E-ISSN 2410-8693
ISSN 2410-3918

Academic Journal of Business, Administration Law and Social Sciences

Vol. 11 No. 3
November, 2025

INTERNATIONAL INSTITUTE FOR PRIVATE COMMERCIAL AND COMPETITION LAW (IIPCCL - AUSTRIA)

Research Article

AI in language learning assessment: Examining the limitations of online testing and the reliability of teacher judgement

Lorena Kapxhiu

DOI: https://doi.org/10.2478/ajbals-2025-0028

Abstract

The rapid integration of AI in language assessment has provoked heated discussions concerning validity, reliability, and academic integrity of remote testing used for ESL/EFL assessment. While the rapid integration of AI in assessment tools improves accessibility, scalability, and efficiency of resources and processes, the challenges of ensuring equitable testing conditions, preventing academic misconduct, and measuring language proficiency in high-stakes settings raise crucial questions. This is a comparison study between online technology-driven AI-based assessment versus teacher-mediated assessment in a British international school in Abu Dhabi. About 60 ESL students contributed to a 3-month study, with AI-administered tests used alongside traditional, in-person teacher assessments. Mixed-method analysis, drawing on quantitative scoring data, teacher observations, and student feedback, reported that AI-based assessments (i.e., CAT4, GL) often yielded inflated scores and were inconsistent at assessing productive and contextual language skills. Despite recent developments in policy such as Cambridge's forthcoming hybrid model of assessment, one that reflects a reliance on the digitalization of testing, the findings indicate that there should still be a human-in-the-loop to support validity, fairness, and the integrity of second-language assessment.

Keywords: AI-based language assessment, Teacher judgement reliability, academic integrity, assessment validity, Blended Learning Models.

1. Introduction

The COVID-19 pandemic disrupted global education, exposing weaknesses in remote assessment and AI limitations. The education system adopted teacher-assessed grades (TAGs) rather than online writing exams, acknowledging that AI lacked the nuance to fairly evaluate extended responses. Though digital exams are planned for 2026, writing components remain excluded due to concerns around algorithmic interpretation of tone, coherence, and creativity. (Cambridge Assessment, 2021; Williamson & Eynon, 2020). Throughout the pandemic, the Cambridge Assessment Directorate observed that moving extended writing assessments online carried unique difficulties. Tasks demanding sustained argumentation and nuanced register simply invite judgement

at levels AI-based scoring still cannot match, even as algorithms evolve. Whereas a multiple-choice question can be scored with a key in milliseconds, extended-essay grading still depends on human calibrations of creativity, coherence, register, and situational appropriateness criteria that matter in high-stakes qualifications such as IGCSE English (Perelman, 2012; Shermis & Burstein, 2013).

The 2026 decision to migrate selected IGCSE, AS, and A-level examinations into a digital format, while deliberately holding back writing-heavy papers, reinforces how both technology and pedagogy remain aligned against the use of AI. Research consistently raises doubts about AI's grasp of discourse-level meaning or rhetorical effect (Hao & Weninger, 2019; Kukulska-Hulme & Lee, 2020), reinforcing the apprehension that led to this policy. The main limitation lies in how current AI systems are designed. Their operation hinges on pattern detection and surface-level grammar repair, yet they fall short in assessing tone, voice, and originality—dimensions that the IGCSE English as a Second Language markers explicitly weigh (Cambridge Assessment International Education, 2023). When rubric descriptors drift into the vague—phrases such as "generally relevant" or "sufficiently developed"—the assessment devolves to a subjective arena where the current algorithms falter (Sadler, 1989; Weigle, 2002).

