Al-Ishlah: Jurnal Pendidikan
Vol.17, 3 (September, 2025), pp. 5615-5625
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v17i3.7655

English Education Students' Perceptions of Automated vs Human Assessment in Spoken English Proficiency

Nur Aeni[1], Muhalim[2], Hasriani Ganteng[3], Muhammad Tahir[4], Ahmad Talib[5]

[1] Universitas Negeri Makassar, Makassar, Indonesia; nur_aeni@unm.ac.id
[2] Universitas Negeri Makassar, Makassar, Indonesia; muhalim@unm.ac.id
[3] Universitas Negeri Makassar, Makassar, Indonesia; hasrianig@unm.ac.id
[4] Universitas Negeri Makassar, Makassar, Indonesia; muhammadtahir@unm.ac.id
[5] Universitas Negeri Makassar, Makassar, Indonesia; ahmadtalib@unm.ac.id

ARTICLE INFO

Keywords:

automated evaluation;
human raters;
spoken English proficiency,
student perceptions

Article history:

Received 2025-05-02
Revised 2025-08-15
Accepted 2025-09-30

ABSTRACT

The increasing use of automated evaluation systems in language assessment raises questions about their acceptance and perceived fairness compared to human evaluation. This study examines how English Education students perceive automated and human assessment of spoken English proficiency, focusing on factors influencing acceptance and preferences for hybrid models. A mixed-methods design was employed with 120 English Education students (80 female, 40 male) from Universitas Negeri Makassar. Quantitative data were collected using a 20-item Likert-scale questionnaire (Cronbach's α = .87) covering six dimensions: Perceived Ease of Use, Perceived Usefulness, Attitude Toward Technology, Self-Efficacy, Behavioral Intention, and Personal Innovativeness. Qualitative data from semi-structured interviews explored students' experiences and preferences regarding automated and human evaluation. Descriptive statistics indicated generally positive perceptions of automated evaluation, with the highest mean scores for "Automated feedback helps improve pronunciation and fluency" (M = 3.9, SD = 0.928) and "I enjoy playing with new technology in language acquisition" (M = 4.0, SD = 1.071). However, the lowest score for "I plan to use automated evaluation frequently" (M = 2.7, SD = 1.071) reflected hesitancy toward regular use. Thematic analysis revealed three main themes: appreciation of efficiency but skepticism about accuracy, preference for human empathy and contextual understanding, and concerns about algorithmic bias, particularly for non-standard accents. Students strongly favored a hybrid approach, endorsing AI for preliminary feedback and routine practice while valuing human evaluation for comprehensive assessment and motivational support. These findings suggest the need for transparent, inclusive AI tools integrated with human oversight to achieve balanced, pedagogically sound evaluation frameworks in English language education.

Corresponding Author:
Nur Aeni
Universitas Negeri Makassar, Makassar, Indonesia; nur_aeni@unm.ac.id

1. INTRODUCTION

The integration of automated evaluation systems into spoken English assessment represents a pivotal development in contemporary language education. As artificial intelligence (AI) technologies become increasingly embedded in educational contexts, their application in language testing offers both promising advancements and critical challenges. One of the central tensions lies in reconciling the efficiency and scalability of AI-driven tools with the pedagogical richness and interpersonal support traditionally provided by human assessors. Particularly in speaking assessments—where intonation, fluency, and pragmatics play a vital role—the balance between automation and human judgment is essential to ensuring both reliability and learner acceptance.

Recent advancements in AI have significantly transformed the landscape of language instruction and assessment. Technologies such as automated speech recognition (ASR), natural language processing, and machine learning have enabled the development of platforms capable of delivering real-time feedback, personalized learning pathways, and scalable testing environments (Fryer & Carpenter, 2022). In the realm of spoken English proficiency, AI-powered tools have been lauded for their consistency, efficiency, and potential to alleviate the subjectivity inherent in human evaluation (Rahayu, 2021). For example, AI-based systems can provide immediate feedback on pronunciation, grammar, and fluency, thereby supporting learner autonomy and continuous improvement. Furthermore, adaptive learning technologies have been shown to enhance student motivation and engagement by aligning task difficulty with individual performance levels (Wang & Liu, 2023; Chapelle & Chung, 2021).

