finally, ChatGPT was trained and asked to rate and provide justifications for each of the 100 essays. This allowed us to systematically compare human and ChatGPT ratings and justifications on the same texts, providing insights into alignment, discrepancies, and the potential of ChatGPT as a supplementary rater. The next section discusses the training provided to ChatGPT.

Training ChatGPT

Previous studies on using ChatGPT to score essays (Mizumoto & Eguchi, 2023; Yancey et al., 2023) recommended that future researchers try various techniques beyond zero-shot prompting to train ChatGPT; therefore, we used multi-turn engagements and employed concepts from chain of thought (CoT) prompting (Wei et al., 2022), in an attempt to teach ChatGPT how to understand the spiraled prompts and how to assess the STANAG 6001 levels. CoT prompting suggests that a LLM can perform more accurately when employing “a coherent series of intermediate reasoning steps that lead to the final answer for a problem” (p.6). To begin, we used the “Customized” feature built into ChatGPT to assign it a role (a tester of STANAG 6001 writing tests) and explain its task of rating. Once ChatGPT confirmed understanding of its role and task, we explained the structure of spiraled prompts one at a time using some of the same examples sent to raters as part of their instructions. Following CoT prompting to include exemplars (Wei et al., 2022), we provided ChatGPT with benchmark samples of responses at each base level. To represent each base level rating of the spiraled prompts, we used two exemplars at Levels 1/2, two exemplars at Levels 2/3, and one exemplar at Level 3. For each benchmark used, we explained to ChatGPT which level the sample represented and provided the rationales written by the human raters justifying the rating. Figure 2 shows how ChatGPT responded after being provided with one of the benchmark essays rated a Level 2 along with the rationale.

Finally, we shared with ChatGPT the same rating tools as were given to the raters. After this training, ChatGPT was asked to read the text, think through the levels first, and then provide a rating. Figure 3 demonstrates the prompt used.

We copied and pasted each of the 100 texts into ChatGPT in one sitting and recorded its response. While ChatGPT followed its instructions to rate and provide a rationale, occasionally it would instead rewrite the passage. When this happened, we informed ChatGPT that it should not rewrite the essays but instead should read, rate, and provide a justification as before. Of the 100 essays, this occurred five times in the process, with ChatGPT being redirected each time and returning to its previous pattern of rating and justifying.

Data Analysis

To address RQ 1 and compare the human and ChatGPT ratings quantitatively, we used Facets (Linacre, 2024) to conduct a Many-Facet

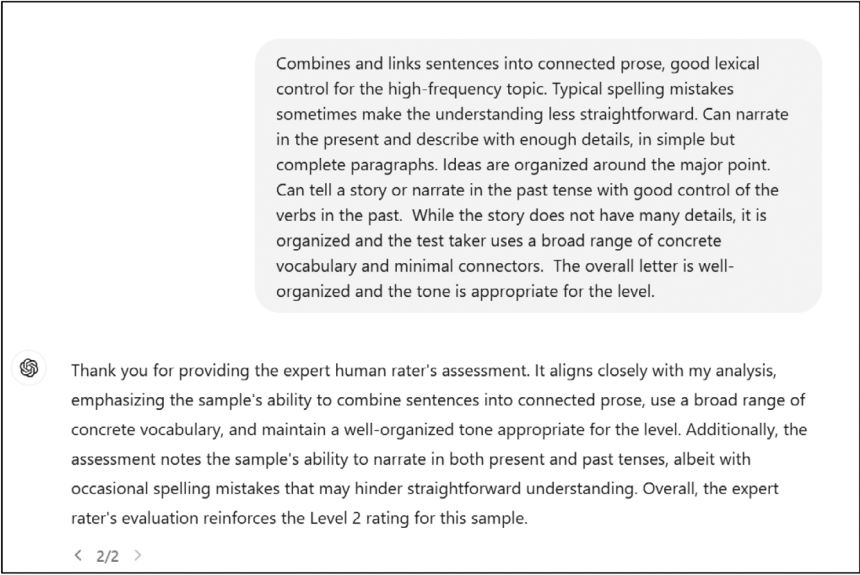


FIGURE 2. ChatGPT demonstrating an understanding of the human rationale at Level 2.

