English Language & Literature International Conference
Vol. 8 No. 1
https://jurnal.unimus.ac.id/index.php/ELLIC/index
p-ISSN: 2579-7549
e-ISSN: 2579-7263

A Systematic Review of AI-Based and Teacher-Based Writing Assessment

Reno Setiyowati,
Viqi Ardaniah
Universitas Airlangga
Surabaya, Indonesia
reno.setiyowati-2024@fib.unair.ac.id

ABSTRACT

Establishing valid and reliable writing assessments in education remains a persistent challenge, especially with the emergence of artificial intelligence (AI) as a tool in evaluation. Although previous studies have investigated both AI-based and teacher-based writing assessments, few have addressed them through a systematic lens, leading to fragmented insights and inconsistent frameworks. This study aims to investigate the current state-of-the-art of AI-based and teacher-based writing assessments and to identify emerging debates and research directions in the field. A total of 258 articles were collected from Scopus, ScienceDirect, JSTOR, Emerald, and Ebscohost, and filtered using PRISMA guidelines. Eight articles met the inclusion criteria. The findings reveal that AI-based assessments offer high consistency and efficiency in evaluating surface-level language features, but struggle with assessing higher-order discourse aspects such as coherence, argumentation, and rhetorical structure. Conversely, teacher-based assessments provide richer, context-aware feedback, yet are limited by issues of subjectivity and scalability. A hybrid model that integrates AI efficiency with human insight emerges as a promising solution to balance reliability and validity in writing assessment. Nevertheless, key debates remain regarding AI's scoring authority, construct validity, ethical concerns and, implementation across diverse educational contexts. This study calls for the development of unified frameworks and teacher training to support equitable and effective AI-human collaboration in writing assessment.

Keywords: AI writing assessment, teacher feedback, hybrid evaluation, EFL, systematic review

p-ISSN: 2579-7549
e-ISSN: 2579-7263

INTRODUCTION

In recent years, writing assessment (WA) has been questioned for its reliability and validity due to lack of clarity in rubrics and subjectivity (Thwaites et al., 2025). Creating reliable and valid writing assessments needs involvement from various educational aspects such as rubric development, assessor training, and quality assurance (Page et al., 2021). For example, higher education institutions spend resources to create detailed rubrics and conduct rater training to ensure scoring consistency (Li et al., 2024). However, one challenge emerges due to rapid technology development through the presence of artificial intelligence (AI), which is now used to assess writing made by learners. Traditional assessments are criticized for being time-consuming and prone to inconsistency and bias (Hand & Li, 2024). Increasing demands for efficiency and scalability have encouraged the exploration of AI-based solutions. This leads to the adoption of AI in writing assessment as an alternative to traditional methods.

AI has been adopted in WA resulting in AI-based writing assessment that uses artificial intelligence to automatically evaluate and score student writing based on linguistic and rhetorical features. AI-based WA is designed to offer efficiency, consistency, and scalability in writing evaluation (Kasih & Putra, 2024; Saleh & Alshulbi, 2025). Studies such as Zhang et al. (2019) and Liu et al. (2020) have shown that AI assessments are capable of providing valid and reliable scores, particularly in large-scale educational settings. However, prior studies have highlighted that relying on AI to assess writing neglects the senses and nuances of the writing itself. The use of AI tends to focus more on surface features like grammar and syntax while lacking the ability to deeply evaluate content development, logical flow, and creativity (Steiss et al., 2024). Xu et al. (2025) suggested that writing assessment should rely on teachers' capability to understand the development of ideas and the context of learners' writing. Hence, there is a need to embed teachers' perspectives in the assessments generated by AI. This study argues that AI based WA and teacher-based WA are combined to get more valid and reliable WA.

AI-based and teacher-based WA refer to two different approaches to assessing student writing. AI-based WA uses algorithms to analyze writing according to a pre-determined rubric or prompt (Lin & Crosthwaite, 2024; Hand & Li, 2024). In contrast, teacher-based WA uses human judgement to evaluate the broader aspects of writing, including both surface features and higher-order

p-ISSN: 2579-7549
e-ISSN: 2579-7263

elements (Jamshed et al., 2024; Li et al., 2024). However, AI-based WA faces challenges such as potential bias in the training data and a lack of adaptability to diverse writing styles (Alsalem, 2024).

