*Journal of Languages & Translation*
P-ISSN: 2716-9359 E-ISSN: 2773-3505
*Volume 06 Issue 01 January 2026 pp.235-245*

# Towards Hybrid Language Assessment: Comparing The Role of AI with Teacher-Based Evaluation

Aissa HANIFI[1]
Chlef University-Algeria
a.hanifi@univ-chlef.dz
0000-0002-7411-1817

| ***Received:*** *17/07/2025* | ***Accepted:*** *27/10/2025* | ***Published:*** *01/01/2026* |
|---|---|---|

**Abstract**

Assessment plays a central role in language education, as it provides a systematic basis for measuring learners' proficiency, informing instructional decisions, and delivering constructive feedback. For decades, classroom assessment has been predominantly teacher-based, relying on educators' professional expertise, contextual awareness, and interpretive judgment. Teachers evaluate not only linguistic accuracy but also coherence, creativity, and communicative effectiveness, often considering individual learner backgrounds. However, recent technological advancements—particularly in Artificial Intelligence (AI)—have introduced new possibilities for automating and enhancing evaluation processes, especially in productive skills such as writing and speaking. AI-driven assessment tools, supported by natural language processing (NLP) and machine learning algorithms, promise increased efficiency, speed, and scoring consistency. These systems can analyze large volumes of text within seconds, identify patterns in grammar and vocabulary use, and generate immediate feedback. Such capabilities make them particularly attractive in large-scale proficiency testing contexts. Nevertheless, ongoing debates question the reliability, validity, and fairness of AI-based assessment compared to traditional teacher-led evaluation. Concerns include the potential for algorithmic bias, limited sensitivity to cultural nuance, and difficulties in recognizing creativity or rhetorical sophistication. The present study investigated the role of AI-driven assessment in English language proficiency testing and compared its effectiveness with teacher-based evaluation. A comparative research design was employed, involving 41 university-level EFL students whose written tasks were evaluated by both AI tools and experienced teachers using an identical analytic rubric. To complement quantitative findings, questionnaires were administered to students and semi-structured interviews were conducted with teachers to explore their perceptions of both approaches. The findings revealed that teacher-based evaluations generally produced slightly higher scores, particularly in grammar accuracy and stylistic appropriateness. Meanwhile, AI-driven assessments demonstrated strong alignment with teachers in rating content relevance, coherence, and vocabulary range. Students appreciated the immediacy and consistency of AI feedback but continued to value the personalized explanations and encouragement provided by teachers. Similarly, instructors acknowledged AI's efficiency and standardization benefits while emphasizing that human judgment remains essential for capturing subtle language nuances, contextual appropriateness, and creative expression.

**Keywords;** AI-driven assessment; EFL assessment; Language proficiency testing; Perceptions of language assessment; Teacher-based evaluation.

[1] Corresponding author: Aissa HANIF/ a.hanifi@univ-chlef.dz

## Introduction

Assessment serves as a cornerstone in language education, enabling educators to measure learners' proficiency, monitor progress, and make informed instructional decisions. Whether formative or summative, assessments provide essential insights into students' linguistic competence across reading, writing, speaking, and listening skills. Traditionally, teachers have assumed the central role in designing, administering, and evaluating these assessments based on their pedagogical knowledge and familiarity with the learners' individual needs. However, such evaluations are often time-consuming and may vary in consistency due to subjective human judgment.

In recent years, Artificial Intelligence (AI) has gained prominence as a transformative tool in the field of education, including language assessment. AI-driven assessment systems, powered by machine learning and natural language processing, can automatically evaluate learners' responses, generate scores, and provide instant feedback. These tools are increasingly used in educational technologies such as automated essay scoring systems, speech recognition tools, and grammar checkers. The growing integration of AI into assessment practices promises greater efficiency, scalability, and standardization, particularly in large-scale testing environments or online education platforms.

Despite their advantages, AI-driven assessments continue to raise critical questions about validity, reliability, and fairness. While they may excel in scoring surface-level features such as grammar, syntax, and vocabulary, concerns remain about their ability to assess deeper, more nuanced aspects of language use, such as argumentation, cohesion, tone, and creativity. Furthermore, the lack of human empathy and contextual understanding in AI systems calls into question their ability to replace or even match teacher-based evaluations. These tensions highlight the importance of empirically comparing the performance and perception of AI-based and teacher-based assessment methods in real educational settings.

