Journal of English Language Teaching and Learning

University of Tabriz

2025, 17, 36

Genre and Rater Variation in IELTS Writing Assessment: A Comparative Perspective

Reihaneh Shoghi (Corresponding Author)

PhD, Department of English Language and Literature, University of Isfahan, Isfahan, Iran.

reihaneshoghi@gmail.com

Mohammad Bahrami

PhD, Department of English Language and Literature, Sheikhbahaee University, Isfahan, Iran.

bahrami_mhm@yahoo.com

Amir Mahshanian

PhD, Department of English Language and Literature, University of Isfahan, Isfahan, Iran.

mshn_amir@yahoo.com

ARTICLE INFO:

Received date:
2025.09.09
Accepted date:
2025.11.17

Print ISSN: 2251-7995
Online ISSN: 2676-6876

Keywords:

Writing Assessment, Rater Background, IELTS Writing, Writing Genre, Native Raters, Non-Native Raters

Abstract

As AI-driven tools gain prominence in informal language assessment, high-stakes proficiency exams such as the International English Language Testing System (IELTS) continue to depend on trained human raters to ensure fairness, genre sensitivity, and construct validity. This study examines how raters' linguistic backgrounds—specifically English L1 and Persian L1—affect holistic scoring across writing genres within the IELTS framework. Some experienced EFL teachers (11 English L1, 11 Persian L1), all trained in writing assessment and familiar with IELTS rating procedures, evaluated 150 argumentative and descriptive essays produced by advanced Iranian learners under time-controlled conditions. Scoring was based on the publicly available IELTS Task 2 band descriptors, ensuring standardization and alignment with institutional criteria set by Cambridge Assessment. Findings revealed that while both groups demonstrated overall scoring consistency, English L1 raters applied stricter standards to organization in argumentative writing, whereas Persian L1 raters were more sensitive to grammatical accuracy across both genres. The results underscore the critical role of human raters in detecting discourse-level features not fully captured by AI-based scoring systems. Implications are offered for rater training, assessment fairness, and genre-specific writing instruction in standardized testing contexts.

Citation: Shoghi, R.; Bahrami, M. & Mahshanian, A. (2025). Genre and Rater Variation in IELTS Writing Assessment: A Comparative Perspective. *Journal of English Language Teaching and Learning*, 17 (36), 319-340. DOI: 10.22034/elt.2025.69090.2819

Introduction

Writing assessment remains a cornerstone of second language (L2) proficiency evaluation, especially in high-stakes standardized tests such as the IELTS (International English Language Testing System) and TOEFL (Test of English as a Foreign Language). Among the four skills assessed in these exams, writing is often considered the most complex and subjective to evaluate (Taylor et al., 2012; Weigle, 2002). Unlike receptive skills, which can be scored objectively, writing tasks require human raters to make interpretive judgments about content, organization, grammar, and vocabulary. These judgments, although guided by analytic or holistic rubrics, are inherently influenced by raters' linguistic and cultural backgrounds (Barkaoui, 2011; Knoch, 2009).

In recent years, the field of language assessment has witnessed a growing interest in the integration of automated essay scoring (AES) systems such as E-rater, IntelliMetric, and more recently, AI-based tools like ChatGPT. These technologies have become especially prevalent in informal testing environments and online IELTS/TOEFL preparation platforms, where scalability, speed, and objectivity are desirable features (Mizumoto & Eguchi, 2023, Uyar & Büyükahıska, 2025). AES systems are praised for reducing labor-intensive marking tasks, ensuring a consistent application of scoring criteria, and promoting scoring objectivity (Hussein et al., 2019). Tools like Grammarly have even been shown to detect more surface-level errors than human raters, though humans tend to assign higher overall scores (Almusharraf & Alotaibi, 2022).

Further research supports the reliability and cost-efficiency of AES systems across various contexts, such as nursing education (Stephen et al., 2021) and primary education (Chen et al., 2022). Nevertheless, these systems exhibit variability in reliability depending on student proficiency levels. Hand-scoring appears more dependable for struggling writers, while AES systems demonstrate greater consistency for proficient ones (Chen & Sun, 2025). Despite notable advancements—such as improved semantic coherence integration in tools like SAGE (Zupanc & Bosnić, 2017) and promising validation results from Chinese-developed AES systems (Chen & Sun, 2025)—the interpretive limitations of AI remain a key concern. As Cotos (2019) and Chan et al. (2023) note, although AES tools may match or exceed inter-rater consistency compared to human scoring, their rhetorical interpretations often fall short of stakeholder expectations in authentic writing assessment contexts.

This view is echoed by Xu et al. (2024)**,** who argue that while AES tools offer promising accuracy in controlled applications, they still require refinement in scalability and interpretability to meet real-world classroom needs. AES also struggles with capturing creativity, practical reasoning, and genre-sensitive discourse structures, especially in tasks that demand higher-order cognitive processing such as argumentative writing (Hussein et al., 2019). These limitations raise concerns about over-reliance on automation in settings where the nuances of writing performance carry significant educational or professional consequences.

Accordingly, while IELTS benefits from fully human-scored writing assessments, TOEFL employs a hybrid model that combines automated scoring with evaluations from certified

human raters. According to ETS, "Writing tasks are scored based on the Writing Scoring Guides (Rubrics) by a combination of AI scoring and certified human raters" (ETS, 2024). Despite the integration of AI in TOEFL scoring, both tests continue to emphasize trained rater input to ensure construct validity and alignment with scoring rubrics. This reliance on human raters is supported by research highlighting their superior capacity to assess discourse-level appropriateness, rhetorical effectiveness, and genre conformity—dimensions where AI systems often underperform or misclassify (Ramezani et al., 2025; Koraishi, 2024). Moreover, studies comparing automated and human scoring show that while numerical scores may statistically correlate, the qualitative feedback and instructional insights generated by human raters often diverge significantly from AI-generated responses (e.g., Ramezani et al., 2025). These findings underscore the ongoing relevance of human judgment in maintaining fairness and validity, particularly in performance-based tasks involving complex genre-specific conventions (Barkaoui, 2010; Lim, 2011).