Several studies have suggested combining both approaches to maximize the benefits. Lin and Crosthwaite (2024) and Hand and Li (2024) found that integrating AI feedback with teacher evaluation leads to better writing revisions and enhances assessment quality. Proponents argue that combining AI and teacher-based WA can reduce teacher workload and improve consistency (Alsalem, 2024), while critics highlight the risk of overreliance on AI and its limitations in understanding deeper writing nuances (Dikli, 2010; Ma & Slater, 2016). Mehdaoui (2024), found that while teachers acknowledge the potential benefits of AI, their resistance to using it is often driven by external challenges**,** such as slow internet connections, lack of proper training, and limited resources (Mehdoui, 2024). These technical and infrastructural issues are key barriers that discourage teachers from adopting AI tools in educational settings.

Although many studies have examined the benefits and limitations of incorporating AI-based WA with teacher-based WA using various methods, including experimental designs, cross-sectional studies, and correlational analyses, a comprehensive and agreed-upon definition of their scope remains elusive. Differing perspectives on key assessment indicators further complicate this issue. An integrated WA concept is urgently needed to address these disparities. Therefore, this study proposes a systematic review to clarify the current state-of-the-art in AI-based and teacher-based WA and to identify new directions for future research. Specifically, this study aims to: a) investigate the current state-of-the-art of AI-based and teacher-based WA; and b) identify future debates to foster further studies.

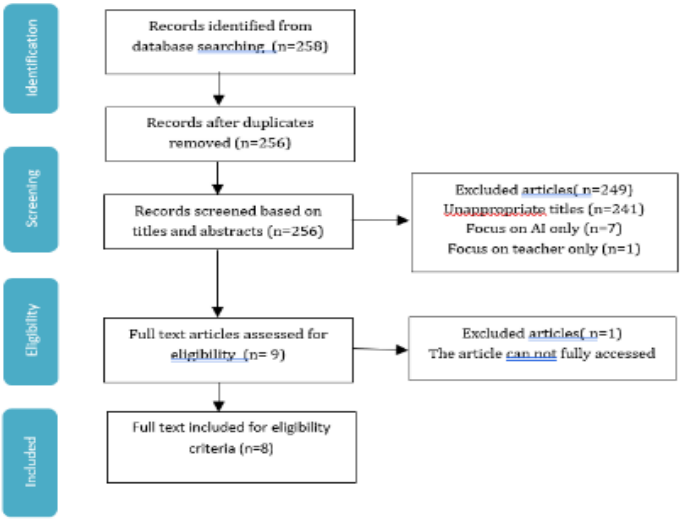
METHOD

This study employed a systematic literature review (SLR) to analyze and summarize the results of studies that compared AI-based writing assessment tools with teacher-based assessment. The review follows PRISMA (Page et al., 2020) guidelines to ensure transparency, reproducibility, and accuracy in the process of article selection, screening, and inclusion. The keywords “writing assessment,” “AI tools,” and “teachers” were employed in five databases: Scopus, Science Direct, JSTOR, Emerald, and Ebscohost. Search limitations were set to include only English language, peer-reviewed, and open-

p-ISSN: 2579-7549
e-ISSN: 2579-7263

access articles. A Boolean search in Scopus retrieved 44 articles; Science Direct yielded 42 articles; Ebscohost, 1 article; and Emerald, 130 articles. In total, 258 articles were obtained from these databases. Through Rayyan.ai, duplicate articles were removed, resulting in 256 articles. Title and abstract screening selected 17 articles for further analysis. Out of these, 8 studies remained relevant according to the inclusion and exclusion criteria. Thus, the 8 articles were included in the units of analysis for this systematic review.

(Figure 1.)



Source: Researcher Analysis

Inclusion Criteria

The following criteria were set to include articles in this systematic review;

- Studies involving student writing performance, particularly in English as a Foreign/Second Language (EFL/ESL) or academic contexts;
- Study type had to be an original research paper with a clear methodology;
- Studies that compare AI-based WA with teacher WA;
- Articles written in English language;
- The studies is fully open accessed.

Quality Assessment

Quality assessment for the selected articles was conducted with

p-ISSN: 2579-7549
e-ISSN: 2579-7263

the help of Rayyan.ai to label and categorize whether the studies combined or compared AI-supported teacher feedback and traditional teacher-written assessment (WA). After a detailed analysis of eight studies, two were found to integrate AI with teacher-driven feedback processes, while the remaining six focused on comparison. One study primarily explored teacher perceptions; however, to measure perceptions, teachers participated in using an AI-supported assessment tool and reviewed its feedback results. Therefore, this study was still included in the analysis, as it compared the performance of AI-supported WA and teacher WA.