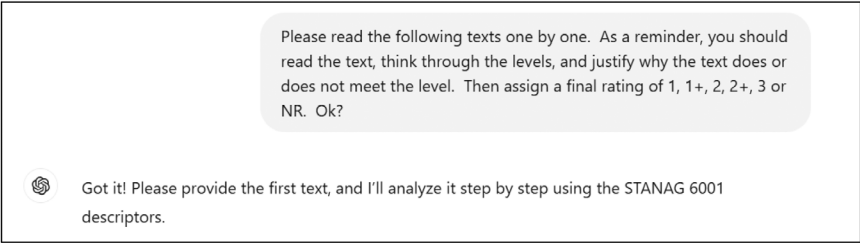


FIGURE 3. Prompt used for ChatGPT to rate and justify its ratings.

Rasch Measurement (MFRM) to analyze the human-rated data and to analyze inter-rater reliability prior to comparing the ratings to ChatGPT. The MFRM analysis in this study employed the Andrich Rating Scale Model (RSM) to evaluate scoring behavior across two facets: students (writing samples) and raters (human and ChatGPT). The Andrich RSM was chosen due to its assumption of uniform thresholds across the STANAG 6001 rubric. Fit statistics (Infit and Outfit MnSq) assessed rater consistency, while the rater separation index and reliability confirmed scoring reliability. Interaction analysis further explored differences in how ChatGPT and human raters applied the rubric across varying student performance levels, providing insights into alignment and discrepancies between the two rating methods. As

the STANAG 6001 rating scale is ordinal, we calculated Spearman's Rho correlation between the two rating methods to evaluate their relationship. In addition, we used Quadratic Weighted Kappa to calculate the reliability between the two methods as it was designed for ordinal rating scales and penalizes disagreements that are further from each other (Matthijs, 2015).

To address RQ 2, whether ChatGPT shares a similar understanding of the rating scale at each level as human raters, we analyzed a total of 80 justifications written for the same 20 previously selected texts. Each of the 30 raters provided justifications for two texts, resulting in 60 human justifications across the 20 texts. Additionally, we extracted ChatGPT justifications for the same 20 texts, bringing the total number of justifications analyzed to 80. These data were analyzed through an iterative cycle of manual coding and content analysis (Saldaña, 2021) to identify descriptive and *in vivo* codes. Two authors collaborated on the coding scheme by discussing 20% of the data from Levels 1 and 1+. After this discussion, the first author coded the rest of the data according to the established coding scheme and provided the second author with 25% of the data ($n = 50$) to independently code. The inter-coder reliability was 94%, which is considered highly reliable; any discrepancies were discussed to further ensure a consistent coding scheme was applied to the data. Table 1 shows the coding scheme with their explanations and exemplars from the data. As an additional measure to determine in which areas both humans and ChatGPT provide feedback, the frequencies of each code were calculated per level for both human raters and ChatGPT. Because the human justifications were three times more than ChatGPT's justifications, we normalized the frequencies by calculating the proportion of comments in each category relative to the total number of comments within each level, for both human raters and ChatGPT. This normalization process adjusted for the difference in sample sizes, enabling a direct comparison of the proportions of comments provided by human raters and ChatGPT in each category.

Inspired by Steiss et al. (2024) and Naismith et al. (2023) who analyzed rationales by ChatGPT as forms of feedback, we analyzed the 20 justifications written by ChatGPT using several criteria of effective feedback in order to answer RQ 3. For the analysis, we examined features that emerged from our data meanwhile using Pearson's (2022) typology for feedback as a springboard. Pearson's typology addresses features of teachers' feedback on L2 writing, which were found to impact the students' revision performances and the quality of revised texts. The categories are shown in Table 2. In particular, we identified if a particular feature was present in the justification and calculated the

TABLE 1
Areas of Human and ChatGPT Writing Feedback

Codes	Explanation	Examples of comments
Sentence structure/grammar	Comments related to grammatical sentence structures, both positive and negative; comments related to use of tenses	Uses short simple sentences and common linking words
Comprehensibility of message	Comments related to ability for the message to be understood by native writers or not; errors causing message to be unclear	Phrases need revision for clarity
Content development/organization	Comments related to development of ideas, paragraphs, and general structure of the writing; well organized or lacking organization	Organized prose with simple paragraphs
Task completion	Comments related to writer’s ability to adhere to the task and complete the whole task	Inconsistent ability to perform the task fully
Cohesion and coherence	Comments related to connectors, cohesive devices, and overall text cohesion	Clear organization and cohesion
Mechanics (spelling, punctuation, capitalization)	Comments related to spelling, punctuation and capitalization errors	Some errors in punctuation, spelling, and capitalization
Lexical complexity and appropriateness	Comments related to type of vocabulary used, whether vocabulary was adequate or not and added to understanding or not	Vocabulary is mostly appropriate
Writing style	Comments related to tone, audience, and formality; comments specifically using the term “style”	Effective writing for the audience
Language command	Comments related to writer’s ability to effectively use language; comments specifically using the term “command of language” or similar wording	Command of written language is firm
Cognitive skills	Comments related to effective and well-thought-out arguments or arguments that were not strong; comments that showed writer’s ability to use rational and logical argumentation	Demonstrates critical thinking and analytical skills

total number and percentage of justifications that exhibited each feature.

