Wanastra: Jurnal Bahasa dan Sastra
Volume 18 No. 1 March 2026
P-ISSN 2086-6151 E-ISSN 2579-3438
https://doi.org/10.31294/wanastra.v18i1

AI and Multimodality-Based Authentic-Innovative Assessment for Evaluating English Speaking Skills

Ibnu Subroto[1], Yanti Rosalinah[2], Cicih Nuraeni[3]

Universitas Bina Sarana Informatika[1,2,3]
e-mail corresponding author: ibnu.isb@bsi.ac.id
email: yanti.yaa@bsi.ac.id, cicih.ccn@bsi.ac.id

Received : 18-02-2026
Revised : 20-03-2026
Accepted : 31-03-2026

Abstract *– This study aimed to develop and validate an AI-Enhanced Multimodal Authentic Assessment model for evaluating EFL students' speaking skills, addressing the limitations of conventional, subjective methods. Employing a mixed-methods approach with a qualitative-dominant design, the research involved 35 university students from a Communication and Language study program. Data were collected through authentic video-based speaking tasks, AI-assisted linguistic analysis (using Google Speech-to-Text and a ChatGPT-based evaluator), detailed multimodal rubric assessments, and student perception questionnaires. Data analysis was conducted through three procedures: multimodal performance analysis using a validated rubric, comparative analysis of AI-generated linguistic metrics, and thematic analysis of questionnaire responses. Quantitative data from AI metrics and Likert-scale questionnaire items were analyzed using descriptive statistics, while qualitative data were analyzed thematically. The findings revealed that multimodal assessment effectively captured verbal, prosodic, visual, and gestural aspects of performance. Concurrently, AI excelled at objectively analyzing micro-linguistic features such as pronunciation, speech rate, and vocabulary. The integration of human and AI evaluation created a comprehensive hybrid model that provided richer, more informative feedback. Furthermore, students expressed positive perceptions regarding the clarity and usefulness of the AI-generated feedback. The study concludes that this integrated assessment model is highly relevant for 21st-century pedagogy and enhances the accuracy and quality of oral performance evaluation. The implications suggest that educators can adopt this framework to create more objective, efficient, and holistic speaking assessments, ultimately fostering better learning outcomes.*

Keywords:Authentic Assessment, Artificial Intelligence, Multimodality, Speaking Skills, EFL.

Abstrak *— Penelitian ini bertujuan untuk mengembangkan dan memvalidasi model Asesmen Otentik Multimodal Berbantuan AI (AI-Enhanced Multimodal Authentic Assessment) untuk mengevaluasi keterampilan berbicara mahasiswa EFL, mengatasi keterbatasan metode konvensional yang subjektif. Dengan menggunakan pendekatan mixed-methods dengan desain kualitatif-dominan, penelitian ini melibatkan 35 mahasiswa dari program studi Komunikasi dan Bahasa di sebuah universitas di Jakarta. Data dikumpulkan melalui tugas berbicara otentik berbasis video, analisis linguistik berbantuan AI (menggunakan Google Speech-to-Text dan evaluator berbasis ChatGPT), penilaian rubrik multimodal yang terperinci, dan kuesioner persepsi mahasiswa. Analisis data dilakukan melalui tiga prosedur: analisis performa multimodal menggunakan rubrik terintegrasi, analisis komparatif terhadap metrik linguistik yang dihasilkan AI, serta analisis tematik terhadap respons kuesioner. Data kuantitatif dari metrik AI dan item kuesioner skala Likert dianalisis menggunakan statistik deskriptif, sementara data kualitatif dianalisis secara tematik. Temuan penelitian mengungkapkan bahwa asesmen multimodal secara efektif menangkap aspek verbal, prosodik, visual, dan gestural dari performa berbicara. Secara bersamaan, AI unggul dalam menganalisis fitur mikrolinguistik secara objektif seperti pelafalan, kecepatan bicara, dan kosakata. Integrasi antara evaluasi manusia dan AI menciptakan model hibrida yang komprehensif, memberikan umpan balik yang lebih kaya dan informatif. Lebih lanjut, mahasiswa menunjukkan persepsi positif mengenai kejelasan dan kebermanfaatan umpan balik yang dihasilkan AI. Penelitian ini menyimpulkan bahwa model asesmen terintegrasi ini sangat relevan dengan pedagogi abad ke-21 dan secara signifikan meningkatkan akurasi serta kualitas evaluasi performa lisan. Implikasinya menunjukkan bahwa pendidik dapat mengadopsi kerangka kerja ini untuk menciptakan asesmen berbicara yang lebih objektif, efisien, dan holistik, yang pada akhirnya mendorong peningkatan hasil belajar yang lebih baik.*