To navigate these limitations, the last few years have witnessed the rise of AI-enhanced environments like Literatu, which merge automated commentary with assessment instruments tightly aligned to the curriculum. Literatu is especially distinctive in its teacher-editable app, permitting instructors to tailor both rubric-driven scoring and the underlying feedback to the nuances of their specific classroom context. This study responds to an unmet need in existing research in the literature by investigating how AI, and Literatu in particular, can help foster equitable writing assessment in ESL contexts. It asks three guiding questions:

1. Why did Cambridge choose to depend on teacher-assessed grades during the pandemic instead of moving to online writing assessments?
2. How does the decision to drop writing from the online exams in 2026 expose present AI shortcomings?
3. To what degree can Literatu evaluate writing in line with IGCSE rubrics while still necessitating human moderation?

Together, these questions navigate the complex interplay between AI capability, assessment dependability, and ethically sound educational practice. This study employed a mixed-methods research design involving 60 ESL learners attending a culturally heterogeneous British international school in Abu Dhabi. Across a five-month intervention, participants progressed through three feedback phases: initial teacher commentary, AI-generated feedback provided through the Literatu platform, and a concluding self- and peer-assessment cycle incorporating the "I Am the Examiner" role-play technique. This sequence allowed for a comparative analysis of the merits and drawbacks of each feedback modality while fostering learners' metacognitive awareness and familiarity with assessment rubrics (Black & Wiliam, 2009; Nicol & Macfarlane-Dick, 2006).

The classroom intervention additionally examined how students assimilate assessment

criteria when exposed to feedback from multiple channels. Portfolios were compiled to document variations in grammatical accuracy, coherence of structure, and depth of argument. These collections were scored against the official Cambridge IGCSE writing rubric, yielding both quantitative and qualitative insights into learners' evolution and the extent to which AI tools concord with ongoing human evaluation (Attali & Burstein, 2006; Hattie & Timperley, 2007).

The paper posits that AI applications such as Literatu can significantly speed up feedback provision and empower students to take charge of their progress, yet their full value emerges only when they are combined into a blended assessment framework. It is only in such a thoughtfully adjusted environment that AI can enhance writing assessment in ways that are fair, substantively rich, and genuinely inclusive, staying true to the core pedagogical principles that make assessment meaningful for every learner.

2. Methodology

This mixed-method research was conducted in three phases which aims to determine the impact of AI technology on writing skills in a high school ESL classroom. Carried out over five months at a British international school in Abu Dhabi, the research comprised 60 learners aged 14–15, in preparation for IGCSE ESL (English as a Second Language) course. The participant cohort was composed of speakers who had exposure, or no exposure, to English as a foreign language and as a first language speakers from different cultural and linguistic backgrounds. This kind of diversity is common in international schools, and it serves as an important background for the study on how well AI-based assessment tools work with different learner identities (García & Wei, 2014).

The key purpose was to find out the extent to which an AI-based platform—Literatu—could provide for the assessment and scaffolding of students' development in writing compared to how teachers would evaluate their learners using the official IGCSE writing rubrics. This narrow window on writing made possible a more detailed examination of the validity of assessment, the nature of automatic feedback, and how far AI tools can be tuned to fit or mimic the finely-balanced judgements of experienced teachers (Bennett, 2011; Warschauer & Grimes, 2008).

A three-phase design is consistent with current education research intervention research best methods, providing a phased but flexible framework for data collection and analysis (Creswell & Plano Clark, 2018). The phases were as follows:

2. 1 Stage 1: Baseline and Preparation of Participants

Phase I of the study provided the ethical, academic, and technical platform for meaningful data collection and intervention. Upon consent, students sat for a baseline assessment through a previous paper of the Cambridge IGCSE English as a Second Language (ESL) examination, the Reading and Writing Paper 1. The assessment was invigilated in exam-type conditions to replicate real external examination

conditions. A composition writing part involving extended responses that focus on content, coherence, precision, and communicative ability was used. These were base skills measured in the official IGCSE marking scheme which was used as a point of comparison for both teacher and AI assessment (Cambridge Assessment International Education, 2023).