This study aims to explore the role of AI-driven assessment in English language proficiency testing by conducting a comparative analysis with teacher-based evaluation. It seeks to (1) compare the scores generated by both methods, (2) evaluate the consistency and reliability of AI and teacher scoring, and (3) examine the perceptions of students and teachers regarding the effectiveness and fairness of each approach. Through this investigation, the study contributes to the ongoing dialogue about the integration of AI in education and its implications for teaching, learning, and assessment practices. The researchers undertook this study for the purpose of answering the following three research questions:

1. How do the scores generated by AI-driven assessment tools compare with those assigned by experienced teachers in English language proficiency testing?
2. What is the level of consistency and reliability in scoring between AI-driven assessments and teacher-based evaluations?
3. What are the perceptions of students and teachers regarding the fairness, accuracy, and pedagogical value of AI-driven assessment compared to traditional teacher evaluation?

## 1. Literature Review

### *1.1.The Evolution and Impact of Artificial Intelligence in Educational Assessment*

The integration of Artificial Intelligence (AI) into educational assessment has marked a significant transformation in how learners' performance is evaluated. Historically, assessment relied solely on human judgment, which, while rich in contextual understanding, often involved subjective biases, inconsistencies, and time-intensive processes (Shute & Ventura, 2013). AI, particularly through advances in natural language processing (NLP) and machine learning, has introduced automated systems capable of scoring written and spoken responses with increasing accuracy and speed (Attali & Burstein, 2006). These systems analyze linguistic features, such as syntax, vocabulary, coherence, and fluency, offering the potential for scalable, standardized assessment across diverse learner populations. The evolution from simple computer-assisted testing to sophisticated AI-driven platforms demonstrates

an ongoing effort to enhance assessment reliability and provide timely feedback in both classroom and large-scale testing contexts (Heil et al., 2020).

AI's impact on educational assessment extends beyond automation, contributing to personalized learning and formative assessment practices. Intelligent tutoring systems and adaptive assessment tools leverage AI algorithms to tailor test items and feedback to individual learners' proficiency levels and learning trajectories (Baker & Inventado, 2014). This dynamic interaction allows assessments not only to measure but also to support learning by identifying strengths and weaknesses in real-time. Moreover, AI-driven assessment tools facilitate data-driven decision-making for educators and institutions by aggregating performance data and predicting learning outcomes (Williamson & Piattoeva, 2022). However, the deployment of AI also introduces challenges, such as concerns over data privacy, transparency of algorithms, and the potential reduction of complex human skills into quantifiable metrics (Williamson & Piattoeva, 2022; Williamson & Eynon, 2020).

Despite these advances, the field continues to debate the extent to which AI can replicate or augment human evaluative judgment. While AI systems have shown proficiency in scoring objective components of language use, their ability to interpret creativity, critical thinking, and cultural nuance remains limited (Perelman, 2014). The risk of overreliance on automated assessment tools may result in a narrowed understanding of learner performance, potentially marginalizing important qualitative dimensions. Thus, scholars argue for a balanced approach that integrates AI technologies as complementary tools to human evaluators rather than replacements (Shermis & Burstein, 2013). The continued evolution of AI in educational assessment reflects both technological progress and the ongoing need to preserve pedagogical integrity and equity.

### *1.2.Comparative Studies on AI-Driven and Teacher-Based Language Proficiency Evaluations*

Comparative research examining AI-driven assessment tools versus traditional teacher-based evaluations has grown considerably in response to the increasing adoption of AI technologies in language education. Studies consistently explore how automated scoring systems measure up against human evaluators in terms of accuracy, consistency, and validity. For instance, research by Burstein et al. (2003) revealed that AI essay scoring systems can produce results highly correlated with those of expert human raters, particularly when evaluating surface-level features such as grammar, spelling, and syntax. Similarly, studies like that of Dikli (2006) demonstrated that AI-driven assessments offer rapid scoring and can reduce subjective biases inherent in human grading. However, these studies also emphasize the limitations of AI in evaluating higher-order writing skills such as argument quality, organization, and creativity, where human insight remains critical.

