A Review of Artificial Intelligence in Language Assessment: Trends, Challenges, and Opportunities

Agung Susilo [1*], Masagus Firdaus [2]

[1] University of PGRI Palembang, Indonesia
[2] University of PGRI Palembang, Indonesia

*Corresponding author's email: susilo4gung@gmail.com

ARTICLE INFO	ABSTRACT
	The impact of AI on language testing has sparked a major upheaval in education thanks to advances in technology and the need for better testing efficiency. The objective of this review is to study and evaluate what is happening at present, what are the big challenges and where there are opportunities in AI-powered language assessment. The research systematically examines existing studies by employing the PRISMA framework, carrying out an extensive search through academic databases to locate, evaluate, and interpret peer-reviewed articles concerning AI in language assessment. It uses qualitative thematic analysis to integrate key patterns, obstacles, and future prospects. While AI has given us new methods for evaluating languages, problems related to bias, maintaining data privacy, and the importance of human role are still major issues. By covering many aspects, the review gives insights into the developing AI in language testing. It draws attention to how AI may improve accuracy, save time and allow for more assessment on condition that certain difficulties are addressed. According to the study, it is essential to put in a balance between using technology and relying on people for language assessment. By proposing recommendations for future research and practical applications, this review provides actionable insights for educators, policymakers, and developers to make informed decisions about integrating AI into language assessment. It emphasizes the importance of designing equitable, transparent, and pedagogically sound AI systems, thereby shaping the responsible and effective use of AI in educational contexts.

Introduction

Language assessment plays a vital role in educational systems globally, shaping curriculum development, teaching strategies, and student performance. Although traditional language assessment methods are fundamental, they often encounter ongoing issues such as scalability, subjectivity, time demands, and limited flexibility to accommodate diverse learner needs. To address these challenges, there has been a rising interest in recent years

in incorporating Artificial Intelligence (AI) into language assessment practices, which offers promising solutions through automation, personalization, and real-time analytics.

A growing body of research has explored various AI applications in educational assessment. Shermis and Burstein (2013) documented the effectiveness of automated essay scoring systems like *e-rater*, which can analyze grammatical accuracy, coherence, and lexical features with strong correlation to human raters. Similarly, studies by (Heffernan & Heffernan, 2014) examined intelligent tutoring systems (ITS) that offer adaptive, feedback-driven learning environments for language learners. More recently, (Guo et al., 2024) emphasized the value of AI-powered large-scale assessments, noting how AI tools can increase efficiency and reduce the costs associated with traditional scoring methods. In addition, AI-driven tools have shown potential in providing multimodal assessments (Chen et al., 2024), where language skills are evaluated through a combination of text, audio, and visual inputs to simulate real-world communication.

Despite these advances, several challenges persist. Existing literature tends to focus on the technical feasibility and scoring reliability of AI-based systems, often overlooking broader issues such as ethical concerns, data privacy, learner inclusivity, and the need for transparency in AI decision-making processes (Floridi et al., 2018). Moreover, most current research is either domain-specific (e.g., writing or speaking only) or lacks a comprehensive synthesis of how AI technologies intersect with pedagogical goals and educational policies in language assessment.

While individual studies have provided valuable insights into specific AI applications in assessment, there remains a lack of integrative reviews that holistically examine trends, implementation challenges, and opportunities for future research. There is also limited exploration of how AI-driven language assessment tools align with broader educational values such as fairness, accessibility, and learner agency. Furthermore, few studies offer practical recommendations for educators and policymakers navigating the ethical and pedagogical complexities of adopting AI in assessment context.

This review aims to address these gaps by systematically synthesizing recent research on AI in language assessment. It identifies key trends, evaluates prevailing challenges, and outlines opportunities for future research and practical implementation. In doing so, it contributes to a more informed and balanced discourse on the role of AI in shaping the future of language education and assessment.