Kata Kunci: *Asesmen Otentik, Kecerdasan Buatan, Multimodalitas, Keterampilan Berbicara, EFL.*

INTRODUCTION

In the landscape of 21st-century education, the ability to communicate effectively in English is paramount. Among the four core language skills, speaking is often considered the most complex, as it requires the simultaneous coordination of linguistic, cognitive, social, and multimodal elements in real-time. Modern communication, particularly in digital formats, extends far beyond the mere production of verbal utterances. As highlighted by contemporary multimodality research, effective oral communication is a "multimodal performance" that integrates verbal channels with crucial nonverbal cues, including gestures, facial expressions, eye contact, and intonation, to construct and convey meaning (Palmour, 2024)(Plough, 2021). This holistic view of communication underscores a significant challenge within English as a Foreign Language (EFL) education: the persistent disconnect between the multifaceted nature of speaking performance and the methods traditionally used to assess it. Technology has also shaped language assessment by shifting assessment practices toward both efficiency and innovation in evaluating language performance (Chapelle, 2016).

Conventional speaking assessment practices in EFL contexts remain largely dominated by traditional models that prioritize micro-linguistic features such as fluency, accuracy, and pronunciation (Luoma, 2004) (Isaacs & Trofimovich, 2016) which often underrepresent multimodal communicative competence and interactional effectiveness (Palmour, 2024). While these elements are undeniably important, such a narrow focus fails to capture the rich, nonverbal dimensions that contribute substantially to a speaker's overall communicative competence and effectiveness. This results in evaluations that are often partial, incomplete, and susceptible to subjectivity and inter-rater variability (Black & Wiliam, 2018)(Huang et al., 2021). In the Indonesian higher education context, these challenges are exacerbated by large class sizes, limited assessment time, and a reliance on instructor intuition, leading to assessment overload and limited formative feedback opportunities in large EFL classrooms (Black & Wiliam, 2018)(Zou et al., 2023). The consequence is an evaluation system that does not accurately reflect a student's true communicative ability nor provide the detailed guidance necessary for meaningful improvement.

In response to the subjectivity of traditional methods, the field has seen a significant rise in the application of Artificial Intelligence (AI) for language evaluation. AI-assisted speaking assessment tools, such as automated speech recognizers and fluency detectors, offer the potential for more objective, consistent, and rapid analysis of linguistic performance (Kasneci et al., 2023)(Shadiev & Feng, 2024). Studies have demonstrated that AI can effectively analyze pronunciation, speech rate, lexical sophistication, and grammatical structures with a high degree of accuracy, providing learners with immediate, data-driven feedback (Zou et al., 2023)(Li et al., 2025). However, this technological solution is not without its limitations. A critical weakness of current AI models is their inability to interpret the multimodal context of communication; they cannot assess the impact of a speaker's gestures, facial expressions, or visual presence on the message being conveyed (Palmour, 2024)(Plough, 2021). As such, AI, while powerful for linguistic analysis, cannot serve as a standalone evaluator for holistic speaking performance. Research on automatic speech recognition further indicates that AI-supported speaking practice can improve pronunciation and speaking performance by offering timely feedback on oral production (Sun, 2023). Automated speech scoring research also shows that speech recognition systems can support consistent evaluation of spoken responses by predicting human scores and generating objective linguistic metrics; however, their validity still depends on construct representation, monitoring procedures, task context, and human interpretation (Zechner et al., 2009)(Xi et al., 2012)(Zechner, K., & Evanini, 2019).

Concurrently, there has been a pedagogical shift towards authentic assessment, an approach that evaluates student performance through tasks mirroring real-world language use (Gulikers et al., 2004) .Authentic tasks, such as video presentations, storytelling, or interviews, provide a naturalistic context for learners to demonstrate their communicative abilities in a more integrated and meaningful way. Research indicates that students tend to exhibit more natural and effective speaking skills when engaged in such authentic tasks compared to highly structured, traditional oral tests (Gulikers et al., 2004)(Huang et al., 2021). Furthermore, authentic assessment aligns perfectly with the principles of multimodal communication, as video-based tasks inherently capture both verbal and nonverbal performance. Yet, while authentic tasks generate rich multimodal data, instructors often lack the systematic tools and frameworks to analyze this data consistently and efficiently, particularly the complex visual and gestural variables. Authentic assessment is also considered valuable because it connects assessment tasks with professional and real-world performance standards, thereby strengthening student engagement, autonomy, and transferable skills(Villarroel et al., 2018). Recent reviews further show that digital technologies can expand authentic assessment practices by supporting task design, evidence collection, feedback, and learner reflection in higher education contexts (Hu et al., 2025).