Equally central to performance-based assessment is the role of genre. In IELTS writing, especially Task 2, candidates are expected to produce argumentative or discursive essays, which demand not only linguistic competence but also genre-specific rhetorical strategies. Genre awareness is thus essential not only for test-takers but also for raters who interpret and score the responses. Research confirms that rater expectations are highly genre-sensitive and that rating behaviors vary across genres, particularly in criteria like logical coherence, organization, and development of argument (Barkaoui, 2010; Zhang & Liu, 2021). For example, studies have shown that raters apply more rigorous standards to argumentative essays due to their inherent demands for evidence-based reasoning and structured logic (Park, 2015). This genre effect has also been supported by Bouwer et al. (2015), who found that only a small portion of variance in writing scores can be attributed to individual writing skill, underscoring the influence of genre and task design. Similarly, Zhang et al. (2021) examined Chinese EFL learners' argumentative and application letter essays and found that despite genre differences, holistic writing scores were significantly influenced by lexical sophistication and syntactic complexity—particularly the use of complex nominals and type-token ratios. Their findings reinforce that genre shapes both writing production and its evaluation, often in systematic ways.

Rater background also remains a major source of variability. For instance, Cumming et al. (2002) showed that genre familiarity and cognitive schema can guide rater interpretation, which raises concerns about construct validity in multilingual and multicultural assessment settings. Schaefer (2008), using multi-faceted Rasch measurement, found that native English-speaking raters exhibited recurring bias patterns depending on the writing trait and student proficiency level—highlighting not just severity differences but also trait- and writer-specific bias. Bejar et al. (2020) advanced this inquiry by developing predictive rater models based on linguistic features, demonstrating that raters exhibit stable and measurable differences in how they interpret essays—particularly as a function of essay length and other features captured by automated engines. These patterns suggest the potential of integrating AI and predictive modeling for better rater quality control, but also highlight the subjective nature of human evaluation.

Training is one approach that has shown promise in minimizing such variation. Attali (2016) demonstrated that even brief training sessions with immediate feedback enabled novice raters to perform comparably to experienced raters in terms of score consistency, variance, and validity coefficients. However, the efficacy of such training depends on whether it includes genre-specific expectations, which many current protocols neglect. While previous work has addressed either rater variation (e.g., Barkaoui, 2010; Lim, 2011) or genre-specific expectations (e.g., Zhang & Liu, 2021; Bouwer et al., 2015), studies like those by Zhang et al. (2021) and Schaefer (2008) underscore that both writer- and rater-side features interact in nuanced ways, affecting scoring validity. Despite this, relatively few studies have examined how genre expectations and rater identity interact simultaneously within test-authentic contexts such as IELTS. This lack of integrated research leaves open important questions regarding scoring consistency and fairness in global assessment environments.

In light of the ongoing evolution in writing assessment practices—including the growing tension between automated evaluation and human judgment—and the limited research integrating both genre sensitivity and rater identity in authentic test contexts, this study addresses a critical gap. Prior research has produced conflicting results: while some studies have found strong alignment between human and automated scoring on surface-level features (e.g., Attali & Burstein, 2006; Mizumoto & Eguchi, 2023), others highlight major discrepancies in discourse-level judgments and genre-specific expectations (Barkaoui, 2011; Cotos, 2023). Likewise, while many studies have explored either rater variation (e.g., Bejar et al., 2020; Schaefer, 2008) or genre effects in isolation (e.g., Bouwer et al., 2015; Zhang & Liu, 2021), few have examined how these two variables interact in high-stakes, test-authentic environments. To address this gap, the current study adopts a rater-comparative quasi-experimental design, aligned with IELTS Task 2 writing formats. It investigates how two distinct groups of trained EFL teachers—one comprising native English-speaking raters and the other non-native Persian-speaking raters—evaluate argumentative and descriptive essays produced by advanced Iranian EFL learners under standardized, time-controlled conditions. The goal is to better understand how rater identity and genre expectations shape holistic judgments, with implications for score validity, fairness, and the training of human raters in global assessment contexts.

Literature Review

The evaluation of L2 writing performance has traditionally involved the interplay of three major components: task characteristics, test-taker ability, and rater judgment (Weigle, 2002). Among these, rater variation has received increasing attention, particularly in light of studies demonstrating that raters differ not only in scoring severity but also in their interpretation of genre expectations and textual coherence (Barkaoui, 2011; Lim, 2011). These discrepancies can pose threats to test validity and score comparability in international assessments such as IELTS.

Seminal studies such as those conducted by Connor-Linton (1995) and Shi (2001) revealed that while raters from different linguistic backgrounds may assign similar scores, their underlying rating strategies and evaluative priorities vary significantly. Connor-Linton (1995), for example, noted that American and Japanese raters differed in their focus on rhetorical

structure versus grammatical accuracy, respectively. Similarly, Shi (2001) found that Chinese L1 raters emphasized content and organization more than English L1 raters, who stressed textual cohesion and coherence.

Kobayashi (1992) extended these findings through both holistic and analytic evaluations, demonstrating that notions like "clarity" and "accuracy" were interpreted through the lens of raters' linguistic training and cultural norms. More recently, Park (2015) emphasized that genre further complicates rater evaluations, with argumentative writing prompting stricter judgment due to expectations of logical reasoning and evidence-based support. Bouwer et al. (2015) also found that only 10% of the variance in writing scores was attributable to individual skill, and that genre significantly influenced generalizability across tasks, underscoring the need for genre-sensitive assessment practices.

Zhang and Liu (2021) offered further evidence on the predictive power of genre in L2 writing assessments. Their study demonstrated that syntactic complexity—especially clausal density—predicted holistic writing scores more effectively in argumentative than in narrative genres. Genre effects were more pronounced under timed conditions, aligning closely with standardized test environments like IELTS. These findings reinforce the notion that writing quality and rater perceptions are deeply intertwined with genre characteristics.

In addition to genre and rater identity, the advent of Automated Essay Scoring (AES) systems has reshaped the landscape of writing assessment. Chan et al. (2023) found that AES systems could achieve scoring consistency comparable to human raters when calibrated through a Many-Facet Rasch Measurement framework. Cohen et al. (2018) showed that although AES systems maintained a similar level of consistency as human raters, they were less valid in capturing nuanced writing quality, thereby separating reliability from construct validity.

Mizumoto and Eguchi (2023) explored the use of GPT-3 for AES on the TOEFL11 corpus and concluded that while AI language models like ChatGPT demonstrated reasonable scoring accuracy, they performed better when linguistic features were explicitly integrated. This supports findings from Kumar and Boulanger (2020), who used deep learning to enhance AES interpretability, showing a high level of agreement (QWK = 0.78) with human ratings. Their study also highlighted the necessity for explainable AI in educational contexts, suggesting that AES systems should mimic human feedback with transparency and rationale.

