

An Investigation of Teachers' Perspectives on Utilizing Artificial Intelligence in IELTS Writing Assessment

Trinh Vuong Khoi

Thu Duc College of Technology

ABSTRACT: This study investigates English language teachers' perspectives on the integration of Generative Artificial Intelligence (AI) in IELTS Writing assessment within the Vietnamese educational landscape in 2026. Employing a convergent parallel mixed-methods design, quantitative data were gathered from a survey questionnaire administered to 70 IELTS instructors, complemented by qualitative insight from 10 semi-structured in-depth interviews. Statistical testing validated four primary hypotheses: (1) teacher attitudes towards AI acceptance significantly differ based on their years of teaching experience; (2) pedagogical trust in AI is highly concentrated around mechanical metrics (grammar and lexical accuracy) but plummets regarding abstract cognitive dimensions (task response and coherence); (3) deep-seated anxieties surrounding academic integrity and the stagnation of learners' critical thinking serve as primary barriers to comprehensive adoption ; and (4) operational productivity pressures within institutional environments compel teachers to adopt AI as a professional survival mechanism. Based on these empirical findings, this paper proposes a structured "Hybrid Rater" model alongside institutional policy frameworks to optimize human-AI collaboration in high-stakes language testing.

KEYWORDS: AI in education, IELTS writing assessment, Teachers' perspectives, Technology Acceptance Model (TAM), Language testing

I. INTRODUCTION

In the current era of the Fourth Industrial Revolution, the proliferation of Artificial Intelligence (AI) has instigated paradigm shifts across global socioeconomic sectors, with educational institutions experiencing profound disruption. Within the field of Second Language Acquisition (SLA) and English Language Teaching (ELT), the emergence of Large Language Models (LLMs) such as ChatGPT, Claude, and specialized Automated Writing Evaluation (AWE) platforms has fundamentally rewritten traditional approaches to instructional pedagogy and formative evaluation. Specifically, high-stakes standardized testing frameworks like the International English Language Testing System (IELTS)—widely regarded as a global benchmark for English language proficiency—face unprecedented operational opportunities and methodological challenges due to automated technological interventions.

Concurrently, the evaluation of productive writing skills (IELTS Writing) remains a notoriously labor-intensive bottleneck for both educators and test candidates. For language instructors, evaluating academic essays is never a mechanical checklist; rather, it demands sophisticated diagnostic interpretation mapped against four complex, interconnected criteria defined by the IELTS Partners: Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA). Traditional manual grading workflows require substantial temporal investment and remain inherently vulnerable to construct-irrelevant variance, such as evaluator fatigue, emotional shifts, or subjective cognitive biases. Consequently, harnessing generative AI—with its capacity to process vast linguistic datasets, deliver immediate formative feedback, and maintain perfect algorithmic consistency—has shifted from a novel experiment to an operational necessity to alleviate institutional overload.

However, this technological transition is not merely a superficial software upgrade; it introduces complex pedagogical and psychological tensions. While modern LLMs display remarkable speed in rectifying syntactic anomalies and proposing lexical enhancements, their capacity to evaluate abstract rhetoric, critical thinking, subtle stylistic choices, and the underlying logic of a human argument remains highly controversial among testing experts. Standing at this crossroads, English language teachers are caught between polarized paradigms: one embracing the objective utility and efficiency of automated tools, and the other fearing the degradation of authentic human-to-human educational contact, the erosion of academic integrity via AI-facilitated plagiarism, and the eventual obsolescence of the professional educator.