This review of the current landscape reveals a critical research gap. While strong bodies of literature exist on

authentic assessment, multimodal communication, and AI-assisted evaluation, these three areas have developed largely in isolation. There is a conspicuous lack of a cohesive framework that integrates authentic tasks, multimodal performance analysis, and AI-generated feedback into a single, unified assessment model. This gap is particularly evident in the Indonesian EFL context, where there is an urgent need for innovative, technology-enhanced evaluation systems that are both pedagogically sound and practically feasible. Therefore, this study was designed to bridge this divide by developing and testing a novel model: the AI-Enhanced Multimodal Authentic Assessment. Therefore, the present study responds to the need for a more integrated model that combines technological efficiency, authentic task design, and multimodal interpretation in speaking assessment.

Accordingly, this study seeks to address the following research questions:

1. How does the implementation of AI-Enhanced Multimodal Authentic Assessment produce a more comprehensive evaluation of EFL students' speaking performance compared to conventional methods?
2. What are students' perceptions regarding the clarity, fairness, and usefulness of the feedback generated through this hybrid assessment model?

The primary objective of this research is to develop and evaluate the effectiveness of this integrated assessment model for evaluating the speaking skills of EFL university students. By addressing these objectives, this research aims to provide a significant contribution to both the theory and practice of language assessment, offering a robust, relevant, and effective model for evaluating speaking skills in the digital age.

RESEARCH METHODOLOGY

This study employed a mixed-methods approach with a qualitative-dominant design to develop and evaluate the AI-Enhanced Multimodal Authentic Assessment model. This methodology was chosen for its suitability in providing an in-depth, naturalistic understanding of a complex phenomenon, such as the integration of technology into performance-based assessment, while also allowing for the collection of quantitative data from AI metrics and questionnaires to support the qualitative findings (Johnson et al., 2007). The design is inherently hybrid, combining the rich, interpretive analysis of qualitative research with the objective, data-driven output of artificial intelligence tools. This approach allows for a comprehensive exploration of not only the effectiveness of the model but also the subjective experiences and perceptions of the learners who engage with it, thereby capturing both the process and its impact.

The research was conducted at the Communication and Language Study Program, Universitas Bina Sarana Informatika (UBSI) Jakarta, Indonesia. This setting was strategically selected for its technological infrastructure and its curriculum, which includes digital presentation tasks relevant to the study's objectives. The participants were 35 students enrolled in a speaking course within this program. A purposive sampling technique was utilized to ensure participants met specific criteria: they had completed a video-based speaking task, possessed basic digital recording skills, and consented to being assessed using both human and AI evaluators. This sample size was deemed appropriate for a mixed-methods study of this nature, as it provided sufficient depth and variation in performance data to allow for thorough analysis while remaining manageable for in-depth, rubric-based assessment and thematic coding.

Data were collected through a multi-pronged approach designed to capture a holistic view of the speaking assessment process. The primary data source was the video recordings of students completing an authentic speaking task, which required them to deliver a presentation on a self-selected topic, simulating a real-world academic or professional scenario. These videos served as the basis for all subsequent analysis. To evaluate these performances, a validated multimodal rubric was developed, drawing on principles from authentic assessment (Gulikers et al., 2004) (Palmour, 2024) and multimodal communication theory (Kress & van Leeuwen, 2020). The rubric was validated through expert judgment by two senior lecturers in applied linguistics and a pilot study with 10 students who were not part of the main sample. This rubric assessed nine key areas across verbal, prosodic, visual, and gestural dimensions, as presented in Table 1. In parallel, two specific AI tools were employed to provide objective linguistic analysis: Google Speech-to-Text (Google STT) was used to generate automatic transcriptions and analyze pronunciation clarity and speech rate, while a custom ChatGPT-based evaluator was used to assess sentence structure, vocabulary use, and provide automated feedback. Finally, upon completion of the assessment cycle, a student perception questionnaire was administered to gather data on learners' experiences with the hybrid model, their perceptions of feedback fairness and usefulness, and their overall acceptance of the technology-enhanced approach. The questionnaire consisted of 12 Likert-scale items and 3 open-ended questions. The use of visual and multimodal indicators in the rubric was also informed by visual grammar theory, which explains how images, gaze, spatial position, and visual presence contribute to meaning-making in communication

(Kress & van Leeuwen, 2020).

