Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 9, 103386 (2026)
https://doi.org/10.29140/lea.2026.103386

A Study on the Reliability of ChatGPT-based English Writing Assessment: Internal and External Perspectives

BOYU WANG[a]
YUNJIA ZHANG[b*]

[a]*School of Foreign Languages, Huazhong University of Science and Technology, Hubei, China*
1197450026@qq.com

[b]*School of English, Sichuan International Studies University, Chongqing, China*
yunjiazhang0418@163.com

Abstract

With the advancement of big data and artificial intelligence, natural language processing (NLP) has been increasingly integrated into educational assessment, facilitating a shift from human to automated scoring in English writing assessment. This study investigates how prompt design, guided by the TELeR taxonomy (Santu & Feng, 2023), and sample training affects ChatGPT-4's scoring reliability. A sample of 120 IELTS-style essays was scored across four analytic dimensions—Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA)—as well as holistically. Internal (intra-rater) reliability was measured via ICC(3,1) across three scoring rounds, while external (inter-rater) reliability was assessed by integrating ChatGPT-4 as an additional rater into a panel of nine human scorers using ICC(2,1). The results revealed that ChatGPT-4's internal consistency was moderate and was markedly improved by detailed prompts on content-related dimensions (TR/CC), but showed minimal gains on linguistic dimensions (LR/GRA). Training with human-scored exemplars under the better prompt reduced discrepancies with human ratings for external reliability, yet slightly decreased internal consistency for most dimensions. These findings position TELeR as a framework linking prompt engineering to psychometric theory and underscore the value of dimension-specific prompt design for reliable AI-based writing assessment.

Keywords: ChatGPT, Human Rating, Writing Assessment, Reliability

Data Availability Statement: *All relevant data are within this paper.*

Introduction

With the rapid development of big data and artificial intelligence, AI technologies are being applied across an expanding range of domains. Natural Language Processing (NLP), as a core area within AI, has seen growing adoption in various language-related applications, particularly in language teaching and assessment, where it demonstrates significant potential. Essay writing, a crucial component of language proficiency assessment, has traditionally relied on human scoring. However, this approach is susceptible to subjective influences, such as the rater's mood and mental state, which can introduce bias into the scoring outcomes. Moreover, manual scoring is time-consuming and labour-intensive, resulting in low efficiency and high operational costs.

In order to improve the reliability of writing assessment, researchers have actively explored and adopted various methods to develop computer-assisted and intelligent essay scoring systems. Early examples of such systems include Criterion, PEG (Project Essay Grade), and the e-rater, among others. The development of automatic scoring systems, capable of replacing manual scoring, has gradually become a major research focus. These systems can efficiently handle large-scale scoring tasks, minimize human errors, and provide standardized and objective assessments. As a result, they reduce subjectivity and improve the consistency and reliability of scores. However, some scholars have pointed out significant limitations in these systems when assessing essays that feature creative ideas, profound insights, or complex argumentation (Ramesh & Sanampudi, 2022).

As one of the most advanced AI models developed by OpenAI, ChatGPT has garnered global attention due to its powerful language generation capabilities, strong learning ability, efficient language processing, and real-time feedback. It has been increasingly applied in language learning and assessment, including the automated scoring of English writing. However, research on the scoring reliability of ChatGPT remains limited, leaving questions about its dependability in writing assessment largely unanswered and in need of empirical investigation. Against this backdrop, this study aims to examine ChatGPT's internal reliability by analyzing consistency across multiple rounds of scoring, and to explore its external reliability by comparing the outcomes of ChatGPT scoring with those of human raters. The findings are expected to provide insights and recommendations for the potential integration of ChatGPT into large-scale English proficiency testing and everyday writing instruction.

Literature Review

Definition and Development of Scoring Reliability

Writing assessment is a systematic process that evaluates proficiency in various dimensions, including vocabulary, grammar, coherence, and rhetorical appropriateness (Bachman & Palmer, 2010). The effectiveness of this assessment in ESL/EFL contexts depends on core qualities such as reliability, validity, and fairness (Huang, 2008). Score reliability is critical, as it reflects the accuracy of evaluating ESL/EFL learners' writing (Huang & Foote, 2010). Without reliability, achieving validity becomes impossible (Mislevy, 2004). Furthermore, if validity is lacking, the fairness of the assessment is compromised (Kunnan, 2000). Therefore, ensuring a high level of score reliability is essential when assessing EFL learners' writing, as it directly influences the perceived credibility and effectiveness of the evaluation process.

In the context of writing assessment, reliability has traditionally been examined from two complementary perspectives: internal (intra-rater) reliability and external (inter-rater) reliability. Internal reliability concerns the extent to which the same rater assigns consistent scores to the same writing sample across multiple evaluation occasions. External reliability, by contrast, concerns the degree of agreement

between different raters (Bachman & Palmer, 2010). The most widely used statistic for quantifying both dimensions of reliability is the intraclass correlation coefficient (ICC) (Koo & Li, 2016; Shrout & Fleiss, 1979). It is based on a variance component model that distinguishes between random errors and systematic errors, making it suitable for evaluating the internal consistency of multiple ratings by the same rater as well as the agreement among different raters (McGraw & Wong, 1996; Koo & Li, 2016). However, a critical methodological distinction, often overlooked in the literature, is between single-measure ICC and average-measure ICC. Single-measure ICC reflects the reliability of a single rater's score (e.g., one human rater), whereas average-measure ICC estimates the reliability of the average score from multiple raters (Koo & Li, 2016). Most studies, however, report ICC values without specifying which type was used, making cross-study comparisons difficult.

The process of writing assessment traditionally relies on human scoring, wherein trained raters apply professional judgment against predefined rubrics (McNamara, 2000). Human scoring offers unique strengths, including sensitivity to nuanced language use, cultural context, and rhetorical devices, as well as the ability to provide personalized feedback (Uludag, 2021). However, it is inherently constrained by factors such as susceptibility to personal bias, substantial time demands, and high operational costs (Wen & Liang, 2024). The scoring method adopted plays a critical role in reliability. Holistic scoring, which assigns a single overall impression score, is efficient for large-scale testing but offers limited diagnostic insight (Huang, 2008; Weigle, 2002). Analytic scoring, which breaks down performance into specific criteria, provides detailed feedback but increases rater cognitive load and may be less practical for mass administration (Weigle, 2002; Choi & Cho, 2018). This trade-off between reliability, practicality, and diagnostic value has been a persistent challenge in human scoring.