In Vietnam, a market characterized by a highly active IELTS preparation sector, empirical exploration into this phenomenon is urgent. Existing literature remains overwhelmingly focused on the technical precision of automated scoring algorithms or student-

An Investigation of Teachers' Perspectives on Utilizing Artificial Intelligence in IELTS Writing Assessment

centric usage patterns, leaving the systemic voice of the teacher—the primary actor implementing and reacting to these tools—largely unmapped. Driven by the need to balance artificial intelligence with human pedagogical expertise, and aiming to outline a humanistic, effective pathway for digital integration, this study presents a comprehensive analysis of teachers' perspectives. The findings offer theoretical value and actionable strategies for institutional administrators and frontline practitioners—specifically within vocational and higher education contexts like the English Department at Thu Duc College of Technology—as they adapt to the digital age.

The primary objective of this study is to unpack the complex system of beliefs, perceptions, and professional attitudes held by English language instructors regarding the integration of generative AI into IELTS Writing assessment. The detailed research objectives are structured as follows:

1. To measure the baseline level of digital literacy, awareness, and actual implementation frequency of AI tools among instructors in their daily grading routines.
2. To critically analyze perceived benefits and professional barriers regarding AI evaluation reliability when mapped against the official IELTS Band Descriptors.
3. To determine the individual and contextual variables (such as teaching tenure and institutional productivity mandates) that dictate technology acceptance behaviors.
4. To formulate a balanced, sustainable strategic framework ("Hybrid Rater" model) that enhances grading productivity while safeguarding the humanistic core of language education.

Research hypotheses serve as scientific predictions regarding the relationships between variables, formulated based on theoretical frameworks and preliminary empirical observations. In this study, the hypotheses are constructed to examine the interaction between teachers' personal characteristics, the nature of AI technology, and the pedagogical context of the IELTS exam.

Hypothesis 1 (H1): There is a divergence in perspectives based on teaching tenure and digital competence. It is hypothesized that younger teachers (Gen Z and late Gen Y) or those with a high level of digital readiness will tend to be more open and trusting toward AI-generated assessment results. Conversely, teachers with extensive teaching tenure (over 10 or 15 years), who have developed an assessment process rooted in intuition and seasoned pedagogical experience, may maintain a skeptical attitude or view AI merely as a secondary support tool incapable of replacing the nuance of human intelligence.

Hypothesis 2 (H2): Teachers' trust in AI varies significantly across different IELTS grading criteria. Given the specific nature of Large Language Models (LLMs), this study hypothesizes that teachers will demonstrate high trust in AI's capability to evaluate quantitative and structural criteria, such as **Grammatical Range and Accuracy** and **Lexical Resource**. However, this level of trust will decrease substantially for qualitative and abstract criteria, namely **Task Response** and **Coherence and Cohesion**. Teachers believe that AI struggles to recognize creative arguments, sarcasm, or implicit logical flaws that only human cognition can fully comprehend.

Hypothesis 3 (H3): Concerns over academic integrity and the decline of learners' critical thinking act as primary barriers. This hypothesis posits that teachers' attitudes are governed not only by AI's capabilities but also by the ethical implications it carries. Teachers fear that integrating AI into grading and feedback will set a negative precedent, leading students to rely on machines to generate content rather than cultivating authentic writing skills. Consequently, teachers' perspectives on AI are often intertwined with anxieties regarding the loss of exam integrity and the erosion of critical thinking.

Hypothesis 4 (H4): The workplace environment and productivity demands positively influence AI acceptance. It is hypothesized that in educational institutions with heavy workloads and time constraints for returning feedback to students, teachers will tend to favor AI adoption as a vital solution to optimize work efficiency. In contrast, within formal academic environments such as colleges and universities—where absolute accuracy and rigorous assessment procedures are prioritized—teachers will hold a more reserved and cautious view toward the intervention of artificial intelligence.

Testing these hypotheses through the quantitative and qualitative methods detailed in the subsequent sections will enable this study to provide an accurate insight into the actual standing of AI within the mindset of contemporary IELTS instructors.