Table 1. Multimodal Assessment Rubric Categories and Indicators

Dimension	Category	Indicators	Scoring Scale (1-4)
Verbal	Content & Organization	Topic development, coherence, logical structure	1 = Poor, 2 = Fair, 3 = Good, 4 = Excellent
	Vocabulary & Grammar	Lexical range, accuracy, sentence complexity	
Prosodic	Pronunciation	Phoneme accuracy, intelligibility	
	Fluency & Speech Rate	Pacing, naturalness, smoothness	
	Intonation & Stress	Sentence stress, pitch variation	
Visual	Eye Contact	Engagement with audience/camera	
	Posture & Presence	Confidence, physical positioning	
Gestural	Facial Expressions	Alignment with message, appropriateness	
	Hand Gestures	Purposefulness, reinforcement of verbal message	

The data analysis was conducted systematically through three interlinked procedures. First, a qualitative multimodal performance analysis was performed on all 35 videos. Following thematic analytical procedures proposed by (Braun & Clarke, 2021), this involved data reduction, where each performance was coded according to the nine criteria in the multimodal rubric; data display, where findings were organized into a matrix for comparison; and conclusion drawing, where patterns of strengths and weaknesses across the cohort were identified. Second, the outputs from the AI tools were analyzed. Quantitative data from AI metrics, including pronunciation accuracy percentages, speech rate (words per minute), and grammatical accuracy scores, were analyzed using descriptive statistics (means, standard deviations, minimum and maximum values). These metrics were then systematically compared and contrasted with the scores from the human-led rubric assessment to identify areas of convergence and divergence, with correlation coefficients calculated to examine the relationship between human and AI evaluations. Third, the questionnaire data were analyzed using a mixed approach: descriptive statistics were used to summarize responses to closed-ended questions, while thematic analysis was applied to open-ended responses to identify recurring themes related to student perceptions. The findings from all three analytical strands were then integrated through a process of triangulation to build a comprehensive and robust understanding of the model's effectiveness and its implications for pedagogical practice.

RESULTS AND DISCUSSION

This section presents the results of the data analysis, structured to answer the two primary research questions. The findings are derived from the integration of the multimodal rubric assessment, AI-assisted linguistic analysis, and student perception questionnaires, offering a comprehensive view of the effectiveness and reception of the AI-Enhanced Multimodal Authentic Assessment model.

1. Comprehensive Evaluation of Speaking Performance

To address the first research question regarding the model's ability to produce a more comprehensive evaluation, the data clearly demonstrate the synergistic relationship between human-led multimodal assessment and AI-assisted linguistic analysis. The analysis of student performances using the multimodal rubric revealed a rich spectrum of communicative abilities that would have been invisible to a conventional assessment. For instance, while some students demonstrated high linguistic accuracy, their performance was often undermined by limited eye contact, static posture, or gestures that did not reinforce their verbal message. Conversely, other students with less grammatical precision were able to convey their message effectively through strong visual presence, expressive facial cues, and purposeful gestures. This highlights the critical importance of assessing performance holistically, as verbal and nonverbal channels work in concert to create meaning.

The quantitative distribution of scores from the human-led multimodal rubric provides a clear picture of student performance across the nine assessed categories. Table 2 presents the mean scores, standard deviations, and ranges for each category.

Table 2. Distribution of Student Performance Scores (Human Multimodal Rubric)

Dimension	Category	Mean Score	SD	Min	Max

Verbal	Content & Organization	3.12	0.54	2	4
	Vocabulary & Grammar	2.89	0.67	1	4
Prosodic	Pronunciation	2.94	0.62	2	4
	Fluency & Speech Rate	3.05	0.58	2	4
	Intonation & Stress	2.83	0.71	1	4
Visual	Eye Contact	2.71	0.83	1	4
	Posture & Presence	3.18	0.61	2	4
Gestural	Facial Expressions	2.84	0.76	1	4
	Hand Gestures	2.65	0.81	1	4

The data in Table 2 reveal that students demonstrated relatively stronger performance in verbal content and posture (mean scores above 3.1), while areas such as hand gestures and eye contact (Eva, Fachriyah, Berita Mambarasi Nehe, 2025) showed greater variation and lower mean scores. This finding underscores the importance of assessing these nonverbal dimensions, as they represent areas for potential pedagogical intervention. This result is consistent with multimodal public speaking research showing that nonverbal behaviors such as posture, gaze, and gesture can contribute to the perceived quality of oral presentation performance (Wörtwein et al., 2015).