RESULTS

In this section, we present our findings as to the three research questions.

TABLE 2
Features of Quality Feedback and Their Explanations

Categories	Explanation
Length	Word count
Text-specificity	Comments may be generic or tie explicitly to a text through examples in the writing sample
Mode and tone	Neutral/negative/positive; advisory, critical, praising, etc.
Presence of mitigation strategies	Ways to cut down on abruptness of feedback, such as through hedging
Advice for revisions	Feedback may or may not include advice for revisions
Criteria-based	Evidence of wording or paraphrasing from the STANAG 6001 rating document, Writing Factor Grid and Writing Assessment Criteria

Comparing ChatGPT 3.5 to Human Ratings

For RQ1, we found substantial agreement between GPT-3.5’s ratings and human ratings on the STANAG 6001-based writing proficiency tests. Using Many-Facet Rasch Measurement (MFRM), we confirmed the reliability of human ratings through the fair average, which adjusts observed ratings to account for variations in rater severity and leniency. Our analysis found a rater separation index of 3.94, indicating a high level of reliability and the ability to distinguish between different rater severities across the 30 raters involved. This corresponds to approximately 5.61 distinct strata, further confirming strong separation and reliable differentiation among rater scoring tendencies. The separation reliability was calculated at 0.94, demonstrating high consistency across raters. We used Infit and Outfit mean-square (MnSq) statistics to assess rater fit, with thresholds for acceptable fit set between 0.5 and 1.5. Of the 30 raters, only two exhibited MnSq values exceeding 2.0, indicating occasional misfit but overall consistent scoring behavior. While a full report of the MFRM analysis, including bias and interaction effects, is beyond the scope of this article, readers can access the control file and subsequent reports at https://osf.io/t54yb/?view_only=64f67f92b0ca4010a65aa2dd042dbafl.

From our MFRM analysis, we felt there was sufficient evidence to support the use of rounding the Fair Average as a way to represent the human ratings. We found that the correlation was significant between ChatGPT and the Human MFRM Fair Average ($\rho = .84$, $df = 98$, $p < .001$). ChatGPT and the Human MFRM Fair Averages had an exact agreement rate of 50% and an adjacent agreement rate of 91% (See Table 3). There was one instance in which ChatGPT underrated the humans by two levels, and six instances in which it overrated them by

TABLE 3
Contingency Table Between Human MFRM Fair Averages and ChatGPT Ratings

	ChatGPT ratings					NR	Total
	1	1+	2	2+	3		
Human MFRM fair averages							
1	8	8	6	0	0	0	22
1+	1	3	5	0	0	0	9
2	0	5	16	2	0	0	23
2+	0	1	6	12	5	0	24
3	0	0	0	9	10	0	19
NR	0	0	0	1	1	1	3
Total	9	17	33	24	16	1	100

Note. Shaded cells indicate exact agreement.

two levels. The “NR” rating category had one instance in which both were in agreement, but two instances in which ChatGPT issued ratings rather than awarding the NR rating. Note that for Level 1, ChatGPT issued 22 ratings whereas the humans only issued 9, and for Level 1+, ChatGPT had 9 ratings whereas the humans had 17. Thus, there seems to be more discrepancy at the lower categories of the ratings.

Interrater reliability was evaluated using Quadratic Weighted Kappa to assess the consistency between ChatGPT and human raters across 100 subjects. The exact agreement rate was 50%, while the Quadratic Weighted Kappa value was 0.80, indicating substantial agreement. A corresponding *z*-value of 8.16 (*p* < .001) confirmed that the agreement was statistically significant and unlikely due to chance.