In response to these challenges, researchers have increasingly investigated the role of Automated Essay Scoring (AES) systems in writing assessment and education. Early AES systems, such as Project Essay Grade (PEG), relied on regression models that focused on surface features like essay length and word count (Page, 1968). Subsequent systems, including e-rater and IntelliMetric, advanced the assessment process by incorporating natural language processing (NLP) techniques to evaluate grammar, syntax, and overall organization (Attali & Burstein, 2006). The introduction of machine learning, deep learning (such as CNNs and LSTMs), and transformer-based models (like BERT and GPT) has further enhanced the accuracy of scoring (Taghipour & Ng, 2016; Yang et al., 2020). The evolution and application of AES systems represent a significant advancement in the fields of writing assessment and instruction. These systems offer several benefits, including reduced time and costs, as well as alleviating the demands placed on teachers and raters. Additionally, they provide consistency in applying rating criteria, thus promoting fairness in scoring. However, most existing AES systems predominantly utilize feature-based approaches that assess overall writing quality rather than evaluating multiple distinct traits (Perelman, 2014; Ramesh & Sanampudi, 2022). To effectively judge writing quality across different dimensions, AES systems must be trained on various rating traits corresponding to each aspect of writing. This training process requires large amounts of graded data, often consisting of hundreds of essays, along with technical expertise, which can pose significant barriers for many educators (Lee et al., 2024). Although recent developments in ensemble approaches have shown promise (Ramesh & Sanampudi, 2022; Wilson & Huang, 2024), the complexities related to data collection, model development, and evaluation remain substantial obstacles for educators lacking programming skills (Lee et al., 2024).

The application of general-purpose LLMs to the scoring task represents a distinct and promising new direction. These models have been trained on extensive datasets, making them versatile tools for teachers. A key advantage of LLMs lies in their capacity for zero-shot and few-shot learning strategies (Kojima et al., 2022). This means that the models can evaluate responses with little to no prior instruction or rubric detailing how to assess the questions. Due to their wide training, LLMs can engage with

texts on virtually any topic, facilitating the grading process without the necessity for course-specific training. From the perspective of educators and students, what distinguishes ChatGPT from traditional grading methods is its user-friendliness and accessibility; anyone can request an evaluation of an exam across various subjects, and the AI delivers a score in mere seconds. However, empirical findings regarding the reliability of LLMs remain inconsistent. While some studies report high reliability and strong alignment with human evaluators (e.g., Jiang et al., 2023; Pfau et al., 2023), others reveal noteworthy variability (e.g., Khademi, 2023; Yang et al., 2024). This inconsistency underscores that LLM scoring reliability is not an inherent, fixed property but is likely highly sensitive to specific operational conditions.

Factors Influencing the Reliability of LLM-based Scoring

Recent advances in generative LLMs have opened new opportunities for automated essay scoring (Yu et al., 2025). Unlike traditional AES systems that rely on predefined linguistic features and task-specific training, LLMs operate through prompt-based learning. Their transformer-based architecture makes them highly sensitive to input framing, tone, sequencing, constraints, and clarity—all of which shape how a task is interpreted and executed (Tavakoli, 2026). Consequently, what once required feature reengineering or task-specific fine-tuning can now be achieved through carefully designed natural language instructions, making prompt design a decisive factor in LLM performance (Coyne et al., 2023). However, despite its centrality, prompting has remained largely unstructured and inconsistent. Practitioners often rely on intuition, trial and error, or copied templates rather than principled methods, and in the absence of shared standards, output quality varies widely—not only across domains but even within the same user session (Tavakoli, 2026).

From a psychometric perspective, automated scoring with LLMs is fundamentally a measurement process. The prompt functions as the instrument, the model's output as the observed score, and the underlying writing ability as the latent construct (Embretson & Reise, 2000). In this framework, prompt variations constitute a source of systematic measurement error rather than random noise—different prompts do not merely add irrelevant variance but can systematically shift the construct being measured (Messick, 1989). Empirical evidence supports this view. Detailed rubrics with explicit evaluation criteria have been shown to improve agreement with human raters (Mizumoto & Eguchi, 2023), whereas studies using minimal or non-specific prompts have reported poor inter-rater reliability (Khademi, 2023). Task decomposition into explicit sub-tasks (Pfau et al., 2023), domain-specific criteria and worked examples (Luo et al., 2026), and role specification (Stahl et al., 2024) have all been shown to enhance scoring consistency. These findings demonstrate that prompt structure affects scoring reliability, a conclusion consistent with classical test theory, in which instrument standardization is a prerequisite for dependable measurement (Traub, 1997). However, studies define "good" prompting differently, using disparate criteria and terminology. Without a shared standard for describing and comparing prompts, it remains unclear whether performance gains reflect genuine improvements in construct validity or merely idiosyncratic method effects (Tavakoli, 2026), and cross-study synthesis becomes inherently difficult.

To systematize this process, several taxonomies have been proposed. In educational settings, McCormack and Keating (2023) proposed the CRTRF framework, which structures prompts around five stages: Context, Role, Task, Requirements, and Format. At a broader level, Tavakoli (2026) introduced the CASTROFF framework, defining eight dimensions, including Constraints, Audience, Structure, Tone, Role, Output format, Focus, and Function, as a professional standard for prompt engineering across sectors. Another prompt engineering framework was the TELeR taxonomy, which comprises four dimensions, namely Turn, Expression, Role, and Level of Details (Santu & Feng, 2023). Among the various dimensions defined by these taxonomies, several can be interpreted as mapping onto identifiable sources of measurement error in automated scoring. This is particularly true for the Level of Details and

Role dimensions in TELeR, and the corresponding elements in CRTRF and CASTROFF. Role reflects rater framing effects known to introduce systematic bias in human scoring (Myford & Wolfe, 2003). Level of Detail governs rubric granularity, which G-theory identifies as a critical determinant of score consistency across measurement conditions (Cronbach et al., 1972). Expression determines construct representation, aligning with Messick's (1989) validity framework wherein ambiguous instructions produce construct-irrelevant variance. Calibration Guidance operationalizes rater training, an established intervention for improving inter-rater reliability through the use of benchmarked exemplars and feedback on scoring patterns (Elder et al., 2005). Thus, TELeR provides not merely a descriptive taxonomy but a diagnostic framework for identifying how specific prompt features introduce or mitigate measurement error.