Despite these advantages, the integration of AI in language assessment is not without its limitations. A key concern is the potential erosion of human interaction, which remains vital for building learner confidence, interpreting nuanced speech elements, and providing affective feedback (Aeni et al., 2024). Additionally, issues of algorithmic bias, data privacy, and fairness raise important questions about the ethical implications and inclusivity of automated systems, particularly in linguistically and culturally diverse educational settings (Wang & Liu, 2023). These challenges underscore the importance of pursuing hybrid assessment models that combine the technological benefits of automation with the empathetic, contextualized judgment of human evaluators.

While the global body of research on AI in language education continues to expand, there is a noticeable gap in empirical studies that directly compare students' perceptions of automated and human assessment methods within the same institutional context. This gap is particularly evident in non-Western educational settings, where technological infrastructure, cultural expectations, and pedagogical norms may differ significantly from those in more digitally mature regions. In Indonesia, for instance, AI-based language assessment remains an emerging field, and little is known about how students perceive these innovations in terms of fairness, trustworthiness, and educational value (Hummel & Donner, 2023; Yastibas & Yastibas, 2021). The lack of comparative research in this area limits the development of assessment frameworks that are both culturally responsive and pedagogically sound.

To address this gap, the present study explores the perceptions of English Education students at Universitas Negeri Makassar regarding automated and human evaluation in spoken English proficiency assessments. Specifically, it investigates learner attitudes related to trust, perceived fairness, usefulness, and behavioral intentions, with an emphasis on preferences for hybrid models that incorporate both automated and human elements. By providing empirical evidence from an Indonesian higher education context, this research aims to contribute to a more nuanced understanding of learner acceptance and inform the design of integrated assessment systems that leverage the complementary strengths of both technological and human evaluative approaches.

2. METHODS

2.1 Research Design

This study employed a descriptive survey design to explore students' perceptions about the use of automated and human evaluation in assessing their spoken English. This approach enabled the collection of both quantitative data through structured questionnaires and qualitative insights through open-ended responses, providing a comprehensive understanding of students' attitudes and experiences toward different evaluation methods.

2.2 Participants

The study involved 120 undergraduate students from the English Education Program at Universitas Negeri Makassar, selected through purposive sampling. Participants were chosen based on specific inclusion criteria: (1) enrollment in the English Education Program, (2) completion of at least one semester of spoken English courses, and (3) exposure to both automated and human evaluation methods in their coursework. Students who had not experienced both evaluation types and were enrolled in other programs were excluded from the study.

Table 1. Participant Demographics

Characteristic	Category	n	%
Gender	Female	80	66.7
	Male	40	33.3
Academic Year	First Year	30	25.0
	Second Year	35	29.2
	Third Year	32	26.7
	Fourth Year	23	19.2

Note: Mean age = 20 years

The purposive sampling strategy was justified by the need to ensure participants had comparable exposure to both evaluation methods and represented various academic levels, thereby capturing diverse perspectives based on different stages of language learning and assessment experience.

2.3 Instruments

A 20-item Likert-scale questionnaire was employed to investigate students' perceptions of fairness, usability, trust, satisfaction, and the effectiveness of both automated and human evaluation methods. The instrument was adapted from the Technology Acceptance Model (TAM) developed by Davis (1989) and previously validated in language learning contexts—specifically, in mobile-assisted language learning (Kim & Lee, 2016) and in simulation-based educational environments (LeMay et al., 2018)—demonstrating strong construct validity and reliability in measuring learner perceptions toward technology-based evaluation. The questionnaire demonstrated high internal consistency (Cronbach's α = .87).

The questionnaire was structured around six dimensions: Perceived Ease of Use (PEU), Perceived Usefulness (PU), Attitude Toward Technology (ATT), Self-Efficacy (SE), Behavioral Intention to Use (BI), and Personal Innovativeness (PI). Items were measured on a 5-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5).

The questionnaire also included open-ended questions allowing participants to elaborate on their personal experiences with both evaluation methods, focusing on perceived advantages, disadvantages, and preferences.

2.4 Ethical Considerations

Prior to data collection, ethical approval was obtained from the institutional review board. All participants provided written informed consent after being informed about the study's purpose, voluntary participation, confidentiality measures, and their right to withdraw at any time without penalty. Participant anonymity was maintained throughout the study.

2.5 Data Analysis

2.5.1 Quantitative Analysis

Quantitative data from the Likert-scale items were analyzed using descriptive statistics, with means and standard deviations calculated to identify general trends in students' responses and to facilitate clearer interpretation through tabular presentation.