Concurrently, the AI-assisted linguistic analysis provided objective and granular data on aspects of performance that are often subjectively evaluated. Google Speech-to-Text effectively measured speech rate and identified specific phonemes that were consistently mispronounced across the cohort. The ChatGPT-based evaluator offered detailed feedback on grammatical structures, such as the overuse of simple sentences or inconsistencies in verb tense, providing a level of linguistic precision that is difficult for human raters to maintain consistently, especially with large groups. However, these tools could not assess the visual or gestural components, nor could they interpret the pragmatic or rhetorical effectiveness of the message. Similar findings have been reported in recent AI-speaking assessment research, which emphasizes that AI tools can provide useful information on score accuracy, perceived validity, and feedback quality in speaking classrooms (Liu et al., 2025).

Table 3 presents a comparison between human rubric scores and AI-generated metrics for selected aspects of performance, revealing moderate to strong correlations that suggest alignment between the two evaluation approaches while also highlighting their distinct contributions.

Table 3. Comparison between Human Rubric Scores and AI-Generated Metrics

Aspect	Human Rating (Mean)	AI Metric (Mean)	Correlation
Pronunciation	2.94	87.3% (accuracy)	0.72
Speech Rate (words/min)	3.05	142.4 wpm	0.68
Grammatical Accuracy	2.89	84.6% (error-free clauses)	0.75
Vocabulary Diversity	2.89	0.48 (type-token ratio)	0.70

The integration of these two evaluative approaches proved to be the core strength of the model, creating a hybrid assessment that was greater than the sum of its parts. The complementary nature of these components is summarized in Table 4. Recent multimodal modeling studies also suggest that speaking assessment should move beyond linguistic accuracy by incorporating multiple performance indicators to represent communicative competence more comprehensively (Mawalim et al., 2025).

Table 4. Complementary Roles of Human and AI Assessment Components

Assessment Component	Strengths & Contributions	Limitations
Human-Led Multimodal Rubric Assessment	Evaluates holistic communicative competence (verbal, nonverbal, pragmatic aspects); Captures visual and gestural performance (eye contact, posture, gestures, facial expressions); Interprets rhetorical effectiveness and coherence of message; Assesses multimodal integration	Subjective and prone to inter-rater variability; Time-consuming, especially with large cohorts; Inconsistent in evaluating micro-linguistic features (e.g., precise pronunciation errors, speech rate)
AI-Assisted Linguistic Analysis	Provides objective, consistent analysis of micro-linguistic features; Identifies specific pronunciation errors and phoneme mispronunciations (via Google STT); Measures speech rate and fluency metrics;	Cannot interpret nonverbal or visual communication elements; Lacks ability to assess pragmatic or contextual appropriateness; May misinterpret

	Analyzes grammatical structures, sentence complexity, and vocabulary use (via ChatGPT-based evaluator); Offers immediate, data-driven, and actionable feedback	prosody, emotion, or intent behind speech; Dependent on accuracy of speech recognition and NLP models
Integrated Hybrid Model	Combines strengths of both human and AI evaluation; Provides comprehensive, multi-dimensional assessment; Enhances objectivity while retaining human interpretive insight; Supports detailed, formative feedback for learners	Requires technological infrastructure and training; May be perceived as complex or intimidating to implement initially; Relies on accuracy and appropriateness of AI tools selected

2. Student Perceptions of the Hybrid Assessment Model

To address the second research question regarding student perceptions, the analysis of the questionnaires revealed a predominantly positive response to the hybrid assessment model. A primary theme was the perceived clarity and objectivity of the AI-generated feedback. Students noted that feedback on specific pronunciation errors or grammatical structures was more precise and actionable than general comments they had received in traditional courses. One student commented, "The AI showed me exactly which words I mispronounced, so I could practice them. It felt less personal and more like a tool to help me improve." This objectivity was frequently linked to a sense of fairness in the evaluation process. Another student stated, "It felt fair because the AI didn't have favorites. It just analyzed my speech."

Furthermore, students found the multimodal nature of the task to be more relevant to their future professional lives than a traditional in-class test. They reported that creating a video presentation felt more authentic and helped them develop skills applicable to online meetings and digital content creation. One participant noted, "Making a video felt like something I'd actually do in a job interview or an online meeting." Another added, "I learned how to present myself on camera, which is useful for online classes and future work."

However, some students did report initial technical challenges and a degree of performance anxiety when being recorded. Despite these hurdles, the overall perception was that the model, while demanding, provided a more valuable and insightful learning experience. As one student summarized, "It was more work, but I learned more about my speaking than ever before." The key themes from the perception analysis are summarized in Table 5.