Despite theoretical links between prompt structure and measurement errors, significant gaps persist in the literature. First, no study has used TELeR to design prompts that test whether the same LLM produces stable scores across repeated runs and agrees with human ratings. Second, the interactions between TELeR-based prompt design and training sample provision remain unexplored; specifically, how zero-shot and few-shot prompting affect scoring consistency has not been examined. Third, whether prompting effects differ between holistic and analytic rating criteria in L2 writing assessment remains unknown. To address these gaps, we conducted an experiment manipulating two factors: prompt quality and training sample provision. Prompt quality was operationalized via TELeR as better versus poorer prompts; training sample provision as with versus without annotated exemplars. This yielded three conditions: better prompt without training, better prompt with training, and poorer prompt with training. This isolated the effect of training samples (comparing the first two conditions) and the effect of prompt quality (comparing the latter two). Temperature was fixed at 0 to eliminate randomness in token generation. Thus, any variation in consistency reflects prompt and training effects, not model randomness. We used the IELTS rubric to produce a holistic band score and four analytic scores: TR, CC, LR, and GRA. We thus examine how prompt quality and training samples affect ChatGPT's internal consistency and external agreement with human raters, for both holistic and analytic scores. The research questions are:

(1) What is the internal reliability of ChatGPT in English writing assessment? Do variations in prompt quality and the provision of training samples significantly affect ChatGPT's internal scoring consistency?
(2) How does the external (inter-rater) reliability of ChatGPT compare to that of human raters in English writing assessment? Do variables such as prompt quality and training sample provision significantly impact the agreement between ChatGPT and human scoring?

Methodology

This study employed a quantitative approach to evaluate the scoring consistency of ChatGPT (GPT-4 version) in comparison with human raters, using the IELTS Writing Task 2 rubric (the public version available at https://takeielts.britishcouncil.org/sites/default/files/ielts_task_2_writing_band_descriptors.pdf) as the assessment framework. The quantitative analysis focused on intra-rater reliability (ChatGPT's self-consistency across multiple scoring rounds) and inter-rater reliability (agreement between ChatGPT and human raters), measured by the intraclass correlation coefficient (ICC).

Participants and Writing Samples

The writing samples were collected from second-year English majors at a comprehensive university in western China. All participants were enrolled in a compulsory academic English writing course, and

their English proficiency was at the intermediate-high level, approximately corresponding to IELTS band 5.5–6.5. The writing task was modelled strictly on IELTS Writing Task 2: an independent argumentative essay requiring test-takers to present a position, support it with relevant ideas, and organise ideas logically. The specific prompt addressed an academic topic concerning whether talent is an important factor in learning a foreign language, requiring students to take a stance, provide supporting reasons, and draw a conclusion. This task format is consistent with common IELTS Task 2 question types. Students completed the essay under authentic examination conditions within 40 minutes, using pen and paper, with no access to dictionaries or electronic materials. A total of 120 essays were randomly selected from the full set of 256 examination scripts. Each handwritten essay was scanned, converted into electronic text, and then line-by-line proofread to ensure accuracy of transcription, including spelling, punctuation and sentence boundaries.

Data Collection Procedures

Nine female raters participated in this study: two university-level English instructors and seven graduate students in English education. The instructors had approximately ten years of teaching experience and had graded national proficiency tests (CET-4/6, TEM-4). The graduate students had strong English proficiency and experience assisting writing assessments. Before formal scoring, all raters attended a calibration session. We sourced 15 sample essays from official IELTS materials, covering Bands 1 through 9. Certified IELTS examiners had previously rated these essays using official band descriptors. Raters studied these exemplars and discussed the rubric to standardise their scoring. After calibration, the nine raters independently scored 120 IELTS-type essays on four dimensions (TR, CC, LR, GRA). The total score was the mean of the four dimension scores. This produced a 9×120 score matrix. We did not implement iterative calibration, consensus discussions, or arbitration, as this was a low-stakes formative assessment.

We also had ChatGPT-4 score the same 120 essays. Temperature was set to 0 and top-p to 0.01 to minimize randomness. ChatGPT-4 produced a holistic score (1–9) and four analytic scores (TR, CC, LR, GRA; each 1–9) for each essay. We used a three-condition design to examine how scoring reliability varies across prompt and training conditions. We made two focused comparisons. First, to isolate prompt quality effects, we compared better and poorer prompts with training. Comparing prompts without training would confound prompt design with the lack of calibration. Second, to isolate training effects, we compared training versus no training under the better prompt. Testing training under a poorer prompt would blur whether gains come from training or from compensating for poor prompt design. We excluded the fourth condition (poorer prompt without training) because prior studies show it produces unreliable, uninterpretable scores (Kim et al., 2024; Poole & Coss, 2024), adding no value to our two core comparisons. Thus, each essay was scored under three conditions: (1) better prompt without training, (2) better prompt with training, and (3) poorer prompt with training.

We designed prompt quality based on the TELeR taxonomy. The better prompt assigned an examiner role, defined the four IELTS criteria, and included calibration instructions with exemplars. The poorer prompt was brief and underspecified, with only a generic role and open-ended tasks. Table 1 shows the key differences. Training consisted of 15 pre-scored sample essays and calibration instructions—the same essays used in the human rater calibration. In training conditions, ChatGPT received these samples to learn the IELTS criteria and human scoring patterns. No-training conditions omitted these materials. We started a new chat session and reset the model's memory before each trial to eliminate carryover effects. All scoring was conducted via the ChatGPT web interface.

Table 1. *Key differences between the better prompt and poor prompt based on the TELeR taxonomy*

Dimension	Better prompt	Poor prompt
Role	"a highly professional IELTS writing examiner with extensive experience"	"a professional IELTS essay grader"
Level of detail	Explicitly lists and defines the four IELTS criteria (TR, CC, LR, GRA); provides structured calibration instructions for 15 sample essays	Only generally references "scoring criteria"; gives open-ended analysis task without structured guidance
Expression	Uses directive language ("study carefully," "internalize," "strictly adhere"); frames task within formal research context	Uses vague instructions ("analyze to identify patterns"); lacks contextual framing
Calibration guidance	Step-by-step instructions to internalize rubric and scoring tendencies from annotated examples	No explicit calibration steps; only encourages pattern recognition

Data Analysis

After processing the data, we imported the scoring results into SPSS for analysis. To evaluate internal consistency (RQ1), we assessed ChatGPT's agreement across three independent trials under each condition using two-way mixed-effects ICCs with absolute agreement: ICC(3,1) for single-measure and ICC(3,k) for average-measure reliability (k = 3). We interpreted values as: > 0.75 high, 0.50–0.75 moderate, < 0.50 poor (Koo & Li, 2016).