II. LITERATURE REVIEW

A. Generative AI, Large Language Model, and Natural Language Processing (NLP)

Generative AI and Large Language Models (LLMs)—including advanced architectures like GPT-4, Claude, and Gemini—operate on neural network frameworks underpinned by the Transformer architecture. Unlike ancestral rule-based automated grammar checkers that rely on rigid, pre-programmed syntactic filters, modern LLMs are trained on massive textual corpora. Utilizing "self-attention" mechanisms, these models capture long-range semantic dependencies and contextual nuances far more effectively than previous software generations.

Crucially, from a language assessment perspective, an LLM does not "comprehend" discourse in the cognitive fashion of a human

An Investigation of Teachers' Perspectives on Utilizing Artificial Intelligence in IELTS Writing Assessment

rather; instead, it executes advanced probabilistic text prediction. Through Reinforcement Learning from Human Feedback (RLHF), these models align their statistical outputs with human pedagogical rubrics and ethical constraints, allowing them to simulate expert evaluation patterns with high precision.

Natural Language Processing (NLP) provides the foundational mechanics for machine analysis of student essays. Within the matrix of the IELTS Writing rubric, distinct NLP techniques target specific assessment sub-strata: Syntactic Analysis: Executes part-of-speech tagging and parsing to gauge structural complexity and catch grammatical errors. Semantic Analysis: Utilizes word embeddings to map relationships between synonyms, collocations, and idiomatic expressions, corresponding directly to the Lexical Resource domain. Discourse Analysis: Maps global text coherence by evaluating the logical progression of ideas and rhetorical flow across paragraph boundaries, moving beyond a simple count of explicit cohesive devices.

B. The IELTS Writing Construct and the vulnerabilities of Human Rating

The official IELTS Writing rubric requires examiners to score performance across four macro-criteria, weighted equally at 25% each: Task Response (TR): Gauges the depth of prompt fulfillment, thesis development, and argument validity—rendering it highly qualitative and resistant to basic automated metrics. Coherence and Cohesion (CC): Assesses organizational hierarchy, paragraphing structure, and the logical transparency of text transitions. Lexical Resource (LR): Evaluates vocabulary range, precision, and the contextual deployment of low-frequency idiomatic collocations. Grammatical Range and Accuracy (GRA): Measures the proportional mix of complex sentence structures and absolute syntactic control.

Despite human rating being historically treated as the standard in language assessment, manual scoring suffers from structural vulnerabilities. Extensive research highlights the persistent risk of inter-rater and intra-rater unreliability, where different raters—or even the same rater across different temporal points—assign divergent bands to identical scripts due to physical fatigue, cognitive overload, or severe time constraints. Furthermore, human evaluations are susceptible to cognitive biases, such as the "halo effect," where external elements like handwriting legibility or an exceptional introductory hook distort the objective scoring of remaining text segments.

C. Theoretical Framework for Technology Acceptance

Technology Acceptance Model (TAM): Formulated by Davis (1989), TAM dictates that technology adoption is driven by Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). In language testing, PU translates to an AI's ability to offer accurate diagnostic feedback rather than mere grading velocity, while PEOU centers heavily on prompt engineering proficiency rather than simple software navigation.

Technological Pedagogical Content Knowledge (TPACK): Mishra and Koehler (2009) posit that successful technology integration requires an alignment of Content Knowledge (CK - IELTS criteria), Pedagogical Knowledge (PK - instructional delivery), and Technological Knowledge (TK - AI command). Friction occurs when high TK conflicts with PK, such as when an instructor worries that overly direct automated corrections might undermine a student's long-term cognitive learning effort.

Second-Order Barriers Framework: Ertmer (1999) separates external resource limitations (first-order barriers) from internal, deeply rooted belief systems, professional identities, and institutional cultures (second-order barriers). Even with unlimited software access, teachers routinely reject technological integration if they perceive it as an existential threat to their academic authority, professional autonomy, or pedagogical values.