The two grading rubrics were used to evaluate the student script independently. The class teacher marked the responses with the official IGCSE criteria and the Literatu platform did automatic marking via natural language processing and AI. At the end of Phase 1, the research project had successfully set a benchmark for writing in the classroom, a personalized AI-supported learning system, and had everyone educative actors (student and teacher) on board for the following phase of integration.

2.2 Phase 2: AI and ITS Integration, Feedback Cycles and Student Self-Evaluation

The second stage of the study focused on embedding the use of the Literatu AI platform in the everyday writing and feedback practices of the classroom, alongside developing students' metacognitive awareness and self-assessment. During the six-week intervention students completed a series of scaffolded writing tasks which were designed to simulate real examination conditions and encourage the development of writing on a broad front of abilities.

Responses were then immediately evaluated (by the teacher and literatu) within the key rubric domains —content, organization, language control, communicative effectiveness—and provided detailed written feedback along with a quantitative score for each of the four strands. This combination of input sources mirrored the result of previous studies showing that hybrid feedback systems, when adequately scaffolded, may improve the quality of writing and learner autonomy (Warschauer & Grimes, 2008; Wang et al., 2021).

One notable characteristic of this phase was the deliberate emphasis on rubric literacy and self-assessment. Students learned the elements of the IGCSE writing rubric through repeated classroom modeling. Teachers supported students to annotate exemplar responses, develop awareness of level descriptors and moderated peer work anonymously. Eventually, students moved from being passive readers of responses to being reflective writers of their own text. This transition is consistent with the principles of assessment for learning (AfL) which promotes student involvement in the evaluation process in order to enhance deeper learning (Black & Wiliam, 2009). At the end of Phase 2, students were able to be independent and critical writers who could revise their work based on a mix of human and AI feedback.

2. 3 Phase 3: Evaluation and Analysis

The last phase of the study was dedicated to summarizing post-intervention responses of the student writing performance. For an evaluation of progress in writing proficiency, the pupils sat a second IGCSE-style Writing Paper presented under examination-style conditions, designed to closely resemble the Writing Paper in baseline testing in Stage 1. The answers were marked independently by the teacher

and the Literatu AI platform with the official IGCSE writing rubric. This provided a direct juxtaposition against human and AI assessments so that we could investigate areas of agreement and disagreement across both types of assessment.

Beyond the summative writing, each student's digital portfolio completed over Phase 2 was analyzed to enable a longitudinal view of the writing across time. To better understand student experiences and behavior throughout the intervention, data analytics from the Literatu platform were collected. To complement the student perspective, post-intervention surveys were conducted to gather data on learner confidence, writing self-efficacy, and overall satisfaction with the feedback received. Additionally, teacher interviews were carried out to explore professional perspectives on the practicality, usability, and perceived reliability of the AI platform in supporting writing instruction. These multiple sources of data enabled a well-rounded understanding of the implementation process and laid the foundation for drawing broader conclusions in the next stage of the research.

2.4 Ethical Considerations

Informed consent, data confidentiality, and student rights were prioritized throughout the study. Only secure platforms were used for storing data, and assessments were independently reviewed to maintain neutrality and fairness. To counteract any potential measurement bias, all writing assessments were subject to independent blind evaluation by educators who had no prior engagement with the participants. These combined safeguards and procedural rigor guaranteed that the data acquired at every stage stood as robust, dependable, and analytically salient, thereby highlighting the contributions of artificial intelligence to writing assessment and the pursuit of equitable educational results across internationally distributed contexts.

3. Discussion

3.1. Understanding the Rubric and Its Limitations

This research was based on the Cambridge IGCSE English as a Second Language (ESL) Writing rubric, which was employed for A and AI evaluation. The rating scale provides up to a total of 15 marks: 6 marks for Content (whether the required skill has been fulfilled, and how fully the task has been addressed) and 9 marks for Language (how the language has been used, eg. appropriateness, accuracy and clarity). Despite being created for the purpose of standardization, much of its success depends on interpretive judgment. The lexis in the rubric—e.g. generally relevant, only partially fulfills—is wide open and vague, requiring subjective interpretation, in which case no artificial intelligence can compete Perelman, L. / 2012 /.