For external reliability (RQ2), we treated ChatGPT as a tenth rater alongside nine humans. Because human raters provided only one score per essay, we randomly selected one of ChatGPT's three trial scores for each essay to maintain parity, yielding ten scores per essay. We calculated two-way random-effects ICCs with absolute agreement: ICC(2,1) for single-rater and ICC(2,k) for average-rater reliability (k = 10). This model treats both ChatGPT and humans as random samples from their respective populations. We applied this to holistic scores and four dimensions (TR, CC, LR, GRA). For each condition, we first computed the ICC for the nine human raters alone as a baseline, then for the full panel of ten. Given that the comparison is based on non-independent human rater data, we adopted a descriptive approach to evaluate the impact of integrating ChatGPT into the scoring panel. The specific criteria for judgment were as follows: (1) a change in the ICC point estimate exceeding 0.1, or a shift across reliability categories (based on the benchmarks proposed by Koo & Li, 2016), was considered practically meaningful; (2) a substantial change was deemed to have occurred if the width of the confidence interval for the 'human+AI' group increased by more than 50% compared to the 'human-only' group, or if the confidence intervals of the two groups exhibited no overlap. Subsequent analyses were interpreted within this framework.

Results

Descriptive Statistics of ChatGPT-4 Ratings and Human Raters

Table 2 presents the mean scores and standard deviations of human raters and ChatGPT-4 under three prompt/training conditions: (1) better prompt without training, (2) better prompt with training, and (3) poorer prompt with training. The ratings are reported for four IELTS analytic dimensions – Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), Grammatical Range and Accuracy (GRA) – and the total score.

Table 2. *Descriptive statistics of ChatGPT-4 ratings under three prompt conditions and human raters*

Dimensions	Better prompt with training		Better prompt without training		Poor prompt with training		Human ratings	
	M	SD	M	SD	M	SD	M	SD
TR	5.530	.397	5.278	.289	5.656	.291	5.544	0.488
CC	5.065	.462	4.811	.272	5.357	.462	5.399	0.401
LR	4.953	.379	4.780	.318	5.103	.289	5.380	0.304
GRA	4.565	.401	4.299	.306	4.778	.294	5.355	0.341
Total Score	5.029	.385	4.791	.290	5.223	.280	5.430	0.354

Note. *M* = mean score; *SD* = standard deviation.

Across all dimensions and the total score, human raters produced higher mean scores (ranging from 5.355 to 5.544) than ChatGPT under any of the three conditions (ChatGPT means ranged from 4.299 to 5.656). For total score, the highest mean was obtained under the poor prompt with the training condition ($M = 5.223$, $SD = 0.280$), followed by the better prompt with training condition ($M = 5.029$, $SD = 0.385$). The lowest total score mean was observed under the better prompt without training condition ($M = 4.791$, $SD = 0.290$). This rank order was consistent across all four individual dimensions and the total score. The better prompt without training condition showed the smallest standard deviations across all dimensions (SD range: 0.272–0.318). In contrast, the better prompt with training condition exhibited the largest standard deviations (e.g., TR: $SD = 0.397$; CC: $SD = 0.462$). The poor prompt with training condition produced moderate standard deviations (SD range: 0.280–0.462), with the highest fluctuation observed for CC ($SD = 0.462$).

For human raters and for all three ChatGPT conditions, the mean scores of the content-related dimensions (TR and CC) were consistently higher than those of the linguistic dimensions (LR and GRA). This pattern was uniform across all conditions. For example, under the better prompt with training condition, TR ($M = 5.530$) and CC ($M = 5.065$) were higher than LR ($M = 4.953$) and GRA ($M = 4.565$). The same ordering was observed under the better prompt without training condition (TR: 5.278, CC: 4.811, LR: 4.780, GRA: 4.299) and under the poor prompt with training condition (TR: 5.656, CC: 5.357, LR: 5.103, GRA: 4.778), as well as in human ratings (TR: 5.544, CC: 5.399, LR: 5.380, GRA: 5.355).

Internal Consistency of ChatGPT Scores Across Prompt and Training Conditions

To assess the intra-rater reliability of ChatGPT, we calculated ICC(3,1) (single-measure) and ICC(3,k) (average-measure) across three rounds under three conditions: (1) better prompt without training, (2) better prompt with training, and (3) poorer prompt with training. Table 3 presents the results with 95% confidence intervals. As shown Table 3, the single-measure ICC values were consistently lower than the corresponding average-measure ICC values across all conditions and dimensions. In practice, however, when ChatGPT is used for essay scoring, it typically produces a single score per response rather than an average of multiple repeated ratings. Therefore, we primarily report the single-measure ICCs as the most realistic indicator of intra-rater reliability. The average-measure ICCs are presented only as a reference for an idealised scenario where the mean of three repeated scores would be used.

Under the training condition, we examined the effect of prompt quality on ICC values. As shown in Table 3, ChatGPT with the better prompt consistently yielded higher single-measure ICCs [ICC(3,1)]

Table 3. *Intra-rater reliability (ICC) of ChatGPT under three conditions*

Dimension	ICC [95% CI]	Better prompt (with training)	Better prompt (without training)	Poorer prompt (with training)
TR	ICC(3,1)	0.547 [0.415, 0.657]	0.608 [0.500, 0.705]	0.388 [0.276, 0.499]
	ICC(3,k)	0.783 [0.681, 0.852]	0.823 [0.750, 0.878]	0.655 [0.533, 0.749]
CC	ICC(3,1)	0.576 [0.472, 0.669]	0.450 [0.307, 0.806]	0.374 [0.251, 0.493]
	ICC(3,k)	0.803 [0.728, 0.859]	0.710 [0.571, 0.806]	0.642 [0.502, 0.744]
LR	ICC(3,1)	0.436 [0.323, 0.545]	0.596 [0.466, 0.705]	0.429 [0.319, 0.537]
	ICC(3,k)	0.699 [0.589, 0.782]	0.816 [0.723, 0.877]	0.693 [0.584, 0.777]
GRA	ICC(3,1)	0.427 [0.288, 0.552]	0.512 [0.389, 0.626]	0.404 [0.289, 0.516]
	ICC(3,k)	0.691 [0.548, 0.781]	0.759 [0.656, 0.834]	0.670 [0.549, 0.762]
Total	ICC(3,1)	0.580 [0.447, 0.688]	0.629 [0.494, 0.735]	0.526 [0.423, 0.624]
	ICC(3,k)	0.805 [0.708, 0.868]	0.836 [0.746, 0.893]	0.769 [0.687, 0.832]

Note. Values are point estimates with 95% confidence intervals in brackets. ICC(3,1) = single measure consistency; ICC(3,k) = average measure consistency across three rounds.

than ChatGPT with the poorer prompt for every criterion, although the magnitude of improvement varied by dimension. For Task Response (TR), the single-measure ICC was 0.547 [95% CI: 0.415, 0.657] for the better prompt and 0.388 [95% CI: 0.276, 0.499] for the poorer prompt. For Coherence and Cohesion (CC), the ICC was 0.576 [95% CI: 0.472, 0.669] for the better prompt and 0.374 [95% CI: 0.251, 0.493] for the poorer prompt. For Lexical Resource (LR), the ICC was 0.436 [95% CI: 0.323, 0.545] for the better prompt and 0.429 [95% CI: 0.319, 0.537] for the poorer prompt. For Grammatical Range and Accuracy (GRA), the ICC was 0.427 [95% CI: 0.288, 0.552] for the better prompt and 0.404 [95% CI: 0.289, 0.516] for the poorer prompt. For the total score, the ICC was 0.580 [95% CI: 0.447, 0.688] for the better prompt and 0.526 [95% CI: 0.423, 0.624] for the poorer prompt.