Yet, caution is warranted. Cotos (2019) argued that AES tools often fail to detect communicative goals and rhetorical strategies, thereby producing feedback that lacks pedagogical clarity. Chen and Sun (2025) echoed this concern in their analysis of Chinese-developed AES tools, noting that while some systems aligned closely with human raters, others inflated scores due to limited sensitivity to linguistic nuance. Similarly, Almusharraf and Alotaibi (2022) found that AES (Grammarly) detected more mechanical errors than human raters but awarded lower scores, revealing a disconnect between surface error detection and holistic writing quality.

Complementing these assessment-oriented studies, Damayanti et al. (2023) approached genre from an instructional perspective. Their research, which implemented a genre-based

pedagogy using the Reading to Learn (R2L) model for IELTS Task 2 preparation, showed that explicit genre instruction improved coherence and organization even among lower-proficiency learners. This pedagogical lens aligns with the current study's emphasis on genre awareness in both test-takers and raters. A related perspective is offered by de Oliveira and dos Santos (2025), who demonstrated how AI-generated mentor texts could be leveraged in the Teaching and Learning Cycle (TLC) to enhance genre-based instruction for L2 learners.

Despite the growing interest in these themes, there is still a lack of empirical studies that simultaneously examine genre effects and rater background within test-authentic contexts. Most existing research isolates either genre or rater identity, limiting the generalizability of findings to operational, high-stakes assessment environments. In contrast, the current study is grounded in high-stakes test design by aligning its tasks, scoring rubrics, and administration protocols with IELTS Task 2. It uniquely explores how two linguistically distinct rater groups—L1 English and L1 Persian—evaluate descriptive and argumentative essays written by advanced EFL learners under timed conditions. The study employs both holistic and analytic scoring to provide a comprehensive analysis of rater judgment across genres, addressing critical gaps in the literature and informing future practices in multilingual language assessment. This investigation is guided by the following research questions:

1. Do native and non-native trained raters assign significantly different scores to IELTS-aligned argumentative writing tasks?
2. Do native and non-native trained raters assign significantly different scores to IELTS-aligned descriptive writing tasks?

Method

Research Design

This study adopted a quasi-experimental comparative design to investigate the influence of rater background (native English L1 vs. Persian L1) and writing genre (argumentative vs. descriptive) on essay evaluation. These two independent variables were systematically integrated to address the above-mentioned research questions. To address study questions, the writing prompts and conditions were modeled on IELTS Task 2, ensuring high ecological validity. Test-takers were given standardized, timed writing sessions (40 minutes) using prompts in both genres (argumentative and descriptive). Essays were produced under conditions mirroring those used in IELTS preparation centers, and all responses were handwritten to simulate authentic test settings.

Scoring employed both analytic and holistic rubrics. The primary evaluation tool was the IELTS Writing Task 2 public band descriptors, which evaluate essays based on four major criteria: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy (Cambridge Assessment English, 2023). These descriptors have been widely used in empirical validation studies and provide a reliable framework for comparing rater judgments across contexts.

Statistical analyses included paired-sample t-tests to assess intra-rater variability across genres and independent-sample t-tests to examine inter-rater group differences (native vs. non-

native raters). This dual-layered analytic approach allowed for a robust examination of how genre and rater background jointly influence writing assessment outcomes.

Participants

Raters

Twenty-two certified EFL instructors participated in this study, consisting of 11 native English-speaking raters and 11 native Persian-speaking raters. All participants were experienced classroom teachers with 9 to 13 years of EFL teaching experience and between 5 to 8 years of IELTS writing assessment experience. To ensure a consistent and test-specific scoring perspective, all raters held formal certification as IELTS instructors through programs approved by Cambridge Assessment. Raters who had any simultaneous certification or teaching experience with TOEFL writing tasks were excluded to avoid potential overlap in rating philosophies.

The raters completed a specialized training program adapted from Knoch (2009), focused on the IELTS Task 2 public band descriptors. This program consisted of four phases: (1) instructional modules on rubric interpretation and rating logic, (2) scoring of benchmark essays with immediate expert feedback, (3) calibration sessions to ensure scoring alignment among raters, and (4) a certification phase requiring interrater agreement at or above a Cohen's kappa of 0.75. The training emphasized rubric alignment and interrater reliability in the context of argumentative and descriptive genres commonly used in IELTS writing tasks. All selected raters successfully passed the calibration phase, ensuring consistency and construct-relevant scoring for the writing samples.

Learners

An initial pool of 32 advanced EFL learners was screened from an academic English course at the C1 CEFR level. Fifteen learners were selected for the final analysis based on a multi-step validation procedure to ensure homogeneity of language proficiency. All learners took the Oxford Placement Test (OPT), and those who achieved scores within the advanced range were shortlisted.

Additionally, a structured oral interview based on IELTS Speaking Part 2 prompts was conducted. Responses were audio-recorded, transcribed, and independently rated by two IELTS-certified instructors using the official IELTS Speaking Band Descriptors (Cambridge Assessment English, 2020). The interrater reliability between the two evaluators was $r = .86$. To further validate the learners' proficiency, the instructor of the course, acting as an expert judge, confirmed the selected students' level based on classroom performance and communicative competence. Only those learners who demonstrated consistent advanced-level proficiency across the written test, oral performance, and expert validation were included in the final sample.

Table 1. *Rater Profile Summary (N = 22)*

Rater	Experience (Years)	Gender	Age	IELTS Rating Experience (Years)	First Language	Texts Rated
R1	11	Male	34	6	English	30
R2	10	Female	30	6	English	30
R3	9	Male	28	5	English	30
R4	12	Female	33	7	English	30
R5	13	Male	35	8	English	30
R6	10	Female	31	6	English	30
R7	11	Male	37	7	English	30
R8	12	Female	39	7	English	30
R9	9	Male	29	5	English	30
R10	10	Female	32	6	English	30
R11	11	Male	36	6	English	30
R12	10	Female	34	6	Persian	30
R13	9	Male	30	5	Persian	30
R14	12	Female	37	7	Persian	30
R15	13	Male	40	8	Persian	30
R16	10	Female	33	6	Persian	30
R17	11	Male	35	7	Persian	30
R18	9	Female	28	5	Persian	30
R19	12	Male	38	7	Persian	30
R20	10	Female	31	6	Persian	30
R21	11	Male	36	6	Persian	30
R22	13	Female	39	8	Persian	30

Note. Each rater assessed a total of 30 essays (15 argumentative and 15 descriptive).