ChatGPT 3.5’s and Human Raters’ Understanding of the Rating Scale

For RQ 2, results of the coding showed that both types of raters commented on the same areas of writing, though frequencies in each category varied per level, as shown in Table 4. For instance, in the area of *content development and organization*, the frequencies were similar or exact in Levels 2+ and 3, respectively, while in Levels 1, 1+, and 2, ChatGPT commented more frequently in that category than humans did. Generally speaking, *task completion* was viewed as important to the human raters more often than ChatGPT, particularly in Level 1, Level 1+, and Level 2+, resulting in significant differences. It was also found that the category *sentence structure/grammar* was assessed relatively similarly in terms of frequency for Level 1 and Level 2, though quite differently in Level 1+, where humans rated this much more frequently, 27%, compared with ChatGPT, 9%.

TABLE 4
Human and ChatGPT Ratings Per Level in Each Category

	Sentence structure/ grammar	Comp. of message	Content dev./org	Task completion	Cohesion and coherence	Mechanics	Lexical complex and app.	Writing style	Language command	Cognitive skills
Level 1										
Human	23%	3%	3%	32%	13%	16%	6%	0%	3%	0%
ChatGPT	20%	6%	29%	12%	16%	8%	6%	4%	0%	0%
Level 1+										
Human	27%	11%	9%	18%	13%	11%	9%	0%	2%	0%
ChatGPT	9%	14%	18%	9%	23%	5%	9%	14%	0%	0%
Level 2										
Human	25%	10%	7%	18%	9%	11%	14%	4%	2%	0%
ChatGPT	24%	5%	19%	19%	5%	5%	19%	10%	0%	0%
Level 2+										
Human	13%	10%	27%	13%	10%	5%	20%	2%	0%	0%
ChatGPT	21%	4%	29%	4%	8%	4%	8%	13%	0%	8%
Level 3										
Human	4%	2%	24%	8%	14%	18%	6%	2%	14%	6%
ChatGPT	8%	0%	24%	11%	18%	8%	11%	8%	5%	8%
NR										
Human	13%	0%	13%	13%	19%	0%	6%	19%	19%	0%
ChatGPT	13%	0%	50%	13%	0%	0%	0%	13%	0%	0%

Also of interest in the data were frequencies for the category *lexical appropriateness and complexity*, with Levels 1 and 1+ resulting in the same frequencies, Levels 2 similar frequencies, and Levels 2+ and 3 showing more differences, with Level 2+ the greatest difference. The category of *cognitive skills* was used by both sets of raters at Level 3, as well as by ChatGPT at Level 2+, where it seemed most appropriate according to the descriptors and demonstrated ChatGPT's ability to use this category sparingly and appropriately in a similar way as the humans did. Level 3 also showed similarities in the areas of *content development and organization*, where frequencies were equal, as well as *comprehensibility of message*, with very few comments by both sets of raters. A look across the levels shows humans rated *mechanics* higher than ChatGPT, with the greatest difference at Level 3. The category of *writing style* was consistently used more frequently in ChatGPT's justifications than humans' overall.

Finally, the Non-ratable (NR) category frequencies show that raters focused their comments on similar areas, with no comments on *comprehensibility of message*, *mechanics*, or *cognitive skills*, and the categories of both *sentence structure/grammar* and *task completion* being cited equally at 13%. The significant difference in this rating category was observed in *content development and organization* which greatly influenced ChatGPT as seen in 50% of its justifications compared with on 13% of human justifications. In addition, human raters justified NR at 19% for *cohesion and coherence* and *language command*, categories ChatGPT did not address, suggesting human justifications may have provided stronger arguments for the NR designation.

Features of ChatGPT 3.5's Justifications

For RQ 3, we analyzed the 20 rationales written by ChatGPT to determine whether they included elements of effective feedback for revision. Results are shown in Table 5. ChatGPT demonstrated the ability to provide detailed justifications, with the length of feedback varying based on proficiency level. Level 3 samples received the longest justifications (272 words), while Level 1 samples received the shortest (191 words). In terms of text specificity, ChatGPT consistently incorporated specific examples from the writing sample, particularly focusing on word choice and grammatical inaccuracies, showcasing its ability to ground feedback in the text. Its mode and tone were generally supportive, with 15 of the justifications containing praise, such as comments like "you do a good job in..." or "your description... is engaging," often paired with constructive criticism. On the other hand, ChatGPT employed mitigation strategies, such as hedging, in 13