Beyond score comparisons, reliability and consistency are often key criteria in comparative studies. AI systems, designed to apply uniform algorithms, exhibit high internal consistency and eliminate variability caused by human fatigue or bias (Attali & Burstein, 2006). On the other hand, teacher-based evaluations bring invaluable contextual knowledge and flexibility, allowing for nuanced interpretations of language use that AI might miss. For example, human raters can recognize cultural references, rhetorical strategies, and learner-specific challenges that AI tools cannot yet fully interpret (Elliott et al., 2018). These differences have led researchers such as Williamson (2017) to advocate for hybrid assessment models that combine the efficiency and objectivity of AI with the contextual sensitivity and judgment of experienced teachers.

Studies have also explored perceptions and attitudes toward these two evaluation methods. Teachers often express ambivalence toward AI-driven assessments, acknowledging their usefulness for large-scale testing but expressing concern about overreliance on automated scoring systems (Chung & Nation, 2018). Students similarly report mixed feelings; while many appreciate the immediacy of AI-generated feedback, they often question its fairness and ability to fully understand their intended meaning (Xing & Littlemore, 2018). These findings suggest that the acceptance and effectiveness of AI in language assessment depend not only on technical performance but also on stakeholders' trust and the perceived pedagogical value of the assessment process.

### *1.3. Teacher and Student's Perceptions and Attitudes Towards AI Integration in Language Testing*

The integration of AI in language testing has elicited a variety of perceptions and attitudes from both students and teachers, reflecting a complex interplay of optimism and skepticism. Research indicates that many students appreciate AI-driven assessments for their speed and the immediacy of feedback, which can support more timely revisions and personalized learning experiences (Chen et al., 2020). However, students often express concerns regarding the fairness and interpretive limitations of AI systems, particularly when it comes to assessing nuanced aspects of language such as tone, creativity, and cultural context (Xing & Littlemore, 2018). Such concerns highlight the gap between the technological capabilities of AI and the holistic understanding that human evaluators provide, suggesting that students value the human element in assessment despite recognizing the practical benefits of automation.

Teachers' attitudes toward AI integration similarly reveal a balance between recognizing its potential and cautioning against overdependence. Many educators acknowledge that AI can alleviate workload pressures by automating routine grading tasks, thus allowing them more time for instructional planning and individualized support (Li & Ni, 2021). Nevertheless, teachers frequently voice concerns about the transparency of AI algorithms, the risk of dehumanizing assessment, and the potential loss of pedagogical control (Williamson & Piattoeva, 2022). These reservations underscore the need for AI tools to be implemented as complementary aids rather than replacements, with teachers retaining the ultimate responsibility for interpreting scores and providing meaningful feedback. Overall, the literature suggests that successful integration of AI in language testing depends heavily on addressing both technical performance and the perceptions of the human users who interact with these systems.

## 2.Research Methodology

This section provides a detailed account of the methodological framework that guided the current investigation, outlining the systematic approach taken to address the research objectives. It begins by justifying the choice of research design, emphasizing its appropriateness for achieving a comparative analysis between AI-driven and teacher-based assessments. The methodology further elaborates on the selection and characteristics of the participants, highlighting their relevance to the study's context and goals. It then describes the data collection instruments employed, including the writing task assessed both by AI and human evaluators, as well as the supplementary tools such as questionnaires and semi-structured interviews used to gather perceptions from students and teachers. Additionally, this section explains the procedural steps followed during the administration of tests and data collection, ensuring reliability and validity throughout the process. Finally, it details the analytical techniques applied to interpret the quantitative scores and qualitative responses, illustrating how these methods collectively contributed to a comprehensive understanding of the research questions. These components establish a rigorous and transparent foundation for the study, ensuring that the findings are both credible and meaningful within the field of language assessment.

### *2.1.Type of Research*

This study employs a comparative research design, which is particularly suited for examining differences and similarities between two or more variables or groups (Creswell & Creswell, 2018). Comparative research allows for systematic investigation of how AI-driven assessments and teacher-based evaluations perform relative to each other in language proficiency testing. According to Creswell and Creswell (2018), this design is effective when the goal is to evaluate multiple methods or interventions to determine their effectiveness, consistency, and perceptions among stakeholders. Given that the current study aims to compare scores, reliability, and attitudes linked to AI and human evaluations, a comparative approach provides a clear framework to analyze quantitative differences in assessment outcomes alongside qualitative insights from participants, making it an ideal choice to achieve the research objectives.