Materials

The study employed a range of materials developed or adapted from validated language assessment resources to ensure alignment with internationally recognized frameworks, particularly IELTS and TOEFL. These materials were used for rater training, learner instruction, writing task administration, and scoring procedures.

Instructional and Preparatory Materials

Materials

The study employed a set of instructional and evaluative materials adapted from validated language assessment resources to ensure alignment with internationally recognized frameworks, particularly the IELTS Academic Writing Task 2 format. These materials supported rater training, learner preparation, task administration, and scoring procedures, thereby maintaining construct validity throughout the study.

Instructional and Preparatory Materials

To ensure genre awareness and task readiness among learners, two instructional booklets were developed—one for argumentative and one for descriptive essay writing. These were adapted from established IELTS preparation materials, specifically *The Official Cambridge Guide to IELTS* (Cullen et al., 2014), and incorporated sample essays, annotated structures, and commonly used lexical bundles found in high-scoring responses. Genre-specific guidance included detailed explanations of purpose, organizational structure, signal words, sentence stems, and paragraph development.

Each booklet was accompanied by structured oral instruction. Learners received four standardized lessons—two for each genre—delivered by course instructors over two consecutive sessions. Each session lasted 45 minutes and emphasized rhetorical purpose, paragraph organization, coherence markers, and lexical choices. This instructional phase ensured consistency in learner exposure and reduced instructional variability, aligning with genre-based pedagogy principles (Hyland, 2007; Lee, 2017).

To validate the instructional content, two PhD-level experts in applied linguistics reviewed the materials for content and face validity. Their feedback led to minor modifications in genre exemplars and cohesion instruction to better reflect IELTS scoring criteria.

Writing Task Prompts

The two writing prompts—one descriptive and one argumentative—were designed to mirror the structure and style of IELTS Task 2. To ensure genre comparability and task fairness, the prompts were piloted with a group of five advanced-level EFL learners (not included in the main study). Three EFL writing instructors evaluated the prompts based on clarity, cognitive demand, and genre alignment. Example prompts included:

- **Descriptive Task**: "Describe a place in your country that is popular among tourists and explain why it is attractive."
- **Argumentative Task**: "Some believe that university education should be free. Others think students should pay for their studies. Discuss both views and give your opinion."

The topics were selected for their cultural neutrality, appropriateness for advanced learners, and relevance to IELTS preparation contexts, in line with the item design principles set out by Bachman and Adrian (2022) and Bachman and Palmer (2010).

Scoring Rubrics and Rater Training Materials

For essay evaluation, the raters used a holistic scoring rubric adapted from the official IELTS Task 2 public band descriptors published by Cambridge Assessment English (2019). These descriptors cover four domains:

1. Task Response
2. Coherence and Cohesion
3. Lexical Resource

4. Grammatical Range and Accuracy

These rubrics are widely recognized in international assessment contexts and have been validated in numerous studies for reliability and construct representation (Knoch, 2009; Barkaoui, 2011).

To support consistent application of the rubric, raters were provided with a comprehensive Rater Training Handbook prepared by the researchers. This handbook included:

- Annotated sample essays at high, mid, and low proficiency levels
- Score justification sheets highlighting domain-level rationale
- A genre comparison guide outlining expectations for descriptive vs. argumentative writing
- A rater self-monitoring checklist

Training involved two 90-minute sessions where raters scored three benchmark essays together. Raters were encouraged to discuss disagreements, and inter-rater reliability was monitored. A threshold of ±0.5 band agreement was set as the benchmark for calibration before official scoring commenced. This procedure followed recommended practices for rater calibration and moderation in language assessment (Weigle, 2002; Lim, 2011).

Procedure

The study followed a structured, multi-phase procedure to ensure methodological rigor and validity within the context of IELTS-style academic writing assessment. The procedure was divided into three phases: participant recruitment and validation, rater training and calibration, and test administration and scoring.

Phase 1: Participant Selection and Validation

Initially, 32 Iranian EFL learners enrolled in an advanced English writing course were considered for participation. To ensure homogeneity in language proficiency, all candidates took the Oxford Placement Test (OPT), and 15 learners were selected based on their scores aligning with the CEFR C1 level. In addition, an oral interview test was administered to all candidates using standardized prompts from the Cambridge English Qualifications Speaking resource, which employs uniform interlocutor frames and task materials to ensure consistency and fairness across administrations (Taylor, 2003). The responses were scored using the IELTS Speaking Band Descriptors. Interviews were recorded, transcribed, and evaluated independently by two trained raters. Inter-rater reliability was established at 0.89 (Cohen's kappa), confirming consistent rating. The learners' original course instructor, a certified IELTS tutor with over 15 years of experience, also validated their proficiency status based on class performance. Candidates who exhibited extreme scores (either very low or very high) or failed to attend the full testing sessions were excluded to minimize statistical noise and maintain comparability across participants.

Phase 2: Rater Selection and Training

Twenty-two experienced EFL teachers (11 L1 English, 11 L1 Persian) were recruited as raters. All raters had at least eight years of EFL instruction and writing assessment experience and had completed a rater training course specifically adapted from the IELTS Task 2 public band descriptors. This training was organized and delivered by the research team and drew on established calibration models (Knoch, 2009; Barkaoui, 2011; Weigle, 2002). The training was delivered in three 2-hour sessions (six hours total) and included the following components:

- Introduction to IELTS Task 2 scoring descriptors
- Side-by-side rubric interpretation and dimension alignment
- Norming with three benchmark essays (high, mid, low) with feedback
- Calibration activities using the Rater Handbook (see Appendix A), which included annotated sample essays, genre feature guides, and scoring justification sheets.

The final calibration step involved raters independently scoring three essays and discussing score discrepancies in guided consensus-building sessions. All raters demonstrated acceptable consistency with a benchmark rate (QWK > 0.75) and were approved for participation. No rater had simultaneous certification in both IELTS and TOEFL to avoid conflated scoring philosophy.

Phase 3: Writing Task Administration and Scoring

Learners were scheduled for two writing sessions (one descriptive, one argumentative), each administered on separate days in a quiet university computer lab supervised by two proctors. Each task was allotted 45 minutes. Learners were seated apart to reduce distractions, and instructions were provided orally and in written form. No additional support or feedback was given during the tasks. Essays were handwritten using standard IELTS response booklets to mimic authentic testing conditions.