Examiner course-trained teachers have access to marked scripts and band descriptors which are of an added benefit in terms of understanding how the marking process works Al-Safinian, et al. (2025). AI tools like Literatu on the other hand rationalize these success criteria through algorithms and pattern recognition but may not have

the contextual depth educators are trained to use. Ifenthaler, D., & Bektik, D. (2022).

3.2. Initial Overreliance on Literatu and Shift Toward Human Feedback

Upon the introduction of Literatu, students readily adopted its instant feedback and gamified interface, lending them a sense of autonomy—much like the common observation of heavy initial reliance on AI tools for mechanical corrections (Amani &Bisriyah, 2025) It is in Week 3 of the term after future LLM students reported starting to notice repetition in the feedback ("generally well organised", "use of appropriate vocabulary") that such limited contextual depth and specificity in the feedback emerged as a problem in legal skills education (Sessler et al., 2025). As a result, by Week5 only about 10% of students used Literatu exclusively, with the remaining 90% using self or peer assessment as well – in line with studies that show that students trust AI/support blends more than AI alone (Zhang et al., 2025.

Table 1: Student Engagement with Feedback Sources Over Time

Week	Students using Literatu (%)	Students using teacher feedback (%)	Students engaged in self-assessment (%)
Week 1	100	0	0
Week 2	100	0	10
Week 3	70	30	20
Week 4	20	20	80
Week 5	10	20	90

3.3 Implementing "I Am the Examiner" Role-Play Strategy

To deepen rubric comprehension and foster assessment literacy, we introduced the "I Am the Examiner" role-playing approach. Learners examined anonymized scripts representing high, middle, and low achievement—echoing the practice of leveraging genuine student work to clarify rubric dimensions (King, 2017).

At first, many struggled to translate general descriptors like "generally relevant" or "partially fulfilled" into concrete judgment however the English Teacher Guided them. Through sequenced practice and subsequent reflective discussions, students moved from hesitation to adopting an "examiner mindset". (Cornell CTI, n.d.; Black & Wiliam, 1998). Interestingly, students tended to award marks that were 2 to 4 points lower than the examiners, which in turn raised their own performance thresholds and informed their revision strategies. This finding corroborates the literature suggesting that peer appraisal not only deepens investment but also raises the bar for learner expectations (Black & Wiliam, 1998).

3.3.1 Case Study – Specimen Report and Student Examiner Judgement

The role-play strategy "I Am the Examiner" was implemented with a practical case study using a specimen report provided by Cambridge. This sample was presented as part of the classroom task where students took on the role of examiners and assessed it independently, based on the rubric criteria. Below is the sample analyzed:

Question:	**Specimen answer**
Your class recently took part in a one-day sports festival for local schools. The organizers want students' opinions about the sports festival, and you have been asked to write a report. In your report, say what was enjoyable about the sports festival, and suggest how it could be improved. Here are some comments from students in your class: • *It was a chance to meet students from other schools.* • *There weren't many new sports to try.* • *I learned a lot about fitness.* • *We spent too much time listening to instructions.* Now write a report for the organizers of the sports festival. The comments above may give you some ideas, and you should also use some ideas of your own. Write about 120 to 160 words. You will receive up to 6 marks for the content of your report and up to 9 marks for the language used.	A one-day sports festival was held recently for all our local schools. The aim of the organizers was to involve as many children as possible and they certainly achieved this with a large variety of activities on offer. There were team games as well as individual competitions. At the end, the students could sign up for local clubs and continue to practice what they had learned. Many students gratefully seized this opportunity. There was food and drink provided not only for the competitors but also for the many parents and friends who came to watch. The refreshments were excellent and the seating at the tables was well planned so that the competitors could become acquainted with students from other schools. The only negative part was the speeches which were a little long and boring, something which could be improved in the future. Overall, the festival was a great success and the organizers hope to repeat it next year.