Research Methods

The process of reviewing was conducted systematically in accordance with the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), which provide a structured and transparent framework for the identification, screening, and inclusion of relevant literature (Page et al., 2021).

Aiming to compile a comprehensive dataset, a methodical search was executed across three key academic databases: ERIC, JSTOR, and Google Scholar. This search was restricted to peer-reviewed journal articles published between 2010 and 2024. To ensure a broad and

relevant selection of studies, the search utilized the following keywords in various combinations: "Artificial Intelligence", "Language Assessment", "Automated Scoring", "AI in Education", "Natural Language Processing", "AI Feedback Tools", and "AI in Language Testing". Studies were considered for inclusion if they met these criteria: written in English, published in peer-reviewed academic journals, explicitly focused on AI applications in language assessment (whether empirical or theoretical), and published within the 2010 to 2024 timeframe. Studies were excluded if they: addressed only general educational AI without specific relevance to language assessment, were not peer-reviewed (such as blogs, editorials, or opinion pieces), or were duplicates of already selected records.

The search initially retrieved 178 articles. After removing 46 duplicates, 132 records remained for title and abstract screening. During this stage, 81 articles were excluded due to irrelevance or failure to meet inclusion criteria. A total of 51 articles were subjected to full-text analysis. Upon closer examination, 21 articles were excluded due to insufficient relevance to the review's core objectives. Finally, 30 articles were included for qualitative synthesis.

The article selection process is illustrated in Figure 1, following the PRISMA flowchart format.
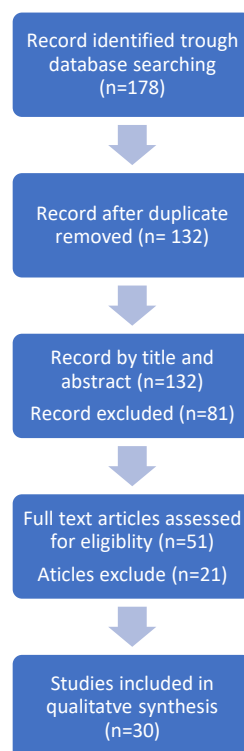

Figure 1. PRISMA chart of the review phase

A qualitative thematic analysis was conducted on the 30 selected articles to discern patterns and recurring themes. This analysis revealed three central themes: 1) Dominant trends in AI-driven language evaluation, 2) Implementation hurdles in educational and ethical frameworks, and 3) Potential avenues for future research and teaching applications.

This methodical process ensured that the review was anchored in high-quality, pertinent, and varied academic sources.

A comprehensive search procedure was employed to ensure the retrieval of pertinent academic sources across several scholarly databases, including ERIC, JSTOR, and Google Scholar. The search process utilized a combination of controlled vocabulary and keyword terms such as "artificial intelligence," "language assessment," "automated scoring," and "AI in education" to capture a wide scope of studies addressing AI in language testing contexts. To ensure the quality and relevance of the included studies, the following inclusion criteria were established: Peer-reviewed journal articles published between 2010 and 2024 and studies with a clear focus on AI applications in language assessment, including both empirical research and conceptual discussions.

The initial search yielded a substantial number of records. Then, the records were first screened by titles and abstracts to assess their relevance. The exclusions administered to articles were duplicates, lacked a direct focus on AI, or did not address aspects of language assessment. The remaining articles were subjected to full-text review to ensure alignment with the inclusion criteria. A final set of eligible studies was selected for data analysis. These articles were examined using qualitative thematic analysis, which facilitated the synthesis of findings into three main thematic categories: 1) prevailing trends in AI applications for language assessment, 2) implementation challenges, and 3) future opportunities for integrating AI into language testing.

Findings

Trends in AI-Driven Language Assessment

AI in language assessment is increasingly popular. Tools like the Intelligent Essay Assessor (IEA) and e-rater have demonstrated strong capabilities in automatically scoring student essays (Shermis & Burstein, 2013). These systems leverage NLP algorithms to analyse grammar, coherence, and persuasiveness, with results showing high correlation to human raters. Furthermore, adaptive tests such as the Duolingo English Test personalize item difficulty based on real-time performance, enhancing test validity and user experience (Duolingo, 2020).