Upon collection, essays were anonymized and randomly ordered before scoring. Each essay was evaluated independently by two raters using a 10-point holistic scale derived from the IELTS Task 2 band descriptors. The scale addressed four dimensions: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. Raters also recorded brief analytic comments for each essay to identify perceived strengths and weaknesses.

Inter-rater reliability was calculated using Cohen's weighted kappa, and disagreements exceeding two band points were resolved through consensus scoring. This dual approach allowed for both quantitative and qualitative insights into rater behavior across genres and linguistic backgrounds.

Data Analysis

All statistical analyses were conducted using SPSS (Version 26). Prior to inferential testing, assumptions of normality and homogeneity of variance were examined. The Kolmogorov–Smirnov test indicated no significant departures from normality for any of the score

distributions ($p > .05$). Additionally, skewness and kurtosis values for all datasets fell within the acceptable range of −2 to +2, confirming that the normality assumption was tenable. Levene's test for equality of variances revealed no significant differences in variance across rater groups ($p > .05$), satisfying the assumption of homogeneity.

Given these results, independent-samples t-tests were employed to compare mean writing scores assigned by native and non-native raters for each genre. Paired-samples t-tests were conducted within each rater group to examine intra-group differences in scoring argumentative versus descriptive essays. Inter-rater reliability within each rater group was estimated using Cohen's Kappa (κ), which yielded values above .75, indicating substantial agreement and aligning with reliability thresholds commonly adopted in writing assessment research (e.g., Barkaoui, 2010; Johnson & Lim, 2009).

Results

Descriptive Statistics

Table 2 presents the descriptive statistics for scores assigned by native (NS) and non-native (NNS) raters across four analytic criteria—content, organization, vocabulary, and grammar—in both argumentative (ARG) and descriptive (DES) essays. Across all criteria, notable differences emerged between rater groups and genres, suggesting differential sensitivity to linguistic and rhetorical features.

For content, non-native raters assigned higher mean scores than native raters in both genres, particularly for descriptive essays (M = 7.82, SD = 0.87) compared to argumentative writing (M = 7.22, SD = 0.75). Native raters demonstrated a slightly more conservative scoring pattern, with argumentative essays receiving the mean score of 6.49 (SD = 1.10) and descriptive essays 6.64 (SD = 0.92). These differences may suggest that non-native raters may have been more positively disposed toward idea development and descriptive elaboration, whereas native raters exhibited greater variation in their evaluations of content.

Organization scores reflected the most pronounced divergence in argumentative writing: non-native raters awarded substantially higher scores (M = 8.00, SD = 0.92) than native raters (M = 6.25, SD = 0.87). In descriptive writing, however, the gap narrowed, with mean scores of 7.62 (SD = 0.93) for non-native raters and 7.15 (SD = 0.77) for native raters. This pattern may suggest that native raters were more stringent in evaluating the logical structuring of arguments, while non-native raters were comparatively lenient, particularly for texts requiring argumentative coherence.

Vocabulary and grammar ratings exhibited different trends. For vocabulary, native raters tended to give higher scores than non-native raters in both genres, most noticeably in descriptive essays (M = 7.72, SD = 1.01 vs. M = 7.32, SD = 0.98). Grammar scores revealed the opposite: native raters consistently awarded higher ratings (ARG M = 8.36, SD = 0.77; DES M = 8.18, SD = 0.75) compared to non-native raters (ARG M = 7.23, SD = 1.00; DES M = 7.27, SD = 1.01), likely indicating that non-native raters were more critical of grammatical accuracy regardless of genre.

Table 2. *Descriptive Statistics of Native and Nonnative Rater Scores Across Argumentative and Descriptive Genres by Four Scoring Criteria*

Scoring Criteria	Genre	Rater	N	Mean	Std. Deviation	Std. Error Mean
Content	ARG	NS	11	6.4910	1.09545	.33029
		NNS	11	7.2182	.75076	.22636
	DES	NS	11	6.6364	.92442	.27872
		NNS	11	7.8182	.87386	.26348
Organization	ARG	NS	11	6.250	.87386	.26384
		NNS	11	8.000	.92442	.27481
	DES	NS	11	7.1462	.77460	.23355
		NNS	11	7.622	.93420	.28167
Vocabulary	ARG	NS	11	7.3022	.92442	.27872
		NNS	11	6.9455	1.12815	.34015
	DES	NS	11	7.7213	1.00905	.30424
		NNS	11	7.3182	.98165	.29598
Grammar	ARG	NS	11	8.3600	.77460	.23355
		NNS	11	7.23.10	1.000	.30151
	DES	NS	11	8.1812	.75076	.22636
		NNS	11	7.2739	1.00905	.30423

Note. ARG = Argumentative Essays; DES = Descriptive Essays; NS = Native Speakers; NNS = Nonnative Speakers

Confidence interval: 95%.

Comparative Analysis of Raters' Scoring Patterns Across Evaluative Criteria

To examine whether rater background influenced scoring performance across genres and analytic criteria, a series of independent-samples t-tests was conducted (see Table 3). These analyses compared mean scores awarded by native (NS) and non-native (NNS) raters for each criterion within both argumentative and descriptive writing tasks. The objective was to determine the extent to which rater linguistic background contributed to variations in judgments of content, organization, vocabulary, and grammar, thereby providing empirical evidence on potential rating biases or sensitivities in standardized writing assessment contexts.

Table 3. *Independent Samples t-Test Results Comparing NS and NNS Rater Scores for Argumentative and Descriptive Genres*

Scoring Criteria	Genre	Equal Variances	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Content	**ARG**	Assumed	.024	-.81818	.40041
		Not Assumed	.024	-.81818	.40041
	DES	Assumed	.006	-1.18682	.31345
		Not Assumed	.006	-1.18682	.31345
Organization	**ARG**	Assumed	.000	-.45455	.38355
		Not Assumed	.000	-.45455	.38355
	DES	Assumed	.250	1.54545	.36590
		Not Assumed	.250	1.54545	.36590
Vocabulary	**ARG**	Assumed	.300	1.81818	.43976
		Not Assumed	.300	1.81818	.43976
	DES	Assumed	.310	1.98236	.42446
		Not Assumed	.310	1.98236	.42446
Grammar	**ARG**	Assumed	.016	1.000	.38139
		Not Assumed	.016	1.000	.38139
	DES	Assumed	.026	.90909	.37921
		Not Assumed	.026	.90909	.37921

Note. ARG = Argumentative Essays; DES = Descriptive Essays

Argumentative Essays

Independent-samples t-test results revealed significant differences between native (NS) and non-native (NNS) raters in their evaluation of argumentative essay content (p = .024). Specifically, NNS raters assigned higher content scores (M = 7.22) than their NS counterparts (M = 6.49), resulting in a mean difference of −0.82. This suggests that NNS raters were more favorable toward idea development and topical coverage in argumentative writing, whereas NS raters may have applied stricter criteria when evaluating conceptual adequacy and argumentative depth. This pattern aligns with previous findings that rater background can influence how raters perceive and reward rhetorical elaboration (e.g., Barkaoui, 2010).