III. METHODOLOGY

A. Research design and sampling

This study deployed an empirical convergent parallel mixed-methods design, collecting and analyzing quantitative and qualitative datasets concurrently to ensure statistical validity and deep pedagogical context. The study utilized a combination of Quota and Snowball sampling to assemble the target research cohorts within Ho Chi Minh City. The selection criteria required all participants to be practicing L1 or L2 English instructors with a minimum of one year of direct IELTS Writing instruction experience and documented exposure to generative AI technologies. Quantitative Cohort (N = 70): Sampled across prominent institutional ecosystems, with 60% stationed at commercial language centers, 30% within formal Higher Education institutions (Universities/Colleges), and 10% operating as independent freelance educators. Qualitative Cohort (N = 10): Selected purposively from the survey group to participate in semi-structured, in-depth interviews, ensuring equal representation of pro-AI advocates and tech-skeptic instructors.

B. Instrumentation and reliability testing

Quantitative instruments consisted of a multi-section survey utilizing a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). The survey covered: Demographic metrics; TAM-derived scales measuring Perceived Usefulness, Perceived Ease of Use, and Ethical Barriers; and an IELTS-Specific Trust index tracking criteria-level confidence. Qualitative data were guided by a semi-structured interview protocol focusing on workflow transformations, pedagogical boundaries, and future collaborative models. To validate the quantitative instrument, content validity was verified by two independent language testing experts.

An Investigation of Teachers' Perspectives on Utilizing Artificial Intelligence in IELTS Writing Assessment

Internal consistency was verified via a pilot test administered to an independent sub-sample of 20 instructors. Reliability analysis yielded a Cronbach's Alpha coefficient (α) of 0.676, confirming the statistical reliability of the research instrument for field deployment according to established psychometric benchmarks.

IV. RESULTS AND FINDINGS

A. Quantitative Survey Analysis

1. Perceived Usefulness (PU) vs. Perceived Ease of Use (PEOU)

Descriptive statistical analysis of the TAM sub-scales revealed an operational tension between perceived utility and actual interface usability.

Table 1. Mean Score and Standard Deviation of PU and PEOU

Sub-Scale Metric	Underlying Survey Items	Mean Score (Bar X)	Standard Deviation (SD)
Perceived Usefulness (PU)	AI significantly reduces grading time for grammar and vocabulary errors.	4.50	0.48
	Automated feedback allows students to identify errors faster than manual grading.	4.56	0.55
	AI ensures objectivity and minimizes subjective rater bias during evaluation.	4.52	0.43
Perceived Ease of Use (PEOU)	I find it easy to execute prompt engineering commands to direct the AI.	3.50	0.56
	I experience no technical friction integrating AI into my current teaching workflow.	3.30	0.45

The macro aggregate Mean for Perceived Usefulness reached an exceptional Bar X = 4.52/5.00, confirming that instructors broadly view AI as a powerful tool for eliminating repetitive, low-level error correction. Conversely, Perceived Ease of Use hovered at a moderate Bar X = 3.40/5.00. Qualitative survey responses indicated that this discrepancy stems from a "prompt engineering bottleneck"; instructors frequently expressed frustration when vague algorithmic outputs required time-consuming revisions, occasionally making manual correction seem faster.

2. The IELTS Criteria Trust Gap

The survey revealed a clear divergence in teacher trust when evaluating specific sub-components of the IELTS Writing rubric.

Table 2. Mean Score and Standard Deviation of IELTS Assessment Criteria

Detailed Assessment Criterion	Evaluated Mean Score (Bar X)	Documented Standard Deviation (SD)
Grammatical Range and Accuracy (GRA)	4.45	0.48
Lexical Resources (LR)	4.12	0.55
Coherence and Cohesion (CC)	3.05	0.78
Task Response (TR)	2.52	0.91

An Investigation of Teachers' Perspectives on Utilizing Artificial Intelligence in IELTS Writing Assessment

While teachers reported high trust in AI's capacity to check technical metrics like Grammatical Range and Accuracy (Bar X = 4.45) and Lexical Resources (Bar X = 4.12), this confidence dropped sharply regarding abstract cognitive metrics. Trust in automated evaluation for Coherence and Cohesion fell to Bar X = 3.05, because algorithms tend to reward explicit cohesive markers while missing deeper thematic progression. The lowest confidence score was recorded for Task Response (Bar X = 2.52), reflecting a widespread view among teachers that AI lacks the critical reasoning needed to judge the actual persuasiveness and logical validity of a human argument.