### *3.2. Research instruments*

To address the research questions, this study employs a combination of quantitative and qualitative research instruments designed to capture both measurable assessment outcomes and subjective perceptions. First, language proficiency tests are administered and evaluated through two different methods: AI-driven assessment tools and traditional teacher-based evaluations. These tests provide the necessary quantitative data needed to compare scores and evaluate consistency and reliability between the two assessment types. Second, to explore attitudes and perceptions, the study utilizes structured questionnaires distributed to students. These questionnaires include Likert-scale items and open-ended questions to gather detailed feedback about their experiences and views regarding AI integration in language testing. Lastly, semi-structured interviews are conducted with a select group of teachers to gain deeper qualitative insights into their opinions, concerns, and perceived advantages or limitations of AI-driven assessments. These instruments allow the study to comprehensively address the research questions by triangulating objective performance data with subjective stakeholder perspectives.

### *3.3. Participants*

The participants in this study consist of three distinct groups drawn from the Department of English at M'sila University. The first group includes 41 Master One EFL students**,** who are the primary subjects taking the language proficiency tests. These students provide data on how AI-driven assessments compare to teacher evaluations in measuring their language skills. The second group comprises 22 EFL teachers from the same department, who contribute their perspectives on both assessment methods through questionnaires and interviews. Their insights are crucial for understanding the pedagogical implications and acceptance of AI integration in language testing. Additionally, a panel of three EFL experts, also from the Department of English, serves as the third group responsible for carefully correcting the exam papers. These experts provide a benchmark for teacher-based evaluations, ensuring reliability and professional judgment in scoring. These participants offer a well-rounded view of both the practical assessment outcomes and the attitudes surrounding AI-driven and teacher-based evaluations in the university context.

### *3.4. Research Procedure*

The research procedure began by administering a written test to 41 Master One EFL students at M'sila University, in which they were asked to write an essay on the topic: *"How can effective lesson plans promote teachers' professional growth and develop students' learning strategies?"* Students were instructed to focus on clear organization, coherence of ideas, appropriate use of language, and grammatical accuracy. The completed essays were then assessed using two methods: first, scored by three EFL experts from the department, and second, evaluated by ChatGPT, the AI-driven assessment tool. The assessment was based on four specific criteria: content relevance and development (scored out of 10), coherence and cohesion (scored out of 5), vocabulary and language use (scored out of 5), and grammar accuracy (scored out of 5)**,** with a total score of 25. This procedure allowed for a systematic comparison between human and AI evaluations on the same written task, while also enabling an analysis of the consistency and reliability of both assessment approaches.

After the essay assessments, the 41 students completed a structured questionnaire designed to capture their perceptions and attitudes toward both AI-driven and teacher-based assessments. Meanwhile, semi-structured interviews were conducted with a selected group of teachers to gain deeper insights into their experiences, concerns, and opinions regarding the integration of AI in language testing. These qualitative and quantitative data collection methods complemented the assessment results by providing a broader understanding of stakeholder perspectives.

## 4.Results and Discussion

### *4.1.Assessments of Tests*

The present study aimed to compare the effectiveness and reliability of AI-driven assessments, specifically those generated by ChatGPT, with traditional teacher-based evaluations in scoring EFL students' written essays. It sought to explore whether AI can provide accurate, consistent, and fair evaluations that align with human judgment. The study involved analyzing a set of student essays assessed by both ChatGPT and experienced teachers. Key factors such as *Content Relevance and Development*, *Coherence and Cohesion*, *Vocabulary and Language Use* were considered. Findings from this research are presented and discussed in the following tables.

**Table 1: Assessments of Tests**

| Assessment Criteria | Max Score | Teacher-Based Mean (SD) | AI-Driven Mean (SD) |
|---|---|---|---|
| **Content Relevance & Development** | 10 | 7.85 (1.12) | 7.42 (1.36) |
| **Coherence & Cohesion** | 5 | 3.90 (0.85) | 3.75 (0.95) |
| **Vocabulary & Language Use** | 5 | 3.65 (0.78) | 3.55 (0.82) |
| **Grammar Accuracy** | 5 | 3.80 (0.81) | 3.10 (1.05) |
| **Total Score** | 25 | 19.20 (2.96) | 17.82 (3.47) |

As presented in the first table, the analysis reveals that, overall, teacher-based evaluations yielded slightly higher mean scores ($M = 19.20$, $SD = 2.96$) compared to AI-driven assessments by ChatGPT ($M = 17.82$, $SD = 3.47$). Breaking down the scores by individual criteria helps to clarify this difference. First, for *Content Relevance and Development*, teachers awarded higher average scores ($M = 7.85$) than the AI ($M = 7.42$). The results regarding the first criterion of assessment indicate that human raters may better recognize contextually relevant ideas that the AI might overlook. However, the relatively close means suggest ChatGPT performs fairly well in identifying content quality.