Table 2: Scoring Comparison Table: Specimen Report

Evaluation Source	Content (6)	Language (9)	Total (15)	Comments
Literatu (AI)	6	9	15	Perfect score; rubric matched well with AI pattern recognition, rewarded passive structures and linking devices

Student (Role-Play)	5	6	11–12	More critical: noted semi-formal tone, inconsistent paragraphing, minimal elaboration on weaknesses, reliance on the ideas given in the question.
Cambridge Examiner	6	8	14	Full task fulfilment; language generally excellent with minor spelling and format issues

3.3.2 Summary of Observations and Pedagogical Value

Rubric Application Gaps: When students used the rubric for the first time, they often assigned lower scores than the official examiner, particularly for language. The gap between their averages and the examiner's results was usually between 2 and 4 marks. Such conservative marking certainly showed their ambition, yet it also indicated a hesitance to accept semi-formal choices or stylistic variations as legitimate.

Critical Awareness: Students backed their scores with relevant details—pointing out uneven paragraph lengths, cursory mentions of critical feedback, and a casual style. Such comments reflected a developing sensitivity to the organizational clarity and level of professionalism called for in formal reporting.

Developing Assessment Literacy: The activity helped boost both their confidence and their ability to think critically. Learners no longer limited their feedback to obvious mistakes; instead, they began noticing how well their peers were attuned to their audience, how formal the tone needed to be, and how smoothly the ideas connected. This aligns with formative assessment theory, which stresses the value of student involvement in the evaluation process (Black & Wiliam, 2009; Sadler, 1989).

AI Limitations: Literatu awarded full marks for the piece, but failed to critique stylistic tone, audience alignment, or unequal paragraph development—further supporting the argument that AI, while efficient, lacks the ability to make nuanced evaluative judgments (Perelman, 2012; Warschauer & Grimes, 2008).

Feedback Awareness: When students matched their scores against the official examiner's, they acknowledged they had graded themselves a little too harshly, yet they were pleased to find their comments were still on target. This insight reshaped their writing process: instead of fearing feedback, they began to trust it. Consequently, 60% of the class recorded measurable improvement by weaving together peer reviews, self-reflection, and guidance from the teacher.

This single-report case reveals the ways assessment literacy shifts when students become evaluators. Existing literature highlights that when learners act as peer assessors, their capacity to read arguments critically and to wield evaluation rubrics improves (Li & Bektik, 2022; Sun et al., 2014).

3.4 Progress Across the Phases

The research utilized a three-phased intervention framework to assess students' development in writing competency and applied the Cambridge IGCSE rubric as the standardizing tool at each stage. This uniform rubric would align the measurement

of aspects that characterize good writing (task realization, content, grammatical accuracy, range of vocabulary, coherence) from elementary through advanced levels. The progression was a qualitative and a quantitative one, one from task data awareness to rhetorical awareness, from tone control and subtlety.

Table 3: Student Writing Progress Across Phases

Phase	Avg Content Score (out of 6)	Avg Language Score (out of 9)	% Meeting Target (11+/15)	Most Common Feedback Focus
Phase 1: Baseline	3	4	10%	Task fulfilment and clarity
Phase 2: Mid-Intervention	5	6	35%	Lexical accuracy and cohesion
Phase 3: Post-Intervention	6	7	65%	Development of ideas and tone

The huge rise in post-intervention performance shows the compound effects of multilayered feedback. It also demonstrates that combining AI tools with active learning tactics like peer evaluation and role play can make students assessment literary. These results are in line with the work of the Human-in-the-Loop (HITL) school of thought. This view emphasizes that teachers cannot be replaced by AI, especially in interpreting outputs and fostering deeper learning (Holstein et al., 2019; Pérez-Paredes, 2022). In sum, the three-stage model not only helped students develop linguistically, but also promoted a shift from passive authors to active readers and critics of their own work. This model thus points out the importance offeedback integrates AI-driven feedback into its pedagogy, with the hope of promoting metacognition skills, reflexiveshort-term, formal instigation and creation of skills that are readily taken for granted by successful writers.