TABLE 5
Instances of Quality Feedback Features in ChatGPT Justifications

Feedback criteria	Number of justifications with the feature present (total $N = 20$)	Percentage of total justifications
Text-specificity	19	95%
Mode and tone (supportive)	15	75%
Presence of mitigation strategies	13	65%
Advice for revisions	12	60%
Criteria-based	14	70%

justifications. Phrases like “you might want to . . .” or “language could benefit from . . .” were used to soften suggestions for improvement, creating a balance between directness and encouragement. In terms of revision advice, ChatGPT provided actionable suggestions in 12 justifications, offering concrete recommendations for improvement, including areas such as organization, language accuracy, and professionalism, all aimed at helping learners achieve higher ratings. Moreover, ChatGPT demonstrated a connection to the rating criteria in 14 justifications, often paraphrasing descriptors and embedding them within elaborated sentences. For instance, in its justification for a Level 2 text, ChatGPT wrote,⁷ “**language used** is **generally** precise and **appropriate** for the intended audience.” Compare this with the STANAG 6001 descriptor for Level 2: “*Vocabulary use is appropriate for high frequency topics, with some circumlocutions . . . the individual writes in a way that is generally appropriate for the occasion.*” This nuanced and contextually relevant integration of criteria-based language provides evidence of ChatGPT’s capacity to offer meaningful and tailored feedback.

Finally, Figure 4 shows a color-coded example of a justification on a sample rated Level 2, demonstrating ChatGPT’s ability to incorporate the previously discussed features of effective feedback into its justification.

DISCUSSION

This study compared human and ChatGPT ratings on a NATO-based writing proficiency test, analyzing both rating alignment and ChatGPT’s potential for writing support. Using Williamson

⁷ ChatGPT-generated text has been bolded for emphasis.

Key:

Text specificity Supportive tone Hedging Advice for revision

This blog post effectively addresses the tasks by comparing and contrasting communication methods used in the military both in the past and present, and speculating on how new technological advances might affect future military communications.

Here's the rationale for the rating:

1. Organization and Logical Development: The post is well-organized, with a clear structure that separates the discussion of past communication methods from the modern ones. Each section flows logically into the next, providing a cohesive narrative. However, there could be a smoother transition between discussing past methods and introducing modern ones.
2. Complexity of Writing: The writing demonstrates a moderate level of complexity, with a mix of simple and compound sentences. There's a good attempt at explaining complex concepts such as the evolution of communication technologies in the military. However, some sentences could be more refined for clarity and precision.
3. Richness and Precision of Vocabulary: The vocabulary used is adequate for the topic, covering terms related to military communication. However, there is room for improvement in terms of variety and precision. For example, using more specific terms for communication technologies could enhance the richness of the vocabulary.
4. Accuracy of Grammar and Spelling: There are some grammatical errors and awkward phrasings throughout the post, such as "methode" instead of "method," "communications" instead of "communications," and "opperations" instead of "operations." These errors detract from the overall clarity and professionalism of the writing.
5. Suitability for the Intended Audience: The post is suitable for blog readers interested in military technology and communication. It provides relevant information in a digestible format, making it accessible to a general audience.

Overall, this blog post demonstrates a solid understanding of the topic and effectively addresses the tasks outlined in the prompt.

FIGURE 4. ChatGPT Feedback Color-Coded for Features of Effective Feedback.

et al.'s (2012) framework for evaluating automated scoring systems, we focused on Construct Relevance and Representation and Empirical Performance. Exact agreement between human raters and ChatGPT was slightly lower (50%) than previous studies. For instance, Kim et al. (2024) had between 55% and 57% exact agreement depending on whether the prompt and source text were used or not. These are similar to the findings by Mizumoto and Eguchi (2023) whose exact agreement rate was 54.33%. However, our adjacent-level agreement was high (91%), which was comparable to these prior studies. These findings underscore that while exact alignment between human and ChatGPT is not easily obtained, there is a significant degree of consistency in the rating process. As viewed in the context of Empirical Performance (Williamson et al., 2012), this pattern of adjacent-level agreement suggests that ChatGPT is capturing aspects of writing proficiency in a similar manner to human raters, even if discrepancies remain. Notably, ChatGPT showed significant discrepancies in

assigning Level 1 and 1+ ratings, suggesting inconsistencies in its interpretation of the scale, particularly at lower levels. This finding is critical for addressing Construct Relevance and Representation, where it is essential to evaluate whether the automated system adequately represents the intended construct. Content analysis revealed variations in category frequency, particularly in *sentence structure, content development, and organization, and task completion*. Adjusting ChatGPT's weighting of these areas at the lower levels could improve its alignment with human raters.