Differences were also observed in grammar scoring for argumentative essays, where NS raters assigned significantly higher scores than NNS raters (p = .016), with a mean difference of 1.00. This indicates that NNS raters tended to be more critical of grammatical accuracy, a trend that may reflect heightened sensitivity to surface-level linguistic errors among raters with an L2 background. In contrast, no statistically significant differences emerged in organization (p = .300) or vocabulary scores (p = .300) for argumentative writing, suggesting greater alignment between rater groups in their assessment of textual coherence and lexical resources in this genre.

Descriptive Essays

In descriptive writing, the content scores once again reflected significant differences between rater groups (p = .006). As in argumentative writing, NNS raters gave higher scores for content (M = 7.82) than NS raters (M = 6.64), yielding a mean difference of −1.19. This finding

suggests that NNS raters may have been more appreciative of descriptive elaboration and topical detail, possibly valuing explicit content realization over implicit or nuanced thematic development often preferred by NS raters. Organization scores, however, did not significantly differ (p = .250), indicating that both rater groups shared similar perspectives on structural sequencing and overall text organization in descriptive essays.

Regarding language use, grammar scores in descriptive writing also showed a significant difference (p = .026), with NS raters assigning higher ratings (M = 8.18) than NNS raters (M = 7.27), resulting in a mean difference of 0.91. This pattern mirrors findings in argumentative essays, where NNS raters consistently applied stricter criteria to grammatical accuracy, suggesting a potential bias toward formal correctness. Vocabulary ratings did not differ significantly (p = .310), implying converging perceptions between rater groups in evaluating lexical appropriacy and range within descriptive texts. Together, these results highlight that while raters shared common ground in evaluating some linguistic and organizational features, systematic differences persisted in how they judged content richness and grammatical accuracy across genres.

Paired-Samples Analysis

To determine whether rater groups exhibited genre-specific scoring tendencies, paired-samples t-tests were conducted separately for native (NS) and non-native (NNS) raters (see Table 4). This analysis compared scores awarded for argumentative (ARG) and descriptive (DES) essays within each analytic criterion—content, organization, vocabulary, and grammar. The primary purpose was to identify whether raters of different linguistic backgrounds applied distinct evaluative standards when rating different genres, thus revealing potential within-group genre sensitivity in scoring behavior.

Table 4. *Within-Group Comparison of NS and NNS Rater Scores Across Argumentative and Descriptive Genres: Paired Samples t-Test Results*

Scoring Criteria	Genre	Raters	Paired Differences	t	df	Sig. (2-tailed)
Content	ARG.DES	NS	1.45820	.740	10	.476
		NNS	.79490	.000	10	1.000
Organization	ARG.DES	NS	.07399	-2.043	10	.030
		NNS	2.02188	3.135	10	.070
Vocabulary	ARG.DES	NS	.59931	-.841	10	.420
		NNS	.77197	-.582	10	.574
Grammar	ARG.DES	NS	.47766	-.614	10	.553
		NNS	.93245	-.504	10	.6255

Note. ARG = Argumentative Essays; DES = Descriptive Essays; NS = Native Speakers; NNS = Nonnative Speakers

Native Raters (NS)

For NS raters, the paired-samples t-test results indicated no significant differences in content scores between argumentative and descriptive essays, t (10) = 0.74, p = .476, suggesting that NS raters maintained consistent evaluations of content development regardless of genre. Vocabulary (t (10) = −0.84, p = .420) and grammar (t (10) = −0.61, p = .553) scores also did

not differ significantly, indicating that lexical and grammatical evaluations were stable across genres for NS raters.

However, organization scores revealed a statistically significant difference, $t(10) = -2.04$, $p = .030$, indicating that NS raters rated organization higher in descriptive essays compared to argumentative essays. This pattern suggests that NS raters may have perceived descriptive writing as more effectively structured or more straightforward to evaluate in terms of coherence and sequencing than argumentative writing, which demands more complex rhetorical structuring. Overall, NS raters exhibited selective genre sensitivity, with organizational aspects being the only dimension where genre influenced scoring behavior.

Non-Native Raters (NNS)

For NNS raters, no significant differences emerged across genres for any of the four analytic criteria. Content scores (t (10) = 0.00, p = 1.000) were virtually identical between argumentative and descriptive writing, indicating complete consistency in evaluating topical development and idea elaboration. Similarly, no significant differences were observed for organization (t (10) = 3.14, p = .070), vocabulary (t (10) = −0.58, p = .574), or grammar (t (10) = −0.50, p = .626). Although organization approached significance (p = .070), the difference did not meet the conventional threshold, suggesting that NNS raters did not display strong genre-dependent variation in organizational evaluation.

These findings indicate that NNS raters tended to apply more uniform standards across genres compared to their NS counterparts, who exhibited sensitivity to organizational differences between argumentative and descriptive writing. The absence of genre effects among NNS raters may suggest either a stricter adherence to generalized scoring principles or less awareness of genre-specific rhetorical demands. Taken together, these results highlight that rater background not only influences how criteria are weighted across groups but also whether genre differences are recognized and factored into holistic evaluations.

Discussion

This study provides critical insights into the complex interplay between rater linguistic background and genre-specific evaluative practices in standardized writing assessment. The findings substantiate that rater characteristics significantly shape scoring patterns, extending and refining prior research on rater bias and genre sensitivity (e.g., Barkaoui, 2010; Park, 2015). Notably, the differential scoring tendencies observed between native (NS) and non-native (NNS) raters reveal nuanced divergences in how rhetorical and linguistic features are prioritized and interpreted across argumentative (ARG) and descriptive (DES) essay genres.

Consistent with extant literature on rater variability, NNS raters demonstrated a more lenient stance toward content in both genres, awarding significantly higher scores than their NS counterparts. This tendency aligns with the notion that raters with L2 backgrounds may emphasize explicit idea elaboration and topical coverage, potentially reflecting their pedagogical experiences or differing conceptualizations of content adequacy (Barkaoui, 2010). Supporting this interpretation, Winke et al., (2013) utilized eye-tracking methodology to reveal that raters' linguistic backgrounds influence their attentional deployment and processing strategies during rating, with NNS raters focusing more on overt textual features. This cognitive

perspective complements our behavioral findings, suggesting distinct evaluative heuristics tied to linguistic experience.