Table 5. Key Themes from Student Perception Analysis

Theme	Description	Representative Student Comment(s)
Clarity & Objectivity of AI Feedback	Students appreciated the precise, data-driven nature of AI-generated feedback, which they perceived as fair and non-judgmental	"The AI showed me exactly which words I mispronounced, so I could practice them. It felt less personal and more like a tool to help me improve."
Perceived Fairness	The objectivity of AI analysis contributed to a sense of transparency and equity in the assessment process	"It felt fair because the AI didn't have favorites. It just analyzed my speech."
Authenticity & Relevance of Video Task	Students found the video-based speaking task more relevant to real-world communication, especially in digital and professional contexts	"Making a video felt like something I'd actually do in a job interview or an online meeting."
Skill Transferability	Learners recognized the task as developing competencies useful beyond the classroom, such as digital presentation and online communication skills	"I learned how to present myself on camera, which is useful for online classes and future work."
Technical & Performance Challenges	Some students reported initial difficulties with recording technology and increased anxiety when being recorded	"I was nervous knowing I was being recorded and assessed by AI."
Overall Positive Engagement	Despite challenges, the majority viewed the hybrid model as a valuable, insightful, and modern learning experience	"It was more work, but I learned more about my speaking than ever before."

DISCUSSION

This study was designed to address the persistent challenges of conventional speaking assessment by developing and evaluating an AI-Enhanced Multimodal Authentic Assessment model. The research conclusively

demonstrates that this integrated approach offers a promising alternative to traditional methods. The primary findings indicate that the synergy between human-led multimodal assessment and AI-assisted linguistic analysis yields an evaluation that is significantly more comprehensive, objective, and informative. The human rater effectively captures the holistic, nonverbal, and pragmatic dimensions of communication, while the AI tools provide precise, consistent, and scalable analysis of micro-linguistic features. Furthermore, the model was met with positive perceptions from students, who valued the objectivity of the AI feedback and the authenticity of the video-based task.

The most significant finding of this research is the demonstrable effectiveness of the hybrid assessment model, which synergistically combines human-led multimodal evaluation with AI-assisted linguistic analysis. This directly addresses the critical research gap identified in the literature, where authentic tasks, multimodal analysis, and AI technology have been treated as separate domains rather than integrated components of a cohesive system (Black & Wiliam, 2018)(Huang et al., 2021). The results of this study demonstrate that these elements are not merely compatible but are, in fact, mutually reinforcing. The human rater, using the multimodal rubric, excels at interpreting the holistic, pragmatic, and nonverbal aspects of communication—the very elements that traditional and AI-only models consistently overlook (Palmour, 2024)(Plough, 2021). This aligns with the view that communication is a "multimodal performance," and its assessment must therefore be performance-based. The AI component, in turn, provides a level of objective, consistent, and granular analysis of micro-linguistic features that is practically impossible for a human rater to maintain, especially with large student cohorts (Kasneci et al., 2023)(Zou et al., 2023). This validates the argument by Holmes et al. (2019) that AI is most effective not as a replacement for human instructors, but as a powerful complementary tool that enhances objectivity and provides a robust data foundation for evaluation. This interpretation is also supported by technology-enhanced pronunciation research, which highlights the pedagogical value of combining digital feedback with teacher-led interpretation in classroom-based language learning (Toyama & Hori, 2025).

The moderate to strong correlations observed between human ratings and AI metrics (ranging from 0.68 to 0.75) further support the validity of both assessment approaches, while the discrepancies in individual cases highlight the unique contributions of each. For instance, a student with strong nonverbal communication skills but lower grammatical accuracy might receive a higher overall evaluation from the human rather than the AI alone would suggest, reflecting the holistic nature of communicative competence. The new understanding that emerges from this is a redefinition of the rater's role in a technology-enhanced environment. The instructor is no longer solely the "judge" of performance but also an "interpreter" of a rich dataset, part of which is generated by AI. This model effectively mitigates the subjectivity and inter-rater variability that plague conventional speaking assessments, as noted by (Luoma, 2004) and (Isaacs & Trofimovich, 2016). By providing objective data on pronunciation, speech rate, and grammatical structures, the AI tools create a common ground for evaluation, allowing the human rater to focus their cognitive energy on the more complex, interpretive aspects of performance, such as coherence, persuasiveness, and the effectiveness of multimodal delivery.