Five major developments characterize the integration of AI within the language assessment:

1) Automated Scoring Systems: AI, particularly NLP, analyses written and spoken responses by assessing syntax, vocabulary, and discourse coherence, providing immediate, consistent feedback (Merkviladze, 2024).
2) Intelligent Tutoring Systems (ITS): AI adapts to learner needs, offering personalized exercises and instant feedback to enhance both writing and speaking skills (Heffernan & Heffernan, 2014).

3) Multimodal Assessment: Incorporating text, audio, and visual inputs allows for more holistic language evaluation, simulating authentic communication scenarios (Chen et al., 2024).

4) Machine Translation and Speech Recognition: Tools like Google Translate and speech-to-text apps support language learners by providing translation, pronunciation checks, and feedback (Lai & Zheng, 2018).

5) Virtual Reality and Augmented Reality (VR/AR): Immersive environments provide contextual language practice, boosting engagement and practical learning (Godwin-Jones, 2024).

Challenges in Implementation

Besides its benefits, AI integration in language assessment encountering several obstacles. The obstacles vary as:

1) Bias and Fairness: Algorithms of AI may arise racial, gender, or socioeconomic biases present in their training data, which can skew results and create inequity (Blodgett et al., 2020).

2) Explainability and Transparency issue: A great number of AI systems operate as "black boxes," lacking clarity in their decision-making processes, which undermines trust and accountability (Selwyn, 2019).

3) Data Privacy: AI systems mostly require demand to sensitive learner data, raising concerns over data protection and ethical use (Jose, 2024).

4) Technical Limitations: AI tools struggle with non-standard language features such as idioms, dialects, and higher-order thinking skills like creativity (Floridi et al., 2018).

5) Reliability and Validity: While some AI scoring aligns with human evaluation, questions remain about its consistency across complex language domains (Ifenthaler, 2022).

6) Integration into Educational Systems: Implementing AI tools requires compatibility with existing infrastructures and alignment with curricular goals (Luckin et al., 2016).

7) Infrastructure and Access: Unequal access to AI-powered technologies is hampered by regional technological differences (UNESCO, 2021).

8) Teacher Training and Acceptance: Effective use of AI depends on teachers' readiness and willingness to engage with emerging technologies (Zawacki-Richter et al., 2019).

9) Academic Integrity: Advanced generative AI makes it harder to detect plagiarism or misuse, posing challenges for maintaining assessment credibility (Cotton et al., 2024).

10) Cost and Maintenance: The development and upkeep of AI systems demand significant financial and technical investment (Cottier et al., 2024).

Summary Table of Challenges

Challenge	Description

Bias and Fairness	Risk of perpetuating bias from training data
Transparency	Lack of interpretability in AI-generated decisions
Data Privacy	Risk of misuse of student data
Technical Limitations	Inability to process idioms, creativity, or non-native speech
Reliability and Validity	Concerns about AI's scoring accuracy
Integration	Difficulty in merging AI into existing systems
Infrastructure and Access	Limited access to resources in underfunded areas
Teacher Training	Need for professional development
Academic Integrity	Risk of cheating using AI tools
Cost and Maintenance	High setup and operational costs

Addressing these challenges requires a collaborative, ethical, and data-driven strategy, combining technological innovation with educational best practices to ensure AI-enhanced language assessment is equitable and effective.

Opportunities for Future Research

AI presents numerous avenues for further exploration. Integrating AI with traditional assessment can identify best practices for blended evaluation (Owan et al., 2023). AI-based assessments offer adaptive, personalized learning pathways and open up new research into efficacy and learner engagement (Kukulska-Hulme, 2020). Future studies should also focus on ethical AI design, especially concerning fairness, data governance, and accountability (Floridi et al., 2018). Longitudinal studies tracking learning outcomes over time would be valuable in evaluating the sustained impact of AI in assessments.