Moreover, under the training condition, the better prompt led to a marked increase in intra-rater reliability for the content-related dimensions (TR and CC) compared to the poorer prompt. Specifically, under the poorer prompt condition, the ICC(3,1) values for the content-related dimensions were lower (TR: 0.388 [0.276, 0.499]; CC: 0.374 [0.251, 0.493]) than those for the linguistic dimensions (LR: 0.429 [0.319, 0.537]; GRA: 0.404 [0.289, 0.516]). Under the better prompt condition, however, reliability improved across all dimensions, and the increase was particularly pronounced for the content-related dimensions (TR: 0.547 [0.415, 0.657]; CC: 0.576 [0.472, 0.669]), such that their ICC values became higher than those for the linguistic dimensions (LR: 0.436 [0.323, 0.545]; GRA: 0.427 [0.288, 0.552].

Under the better prompt condition, we examined the effect of sample training (with vs. without) on ICC values. As shown in Table 3, for most dimensions, the no-training condition yielded higher single-measure ICCs [ICC(3,1)] than the training condition, with the exception of CC, where training actually improved consistency. For the two linguistic features (LR and GRA) the no-training condition consistently produced higher ICCs. The difference was more pronounced for LR: the ICC decreased from 0.596 [95% CI: 0.466, 0.705] without training to 0.436 [95% CI: 0.323, 0.545] with training. Similarly, GRA dropped from 0.512 [95% CI: 0.389, 0.626] to 0.427 [95% CI: 0.288, 0.552]. For the content-related dimensions, the pattern was mixed. Task Response (TR) followed the same trend as the linguistic features, with the no-training condition outperforming the training condition (0.608 [0.500, 0.705] vs. 0.547 [0.415, 0.657]). In contrast, Coherence and Cohesion (CC) was the only dimension where training improved consistency: the ICC rose from 0.450 [0.307, 0.806] without training to 0.576 [0.472, 0.669] with training.

External Reliability of ChatGPT Scores Across Prompt and Training Conditions

We examined external reliability by comparing mixed human-AI panels with a pure human baseline. Three conditions were tested: poorer prompt with training, better prompt with training, and better prompt without training. We compared dependent ICC estimates using 95% confidence interval overlap. Overlap of less than 50% indicated a significant difference at approximately $\alpha = .05$ (Cumming & Finch, 2005). For the pure human rater panel (Table 4), single-measure ICC(2,1) values ranged from 0.162 to 0.379. Specifically, the lowest value was 0.162 for GRA (95% CI [0.092, 0.242]), and the highest was 0.379 for TR (95% CI [0.270, 0.489]). Average-measure ICC(2,k) values ranged from 0.635 to 0.846. For GRA, the average-measure ICC(2,k) was 0.635 (95% CI [0.477, 0.742]); for TR, it was 0.846 (95% CI [0.769, 0.896]). Our design focused on the change in inter-rater reliability after adding one AI score. We were interested in the consistency of any single rater (including the AI) with the whole group, rather than the reliability of the group average. Therefore, following methodological guidelines (Koo & Li, 2016; Shrout & Fleiss, 1979), we primarily report single-measure ICC(2,1) to capture individual rater effects and sensitivity to changes in ICC values when an AI rater is added.

When a ChatGPT rater using the poorer prompt with training was added to form a 10-rater team, mixed effects were observed: content-related dimensions (TR, CC) decreased, while linguistic features (LR, GRA) increased. For TR, ICC decreased from 0.379 (95% CI [0.270, 0.489]) to 0.348 (95% CI [0.253, 0.448]). The 95% CIs overlapped by 0.178, approximately 86% of the average interval length, indicating no statistically significant difference. For CC, ICC decreased from 0.282 (95% CI [0.198, 0.376]) to 0.262 (95% CI [0.188, 0.348]), with an overlap of 0.150 (approximately 89%), indicating a non-significant difference. In contrast, LR increased from 0.163 (95% CI [0.090, 0.251]) to 0.174 (95% CI [0.102, 0.260]), with an overlap of 0.104 (approximately 62%), indicating a non-significant difference. For GRA, ICC increased from 0.162 (95% CI [0.092, 0.242]) to 0.224 (95% CI [0.138, 0.322]), with an overlap of 0.104 (approximately 62%), indicating a non-significant difference. At the total score level, ICC declined from 0.297 (95% CI [0.183, 0.416]) to 0.282 (95% CI [0.180, 0.392]), with 92% CI overlap.

When a ChatGPT rater using the better prompt with training was added, the pattern mirrored the poorer prompt with training condition: content-related dimensions (TR, CC) decreased, while linguistic dimensions (LR, GRA) increased. However, ICC values were closer to the human-only baseline, and CIs showed greater overlap. For TR, ICC decreased from 0.379 (95% CI [0.270, 0.489]) to 0.361 (95% CI [0.264, 0.463]), with approximately 92% CI overlap, indicating no significant difference. For CC, ICC decreased from 0.282 (95% CI [0.198, 0.376]) to 0.277 (95% CI [0.199, 0.366]), with approximately 97% CI overlap, indicating no significant difference. For LR, ICC increased from 0.163 (95% CI [0.090, 0.251]) to 0.172 (95% CI [0.098, 0.259]), with approximately 95% CI overlap, indicating a non-significant difference. For GRA, ICC increased from 0.162 (95% CI [0.092, 0.242]) to 0.199 (95% CI [0.110, 0.300]), with approximately 78% CI overlap, indicating a non-significant difference. At the total score level, ICC declined slightly from 0.297 (95% CI [0.183, 0.416]) to 0.291 (95% CI [0.186, 0.402]), with approximately 96% CI overlap, indicating no statistically significant difference.