The second major finding concerns the predominantly positive perceptions of students, which carries significant implications for pedagogical practice. The students' appreciation for the objectivity and clarity of AI-generated feedback provides a direct counterpoint to the subjectivity often lamented in Indonesian EFL contexts (Black & Wiliam, 2018)(Shadiev & Feng, 2024). The data-driven nature of the AI feedback was perceived as fair and impersonal, which in turn increased students' trust in the evaluation process. This supports the principles of "assessment for learning," where the primary goal of evaluation is to provide clear, actionable feedback to guide improvement (Black & Wiliam, 2018). The AI component of the model excelled in this regard, offering specific, non-judgmental advice that students could immediately apply.

Furthermore, the students' positive response to the authentic, video-based task confirms the theoretical benefits outlined by proponents of authentic assessment (Gulikers et al., 2004).The task was not merely an assessment but a learning experience that developed skills directly transferable to modern professional communication. This aligns with previous studies suggesting that authentic video-based speaking tasks encourage more natural communication and better reflect real-world communicative demands (Gulikers et al., 2004) who noted that video tasks elicit more natural communication. The current study extends this by demonstrating that when these authentic tasks are paired with a robust, technology-enhanced assessment framework, the learning experience is not only more authentic but also more insightful and empowering for the learner.

The implications of these findings are both theoretical and practical. Theoretically, this study provides a strong empirical argument for a new paradigm in language assessment—one that is integrative, technology-informed, and grounded in a multimodal understanding of communication. It challenges the field to move beyond siloed approaches and develop frameworks that reflect the complex reality of language use. Practically, the study offers

a concrete, replicable model for educators and institutions seeking to modernize their assessment practices. It provides a blueprint for designing assessments that are not only more valid and reliable but also more engaging and beneficial for student learning.

CONCLUSION

This study successfully developed and evaluated an AI-Enhanced Multimodal Authentic Assessment model for evaluating English speaking skills in the EFL context. The findings suggest that the integration of human-led multimodal assessment and AI-assisted linguistic analysis creates a synergistic evaluation framework that addresses the fundamental limitations of conventional speaking assessments. The model captures the holistic nature of communicative competence by evaluating verbal, prosodic, visual, and gestural dimensions through the multimodal rubric, while simultaneously leveraging AI tools to provide objective, consistent, and granular analysis of micro-linguistic features such as pronunciation accuracy, speech rate, and grammatical structures. The research reveals three key contributions to the field of language assessment. First, the hybrid model successfully bridges the gap between authentic task design and systematic evaluation, enabling educators to assess speaking performance in a manner that reflects real-world communication demands. Second, the integration of AI tools enhances the objectivity and reliability of assessment while reducing the subjectivity and inter-rater variability that have long plagued traditional speaking evaluations. Third, the positive student perceptions regarding feedback clarity, fairness, and task authenticity confirm that this model may improve both assessment quality and the learning experience.

The practical implications of this research are substantial. For educators, the model provides a replicable framework for designing technology-enhanced assessments that yield richer, more actionable feedback for learners. For institutions, it underscores the strategic importance of investing in AI technologies and professional development programs that prepare instructors to implement innovative assessment practices. Theoretically, this study contributes to the evolving discourse on multimodal communication and technology-mediated assessment, offering empirical evidence for the effectiveness of integrated approaches. Nevertheless, this research acknowledges certain limitations. The study was conducted with a specific group of students at a single university and utilized only two specific AI tools. Given the context-specific nature of this study and the absence of statistical generalization, the findings should be interpreted as preliminary. The model's effectiveness may vary across different cultural contexts, proficiency levels, or with different AI platforms. Furthermore, the study focused solely on speaking skills; future research could explore the applicability of this integrated model to other language skills, such as writing or interactive listening. Longitudinal studies are also needed to determine the long-term impact of this assessment model on students' speaking development and motivation. Further research with larger and more diverse samples is recommended to validate the effectiveness of this model across different educational settings and proficiency levels. Therefore, future implementation should consider the accuracy of automated speech recognition, ethical use of learner data, and the pedagogical readiness of both teachers and students when adopting AI-enhanced assessment tools. Despite these limitations, this study provides a significant contribution to the advancement of language assessment practices, offering a robust, relevant, and forward-thinking model for evaluating speaking skills in the digital age.

REFERENCES

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy \& Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Braun, V., & Clarke, V. (2021). *Thematic Analysis: A Practical Guide*. SAGE Publications. https://books.google.co.id/books?id=eMArEAAAQBAJ

Chapelle, C. A. (2016). *20 YEARS OF TECHNOLOGY AND LANGUAGE ASSESSMENT IN LANGUAGE LEARNING & TECHNOLOGY*. *20*(2), 116–128.