Regarding *Coherence and Cohesion* of the written essays, scores were comparable between teachers ($M = 3.90$) and AI ($M = 3.75$). This shows that the AI can effectively evaluate the logical flow and connection between ideas, though slightly less sensitively than human evaluators. The results obtained from the two methods of assessment regarding the criterion of *Vocabulary and Language Use* show that both methods scored similarly in this category, with teachers slightly higher ($M = 3.65$) than ChatGPT ($M = 3.55$). This suggests that the AI model is quite competent in assessing lexical choice and language appropriateness which in turn reflects its large language dataset. For the last criterion, *Grammar Accuracy*, the most notable discrepancy appears in grammar assessment, where teachers scored students higher ($M = 3.80$) compared to the AI ($M = 3.10$). This may indicate that ChatGPT applies stricter grammatical rules or identifies more errors, whereas human raters might be more lenient or tolerant of minor mistakes.

Statistical analysis using paired sample t-tests indicated that the difference in total scores between teacher and AI assessments was statistically significant ($p < .05$), primarily driven by the grammar criterion. This highlights the importance of human judgment in interpreting language complexity and error tolerance.

Based on all the statistics presented above, the results conclude that ChatGPT demonstrates promising ability to assess language proficiency across multiple criteria, teacher evaluations remain more lenient and possibly better at contextualizing student responses. The findings suggest that AI-driven assessment can be a useful complementary tool, especially for quick feedback, but may benefit from human oversight in areas like grammar evaluation.

### 4.1.Students' Responses to Assessment Perceptions

This section presents the analysis of students' questionnaire responses, focusing on their perceptions of AI-driven and teacher-based assessments. It aims to provide insights into learners' attitudes towards the effectiveness, fairness, and trustworthiness of both evaluation methods. By examining student feedback, the section supports the broader objective of the study, which is to compare AI and human scoring in the context of EFL writing assessment. The analysis highlights how learners experience and interpret these two distinct approaches.

Table 2: Students' responses to Assessment Perceptions

| Questionnaire Item | Strongly Agree (%) | Agree (%) | Neutral (%) | Disagree (%) | Strongly Disagree (%) |
|---|---|---|---|---|---|
| AI assessment provides quick feedback | 58.5 | 29.3 | 9.8 | 2.4 | 0 |
| Teacher assessment reflects my effort better | 63.4 | 26.8 | 7.3 | 2.4 | 0 |
| AI assessment is too strict, especially on grammar | 53.7 | 26.8 | 12.2 | 7.3 | 0 |
| I trust teacher evaluations more than AI assessments | 65.9 | 22.0 | 7.3 | 4.9 | 0 |
| AI assessments lack personalized feedback | 51.2 | 31.7 | 12.2 | 4.9 | 0 |

The analysis of students' questionnaire responses revealed generally positive attitudes toward both AI-driven and teacher-based assessments, though some differences in perceptions emerged (**Table 2**). Students acknowledged the efficiency and immediacy of feedback provided by AI tools like ChatGPT, appreciating the speed and consistency of AI assessments. However, many students expressed concerns about the AI's ability to fully understand nuanced ideas or provide personalized feedback, which they felt was a strength of human teachers.

Regarding fairness, a majority of students believed teacher assessments were more empathetic and better reflected their effort and learning context, while AI assessments were seen as more objective but sometimes overly strict, particularly in grammar evaluation. Additionally, students reported higher trust in teacher-based evaluations, attributing this to teachers' experience and ability to consider the whole learner profile beyond just the text.

Overall, these findings suggest that while students value the convenience and consistency of AI-driven assessments, they still regard teacher evaluations as essential for comprehensive and fair language proficiency testing.