When an untrained ChatGPT with the better prompt was added, inter-rater reliability declined consistently across all dimensions, with the largest drops observed for content-related dimensions (TR, CC). For TR, ICC decreased from 0.379 (95% CI [0.270, 0.489]) to 0.295 (95% CI [0.200, 0.402]), with approximately 63% CI overlap, indicating no significant difference. For CC, ICC decreased from 0.282 (95% CI [0.198, 0.376]) to 0.193 (95% CI [0.123, 0.280]), with approximately 49% CI overlap, suggesting a difference approaching statistical significance. For LR, ICC decreased from 0.163 (95% CI [0.090, 0.251]) to 0.120 (95% CI [0.063, 0.195]), with approximately 72% CI overlap, indicating a

Table 4. *Inter-rater reliability of mixed human-AI teams under three configurations*

	Pure human (9 raters)		+ ChatGPT (PPWT)		+ ChatGPT (BPWT)		+ ChatGPT (BPNT)	
	ICC(2,1)	**ICC(2,k)**	**ICC(2,1)**	**ICC(2,k)**	**ICC(2,1)**	**ICC(2,k)**	**ICC(2,1)**	**ICC(2,k)**
TR	0.379 [0.270, 0.489]	0.846 [0.769, 0.896]	0.348 [0.253, 0.448]	0.842 [0.772, 0.890]	0.361 [0.264, 0.463]	0.850 [0.782, 0.896]	0.295 [0.200, 0.402]	0.807 [0,714, 0.870]
CC	0.282 [0.198, 0.376]	0.780 [0.690, 0.844]	0.262 [0.188, 0.348]	0.780 [0.698, 0.842]	0.277 [0.199, 0.366]	0.793 [0.713, 0.852]	0.193 [0.123, 0.280]	0.705 [0.584, 0.795]
LR	0.163 [0.090, 0.251]	0.637 [0.470, 0.751]	0.174 [0.101, 0.260]	0.678 [0.530, 0.778]	0.172 [0.098, 0.259]	0.674 [0.521, 0.778]	0.120 [0.063, 0.195]	0.578 [0.402, 0.708]
GRA	0.162 [0.092, 0.242]	0.635 [0.477, 0.742]	0.224 [0.138, 0.322]	0.743 [0.616, 0.826]	0.199 [0.110, 0.300]	0.713 [0.554, 0.811]	0.142 [0.074, 0.229]	0.624 [0.445, 0.748]
Total	0.297 [0.183, 0.416]	0.792 [0.668, 0.865]	0.282 [0.180, 0.392]	0.797 [0.686, 0.866]	0.291 [0.186, 0.402]	0.804 [0.695, 0.871]	0.193 [0.106, 0.297]	0.705 [0.542, 0.809]

Note. Values are single measure ICC(2,1) with 95% confidence intervals. Values are average measure ICC(2,k) with 95% confidence intervals. Higher values indicate greater agreement among raters. PPWT = poorer prompt, with training; BPWT = better prompt, with training; BPNT = better prompt, no training.

non-significant difference. For GRA, ICC decreased from 0.162 (95% CI [0.092, 0.242]) to 0.142 (95% CI [0.074, 0.229]), with approximately 90% CI overlap, indicating a non-significant difference. For total scores, ICC decreased from 0.297 (95% CI [0.183, 0.416]) to 0.193 (95% CI [0.106, 0.297]), with approximately 54% CI overlap, indicating a non-significant difference.

Discussion

This study evaluates ChatGPT's reliability in scoring IELTS Writing Task 2 essays. We examined intra-rater consistency across three rounds and inter-rater agreement with humans for both holistic and analytic scores. A systematic asymmetry emerged. Prompt quality and calibration training affected dimensions differently, and this carries implications for LLM use in automated writing assessment.

Construct-dependent Asymmetry in Internal Consistency

Under the training condition, the better prompt produced markedly higher internal reliability for content-related dimensions (TR and CC) than for linguistic dimensions (LR and GRA). For TR, the ICC rose from 0.388 to 0.547; for CC, from 0.374 to 0.576. In contrast, the gains for LR (0.429 to 0.436) and GRA (0.404 to 0.427) were negligible. This asymmetry inverts the typical profile of traditional AES systems, which excel at surface linguistic features such as vocabulary and grammar but struggle with discourse-level constructs (Perelman, 2014; Ramesh & Sanampudi, 2022).

The discrepancy can be explained by considering how the transformer architecture processes text—and the asymmetry operates in both directions. On one hand, the multi-head self-attention mechanism is intrinsically suited to tracking long-range dependencies and holistic textual coherence (Vaswani et al., 2017), making it inherently responsive to macro-structural features such as argument development and logical progression. When the better prompt explicitly operationalises TR and CC criteria through detailed descriptors and exemplars, it functions as a top-down attentional template (Duncan & Humphreys, 1989), directing the model's representations toward these discourse-level features and suppressing reliance on surface proxies. Without such scaffolding, this content advantage disappears: under the poorer prompt, linguistic dimensions showed slightly higher ICCs (LR: 0.429; GRA: 0.404) than content dimensions (TR: 0.388; CC: 0.374), suggesting that the model defaults to surface cues when not guided otherwise. On the other hand, grammatical and lexical evaluation requires fine-grained, rule-bound discrimination, such as article usage and collocation precision, that depends on genuine syntactic understanding rather than surface form manipulation (Bender et al., 2021). Because LLMs process language probabilistically without access to communicative intent or symbolic grammar, they cannot apply explicit rules for grammatical judgment. Instead, they rely on distributional co-occurrence patterns, which are inherently less stable for borderline constructions where adjacent score probabilities are nearly indistinguishable. This architectural constraint explains why traditional AES systems, which rely on deterministic feature counts (Attali & Burstein, 2006; Ramesh & Sanampudi, 2022), achieve stability on these dimensions—though at the cost of construct depth. In Messick's (1989) terms, this pattern captures both sides of construct representation. Well-structured prompts enhance the representation of features aligned with the architecture (TR/CC), while failing to overcome construct underrepresentation for features that are architecturally mismatched (LR/GRA). The content advantage is therefore not an inherent property of the model but emerges from the interaction between prompt design and the transformer's native processing strengths.

Beyond the architectural level, a separate source of variability operates at the inference level. This was evident in the observation that even with temperature set to 0, theoretically the most deterministic configuration, which always selects the highest-probability token, internal consistency did not approach perfect reliability under the better prompt condition. One source of such variability is numerical

non-determinism in GPU computations, where variable execution order in parallel floating-point operations can introduce microscopic rounding errors that propagate through layers (Ouyang et al., 2025). When logits for adjacent score points are nearly equal, as is common with borderline essays, these tiny perturbations can flip the argmax decision. Additionally, platform-specific implementation differences in API or local inference pipelines may prevent bitwise reproducibility across separate calls, even with identical seeds (Akamine et al., 2024). Temperature = 0 is therefore necessary but not sufficient for perfect reproducibility. From a measurement perspective, LLM-based scoring carries an irreducible inference-level noise component that operates independently of prompt design. For high-stakes applications, single-run LLM scoring cannot be treated as perfectly reliable, even under optimal prompting; multiple runs or ensemble approaches may be necessary to stabilise the measurement.