3.5 Comparative Effectiveness of Feedback Modalities: AI, Teacher, and Peer Evaluation

The study compared three main modalities--traditional teacher feedback, AI-powered Literatu (with teacher customization), and the peer/self-assessment strategy through the "I Am the Examiner" role-playing--to find which part of the reinforcer strength lay were. This paper examines each modality in terms of three core areas in line with IGCSE rubric implementation: grammar accuracy (or lack thereof), structure and coherence and vocabulary range. This comparative analysis further emphasizes the importance of diversified feedback sources in supporting ESL learners as they make their way through international high-stakes writing tests.

Table 4: Writing Improvement Based on AI vs. Teacher Feedback

Feedback Type	Grammar Accuracy	Structure & Coherence	Vocabulary Range
Role Play "I am the Examiner"	92%	65%	70%
Literatu (Teacher-Customized AI Feedback)	88%	82%	80%
Teacher Feedback (Traditional)	95%	89%	82%

3.5.1 Role Play: "I Am the Examiner" – Building Rubric Fluency and Grammar Awareness
Peer assessment has led to a remarkable 92% improvement in students' grammar, a key factor contributing to this extraordinary outcome. Researchers attribute this success to the metalinguistic awareness fostered by repeated exposure to results and the required use of a structured rubric. However, the improvement in structure and coherence was only 65%. While students excelled at identifying grammatical issues, they often lacked the extensive vocabulary needed for effective revisions. This aligns with previous studies indicating that peer assessment, although it strengthens error recognition and evaluative feedback, often requires additional support from teachers or technology to transform criticism into constructive improvements (Min, 2005; Liu & Brantmeier, 2019). Integrating AI feedback alongside peer evaluations, particularly in identifying misplaced ideas and linking words, can significantly enhance this improvement rate (Warschauer, 2021; Holstein et al., 2019). Moreover, students achieved a 70% improvement in vocabulary, driven by their increasing confidence in selecting appropriate word choices and employing lexical variation, especially when analyzing model answers or reflecting on their own essays. Although learners may struggle to propose alternatives independently, rubric-based assessment encourages careful consideration of tone and register (Hyland, 2003).

3.5.2 AI-Enhanced Feedback via Literatu – Scalable, Consistent, and Customizable
The findings indicate that the Literatu system, particularly when tailored by educators (adjusting the rubric), has achieved impressive results across three key domains: 88% accuracy in grammar, 82% coherence, and an 80% improvement in vocabulary. These statistics underscore the platform's effectiveness in providing immediate and detailed feedback, grounded in established criteria for accuracy (Cambridge Assessment, 2021).
This paper expands upon the foundational principles presented at last year's ATEE conference in Split, reinterpreting the familiar concept of feedback models. It draws on Wulf Orth's (1988) classification of three types of feedback: "what you want," "what you got," and "how to get from A to B" (Bax, 2018).
Crucially, the integration of AI tools like Literatu significantly alleviates the burden on teachers. By managing low-level feedback tasks, these tools free up valuable teaching time for educators to focus on identifying deeper, more complex errors—achieving this with three times the accuracy of a native speaker working alone

(Shibani et al., 2020). However, as noted by several scholars (Perelman, 2012; Hao and Weninger, 2019), these systems should not operate in isolation. Without contextual understanding, AI may produce writing that is overly formulaic or mechanical, potentially leading to miswriting as much as misreading. This could detract from the quality of otherwise strong texts, reducing them to mere exercises for students with dyslexia.