Tabel 1. Data Extraction Representing Unit Analysis

Research	Title	AI WA	Teacher WA
Hand & Li (2024	Exploring ChatGPT-supported teacher feedback	• ChatGPT (LLM by OpenAI) trained to provide (1) corrective feedback on Ferris's 15 error types and (2) holistic rhetorical feedback. Prompts were crafted for detailed feedback generation. • Accuracy and completeness of AI feedback measured by student revision success. • Reduction in language errors and rhetorical issues based on students' writing revisions after AI-supported feedback across two tasks • AI reduces teacher burden in large classes, provides detailed feedback efficiently, and promotes learner autonomy	• Teachers provided adapted feedback using ChatGPT outputs , coded indirect feedback • Teachers provided adapted feedback using ChatGPT outputs • Teachers ensured accuracy of grammar feedback (Ferris's 15 types) and contextualized rhetorical feedback • Maintains teacher-student interaction while using AI as a supportive tool. Enhances feedback quality and consistency without fully replacing human judgment. • Teachers maintain oversight to ensure AI feedback validity and appropriateness.
Lin & Crosthwaite, 2024	The Grass is not Always Greener: Teacher vs. GPT-assisted Written Corrective Feedback	• ChatGPT (GPT-4) used to generate written corrective feedback (WCF) with prompts. Feedback was predominantly metalinguistic and reformulations with notable inconsistency between sessions even for the same text. • Feedback was often more local (sentence-level) and sometimes redundant. • Limited attention to global writing features like organization and content development. • AI can reduce teacher workload in providing local error correction	• Teachers used a combination of direct and indirect WCF, typically balancing local and global issues with more personalized attention to content, structure, and argument development. • Both local-level (grammar, syntax, word choice) and global-level (organization, coherence, content) writing components were addressed. • Evaluated based on feedback form (direct, indirect, reformulation, metalinguistic), focus (local/global), and judgment of redundancy and accuracy. Consistency across teacher feedback noted. • Teachers provided comprehensive, context-sensitive feedback fostering critical revision and deeper engagement with the text.

p-ISSN: 2579-7549
e-ISSN: 2579-7263

Dhini, Girsang, Sufandi, & Kurniawati, 2023	Automatic Essay Scoring for Discussion Forum in Online Learning Based on	• Automated scoring system computes semantic similarity between student essays and reference answers and keyword similarity against rubrics. No direct human intervention in feedback or scoring. • Focus on sentence-level semantics (meaning similarity) and keyword overlap • Evaluation based on correlation between system-generated and human-assigned scores; strong correlation observed with a Pearson value of 0.65 using combined semantic and keyword metrics. • Potential biases in the model due to training data limitations; risk of reduced fairness without human review	• Teachers manually grade student discussion forum contributions based on predetermined rubrics including keyword and semantic similarity argumentation quality. • Human scoring is based on holistic evaluation guided by rubrics without explicit use of statistical metrics like Pearson correlation. • Human evaluators' scores serve as ground truth for model comparison; manual grading is subject to subjective
Jamshed, Ahmed, Sarfaraj, & Warda, 2024	The Impact of ChatGPT on English Language Learners' Writing Skills: An Assessment of AI Feedback on Mobile	• ChatGPT mobile application (GPT-3.5) used for providing immediate, personalized feedback on grammatical and structural errors based on prompts entered by learners. • Provide corrective feedback with explanations and examples for learning reinforcement. • Demonstrated that AI tools can enhance writing skills through immediate feedback; supports integration of AI in ESL curricula to supplement traditional teaching and boost learner confidence and motivation. • Concerns about the adaptability of AI feedback to different proficiency levels; potential over-reliance on technology; emphasized need for human oversight and further development for personalized learning.	• Teachers provided handwritten feedback manually identifying and explaining grammatical errors. • Common grammatical errors including third-person singular, past tense, progressive, past participle, plural, possessive, comparative, and superlative forms. • Teacher feedback served as a control benchmark for comparison; evaluated against AI feedback on error correction effectiveness and impact on writing skills. • Teacher feedback is slower than ChatGPT's feedback