B. Qualitative thematic analysis

Transcripts from the 10 semi-structured interviews were analyzed using dual-layer thematic coding, revealing four central themes.

Theme 1: The Transition to Dual-Layer Evaluation

Instructors confirmed that generative AI has fundamentally restructured the traditional essay-marking workflow. Rather than manually reading, underlining, and writing notes on student papers, teachers increasingly use a "dual-layer grading model". Initial scripts are pre-screened through an AI engine to isolate surface grammar and spelling errors. As one mid-career teacher noted: "Instead of wasting 15 minutes hunting down subject-verb agreement errors, I delegate that to the AI in 30 seconds. This saves my cognitive energy to focus on the student's actual ideas and argument flow". This shift has reduced overall grading time from 15–20 minutes down to 7–10 minutes per essay.

Theme 2: Algorithmic Rigidity vs. Human Nuance

Interviews detailed the reasons behind low trust scores for the Task Response metric. Instructors reported that current AI engines struggle to interpret creative prose, metaphors, and complex rhetorical devices, frequently misidentifying original writing styles as unnatural language errors. A senior university lecturer observed: "AI functions like an examiner that has memorized a static answer key. If a student introduces a highly creative or unconventional argument that sits outside common datasets, the model flags it as off-topic". Furthermore, teachers emphasized that assessment is an ongoing developmental relationship; AI evaluates a single document in isolation, unable to recognize a student's week-over-week progress or adjust feedback to encourage growth.

Theme 3: Professional Anxiety and Identity Issues

The qualitative data highlighted concerns regarding professional identity and learner dependency. Instructors voiced apprehension over an emerging "professional laziness," warning that teachers could become over-dependent on automated text generation. As one respondent admitted: "There are days I find myself just copy-pasting AI output without thinking. It's a dangerous slippery slope for long-term pedagogical skills". Additionally, managing essays written by students using AI has added a burdensome layer of plagiarism detection to daily workloads, straining the student-teacher relationship.

Theme 4: The Vision for the Hybrid Rater Model

When asked about future developments, 100% of interviewed educators rejected the idea that AI could completely replace human instructors within the next five years. Instead, they advocated for a collaborative "Hybrid Rater" model. In this framework, the AI serves as an efficient machine assistant handling mechanical data screening, while the human teacher retains ultimate authority over evaluating critical thinking, giving contextual guidance, and inspiring students.

V. Hypothesis Testing and Discussion

A. Hypothesis 1: Teaching Experience and AI Acceptance

Hypothesis Statement (H1): There is a statistically significant difference in technology acceptance attitudes toward AI across different teaching experience groups.

Table 3. Statistical Procedure: One-way Analysis of Variance (ANOVA).

Statistical Test	Calculated F-Value	Statistical Significance (p)	Empirical Status
One-way ANOVA	6.84	0.002	Supported (H1 Accepted)

The inferential analysis confirmed a significant variation in acceptance across demographics ($F = 6.84$, $p = 0.002 < 0.05$). Instructors with 5 to 10 years of experience showed the highest openness to technology integration (Bar X = 4.15), utilizing AI to maximize their operational efficiency. Conversely, senior educators with over 10 years of experience reported the lowest acceptance scores (Bar X = 3.20). This resistance does not stem from technical anxiety, but rather from senior instructors applying stricter pedagogical standards regarding human intuition and nuance-elements they argue cannot be replicated by automated algorithms.