Key areas for exploration include:

1) Scalable, automated analytics to personalize instruction (Chen et al., 2024).
2) Multilingual NLP tools for minoritized languages (Gutiérrez, 2023).
3) Explainable AI to foster trust through transparency (Selwyn, 2019).
4) Deep learning for nuanced and equitable assessment models (Borna et al., 2024).

Discussion

The application of artificial intelligence (AI) in language testing has brought to light new trends, major challenges, and promising opportunities. These elements are interconnected and play a crucial role in how AI is implemented, perceived, and refined in educational environments.

Trends

An emerging trend in language assessment is the growing reliance on automation and customization. AI technologies like e-rater and IEA have demonstrated a high degree of alignment with human evaluators, significantly boosting the efficiency and reliability of scoring processes (Shermis & Burstein, 2013). Additionally, the advent of adaptive testing

systems, exemplified by the Duolingo English Test, represents a move towards real-time adjustment of test difficulty in response to learner input, thereby enhancing both test validity and user experience (Duolingo, 2020). The integration of multimodal assessment, which combines visual, auditory, and textual components, indicates a shift towards a more holistic evaluation of communicative skills (Chen et al., 2024). Moreover, the increasing use of virtual and augmented reality (VR/AR) environments provides immersive and contextually rich language learning opportunities. Collectively, these developments signify a shift from traditional, uniform assessments to more dynamic, learner-centered approaches.

Challenges

Despite technological advancements, numerous obstacles still hinder the broad and fair implementation of AI in language:

- **Bias and Fairness**: AI systems trained on biased datasets can perpetuate systemic inequities, particularly affecting underrepresented groups (Blodgett et al., 2020).
- **Transparency and Explainability**: Many AI tools function as "black boxes," lacking interpretability, which undermines user trust and limits pedagogical insight (Selwyn, 2019).
- **Privacy and Data Governance**: AI systems demand large volumes of sensitive learner data, raising ethical and legal concerns around data protection (Jose, 2024).
- **Technical Limitations**: Current AI struggles with understanding idiomatic expressions, creativity, and dialectal variations, which are essential in nuanced language assessment (Floridi et al., 2018).
- **Implementation Barriers**: Integrating AI into curricula requires infrastructural readiness, teacher training, and ongoing support, especially in under-resourced regions (UNESCO, 2021; Zawacki-Richter et al., 2019).
- **Academic Integrity**: The rise of generative AI complicates plagiarism detection and authenticity verification in assessments (Cotton et al., 2024).

These issues indicate that while AI brings efficiencies, its deployment must be critically managed to ensure fairness, validity, and trustworthiness.

Opportunities

Amidst these challenges, significant opportunities arise for both research and practice:

- **Hybrid Assessment Models**: Combining human judgment with AI-driven tools can yield more balanced and robust evaluation systems that preserve pedagogical context while benefiting from technological precision (Zhang & Aslan, 2021).
- **Ethical AI Design**: There is a growing need for explainable and equitable AI systems that align with educational values such as inclusivity and learner agency (Floridi et al., 2018).
- **Personalized Learning Analytics**: AI can generate fine-grained insights into learner performance, enabling tailored instruction and targeted interventions (Hooda et al., 2022).

- **Support for Multilingual and Minoritized Languages**: Research into multilingual NLP can promote linguistic diversity and support learners from a range of backgrounds (Gutiérrez, 2023).
- **Longitudinal Impact Studies**: Future research should explore the long-term educational effects of AI-driven assessments on learning outcomes, motivation, and teacher practices.

These opportunities suggest that with the right frameworks and collaborations, AI can significantly enhance the quality and equity of language assessment.