p-ISSN: 2579-7549
e-ISSN: 2579-7263

Li, Huang, Wu, & Whipple, 2024	Evaluating the Role of ChatGPT in Enhancing EFL Writing Assessments in Classroom Settings: A Preliminary Investigation	• ChatGPT versions 3.5 and 4 were used. Both provided holistic scoring based on the College English Test Band 4 (CET-4) rubric and qualitative feedback on language, content, and organization aspects. • Version 4 showed higher scoring reliability and provided more relevant feedback compared to teachers. • Evaluation based on generalizability (G-) theory analysis of reliability coefficients (G-coefficients and Phi-coefficients) for scoring,. • Higher G-coefficients for ChatGPT4 (0.89) versus teacher (0.80) and ChatGPT3.5 (0.66) indicated better reliability. • ChatGPT4 also provided a larger number of relevant feedback comments across all writing aspects. • The adoption of AI should be navigated with attention to ethical consideration	• Four experienced college English teachers manually scored essays holistically using the CET-4 rubric and provided qualitative feedback on language, content, and organization. • Teacher feedback generally emphasized grammatical corrections and basic feedback on content and structure. Feedback was less consistent and less comprehensive across multiple writing domains compared to ChatGPT. • Teachers showed lower reliability and fewer detailed feedback comments. • Teachers showed consistency in language-related feedback but gaps in depth and coverage of content and organization aspects.
Hong Ma & Tammy Slater, 2016	Connecting Criterion Scores and Classroom Grading Contexts: A Systemic Functional Linguistic Model for Teaching and Assessing Causal	• Criterion AWE system (developed by ETS) was used. It provided holistic scoring using the E-rater engine and diagnostic feedback via the Critique function, focused on grammar, usage, mechanics, vocabulary, style, and discourse elements. • Automated scoring based on a large corpus of human-rated essays; • Risk of students gaming the system (e.g., lengthiness, overuse of transitions) if not mediated by teacher instruction; automated	• Manual scoring by classroom instructors based on holistic rubrics emphasizing general writing quality (grammar, coherence, content, and organization) • SFL raters focus on causal discourse: language use (grammar, word choice), causal relationships (cause-effect expressions), coherence, cohesion, logical structure, and rhetorical organization analyzed by Developmental Path of Cause • Teachers' intuitive judgments align closely with the Developmental Path of Cause. • Teachers graded primarily

p-ISSN: 2579-7549
e-ISSN: 2579-7263

			• Teachers' rubric-based scores served as a benchmark to highlight differences in evaluation focus compared to Criterion and SFL model-based assessments. • Traditional human grading faces challenges such as inconsistency, bias, and rubric limitations
Semire Dikli, 2010	The Nature of Automated Essay Scoring Feedback	• MY Access! (Version 6.0) based on Vantage Learning's IntelliMetric engine (focus and unity, content and development, organization, language use and style, mechanics and conventions). • Strengths: Fast scoring and feedback, consistent and systematic, handles large numbers without fatigue. • Weakness: Overwhelming for low-proficiency ESL learners; Same feedback repeated even after revisions; Lacked specificity to individual essays; Students couldn't clarify or discuss; In some cases, program failed to generate feedback for short, repetitive, or grammatically poor essays; Wrong advice, especially for ESL errors like prepositions or pronouns.	• Manual written feedback provided by an ESL instructor using anlytic and holistic rubrics generated by the MY Access! • Strengths: Shorter and more specific; Built on previous drafts, addressed consistent patterns of errors; Praises even for small improvements; Students could ask questions and get clarifications. • Weakness: Time consuming, subjective and inconsistent, limited scalability • AES can supplement teacher feedback but cannot replace human evaluation, especially for lower-proficiency students. • More independent research is needed on AES feedback, especially for ESL/EFL learners.

p-ISSN: 2579-7549
e-ISSN: 2579-7263

Mashael Salem Alsalem, 2024	EFL Teachers' Perceptions of the Use of an AI Grading Tool (CoGrader) in English Writing Assessment at Saudi	• CoGrader, an AI grading tool providing grades and detailed feedback on student essays based on rubrics set by teachers. • AI grading saves teachers' time and more objective than human grading. • CoGrader lacks depth in personal feedback and faces challenges with nuanced assessment. • CoGrader is lack of ability to evaluate the content and organization. • Concerns about fairness, cultural appropriateness, and lack of nuanced feedback • Overreliance on AI	• Teachers uploaded student essays and input the pre-determined rubric by the department into CoGrader. CoGrader generated grades and feedback based on that rubric. Teachers reviewed the AI-generated scores and feedback. • Teachers' feedback perceived as more context-sensitive and capable of addressing individual learning differences • Using AI grading requires professional training because how the AI works depending on the rubric/prompt input by teachers. • Human grading risks include subjective bias, fatigue, inconsistency, and time constraints; however, teachers' adaptability allows for context-sensitive assessment unavailable to AI systems.