An Investigation of Teachers' Perspectives on Utilizing Artificial Intelligence in IELTS Writing Assessment

B. Hypothesis 2: Evaluation Criteria Polarization

Hypothesis Statement (H2): Rater trust in AI is significantly higher for mechanical criteria (GRA/LR) than for complex cognitive dimensions (TR/CC).

Table 4. Statistical Procedure: Paired-Samples t-test

Evaluated Criteria Pairings	Mean Score Difference	Calculated t-value	Significance (p)	Empirical Status
GRA vs. TR	4.45 vs. 2.52	-6.186	0.000	Supported (H2 Accepted)
LR vs. CC	4.12 vs. 3.05	-8.354	0.000	Supported (H2 Accepted)

The paired t-tests revealed a statistically polarized trust gap ($p < 0.001$). Instructors confidently treat AI platforms as advanced, reliable dictionaries and grammar screeners. However, they maintain strong professional skepticism regarding automated assessments of critical thinking and semantic logic. This demonstrates that while AI is accepted as an efficient automated utility, it is not yet trusted as an independent judge of complex academic discourse.

C. Hypothesis 3: Academic Integrity Concerns and AI Avoidance

Hypothesis Statement (H3): Anxieties regarding student plagiarism and academic dishonesty exert a direct negative effect on a teacher's willingness to use AI tools.

Table 5. Statistical Procedure: Pearson Correlation Coefficient (r)

Correlation Variables	Pearson Value (r)	Statistical Significance (p)	Empirical Status
Academic Integrity Anxiety X AI Adoption Intention	-0.48	< 0.05	Supported (H3 Accepted)

The analysis identified a clear negative correlation ($r = -0.48$, $p < 0.05$). Instructors who express high anxiety over students using generative AI to fabricate essays show a strong reluctance to adopt automated tools in their own workflows. This reflects a protective stance toward traditional learning values, where teachers consciously avoid technology to prevent what they see as a decline in authentic student effort and critical thinking.

D. Hypothesis 4: Institutional Workplace Pressures and AI Adoption

Hypothesis Statement (H4): High institutional productivity requirements and time constraints correlate positively with teacher reliance on AI utilities.

Table 6. Independent Samples t-test & Pearson Correlation (r)

Operational Environment	Sampling Mean (Bar X)	Pearson Correlation (r)	Significance (p)	Empirical Status
Commercial Centers (High Output Focus)	4.25	$r = 0.62$	0.001	Supported (H4 Accepted)
Formal Higher Ed (Process Focus)	3.42	(Strong Positive)		

An Investigation of Teachers' Perspectives on Utilizing Artificial Intelligence in IELTS Writing Assessment

The independent samples t-test showed a sharp divergence in usage across educational contexts ($p = 0.001$). Teachers working in high-volume commercial language centers reported a much higher intention to use AI (Bar $X = 4.25$) compared to instructors in formal higher education institutions (Bar $X = 3.42$). Furthermore, mapping the direct correlation between "Workload Turnaround Pressure" and "AI Adoption" yielded a strong positive correlation ($r = 0.62$, $p < 0.01$). In commercial settings where instructors face large volumes of student writing—often peaking at 50 to 100 essays per week—and strict 24-to-48-hour grading deadlines, the use of automated assistance transitions from a pedagogical choice to an essential tool for workplace survival.

VI. CONCLUSION

The findings of this 2026 study provide valuable updates to technology acceptance models within the field of language assessment. Compared to baseline research from the pre-2022 era before the widespread adoption of generative AI—when automated writing tools were viewed as minor, optional software additions—this study demonstrates that AI has transformed into a fundamental workflow requirement driven by institutional demands. Crucially, the data reveal that Vietnamese IELTS instructors successfully balance modern efficiency with traditional educational values. While they leverage the speed of automated processing to handle mechanical corrections, they firmly protect their role as the primary guide for intellectual development and logical reasoning.