3.5.3 Traditional Teacher Feedback – Depth, Personalization, and Contextual Nuance

The most anticipated outcome among the findings was that traditional teacher feedback resulted in the most significant improvement in students' scores. Specifically, grammar scores improved by 95%, while structure and coherence increased by 89%, and vocabulary use rose by 82%. There are several distinct advantages to receiving feedback directly from teachers:

- *Personalized Scaffolding:* Feedback is tailored based on classroom observations and individual student histories, leading to a more nuanced understanding of often vague rubric language.
- *Integration of Cultural and Rhetorical Considerations:* Teacher feedback incorporates elements of cultural creativity and rhetorical awareness.

These findings align with Hattie and Timperley's (2007) theory on effective feedback, which poses three essential questions: Where am I going? How am I doing? What's next? At these critical junctures, no other feedback delivery method can match the impact of teacher responses.

Moreover, teacher feedback can facilitate dialogue: Students gain clarification and, in turn, contribute new ideas about the material. Through interactions with instructors and peer discussions, learners create a web of meanings that may not have existed prior to the intervention (Brookhart, 2008; Bitchener & Ferris, 2012). This developmentally oriented teaching approach is particularly crucial when launching an engagement campaign for high-stakes exams, where awareness of the audience, tone, and genre appropriateness are vital variables for success.

4. Conclusion

This research highlights that effective writing instruction in ESL contexts cannot rely solely on one type of feedback, especially when preparing students for Cambridge IGCSE examinations. AI-powered tools, such as Literatu, fall short as they often fail to capture the nuanced messages inherent in academic writing—specifically the interplay of tone, purpose, and audience (Perelman 2012; Hao & Weninger 2019). Implementing peer-assessment strategies, particularly the "I Am the Examiner" role-play approach, has proven beneficial for students. This method allows them to internalize rubric criteria and engage in metacognitive reflection on their writing, resulting in noticeable improvements in coherence, grammar accuracy, and lexical sophistication (Black & Wiliam 2009; Andrade & Valtcheva 2009).

Crucially, the most significant advancements stem from a combination of human, peer, and AI feedback. The shift from reliance on AI to student-led analysis emphasizes

the urgent need for schools to adapt, particularly as many students face challenges due in part to limited access to high-quality educational resources. Enhancing students' interactions with AI-generated feedback empowers them to take greater responsibility for their learning, fostering autonomy and allowing them to benefit from high-quality assessment without the accompanying anxiety (Sadler 1989; Nicol & Macfarlane-Dick 2006).

The three-phase intervention model—diagnostic stage, treatment, and post-treatment—demonstrated how iterative feedback could enhance learners' writing proficiency. By the third stage, a majority not only performed better on their tests but also developed a deeper understanding of genre conventions, coherence, and tonal modulation (Cambridge Assessment 2023; Hyland 2003). The study underscores the necessity of integrating AI within a pedagogically sound, teacher-guided framework to achieve effective, equitable, and authentic writing instruction. It also suggests that the 'Human-in-the-Loop' (HITL) model offers a promising long-term solution. More than any alternative currently available, it represents our best hope for enabling diverse voices to engage in meaningful dialogue in the future (Holstein et al. 2019; Pérez-Paredes 2022).

4.1 Final Pedagogical Implications and Future Considerations

In light of the findings from this study, future research should focus on long-term analyses of test-specific feedback practices, particularly those aligned with established rubrics exams. While the assistance provided by AI tools and peer-assessment strategies is undoubtedly valuable, this inquiry reaffirms that teacher feedback remains indispensable for crafting nuanced writing that meets examination standards. Educators possess interpretive expertise, contextual awareness, and developmental insights that cannot be replicated by any AI system (Perelman, 2012; Hattie & Timperley, 2007; Brookhart, 2008). Therefore, future studies should move beyond merely assessing the efficacy of feedback and instead explore how teacher feedback, aligned with examination requirements, influences student performance.