FINDINGS AND DISCUSSION

This review analyzed eight empirical studies comparing AI-based and teacher-based WA. A summary of these papers is presented in Table 1.

Major Findings : AI-Based vs Teacher-Based WA

AI-based tools such as ChatGPT**,** Criterion**,** CoGrader, and MY Access! are increasingly utilized in writing assessment (WA), particularly for evaluating surface-level features. These tools often focus on grammar, vocabulary, syntax, and mechanics (Jamshed et al., 2024; Lin & Crosthwaite, 2024). Unfortunetely, the use of AI has shed a diverse range of perspectives on the role of AI in classroom writing assessment. Some, like those by Dhini et al. (2023) and Jamshed et al. (2024), use AI mainly for surface-level corrections such as grammar and spelling. Others take a more collaborative approach, for instance, Hand and Li (2024) explore how AI can enhance teacher feedback, while Li et al. (2024) assess how reliable and comprehensive AI feedback is, especially on language, content, and organization. These varying perspectives highlight an important gap: there's still no clear agreement on what AI should actually do in writing assessment. This lack of shared understanding makes it harder to compare results

p-ISSN: 2579-7549
e-ISSN: 2579-7263

across studies and points to the fact that AI in education is still very much in an exploratory stage.

Meanwhile, teacher-based writing assessment typically employs rubric-based methods, both holistic and analytic, as well as models like the *Developmental Path of Cause* for evaluating causal writing. Across the studies, teachers serve not only as benchmarks against which AI tools are validated but also as critical agents in ensuring contextual relevance, discourse-level interpretation, and ethical oversight. These approaches are praised for their capacity to account for context, learner intent, and broader discourse-level features that AI systems often overlook.

Strength & Weakness: AI based WA vs Teacher based WA

AI-based writing assessment have demonstrated increasing sophistication in evaluating surface-level linguistic features (Alsalem, 2024; Hand & Li, 2024; Jamshed et al., 2024). These systems offer fast, consistent, and scalable feedback, with several studies showing positive effects on student revision quality and grammatical accuracy (Jamshed et al., 2024; Lin & Crosthwaite, 2024). Notably, ChatGPT4 outperformed teacher raters in terms of scoring reliability and comprehensiveness of feedback across domains. For instance, Li et al. (2024) reported that ChatGPT4 yielded a G-coefficient of 0.89 in scoring essays based on CET-4 criteria higher than both ChatGPT3.5 (0.66) and human raters (0.80). Similarly, Dhini et al. (2023) demonstrated strong alignment between automated scores and human scores using Pearson correlation coefficients ($r = 0.65$), supporting the statistical reliability of semantic and keyword-based scoring models. This consistency is particularly advantageous in large-scale testing contexts where uniformity across raters is essential.

However, they tend to struggle with global aspects of writing such as content development, coherence, and rhetorical structure (Alsalem, 2024; Ma & Slater, 2016). When it comes to things like how ideas are organized, how arguments are built, or how content flows, these systems often fall short. The feedback does not reflect cultural or educational context awareness, which may be critical in EFL assessment. Several studies also raise concerns about inconsistent and sometimes confusing feedback. For example, Lin and Crosthwaite (2024) found that ChatGPT gave different feedback on the same writing input, calling its reliability into question. Zhang and Zou

p-ISSN: 2579-7549
e-ISSN: 2579-7263

(2024) also point out that AI models like GPT operate as "black boxes"—their decision-making processes are not transparent and can't easily be justified from a teaching perspective. Without standardized scoring rubrics, consistent prompts, or clear training guidelines, students can end up getting feedback that's not only inconsistent but also pedagogically unsound. These issues are especially problematic in high-stakes settings where feedback accuracy really matters.