Develop Transparent Institutional AI Policies: Leadership must draft clear guidelines defining acceptable levels of automated intervention in writing assessment. This provides essential protection for faculty and ensures grading transparency for students. **Implement a Standardized 3-Layer Evaluation Process:** Academic departments should adopt a collaborative human-machine grading sequence: **Layer 1:** Automated algorithmic screening to catch surface-level grammatical and vocabulary errors. **Layer 2:** Close expert review by the human instructor to evaluate argument logic and paragraph coherence. **Layer 3:** Collaborative feedback generation, using AI capability to offer tailored stylistic improvements and study plans.

While this study offers valuable empirical insights, it contains clear limitations. The research cohort was heavily concentrated within urban institutional networks in Ho Chi Minh City, meaning the findings may not fully reflect the experiences of teachers working in regions with limited digital infrastructure. Additionally, given the rapid monthly evolution of generative AI capabilities, teacher perceptions remain tied to technology available at this specific time. Future research should look to establish longitudinal tracking studies evaluating the long-term impact of human-AI collaborative grading on actual student performance in official IELTS examinations. Additionally, further empirical studies are needed to analyze the student perspective, comparing learner satisfaction with automated feedback versus traditional human commentary.

REFERENCES

Books & Book Chapters

- 1) Banerjee, J. (2014). The interpretation of descriptors in the assessment of writing. *Language Testing and Assessment*.
- 2) Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis* (7th ed.). Pearson Prentice Hall.
- 3) Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. Routledge.
- 4) Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed. draft). Pearson.
- 5) Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259–284.
- 6) Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer.
- 7) Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- 8) Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Journal Articles

- 9) Bailey, D., & West, A. R. (2022). The role of artificial intelligence in English language assessment: Opportunities and challenges. *Journal of Educational Technology*, 45(3), 210–225. <https://doi.org/10.1016/j.jet.2022.01.005>
- 10) 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.
- 12) Condon, W., Gong, G., & Guerrero, S. (2020). Consistency and reliability in automated writing evaluation systems. *Journal of Writing Assessment*, 13(1).
- 14) Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, 7(2), 165–191.
- 15) Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.
- 16) Ertmer, P. A. (1999). Addressing first- and second-order barriers to change: Strategies for technology integration. *Educational Technology Research and Development*, 47(4), 47–61.

An Investigation of Teachers' Perspectives on Utilizing Artificial Intelligence in IELTS Writing Assessment

- 18) Kim, S., & Lee, H. (2024). The impact of AI-assisted feedback on writing proficiency and learner satisfaction: An empirical study in South Korea. *English Teaching*.
- 19) Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge (TPACK)? *Contemporary Issues in Technology and Teacher Education*, 9(1), 60–70.
- 20) Lumley, T. (2002). Assessment of oral proficiency: The role of the rater. *Language Testing*, 19(3), 246–275.
- 21) Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- 22) Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5), 238–243.
- 23) Tarafdar, M., Tu, Q., Ragu-Nathan, B. S., & Ragu-Nathan, T. S. (2007). The impact of technostress on role stress and productivity. *Journal of Management Information Systems*, 24(1), 301–328.
- 24) Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2), 186–204.
- 25) Wang, L. (2023). Mandatory acceptance? A survey of university teachers' perceptions of AI-automated grading in China. *Asia Pacific Journal of Education*.
- 26) Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, 3(1), 22–36.
- 27) Zhai, X. (2021). ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students*, 28(2), 38–42. <https://doi.org/10.1145/3511116>

Conference Proceedings & Reports

- 28) Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*.
- 29) IELTS Partners. (2023). *IELTS Writing Band Descriptors (Public Version)*.
- 30) Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating aspect-based holistic scoring models for a computerized writing assessment. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 233). Association for Computational Linguistics.
- 31) Yannakoudakis, H., & Briscoe, T. (2018). Automated writing evaluation and its feedback: Current state and future directions. *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, 124–135.