Conversely, NS raters' stricter evaluation of content—particularly in argumentative essays—likely stems from heightened expectations regarding the depth of conceptual reasoning and coherence, echoing Zhang and Elder (2014) on genre-related cognitive demands. Further validating genre sensitivity, Knoch and Chapelle's (2018) investigations found that raters with strong genre knowledge apply more differentiated scoring criteria, particularly attending to rhetorical conventions in argumentation. This supports our result that NS raters prioritize argumentative organization more conservatively, indicative of their rigorous application of discourse-level criteria.

The pronounced disparity in organization scores for argumentative writing underlines this genre-specific sensitivity. NS raters' more conservative ratings suggest a rigorous application of rhetorical coherence criteria, consistent with theoretical models emphasizing the complexity of argument structure (Bui & Barrot, 2024; Li & Huang, 2022; Ruegg & Sugiyama, 2013). The relative leniency of NNS raters on this criterion may indicate less familiarity with argumentative discourse intricacies or a reliance on generalized scoring frameworks (Lumley, 2005).

Our findings also highlight distinct trends in language-related criteria. NS raters consistently awarded higher grammar scores, whereas vocabulary ratings showed less pronounced group differences. This pattern suggests NS raters have heightened sensitivity to morphosyntactic accuracy, while lexical evaluations converge between rater groups, possibly reflecting shared standards of lexical appropriacy in L2 contexts. Aligning with this, research in second language acquisition and rater cognition (Chiang, 2003; Ghanbari, 2024; Marefat & Heydari, 2016) attributes NNS raters' heightened grammatical criticality to their metalinguistic awareness.

Paired-samples analyses demonstrated NS raters' selective genre sensitivity, particularly in organizational scoring, with a tendency to rate descriptive essays higher. This pattern is consistent with Bouwer et al. (2015), who reported that descriptive genres' straightforward rhetorical structures are perceived as less cognitively demanding, simplifying evaluation. NNS raters' more uniform scoring across genres suggests stable, rubric-driven evaluative standards less influenced by genre demands. This dichotomy corroborates Friginal et al.'s (2014) conceptualization of rating behavior as influenced by both cognitive schemas and cultural-linguistic backgrounds.

The present results also bear significant implications for language assessment practice. The documented rater discrepancies underline the necessity of rigorous rater training protocols emphasizing genre-specific expectations to mitigate subjective bias and enhance inter-rater reliability (Weigle, 2002; Hamp-Lyons & Condon, 2000).

Moreover, the differential weighting of analytic criteria across rater groups suggests assessment frameworks should incorporate explicit guidance balancing surface-level linguistic accuracy with higher-order discourse features, especially in diverse rater populations. This aligns with calls for multi-dimensional construct validity in writing assessment (Bachman & Palmer, 2010).

Finally, the nuanced interplay of rater background and genre sensitivity highlights potential avenues for leveraging automated scoring and artificial intelligence tools. While automated essay scoring (AES) systems show considerable consistency in detecting surface-level errors such as grammar (Attali & Burstein, 2006; Shermis & Burstein, 2013), their limitations in capturing complex rhetorical features reaffirm the indispensable role of skilled human raters, particularly those proficient in genre-specific evaluation. Integrating human judgment with AES in hybrid models may offer a scalable, valid solution balancing precision with interpretive depth.

In conclusion, this study advances understanding of how rater linguistic background interacts with genre demands to shape analytic scoring in L2 writing assessment. By elucidating distinct evaluative priorities and sensitivities of NS and NNS raters across argumentative and descriptive genres, it contributes to ongoing efforts to refine rating validity and fairness. Future research should further explore rater cognition through qualitative and process-oriented methods and investigate training interventions aimed at harmonizing rating standards across diverse rater populations.

Conclusion and Implications

This study provides robust evidence that both rater background and genre type significantly influence writing assessment in standardized English proficiency tests. While AES technologies continue to evolve, the unique interpretive capacity of trained human raters remains indispensable, particularly for capturing genre-specific textual features such as argumentative coherence and descriptive elaboration. The findings underline the need for comprehensive rater training programs that include genre-awareness modules and task-specific calibration to ensure fairer and more reliable assessments across rater populations.

In addition, our results advocate for greater alignment between assessment practices and instructional approaches. Genre-based writing instruction, as demonstrated by Damayanti et al. (2023), and the integration of GenAI tools for mentor text generation (de Oliveira & dos Santos, 2025), can support learners in navigating the complex expectations of tasks like IELTS Task 2 and TOEFL Independent Writing. These innovations promise more equitable outcomes when coupled with informed human judgment in scoring.

Limitations and Future Directions

Although the study was carefully designed to mirror authentic testing conditions and employed dual-mode scoring (holistic and analytic), it was limited by its sample size of raters and genre scope. Expanding the study to include integrated genres, such as TOEFL Integrated Tasks, or other language backgrounds among raters could yield broader generalizations. Furthermore, additional research is warranted on how AES tools and GenAI models might assist raters during calibration or post-scoring validation processes without compromising interpretive accuracy.

Future studies should explore the integration of human-AI hybrid scoring systems that preserve rater sensitivity while leveraging the efficiency of AI. A human-in-the-loop framework, as advocated by Kumar and Boulanger (2020), may enhance both validity and scalability in writing assessment.

Acknowledgements

The authors gratefully acknowledge the participants and contributors whose support was essential for this study. Appreciation is also extended to the editorial board and reviewers for their valuable insights and contributions to this manuscript.

References

Almusharraf, N., & Alotaibi, H. (2023). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology, Knowledge and Learning*, *28*(3), 1015-1031. https://doi.org/10.1007/s10758-022-09592-z

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Bachman, L., & Adrian, P. (2022). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press. We can refer to this work instead of the next. This one is the up-dated version of the Bachman and Palmer 2010.

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.

Bejar, I. I., Li, C., & McCaffrey, D. (2020). Predictive Modeling of Rater Behavior: Implications for Quality Assurance in Essay Scoring. *Applied Measurement in Education*, 33(3), 234-247.