By contrast, teacher-based assessments are often considered to have higher construct validity, as teachers incorporate a broader range of discourse features and sociocultural awareness into their evaluation. Teachers can recognize rhetorical strategies, content appropriateness, and learner intention, aligning their feedback with curricular goals and developmental stages (Hand & Li, 2024; Lin & Crosthwaite, 2024). However, the reliability of teacher-based assessment is frequently challenged due to inter-rater variability, fatigue, and implicit biases (Dikli, 2010). The same essay may receive different scores depending on the teacher's interpretation of rubrics, background knowledge, or even emotional state reducing scoring consistency. Although teacher feedback can vary in depth and consistency, particularly under workload constraints, it consistently outperforms AI in handling nuanced features of writing such as argument quality, coherence, and learner intent.

Some studies attempt to reconcile this tension through hybrid approaches. For example, Hand and Li (2024) reported improved scoring reliability and feedback completeness when teachers reviewed and adapted AI-generated feedback. When properly fine-tuned or used with prompt engineering, AI such as GPT-4 models can produce scoring outputs that correlate highly with expert raters. This suggests that AI–teacher collaboration models could offer a pathway to achieving both high reliability and contextual validity with appropriate training of human raters.

Human-AI Collaboration: A Hybrid Approach to WA

A growing body of research supports the potential benefits of integrating AI-based feedback with teacher-mediated assessment, forming a hybrid model that combines the strengths of both approaches. In their study (such as Hand & Li, 2024), teachers used AI-generated feedback as a foundation but adapted and refined it to better align with student needs, ensuring accuracy and instructional

p-ISSN: 2579-7549
e-ISSN: 2579-7263

relevance. This model reduced teacher workload while preserving the pedagogical depth of human judgment.

The hybrid approach addresses the long-standing trade-off between reliability and validity. AI systems offer consistent, scalable scoring and immediate feedback, making them suitable for addressing local-level errors and improving learner autonomy. When integrated, these two modes can compensate for each other's limitations AI ensuring scoring consistency and coverage, and human oversight ensuring communicative relevance and developmental appropriateness.

Nonetheless, successful implementation of hybrid models requires careful planning. Teachers must be adequately trained to interpret and mediate AI outputs, and AI tools must be transparent and adaptable to various learning contexts. Also, hybrid feedback models should be designed with clearly defined roles for AI and human instructors. A multi-phase assessment model could be developed, for example: (1) AI provides initial feedback on language and structure, (2) teachers validate and supplement with content-focused comments, and (3) students revise based on a blended feedback approach.

Future research should explore how hybrid feedback models impact long-term writing development, learner trust, and teacher agency. As writing classrooms increasingly adopt digital tools, the human-AI collaborative framework offers a promising direction toward more effective, scalable, and pedagogically grounded assessment systems.

Emerging Gaps and Future Debates in Writing Assessment

Despite the promising results of hybrid models that combine AI-generated feedback with teacher mediation, several unresolved debates and challenges remain. One major issue concerns the distribution of authority and responsibility between human raters and AI systems. While AI provides rapid, consistent feedback, questions arise regarding how much influence it should exert in shaping final scores and revisions. In practice, teachers often serve as corrective agents modifying or filtering AI suggestions but it is unclear whether this enhances assessment quality or merely compensates for AI's limitations. Another challenge lies in maintaining construct validity within these hybrid settings. Though human oversight is

p-ISSN: 2579-7549
e-ISSN: 2579-7263

assumed to improve validity, the extent to which AI-generated input truly supports deeper discourse-level assessment remains underexplored. Furthermore, the presence of AI in feedback processes can shift learner and teacher perceptions. Students may perceive AI feedback as more "objective," potentially undermining the authority of teacher guidance, while teachers may experience reduced agency or feel compelled to align their judgments with algorithmic outputs.

The socio-cultural applicability of hybrid models also demands scrutiny. Most studies are conducted in well-resourced contexts with experienced educators and reliable digital infrastructure. In lower-resourced or culturally diverse environments, the adaptability of AI systems and the readiness of teachers to critically engage with them are not guaranteed. This raises concerns about equity, especially when AI systems embed biases from their training data. Finally, hybrid assessment models introduce a new layer of opacity: when both AI and human input contribute to scoring and feedback, tracing the rationale behind final evaluations becomes increasingly complex. This ambiguity has implications for student trust, grading transparency, and institutional accountability especially in high-stakes or large-scale assessment settings. Addressing these challenges requires future research that not only refines technological tools but also redefines pedagogical roles and ethical standards within writing assessment practices.