Baseline-dependent Asymmetry in External Agreement

When ChatGPT joined the human panel as a tenth rater, a pattern emerged that at first appears paradoxical: external consistency decreased for content dimensions (TR and CC) but increased for linguistic dimensions (LR and GRA). This asymmetry, however, is not a contradiction. Rather, it reflects the different baseline characteristics of human agreement across dimensions.

For LR and GRA, human raters showed low single-measure ICCs (0.163 and 0.162), indicating that individual raters are unreliable on these dimensions. When aggregated across nine raters, however, their average-measure ICCs reached acceptable levels. This pattern is consistent with Weigle (2002) and Eckes (2012), who showed that individual raters apply idiosyncratic standards—differing in error gravity sensitivity, acceptability thresholds, and attentional fatigue. The human panel's error structure is therefore dominated by random errors rather than shared systematic variance. In G-theory terms, the rater variance component is maximised (Cronbach et al., 1972). When a calibrated ChatGPT joins the panel, it does not add systematic deviation to an already noisy group. Instead, it provides a stable reference point amid high random scatter. By reducing random-error dominance, ChatGPT raises the overall ICC. Under the no-training condition, however, ChatGPT lacks this stability and becomes an additional source of random error, explaining the ICC decline (LR: 0.163 to 0.120; GRA: 0.162 to 0.142).

For TR and CC, human raters also showed low inter-rater agreement (ICC = 0.379 and 0.282), slightly higher than for LR and GRA. When averaged across nine raters, however, the ICCs reached good levels, indicating that the mean of multiple human raters provides a reliable benchmark. The human panel operates with relatively low random error but shared systematic variance around the construct. ChatGPT's judgments are internally stable, but they deviate systematically from this human consensus. The model weights argument components and cohesion cues differently than human rhetorical expectations, introducing construct-irrelevant variance (Messick, 1989) into the panel. Adding such a rater to a panel with low existing variance inflates between-rater variance and depresses ICC. The attenuated decline under training conditions confirms that calibration functions as bias reduction: the drop was smallest with the better prompt with training, larger with the poorer prompt with training, and largest with the better prompt without training. By anchoring the model to human-scored exemplars, training reduces systematic deviation and shifts the model's error structure from systematic toward random.

This baseline-dependent pattern has two implications for human-AI assessment research. First, rater calibration may need to be more intensive before AI benchmarking can be meaningful. Second, benchmarking AI against aggregated human means may be more defensible than adding AI as an extra rater in a noisy panel, because the mean represents the best estimate of the true score, whereas individual raters are themselves unreliable instruments.

Bias-variance Trade-off in Calibration Training

Calibration brought ChatGPT's mean scores closer to human norms yet simultaneously reduced internal consistency across most dimensions. This trade-off, improved criterion alignment at the cost of measurement stability, reflects the bias-variance decomposition (Bishop, 2006).

The untrained model relies on a shallow heuristic, using surface-level indicators such as lexical complexity that are easy to execute consistently but construct-poor. Training induces deeper processing; the model must weigh conflicting evidence—such as sophisticated vocabulary against incorrect collocation—creating a more complex decision boundary that is sensitive to input variations and minor logit perturbations. As model complexity increases, variance rises while bias falls. In measurement terms, training shifts the model from systematic bias toward random error. This pattern echoes findings in rater cognition research. Zhang (2016) found that more accurate human raters are better at integrating information from target essays and more self-conscious about their rating accuracy, suggesting that effective rating requires active synthesis rather than rigid rule application. Training an LLM with human-scored exemplars appears to induce a similar form of expertise: it replaces oversimplified heuristics with a more nuanced, evidence-based decision process that, like human expertise, comes at the cost of stability. The observed pattern is therefore not a failure but a reconfiguration of the error structure from bias-dominated to variance-dominated, where better alignment with human means coexists with lower internal ICC. Among the four analytic dimensions, CC was the exception to this pattern: training improved its consistency from 0.450 to 0.576. This may be because CC evaluates logical organisation and cohesion—constructs that align with the transformer's relational processing strengths. Unlike TR, where training introduced mismatched complexity, or LR/GRA, where the mismatch between probabilistic pattern matching and rule-bound judgment persisted, the CC case suggests that training effects depend on the fit between the construct and the architecture. Training with human-scored exemplars thus reinforces an existing architectural strength rather than introducing complexity that the architecture cannot easily accommodate.

These findings extend the TELeR framework by showing that prompt dimensions affect reliability through distinct mechanisms that interact with the construct being assessed. The Level of Details dimension matters most for TR and CC because it captures the discourse-level features that the transformer architecture processes most effectively. For LR and GRA, the same dimension has weaker effects because fine-grained grammatical judgment cannot be scaffolded through prompt design alone. The Role dimension may also contribute: assigning a specific rater persona provides a stable interpretive frame that reduces variability from inconsistent judgmental stances across runs, an effect particularly consequential for content dimensions where holistic interpretive frames matter more than for fine-grained grammatical judgments. This cautions against treating prompt quality as unidimensional. Its effects are construct-specific—a principle that parallels facet design in G-theory, where measurement procedures are optimised separately for each variance component (Cronbach et al., 1972). These findings thus link prompt engineering to psychometric theory: well-structured prompts improve measurement quality for discourse-level features (TR and CC) but cannot resolve the fundamental signal-to-noise problem for micro-linguistic features (LR and GRA). The training-induced trade-off between mean-score alignment and internal consistency exemplifies the bias-variance dilemma inherent in complex measurement instruments, extending the classical reliability-validity tension into the domain of LLM-based assessment.

Conclusion

This study evaluated ChatGPT's reliability in scoring IELTS Writing Task 2 essays. We examined intra-rater consistency across three scoring rounds and inter-rater agreement with a panel of nine

human scorers for both holistic and analytic scores. Three findings emerged from this investigation. First, ChatGPT's internal consistency is not uniform across dimensions. It is higher for content-related dimensions (TR and CC) than for linguistic dimensions (LR and GRA) when the prompt provides discourse-level scaffolding. Second, this internal consistency does not translate equally into external agreement. When ChatGPT is added to a human panel, agreement decreases for TR and CC but increases for LR and GRA. Third, training with human-scored exemplars improves alignment with human raters but simultaneously reduces internal consistency—a bias-variance trade-off that echoes patterns observed in human rater cognition. Taken together, these three patterns suggest that the reliability of LLM-based scoring is not a single property but a function of the interaction between the construct being assessed, the prompt design, and the reference standard against which the model is compared.