Conclusion

This review emphasizes that AI is transforming language assessment by incorporating automation, personalization, and multimodal evaluation, which enhances efficiency and scalability while often aligning with human scoring. Nonetheless, issues such as algorithmic bias, transparency deficits, data privacy concerns, and infrastructural challenges highlight the necessity for careful and ethical implementation. The study suggests that incorporating AI into language assessment should be balanced with human oversight, educator training, and equitable access to ensure fairness and educational relevance. As AI tools become more integrated into assessment practices, future research should concentrate on creating explainable AI systems, assessing long-term learning effects, and broadening support for diverse linguistic and cultural contexts to foster responsible and inclusive language education.

References

Blodgett, S. L., Barocas, S., Daumé, H., & Wallach, H. (2020). *Language (Technology) is Power: A Critical Survey of "Bias" in NLP*. http://arxiv.org/abs/2005.14050

Borna, M. R., Saadat, H., Hojjati, A. T., & Akbari, E. (2024). Analyzing click data with AI: implications for student performance prediction and learning assessment. *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1421479

Bulut, O., & Beiting-Parrish, M. (2024). The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, *5*(3). https://doi.org/10.59863/MIQL7785

Chen, C., Tang, M., & Yang, A. (2024). Progress in a multimodal learning evaluation based on artificial intelligence. *Frontiers in Educational Research*, *7*(7). https://doi.org/10.25236/fer.2024.070704

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). *The rising costs of training frontier AI models*. http://arxiv.org/abs/2405.21015

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Godwin-Jones, R. (2024). *AI and VR converge: The future of language learning in an emerging metaverse*.

Guo, H., Johnson, M. S., Ercikan, K., Saldivia, L., & Worthington, M. (2024). Large-Scale Assessments for Learning: A Human-Centred AI Approach to Contextualizing Test Performance. *Journal of Learning Analytics*, *11*(2), 229–245. https://doi.org/10.18608/jla.2024.8007

Gutiérrez, L. (2023). Artificial Intelligence in Language Education: Navigating the Potential and Challenges of Chatbots and NLP. *Research Studies in English Language Teaching and Learning*, *1*(3), 180–191. https://doi.org/10.62583/rseltl.v1i3.44

Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S. (2022). Artificial Intelligence for Assessment and Feedback to Enhance Student Success in Higher Education. *Mathematical Problems in Engineering*, *2022*. https://doi.org/10.1155/2022/5215722

Ifenthaler, D. (2022). Automated Essay Scoring Systems. In *Handbook of Open, Distance and Digital Education* (pp. 1–15). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0351-9_59-1

Jose, D. (2024). Data Privacy and Security Concerns in AI-Integrated Educational Platforms. *Recent Trends in Management and Commerce*, *5*(2), 87–91. https://doi.org/10.46632/rmc/5/2/19

Lai, C., & Zheng, D. (2018). Self-directed use of mobile devices for language learning beyond the classroom. *ReCALL*, *30*(3), 299–318. https://doi.org/10.1017/S0958344017000258

Luckin, R., Holmes, W., Griffiths, M., & Pearson, L. B. F. (2016). *Intelligence Unleashed An argument for AI in Education*.

Merkviladze, M. (2024). Integration of Automated Feedback Tools in EFL Academic Writing Classes: Teachers' Perspective. In *Journal of Education in Black Sea Region* (Vol. 10).

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. In *Eurasia Journal of Mathematics, Science and Technology Education* (Vol. 19, Issue 8). Modestum LTD. https://doi.org/10.29333/ejmste/13428

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., … Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. In *The BMJ* (Vol. 372). BMJ Publishing Group. https://doi.org/10.1136/bmj.n160

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

Wei, L. (2023). Artificial intelligence in language instruction: impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1261955

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? In *International Journal of Educational Technology in Higher Education* (Vol. 16, Issue 1). Springer Netherlands. https://doi.org/10.1186/s41239-019-0171-0

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. In *Computers and Education: Artificial Intelligence* (Vol. 2). Elsevier B.V. https://doi.org/10.1016/j.caeai.2021.100025