ChatGPT demonstrated a comparable understanding of the rating scale across many categories, though variation existed by level. *Task completion*, for example, was rated similarly at Levels 1 and 2 but differed at Level 1+. It should be noted that the differing total of justifications, with ChatGPT's justifications representing a significantly smaller sample size, may introduce variability and is therefore a limitation of this study. Nevertheless, these findings further highlight the Construct Relevance and Representation of the system, revealing that ChatGPT can process some elements of writing that are consistent with human raters' expectations, but there are nuances in its approach. As Naismith et al. (2023) suggested in their findings, discrepancies could be used to prompt ChatGPT to weigh certain categories more or less to help align it with human raters. For example, our study would suggest that ChatGPT should be prompted to weigh the category of *mechanics* greater across the levels, as humans factored this into their ratings more frequently than ChatGPT, and that the idea of *task completion* should be clarified particularly at Levels 1, 1+, and 2 +.

However, the frequency differences did not always demonstrate ChatGPT's deficiencies in analyzing a text. ChatGPT more frequently commented on both *content development and organization* as well as *writing style* than human raters. A closer look at the data suggests that in commenting more on these areas, ChatGPT had a thorough understanding of more aspects of the STANAG 6001 criteria. Implications of these results are that ChatGPT may be useful in highlighting areas of the scale that tend to be neglected by raters, particular "blind spots" in their rationales. This insight is in line with the Empirical Performance element, which calls for comparing the automated tool's performance with human ratings. Yet, this should be used judiciously as our quantitative analysis shows that ChatGPT was not as accurate with rating the lower-level essays as the human raters were. Furthermore, we concur with Naismith et al. (2023) who caution that the rationales should not be used to provide "insight into the *process of generating* ratings, even when they provide true and relevant information about the response" (p.399).

Our analysis of justifications based on five criteria found that ChatGPT included many features of effective feedback, indicating its strong potential as an AWE tool, a finding similar to results from Steiss et al. (2024). For the first criteria of length, ChatGPT was able to provide very long justifications. Length of justifications does not necessarily imply better quality, and at times the longer justifications focused disproportionately on just a few aspects of the STANAG 6001 descriptor or the connections to the STANAG 6001 descriptor became “lost” in the writing. However, the longer justifications often indicated that ChatGPT was able to write thorough explanations using criteria from the scale to justify its rating, a finding supported by research that shows longer feedback has been found to provide greater support to writers in their revision process (Ferris, 1997). Additionally, ChatGPT’s feedback was text-specific, which has been found to lead students to more revisions (Ferris & Hedgcock, 2014) and used hedging, a technique that has been found to have a positive impact on the revision process (Nurmukhamedov & Kim, 2009). ChatGPT’s evaluation also used a supportive, motivating tone which is highly valued by learners (Weaver, 2006;). In addition, the tone struck a balance of praise and criticism which has been shown to have a positive effect on learners’ ability to revise their text (Andrade & Evans, 2012), and it provided recommendations and advice for revisions which are generally supported by learners (Hyland & Hyland, 2006). Unlike Steiss et al. (2024) who prompted ChatGPT to use “a friendly and encouraging tone” (p.4) and to give advice on ways to improve the writing, our study found that ChatGPT did this without any special prompting to do so. These automatically generated features in ChatGPT evaluations could help students engage with the AI feedback and incorporate it into revisions.

CONCLUSION

Our study presented ChatGPT as a higher level AWE tool to aid learners and teachers in the writing classrooms by helping them understand the assessment criteria for which they are measured in writing proficiency tests. While ChatGPT shows promise in evaluating writing at higher proficiency levels and can provide useful feedback for revision, further refinements are necessary to strengthen its understanding and application of the rating scale. This study suggests that with targeted adjustments to its training and additional data, ChatGPT can become a viable tool for supporting writing assessments. Research on feedback in the L2 classroom has provided much insight into its importance for learners’ writing development (Ferris &