Several limitations warrant acknowledgement. First, the incomplete 2×2 design—omitting the poorer prompt without training condition—prevents a full factorial analysis; we cannot disentangle the main effects of prompt quality from those of calibration training or test for interaction effects. Our conclusions about prompt quality and training therefore apply only to the specific combinations examined. Second, the sample comprised 120 essays from second-year English majors at a single university in western China (approximately IELTS band 5.5–6.5), all responding to a single Task 2 prompt. This restricted proficiency range narrows score variance, attenuates ICC estimates, and limits generalisability. Third, the low human single-measure ICC constrains the validity of using individual raters as benchmarks. Future studies might benchmark AI against aggregated human means, or adopt calibration protocols (e.g., iterative calibration, consensus, arbitration) modelled on high-stakes assessment. Fourth, we did not manipulate model parameters or collect qualitative data (e.g., scoring justifications), potentially losing insights into why human consistency was low or how training affected ChatGPT's internalisation of scoring standards. The findings also derive from a single LLM (ChatGPT-4); whether they generalise to other architectures (e.g., Claude, LLaMA) remains an open question.

Despite these limitations, the study offers both theoretical and practical contributions. Theoretically, it positions prompt engineering as inherently measurement-relevant, demonstrating that prompt dimensions affect reliability differently depending on the construct being assessed. This extends the TELeR taxonomy from a descriptive framework to a diagnostic one, linking prompt design to established psychometric theory. Practically, the findings highlight that a one-size-fits-all approach to prompt design is insufficient. Effective prompting may be dimension-specific, tailored to the cognitive demands of each scoring criterion. For low-stakes formative assessment, where consistency across submissions is paramount, simpler prompts without calibration may suffice. For high-stakes summative assessment, where alignment with human judgment is critical, the marginal gains in mean-score convergence must be weighed against increased within-model variability.

Declarations

In the preparation of this manuscript, the authors used DeepSeek (provided by Hangzhou DeepSeek Artificial Intelligence Co., Ltd., China) solely for language polishing and proofreading to improve the clarity and readability of the text. No AI tools were used for data analysis, interpretation, or the generation of research findings or conclusions. The authors assume full responsibility for the originality, accuracy, and integrity of all content presented in this manuscript.

Funding

基于AI生成式命题技术的外语学业测试命题模式探索与实践[257033].

References

Akamine, A., Hayashi, D., Tomizawa, A., Nagasaki, Y., Akamine, C., Fukawa, T., ... & Odate, Y. (2024). Effects of temperature settings on information quality of ChatGPT-3.5 responses: A prospective, single-blind, observational cohort study. *medRxiv*, 2024–06. https://doi.org/10.1101/2024.06.11.24308759

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment, 4*(3).

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). Springer.

Choi, I., & Cho, Y. (2018). The impact of spelling errors on trained raters' scoring decisions. *Language Education & Assessment, 1*(2), 45–58. https://dx.doi.org/10.29140/lea.v1n2.61

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction. *arXiv*. https://doi.org/10.48550/arXiv.2303.14342

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* Wiley.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological review*, *96*(3), 433.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Elder, C., Knoch, U., Barkhuizen, G., & Von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly: An International Journal*, *2*(3), 175–196. http://dx.doi.org/10.1207/s15434311laq0203_1

Embretson, S. E. (2000). Item response theory for psychologists. *Lawrence E*.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing, 13*(3), 201–218. https://doi.org/10.1016/J.ASW.2008.10.002

Huang, J., & Foote, C. J. (2010). Grading between the lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly, 7*(3), 219–233.

Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages, 8*(4), 247.

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning & Teaching*, *6*(1), 75–80. https://doi.org/10.37074/jalt.2023.6.1.28

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. *Exploring artificial intelligence in applied linguistics*, 73–95. https://doi.org/10.31274/isudp.2024.154.06

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199–22213. http://dx.doi.org/10.5555/3600270.3601883

Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (No. 9). Cambridge University Press.

Lee, Y. J., Davis, R. O., & Lee, S. O. (2024). University students' perceptions of artificial intelligence-based tools for English writing courses. *Online Journal of Communication and Media Technologies, 14*(1), e202412.

Luo, N., Wang, Y., Zhang, Z. V., Zhou, Y., & Zhao, R. (2026). Impact of prompt sophistication on ChatGPT's output for automated written corrective feedback. *ReCALL*, *38*(2), 240–254. https://doi.org/10.1017/S095834402510044X

McCormack, M. P., & Keating, J. G. (2023). *Investigating the impact of large language model technologies on syllabi and assessment design.*

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30.

McNamara, T. (2000). *Measuring second language performance.* Routledge.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Mislevy, R. J. (2004). Can there be reliability without "reliability?". *Journal of Educational and Behavioral Statistics, 29*(2), 241–244.

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

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2025). An empirical study of the non-determinism of ChatGPT in code generation. *ACM Transactions on Software Engineering and Methodology*, *34*(2), 1–28. https://doi.org/10.1145/3697010

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, 210–225.

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

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

Poole, F. J., & Coss, M. D. (2024). Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The Devil is in the Prompt (s). *Journal of Technology & Chinese Language Teaching*, *15*(1).

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

Santu, S. K. K., & Feng, D. (2023). Teler: A general taxonomy of LLM prompts for benchmarking complex tasks. *arXiv*, 2305.11430.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.

Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024, June). Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th workshop on innovative use of NLP for building educational applications (BEA 2024)* (pp. 283–298).

Taghipour, K., & Ng, H. T. (2016, November). A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882–1891).

Tavakoli, H. (2026). *Prompt engineering for everyone: A self-taught, human-centered approach to AI programming*. Apress.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, *16*, 8–13.

Uludag, P. (2021). Exploring EAP students' perceptions of integrated writing assessment. *Language Education & Assessment, 4*(3), 81–104. https://doi.org/10.29140/lea.v4n3.507

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Weigle, S. C. (2002). *Assessing writing.* Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wilson, J., & Huang, Y. (2024). Validity of automated essay scores for elementary-age English language learners: Evidence of bias? *Assessing Writing, 60*, 100815.

Wen, Q. F., & Liang, M. C. (2024). Human-AI interactive negotiation competence: ChatGPT and foreign language education. *Foreign Language Teaching and Research, 56*(2), 286–296.

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020, November). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1560–1569).

Yang, Y. (2024). The reliability of using ChatGPT in rating EFL writings. *Shanlax International Journal of Education, 12*(4), 49–59. https://doi.org/10.34293/education.v12i4.7855

Yu, S., Androsov, A., & Yan, H. (2025). Exploring the prospects of multimodal large language models for Automated Emotion Recognition in education: Insights from Gemini. *Computers & Education*, *232*, 105307.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, *27*, 37–53. https://doi.org/10.1016/j.asw.2015.11.001