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, *27*(4), 515-535.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279-293. https://doi.org/10.1080/0969594X.2010.526585

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83-100. https://doi.org/10.1177/0265532214542994

Bui, M., & Barrot, J. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Educ. Inf. Technol.*, 30, 2041-2058. https://doi.org/10.1007/s10639-024-12891-w.

Chen, T., & Sun, S. (2025). Evaluating automated evaluation systems for spoken English proficiency: An exploratory comparative study with human raters. *PLOS One*, 20. https://doi.org/10.1371/journal.pone.0320811.

Chan, K. K. Y., Bond, T., & Yan, Z. (2023). Application of an automated essay scoring engine to English writing assessment using many-facet Rasch measurement. *Language Testing*, *40*(1), 61-85. https://doi.org/10.1177/02655322221076025

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System*, 31, 471-484. https://doi.org/10.1016/J.SYSTEM.2003.02.002.

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education*, *31*(3), 241-250. https://doi.org/10.1080/08957347.2018.1464450

Connor-Linton, J. E. F. F. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes*, *14*(1), 99-115. https://doi.org/10.1111/j.1467971X.1995.tb00343.x

Cotos, E. (2019). Understanding the 'Black-Box'of Automated Analysis of Communicative Goals and Rhetorical Strategies in Academic Discourse. *Center for Communication Excellence Proceedings and Presentations. 1*. https://lib.dr.iastate.edu/communicationexcellence_conf/1

Cullen, P., French, A., & Jakeman, V. (2014). *The official Cambridge guide to IELTS for academic & general training*. Cambridge: Cambridge University Press.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1), 67-96.

Damayanti, I. L., Hamied, F. A., Kartika-Ningsih, H., & Dharma, N. S. (2023). Building knowledge about language for teaching IELTS writing tasks: A genre-based approach. *Studies in English Language and Education*, *10*(2), 756-776. https://doi.org/10.24815/siele.v10i2.26957

de Oliveira, L. C., & dos Santos, A. E. (2025). Using AI-text generated mentor texts for genre-based pedagogy in second language writing. *Journal of Second Language Writing*, *67*, 101184. https://doi.org/10.1016/j.jslw.2025.101184

Friginal, E., Li, M., & Weigle, S. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing*, 23, 1-16. https://doi.org/10.1016/J.JSLW.2013.10.001.

Ghanbari, N. (2024). Academic writing assessment in the Iranian expanding circle context: any trust in local criteria? *Asian Englishes*, 27, 265 - 285. https://doi.org/10.1080/13488678.2024.2386481.

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory and research.* Hampton Press.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the Dynamics of Second Language Writing* (pp. 162–189). Cambridge University Press.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *Peer Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of second language writing*, *16*(3), 148-164. https://doi.org/10.1016/j.jslw.2007.07.005

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505. https://doi.org/10.1177/0265532209340186

Knoch, J. (2009). Optimizing tunnel FET Performance-Impact of device structure, transistor dimensions and choice of material. In *2009 International Symposium on VLSI Technology, Systems, and Applications* (pp. 45-46). IEEE. https://doi.org/10.1109/VTSA.2009.5159285

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477-499.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, *26*(1), 81-112. https://doi.org/10.2307/3587370

Koraishi, O. (2024). The Intersection of AI and Language Assessment: A Study on the Reliability of ChatGPT in Grading IELTS Writing Task 2. *Language Teaching Research Quarterly*, *43*, 22-42. https://doi.org/10.32038/ltrq.2024.43.02

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in Education* (Vol. 5, p. 572367). Frontiers Media SA.

Lee, I. (2017). *Classroom Writing Assessment and Feedback in L2 School Contexts*. Singapore: Springer Singapore.

Li, J., & Huang, J. (2022). The impact of essay organization and overall quality on the holistic scoring of EFL writing: Perspectives from classroom English teachers and national writing raters. *Assessing Writing*. https://doi.org/10.1016/j.asw.2021.100604.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560. https://doi.org/10.1177/0265532211406422

Lumley, T. (2005). *Assessing Second Language Writing: The Rater's Perspective* (Vol. 3). P. Lang.

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

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing*, 27, 24-36. https://doi.org/10.1016/J.ASW.2015.10.001.

Park, S. K. (2015). The Interplay of Task, Rating Scale, and Rater Background in the Assessment of Korean EFL Students' Writing. *English Teaching*, *70*(2).

Shi, L. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), 303-325.https://doi.org/10.1177/026553220101800303

Ramezani, A., Bijani, H., & Oroji, M. R. (2025). Comparative Analysis of AI vs. Human Feedback Effects on IELTS Candidates' Writing Performance. *Journal of Foreign Language Teaching and Translation Studies*, *10*(1), 17-40. https://doi.org/10.22034/efl.2025.493559.1334

Ruegg, R., & Sugiyama, Y. (2013). Organization of ideas in writing: what are raters sensitive to? *Language Testing in Asia*, 3, 1-13. https://doi.org/10.1186/2229-0443-3-8.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493.

Shermis, M. D., & Burstein, J. (2013). *Handbook of Automated Essay Evaluation*. Routledge.

Stephen, T. C., Gierl, M. C., & King, S. (2021). Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation. *Nurse Education in Practice*, *54*, 103085. https://doi.org/10.1016/j.nepr.2021.103085

Taylor, L. (2003). The Cambridge approach to speaking assessment. *Research Notes*, 13, 2-4.

Taylor, L. B., Taylor, L., & Weir, C. J. (Eds.). (2012). *IELTS collected papers 2: Research in Reading and Listening Assessment* (Vol. 2). Cambridge University Press.

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education*, *12*(1), 20-32. https://doi.org/10.21449/ijate.1517994

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

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231-252.

Xu, W., Mahmud, R., & Hoo, W. L. (2024). A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? *IEEE Access*, *12*, 77639-77657. https://doi.org/10.1109/ACCESS.2024.3399163

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306-325. https://doi.org/10.1080/0969594X.2013.845547

Zhang, L., & Liu, H. (2021). Genre effect on L2 syntactic complexity and holistic rating for writing quality of intermediate EFL learners. *Chinese Journal of Applied Linguistics*, *44*(4), 451-469. https://doi.org/10.1007/s11145-007-9107-5

Zhang, X., & Lu, X. (2022). Revisiting the predictive power of traditional vs. fine-grained syntactic complexity indices for L2 writing quality: The case of two genres. *Assessing Writing*, *51*, 100597. https://doi.org/10.1016/j.asw.2021.100597

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, *120*, 118-132. https://doi.org/10.1016/j.knosys.2017.01.006