This focus is particularly crucial in global educational systems, where academic success is closely tied to performance on standardized tests. In both British and American curricula, exam-style rubrics are integral to pedagogy, necessitating that instructional feedback aligns with the criteria used in actual examinations to ensure student success (Cambridge Assessment, 2023; Sadler, 1989). Longitudinal research investigating how teacher feedback, closely linked to test descriptors, shapes learner outcomes over time—especially in contexts where students are learning English as a second or foreign language—will enhance our understanding of the real-world effects of formative assessment. Moreover, such research could explore ways to improve AI tools so they can complement rather than replace the essential feedback loop provided by teachers (Perez-Paredes, 2022; Nicol & Macfarlane-Dick, 2006). Ultimately, teacher judgment, rooted in assessment literacy and situational awareness, should be the foundation of any feedback model aimed at preparing students for success in standardized academic environments.

References

Attali, Y., & Burstein, J. (2006). *Automated essay scoring with e-rater® V.2*. Journal of Technology, Learning, and Assessment, 4(3), 1–30.
Bax, S. (2018). *Discourse and evaluation in language learning: A critical review*. Applied Linguistics, 39(4), 438–457.
Bennett, R. E. (2011). *Formative assessment: A critical review*. Assessment in Education: Principles, Policy & Practice, 18(1), 5–25.
Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. Phi Delta Kappan, 80(2), 139–148.
Black, P., & Wiliam, D. (2009). *Developing the theory of formative assessment*. Educational Assessment, Evaluation and Accountability, 21(1), 5–31.
Brookhart, S. M. (2008). *How to give effective feedback to your students*. ASCD.
Cambridge Assessment (2021). *Understanding assessment: A guide for teachers*. Cambridge University Press.
Cambridge Assessment International Education. (2023). *IGCSE English as a Second Language (0510/0991): Syllabus and Support Material*.
Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.
Hao, Y., & Weninger, C. (2019). *Automated writing evaluation and its pedagogical implications for second language learners*. ELT Journal, 73(2), 144–153.
Hattie, J., & Timperley, H. (2007). *The power of feedback*. Review of Educational Research, 77(1), 81–112.
Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). *Improving fairness in machine learning systems: What do industry practitioners need?* Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–16.
Hyland, K. (2003). *Second language writing*. Cambridge University Press.
Ifenthaler, D., & Bektik, D. (2022). *AI-supported assessment in education: Opportunities and challenges*. (Publication info needed if available)
Kukulska-Hulme, A., & Lee, H. (2020). *Mobile and AI-supported language learning: Affordances and limitations*. Language Teaching, 53(2), 157–182.
Liu, J., & Brantmeier, C. (2019). *Second language writing development and automated feedback*. Journal of Second Language Writing, 45, 1–8.
Min, H-T. (2005). *Training students to become successful peer reviewers*. System, 33(2), 293–308.
Nicol, D. J., & Macfarlane-Dick, D. (2006). *Formative assessment and self-regulated learning*. Studies in Higher Education, 31(2), 199–218.
Perelman, L. (2012). *The problem with automated essay scoring*. MIT Media Lab Working Paper.
Pérez-Paredes, P. (2022). *AI in language education: Where we are and where we're going*. CALICO Journal, 39(1), 23–46.
Sadler, D. R. (1989). *Formative assessment and the design of instructional systems*. Instructional Science, 18(2), 119–144.
Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation*. Routledge.
Shibani, A., Knight, S., & Buckingham Shum, S. (2020). *Educator customization of automated*

feedback. International Journal of Artificial Intelligence in Education, 30(3), 537–580.
Sun, Y., et al. (2014). *Peer feedback in ESL writing*. Journal of Second Language Writing, 25, 1–15.
Wang, Y., Tai, Y., & Zhou, Z. (2021). *Hybrid feedback models in L2 writing classrooms*. Language Learning & Technology, 25(3), 1–18.
Warschauer, M. (2021). *The need for human-centered AI in language education*. Language Learning & Technology, 25(2), 2–5.
Warschauer, M., & Grimes, D. (2008). *Automated writing assessment in the classroom*. Pedagogies, 3(1), 22–36.
Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.