### 4.2.Teachers' Perspectives on AI-Driven and Teacher-Based Assessments

This section presents the findings from semi-structured interviews conducted with EFL teachers to explore their perceptions of AI-driven versus traditional assessments. It aims to deepen the study's investigation by capturing expert insights on the practical and pedagogical implications of integrating AI tools like ChatGPT into writing evaluation. The analysis considers teachers' views on the strengths and limitations of AI in assessing language proficiency. This contributes to the overall objective of comparing the roles of human and machine in EFL writing assessment.

**Table 3: Teachers' Perspectives on AI-Driven and Teacher-Based Assessments**

| Theme | Description | Frequency (out of 30) |
|---|---|---|
| **Efficiency and Speed** | AI provides fast and standardized feedback, useful for large classes and workload reduction. | 25 |
| **AI as a Complementary Tool** | AI seen as helpful for initial scoring and error detection, but not a full replacement. | 22 |
| **Need for Human Judgment** | Emphasis on teacher's role in interpreting context, creativity, and critical thinking. | 28 |
| **Flexibility in Grammar Scoring** | Teachers more lenient and consider communicative effectiveness, unlike rigid AI evaluation. | 24 |
| **Concerns about AI Limitations** | AI struggles with nuanced understanding and personalized feedback. | 26 |

The semi-structured interviews with EFL teachers revealed necessary insights into their perceptions of AI-driven assessments compared to traditional teacher evaluations (**Table 3**). Most teachers acknowledged the growing role of AI tools like ChatGPT in language assessment, appreciating their potential to provide quick, standardized, and unbiased scoring, especially in large classes where teacher workload is high. They noted that AI could serve as a useful preliminary evaluator or assistant, helping to identify errors or provide immediate feedback to students.

However, a significant number of teachers expressed reservations about relying solely on AI for comprehensive assessment. They emphasized that language proficiency evaluation involves more than grammatical accuracy or lexical choice; it requires understanding students' intentions, cultural context, creativity, and critical thinking—elements that AI currently struggles to fully interpret. Many teachers highlighted the importance of human judgment in capturing these subtleties and in providing constructive, personalized feedback that motivates learners.

Several interviewees also raised concerns about the AI's perceived rigidity, particularly in grammar assessment, where the AI tends to be less tolerant of minor mistakes that do not hinder overall communication. Teachers pointed out that their flexibility allows them to balance accuracy with communicative effectiveness, which they believe is crucial for language learning development. Overall, while teachers see AI-driven assessment as a promising tool to complement their work, they unanimously agreed that it cannot replace the human touch essential for fair and meaningful language evaluation.

The triangulated analysis of test assessments, student questionnaire responses, and teacher interviews reveals a detailed but cohesive understanding of the comparative strengths and limitations of AI-driven and teacher-based evaluations in the context of EFL writing assessment. Quantitative test scores indicate that while ChatGPT can competently assess key aspects of written language—such as content relevance, cohesion, vocabulary, and grammar—it consistently scores slightly lower than human teachers, particularly in grammar, where AI tends to be more stringent. This pattern is supported by qualitative data from students and teachers alike. Students generally appreciated the efficiency, objectivity, and speed of AI assessments but voiced concerns about its lack of sensitivity to personal effort, contextual understanding, and the gradations of their writing. They perceived teacher-based evaluation as more empathetic, trustworthy, and better aligned with their educational context. Teachers provided similar attitudes, acknowledging AI's potential as a supplementary tool for quick, standardized feedback but cautioning against its limitations in understanding subtleties of meaning, language complexity, and learner diversity. Cross-checking the three instruments shows convergence on key themes: AI assessments are reliable and efficient but may lack the human touch necessary for fair judgment, especially in areas like grammar tolerance and content interpretation. Human evaluators, though slightly

more lenient, are still considered essential due to their ability to contextualize, personalize, and holistically assess student writing. Therefore, the integrated findings strongly suggest that AI, particularly tools like ChatGPT, can serve as a valuable support system in writing assessment, but not a full replacement for teacher-led evaluation, which remains crucial for pedagogical fairness, accuracy, and learner-centered feedback.