Eva, Fachriyah, Berita Mambarasi Nehe, A. H. (2025). Harnessing video reaction-based tasks to foster speaking fluency and critical thinking in English: A mixed-method study. *Jeltim*, *7*(2), 182–205.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A Five-Dimensional Framework for Authentic Assessment. *Educational Technology Research and Development*, *52*(3), 67–86. https://doi.org/10.1007/BF02504676

Hu, Anjin, Liu, Qian, & Daniel, Ben. (2025). Digital Technologies in Authentic Assessment in Higher Education: A Systematic Literature Review and Narrative Synthesis. *Sage Open*, *15*(3), 21582440251357200. https://doi.org/10.1177/21582440251357198

Huang, Becky H, Bailey, Alison L, Sass, Daniel A, & Shawn Chang, Yung-hsiang. (2021). An investigation of the validity of a speaking assessment for adolescent English language learners. *Language Testing*, *38*(3), 401–428. https://doi.org/10.1177/0265532220925731

Isaacs, T., & Trofimovich, P. (2016). *Second Language Pronunciation Assessment*. Multilingual Matters. https://doi.org/10.21832/ISAACS6848

Johnson, R. Burke, Onwuegbuzie, Anthony J, & Turner, Lisa A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, *1*(2), 112–133. https://doi.org/10.1177/1558689806298224

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/https://doi.org/10.1016/j.lindif.2023.102274

Kress, G., & van Leeuwen, T. (2020). *Reading Images: The Grammar of Visual Design*. Taylor \& Francis. https://books.google.co.id/books?id=zmsJEAAAQBAJ

Li, Junfei, Huang, Jinyan, & Sheeran, Thomas. (2025). ChatGPT4o as an AI Peer Assessor in EFL Speaking Classrooms: Examining Scoring Reliability and Feedback Effectiveness. *Sage Open*, *15*(3), 21582440251369936. https://doi.org/10.1177/21582440251369938

Liu, X. J., Wang, J., & Zou, B. (2025). Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback as feedback enhancement. *Journal of English for Academic Purposes*, *75*, 101505. https://doi.org/https://doi.org/10.1016/j.jeap.2025.101505

Luoma, S. (2004). *Assessing Speaking*. Cambridge University Press. https://doi.org///doi.org/10.1017/CBO9780511733017

Mawalim, C. O., Leong, C. W., Sivan, G., Huang, H.-H., & Okada, S. (2025). Beyond accuracy: Multimodal modeling of structured speaking skill indices in young adolescents. *Computers and Education: Artificial Intelligence*, *8*, 100386. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100386

Palmour, L. (2024). Assessing speaking through multimodal oral presentations : The case of construct underrepresentation in EAP contexts. *Lamguage Testing*, *41 (1)*(X), 9–34. https://doi.org/10.1177/02655322231183077

Plough, I. (2021). 3 A Case for Nonverbal Behavior: Implications for Construct, Performance and Assessment. In M. R. Salaberry & A. R. Burch (Eds.), *Expanding the Construct and its Applications* (pp. 50–70). Multilingual Matters. https://doi.org/doi:10.21832/9781788923828-004

Shadiev, R., & Feng, Y. (2024). Using automated corrective feedback tools in language learning: a review study. *Interactive Learning Environments*, *32*(6), 2538–2566. https://doi.org/10.1080/10494820.2022.2153145

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation. *Frontiers in Psychology*, *Volume 14-2023*. https://doi.org/10.3389/fpsyg.2023.1210187

Toyama, M., & Hori, T. (2025). Technology-enhanced multimodal approaches in classroom L2 pronunciation training. *Frontiers in Education*, *Volume 10-2025*. https://doi.org/10.3389/feduc.2025.1552470

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, *43*(5), 840–854.

https://doi.org/10.1080/02602938.2017.1412396

Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.-P., Stiefelhagen, R., & Scherer, S. (2015). Multimodal Public Speaking Performance Assessment. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 43–50. https://doi.org/10.1145/2818346.2820762

Xi, Xiaoming, Higgins, Derrick, Zechner, Klaus, & Williamson, David. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, *29*(3), 371–394. https://doi.org/10.1177/0265532211425673

Zechner, K., & Evanini, K. (2019). *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech (1st ed.).* Routledge. https://doi.org/https://doi.org/10.4324/9781315165103

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895. https://doi.org/https://doi.org/10.1016/j.specom.2009.04.009

Zou, Bin, Du, Yiran, Wang, Zhimai, Chen, Jinxian, & Zhang, Weilei. (2023). An Investigation Into Artificial Intelligence Speech Evaluation Programs With Automatic Feedback for Developing EFL Learners' Speaking Skills. *Sage Open*, *13*(3), 21582440231193816. https://doi.org/10.1177/21582440231193818