CONCLUSION

Based on the thematic analysis above, several systematic recommendations emerge. First, a unified evaluation framework is needed for AI-based writing assessment that integrates linguistic, rhetorical, and affective dimensions. Second, future research should prioritize developing culturally and linguistically adaptive AI systems, while also addressing ethical concerns related to AI implementation, especially in EFL and Global South contexts.

Third, hybrid feedback models should be developed with clearly defined roles for both AI tools and human instructors. Teachers play a central role in operating these systems and guiding students in their use, particularly through prompt design and interpretation. They require targeted training to adapt effectively to AI-integrated environments. To ensure comparability and accountability, future research should adopt standardized reporting

p-ISSN: 2579-7549
e-ISSN: 2579-7263

templates that specify feedback domain coverage, reliability metrics, and the extent of teacher involvement. These measures would support the development of a more coherent, transparent, and ethically grounded ecosystem for AI-assisted writing assessment.

These attributes can serve as guiding principles for educators and policymakers seeking to adopt AI-assisted tools without compromising pedagogical depth. The results can support teacher training, inform curriculum design, and encourage the ethical adoption of AI technology in writing assessment. Ultimately, this will lead to more equitable and scalable language education systems.

REFERENCES

Alsalem, M. S. (2024). EFL teachers' perceptions of the use of an AI grading tool (CoGrader) in English writing assessment at Saudi universities: An activity theory perspective. *Computers and Education: Artificial Intelligence, 6*, 100228. https://doi.org/10.1016/j.caeai.2024.100228

Dikli, S. (2010). The nature of automated essay scoring feedback. *Computers and Composition, 27*(3), 195–207. https://doi.org/10.1016/j.compcom.2010.05.002

Hand, B., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *Journal of Writing Research, 16*(1), 88–106. https://doi.org/10.1016/j.jowr.2024.01.005

Jamshed, S., Ahmed, W., Sarfaraj, M., & Warda, M. (2024). The impact of ChatGPT on English language learners' writing skills: An assessment of AI feedback on mobile. *TESOL Journal, 15*(2), e00489. https://doi.org/10.1002/tesj.489

Kasih, E. N. E. W., & Putra, A. V. L. (2024). Artificial intelligence for literature class: Trends and attitude. In *2024 10th International Conference on Education and Technology (ICET)* (pp. 142–148). IEEE. https://doi.org/10.1109/icet60097.2024.103456

Li, A. W., Huang, Y., Wu, Y., & Whipple, M. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *System, 122*, 102878. https://doi.org/10.1016/j.system.2024.102878

p-ISSN: 2579-7549
e-ISSN: 2579-7263

Lin, T., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *Assessing Writing, 50*, 100617. https://doi.org/10.1016/j.asw.2024.100617

Liu, S., Hao, J., & Wang, Y. (2020). AI in EFL writing assessment: Validity and reliability evidence. *Language Testing in Asia, 10*(5), 45–63. https://doi.org/10.1186/s40468-020-00107-3

Ma, H., & Slater, T. (2016). Connecting Criterion scores and classroom grading contexts: A systemic functional linguistic model for teaching and assessing causal language. *Assessing Writing, 30*, 30–43. https://doi.org/10.1016/j.asw.2016.07.003

Mehdaoui, A. (2024). Unveiling Barriers and Challenges of AI Technology Integration in Education: Assessing Teachers' Perceptions, Readiness and Anticipated Resistance. *Futurity Education*, 4(4), 95–108. https://doi.org/10.57125/FED.2024.12.25.06

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Saleh, S., & Alshulbi, A. I. (2025). The role of techno-competence in AI-based assessments: Exploring its influence on students' boredom, self-esteem, and writing development. *Language Testing in Asia, 15*(6). https://doi.org/10.1186/s40468-025-00344-1

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Thwaites, T., Smith, J., & Browne, R. (2025). Reliability and validity issues in writing assessment: An updated perspective. *Assessing Writing, 62*, 100640. https://doi.org/10.1016/j.asw.2025.100640

p-ISSN: 2579-7549
e-ISSN: 2579-7263

Xu, X., Sun, F., & Hu, W. (2025). Integrating human expertise with GenAI: Insights into a collaborative feedback approach in translation education. *System, 129*, 103600. https://doi.org/10.1016/j.system.2025.103600

Zhang, Y., Chen, L., & Zhao, H. (2019). Validity and reliability of automated writing assessment in EFL contexts. *Language Testing, 36*(2), 123–142. https://doi.org/10.1177/0265532218758127