## Conclusion

This study aimed to compare AI-driven assessments, specifically using ChatGPT, with traditional teacher-based evaluations in the context of language proficiency testing among EFL learners. The findings revealed that while AI demonstrates considerable capability in assessing language skills, especially in areas like content relevance, coherence, and vocabulary, teacher evaluations generally provided higher scores and showed greater leniency, particularly in grammar accuracy. This suggests that human evaluators bring an important dimension of flexibility and contextual understanding that AI tools currently lack. Therefore, AI-driven assessment can be considered a valuable complementary resource rather than a full substitute for human judgment in language testing. The perceptions of both students and teachers further underscored the complementary nature of AI and human evaluations. Students appreciated the speed and consistency of AI feedback but trusted teachers more for personalized and empathetic assessment. Likewise, teachers recognized AI's efficiency in managing large workloads and providing standardized feedback, yet they emphasized the indispensability of their role in interpreting subtleties of language use, creativity, and communicative effectiveness. These insights indicate that an integrated assessment model combining AI tools and human expertise could enhance the reliability and fairness of language proficiency testing while optimizing resource use. In conclusion, this study highlights the evolving role of AI in language education assessment and the continuing importance of teacher involvement. While AI-driven assessments are advancing rapidly and show promise in delivering timely and objective evaluation, the holistic judgment of experienced teachers remains crucial. Future applications should focus on developing hybrid assessment approaches that leverage the strengths of both AI and human evaluators to provide comprehensive, accurate, and meaningful feedback to learners, ultimately supporting their language development more effectively.

## References


- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3). https://doi.org/10.1016/j.compedu.2009.07.008
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4
- Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 55-67). Routledge.
- Chen, X., Zou, D., Xie, H., & Wang, F. L. (2020). Impact of automated feedback on student writing performance: A meta-analysis. *Computers & Education, 143*, 103675. https://doi.org/10.1016/j.compedu.2019.103675
- Chung, K., & Nation, I. S. P. (2018). The effect of immediate feedback on EFL learners' writing revision. *Language Testing in Asia, 8*(1), 1-15. https://doi.org/10.1186/s40468-018-0067-7
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment, 5*(1). https://ejournals.bc.edu/index.php/jtla/article/view/1650
- Elliott, S., Haswell, R., & Jang, E. E. (2018). Holistic vs. analytic scoring: What do automated essay scoring systems really measure? *Assessing Writing, 37*, 11-23. https://doi.org/10.1016/j.asw.2018.06.002
- Heil, C. R., Wu, J. S., Lee, J. J., & Schmidt, C. (2020). A review of technology-enhanced language learning: The role of AI and machine learning. *Computer Assisted Language Learning, 33*(3), 280-306. https://doi.org/10.1080/09588221.2018.1514897
- Li, Z., & Ni, X. (2021). Teachers' perceptions of artificial intelligence applications in education: A qualitative study. *International Journal of Educational Technology in Higher Education, 18*(1), 1-18. https://doi.org/10.1186/s41239-021-00284-7
- Perelman, L. (2014). An exploratory essay: Four propositions about the future of automated essay scoring. *Assessing Writing, 19*, 89-97. https://doi.org/10.1016/j.asw.2013.10.001
- Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.
- Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. MIT Press.

- Williamson, B. (2017). Big data in education: The digital future of learning, policy and practice. *SAGE Publications*.
- Williamson, B., & Eynon, R. (2020). Historical perspectives on the datafication of education. In *Data and society: Theories and practices of datafication* (pp. 17-30). Routledge.
- Williamson, B., & Piattoeva, N. (2022). Objectivity as standardization in data-scientific educational governance: Grasping the global through the local. *Educational Philosophy and Theory*, *54*(5), 556-569. https://doi.org/10.1080/00131857.2020.1747200
- Williamson, B., & Piattoeva, N. (2022). Objectivity as standardization in data-scientific educational governance: Grasping the global through the local. *Educational Philosophy and Theory, 54*(5), 556-569. https://doi.org/10.1080/00131857.2020.1747200
- Xing, J., & Littlemore, J. (2018). Perceptions of automated essay scoring: A case study with Chinese learners. *Language Testing in Asia, 8*(9). https://doi.org/10.1186/s40468-018-0069-5


- Xing, J., & Littlemore, J. (2018). Perceptions of automated essay scoring: A case study with Chinese learners. *Language Testing in Asia, 8*(9). https://doi.org/10.1186/s40468-018-0069-5