Hedgcock, 2014), and yet teachers face many challenges as assessors of writing (Hamp-Lyons, 2013). The ability of ChatGPT to provide meaningful, scale-based feedback to learners is of great significance, and instructors can draw on the freely accessible tool to support them in their feedback practices and help their students produce writing in light of the rating scale. Nevertheless, we should be cautioned that the AI engine might not respond well to long pieces of writing on multi-level tests, and a sufficient training on the AI tool in advance would be indispensable. Furthermore, we caution that learners should not rely solely on the tool to assess their writing and that teachers should scrutinize and adapt ChatGPT-generated feedback to ensure the evaluation's accuracy and appropriateness (Han & Li, 2024). In light of these warnings, the use of ChatGPT for teachers and trainees in assessments are noteworthy. ChatGPT can provide new insight into the rating scale and help human raters better understand the levels, something that may be welcomed particularly as even the experienced human raters were at times hesitant and unsure of their rating, which was revealed in this study. ChatGPT could serve as a "new pair of eyes" for raters in initial rating training who are learning how to interpret and apply the rating scale. As research suggests that continuous rater training is necessary to ensure that raters develop and remain consistent in their understanding of rating scales (Yan & Chuang, 2023), trainers could use ChatGPT to generate justifications, and after vetting them for appropriateness, these texts would provide new raters with ample training data that could be used in a variety of ways in training sessions.

Based on our findings, future studies can further investigate discrepancies between human rating and ChatGPT rating to better understand why they occur and whether they can be used either for additional ChatGPT training to improve its abilities or human understanding of the grading criteria. Also, future research, particularly longitudinal studies, can explore how ChatGPT ratings and justifications on students' practice essays will prepare them for writing proficiency tests, drawing on triangulated data sources such as interviews, questionnaire surveys, ChatGPT output, and students' writing products and writing scores. Future research can also explore different stakeholders' perceptions of ChatGPT's potential as an AWE/AES tool. As ChatGPT continues to develop, future studies can further examine ChatGPT as an assessment tool in diverse high-stakes testing settings. While there is still much to consider over the use of AI in language assessment in general, testing experts, teaching practitioners, and learners of all types will increasingly experiment with ChatGPT and other AI technologies. As we anticipate innovative ways AI can help in language instructional and assessment practices, we hope to see future research studies

that provide more insights into how we can leverage the AI tools to contribute to our human understanding, and how humans can continue to collaborate with machines to make achievements.

ACKNOWLEDGMENTS

We would like to thank the language testers in the Bureau for International Language Coordination (BILC) community for their support in this study.

DISCLAIMER

The views and opinions presented herein are those of the authors and do not necessarily represent the views of U.S. Department of Defense (DoD) or its Components. Appearance of, or reference to, any commercial products or services does not constitute DoD endorsement of those products or services. The appearance of external hyperlinks does not constitute DoD endorsement of the linked websites, or the information, products or services therein.

CONFLICT OF INTEREST STATEMENT

There is no conflict of interest in this study.

THE AUTHORS

Andrea Gjorevski is a Ph.D. student in Applied Linguistics at East Texas A&M University. Her research areas include second language learning and teaching, second language writing, and language assessment. She currently works for the US government in the field of language teaching and assessment as an ESL/EFL Specialist promoting linguistic operability in multinational contexts.

Mimi Li, Ph.D., is an associate professor of Applied Linguistics at East Texas A&M University. Her research work includes collaborative writing, computer-mediated teacher/ peer feedback, digital multimodal composing, multimodal pedagogy, and use of AI in language education. She has published widely in peer-refereed journals and authored three books and multiple book chapters. She currently serves as an associate editor of the *Journal of Second Language Writing*.

Troy L. Cox, PhD, has worked at Brigham Young University since 1996 and currently serves as the Associate Director of Research and Proficiency Services at the Center for Language Studies, and the coordinator of the Language Sciences Laboratory. He holds a PhD in Instructional Psychology and Technology, with specialization in educational measurement and learning. His research and publications

focus on language proficiency, assessment, language acquisition, psychology of language learning, and objective measurement.

REFERENCES

- Andrade, M. S., & Evans, N. W. (2012). *Principles and practices for response in second language writing: Developing self-regulated learners*. Abingdon: Taylor & Francis Group.
- Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, 100745. <https://doi.org/10.1016/j.asw.2023.100745>
- Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3), 385–405. <https://doi.org/10.1177/0265532214565386>
- Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals*, 49(2), 224–234. <https://doi.org/10.1111/flan.12201>
- Cotos, E. (2014). Automated writing evaluation. In E. Cotos (Ed.), *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement* (pp. 40–64). London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137333377_3
- Cumming, A. (2009). What needs to be developed to facilitate classroom-based assessment? *TESOL Quarterly*, 43(3), 515–519. <https://doi.org/10.1002/j.1545-7249.2009.tb00247.x>
- Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal*, 63(2), 97–107. <https://doi.org/10.1093/elt/ccn023>
- Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly*, 31(2), 315. <https://doi.org/10.2307/3588049>
- Ferris, D. R., & Hedgcock, J. H. (2014). *Teaching L2 composition: Purpose, process, and practice* (3rd ed.). Abingdon: Routledge.
- Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, 10, 161–184.
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9(2), 113–132. <https://doi.org/10.1080/15434303.2011.642041>
- Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning, and Assessment*, 8(6), 1–42. <http://www.jtla.org>
- Hamp-Lyons, L. (2013). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162–183). Cambridge: Cambridge University Press.
- Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System*, 126, 103502.
- Hyland, F., & Hyland, K. (2001). Sugaring the pill: Praise and criticism in written feedback. *Journal of Second Language Writing*, 10, 185–212.
- Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(2), 83–101. <https://doi.org/10.1017/S0261444806003399>
- Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System*, 93, 102302.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535.

- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18–64). Washington, DC: American Council on Education/Praeger.
- Kim, H., Baghestani, S., Yin, S., 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). Ames, IA: Iowa State University Digital Press. <https://doi.org/10.31274/isudp.2024.154.06>
- Kim, M., & Chon, Y. V. (2025). The impact of self-revision, machine translation, and ChatGPT on L2 writing: Raters' assessments, linguistic complexity, and error correction. *Assessing Writing*, 65, 100950. <https://doi.org/10.1016/j.asw.2025.100950>
- 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.
- Linacre, J. M. (2024). *Facets computer program for many-facet Rasch measurement, version 3.87.0*. Beaverton, Oregon: Winsteps.com.
- Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards better ESL practices for implementing automated writing evaluation. *CALICO Journal*, 31 (3), 323–344.
- Matthijs, J. W. (2015). Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, 5(4), 1–4. <https://doi.org/10.4172/2161-0487.1000197>
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, 394–403. <https://doi.org/10.18653/v1/2023.bea-1.32>
- Nurmukhamedov, U., & Kim, S. (2009). 'Would you perhaps consider ...': Hedged comments in ESL writing. *ELT Journal*, 64(3), 272–282. <https://doi.org/10.1093/elt/ccp063>
- OpenAI. (2022). ChatGPT: Optimizing language models for dialogue. OpenAI. <https://openai.com/blog/chatgpt/>
- Page, E. B. (1966). The imminence of... Grading essays by computer. *The Phi Delta Kappan*, 47(5), 238–243.
- Pearson, W. S. (2022). A typology of the characteristics of teachers' written feedback comments on second language writing. *Cogent Education*, 9(1), 2024937. <https://doi.org/10.1080/2331186X.2021.2024937>
- Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144. <https://doi.org/10.1007/BF00117714>
- Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Thousand Oaks, CA: SAGE.
- Sarrion, E. (2023). *Exploring the power of ChaGPT: Applications, techniques, and implications*. Berkeley, CA: Apress. <https://doi.org/10.1007/978-1-4842-9529-8>
- Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 29–47). New York, NY: Routledge.
- 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. <https://doi.org/10.1016/j.learninstruc.2024.101894>

- Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 57, 100752. <https://doi.org/10.1016/j.asw.2023.100752>
- Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, 7, 100270. <https://doi.org/10.1016/j.caeai.2024.100270>
- Vygotsky, L. S. (1978). In M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press. <https://doi.org/10.2307/j.ctvjf9vz4>
- 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., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157–180. <https://doi.org/10.1191/1362168806lr190oa>
- Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment and Evaluation in Higher Education*, 31(3), 379–394.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models* (arXiv:2201.11903). *arXiv*. <https://doi.org/10.48550/arXiv.2201.11903>
- Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, 16(3), 194–209. <https://doi.org/10.1016/j.jslw.2007.07.004>
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- Wilson, J., & Roscoe, R. D. (2019). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1), 87–125. <https://doi.org/10.1177/0735633119830764>
- Woodworth, J., & Barkaoui, K. (2020). Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal*, 37(2), 234–247. <https://doi.org/10.18806/tesl.v37i2.1340>
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300. <https://doi.org/10.1177/0265532210364643>
- Yan, X., & Chuang, P. L. (2023). How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program. *Language Testing*, 40(1), 153–179.
- Yan, X., Zhang, C., & Fan, J. J. (2018). "Assessment knowledge is important, but ...": How contextual and experiential factors mediate assessment practice and training needs of language teachers. *System*, 74, 158–168. <https://doi.org/10.1016/j.system.2018.03.003>
- Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, 576–584. <https://doi.org/10.18653/v1/2023.bea-1.49>
- Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, 43, 100439.