2.5.2 Qualitative Analysis

Qualitative data from open-ended questionnaire responses and focus group discussions were examined thematically using Miles and Huberman's (1994) qualitative data analysis framework to uncover recurring perceptions, concerns, and patterns regarding both automated and human evaluation methods. The analysis involved three concurrent flows of activity:

a. Data Reduction: Systematically selecting, focusing, simplifying, and abstracting data from written responses and transcripts
b. Data Display: Organizing and compressing information to enable conclusion drawing through matrices and narrative summaries
c. Conclusion Drawing and Verification: Identifying patterns, themes, and meanings from the data while continuously verifying findings against the original data

3. FINDINGS AND DISCUSSION

3.1 Quantitative Findings

Table 1 presents descriptive statistics highlighting trends in students' perceived effectiveness of both automated and human evaluation methods across six theoretical constructs.

Table 1. Descriptive Statistics of Automated and Human Evaluation Perceptions

No	Item	N	Mean	Std. Dev
1	(PEU) Automated evaluation provides accurate assessments	120	3.9	0.928
2	(PEU) Automated evaluation is as fair as human raters	120	3.9	0.858
3	(PEU) Human raters provide more reliable scores	120	3.5	0.926
4	(PEU) Automated evaluation is free from bias	120	3.4	0.941
5	(PU) Automated evaluation provides fast and useful feedback	120	3.5	0.926
6	(PU) Human evaluation is time-consuming but provides valuable insights	120	3.2	0.963
7	(PU) Automated evaluation systems are easy to use	120	3.3	0.978
8	(PU) I feel comfortable using automated evaluation tools	120	3.7	0.830
9	(PU) I prefer using automated tools because they are more accessible	120	3.8	0.609
10	(PU) Automated feedback helps improve pronunciation and fluency	120	3.9	0.928
11	(PU) Automated evaluation tools help identify areas of improvement	120	3.7	0.838
12	(ATT) Using automated evaluation is an excellent concept	120	3.5	0.928
13	(ATT) I am optimistic about employing automated evaluation	120	3.8	0.782
14	(ATT) Using automated evaluation for language learning is enjoyable	120	3.9	0.912
15	(SE) I'm confident in leveraging automated evaluation	120	3.7	0.999
16	(SE) I have the required expertise to use automated evaluation	120	3.2	0.911

17	(BI) I plan to use automated evaluation frequently	120	2.7	1.071
18	(BI) I want to learn more about using automated evaluation	120	3.4	0.996
19	(PI) I enjoy playing with new technology in language acquisition	120	4.0	1.071
20	(PI) Among my peers, I am typically the first to examine new technologies	120	3.3	0.978

Note: PEU = Perceived Ease of Use; PU = Perceived Usefulness; ATT = Attitude Toward Technology; SE = Self-Efficacy; BI = Behavioral Intention; PI = Personal Innovativeness

The results are organized by the six theoretical constructs to provide a clearer understanding of students' perceptions.

3.1.1 Analysis of Construct-Based Perceptions

The descriptive analysis reveals distinct patterns across the six theoretical constructs examined. Students demonstrated strong positive attitudes toward automated evaluation systems, with Personal Innovativeness showing the highest mean score (M = 4.0 for item 19), indicating genuine enthusiasm for technological integration in language learning. This finding aligns with Technology Acceptance Model predictions, where personal innovativeness serves as a key predictor of technology adoption (Davis, 1989).

Perceived Ease of Use and Usefulness: Items measuring PEU and PU consistently scored above the neutral point (3.0), with automated assessment accuracy (M = 3.9) and pronunciation feedback effectiveness (M = 3.9) receiving particularly high ratings. However, the moderate score for bias-free evaluation (M = 3.4) suggests students recognize potential algorithmic limitations, particularly regarding accent and dialect variations.

Attitude and Self-Efficacy Patterns: Students expressed positive attitudes toward automated evaluation (ATT items M = 3.5-3.9) but showed lower confidence in their technical expertise (SE item 16, M = 3.2). This gap between positive attitude and self-perceived competence may influence actual technology adoption.

Critical Implementation Gap: The most significant finding emerges in Behavioral Intention scores, where item 17 ("I plan to use automated evaluation frequently") recorded the lowest mean (M = 2.7, SD = 1.071). This substantial gap between conceptual acceptance and intended usage suggests implementation barriers that warrant further investigation. The high standard deviation (1.071) indicates considerable variance in student intentions, pointing to individual differences in technology acceptance.

3.1.2 Implications for Technology Integration

These quantitative findings suggest that while students appreciate automated evaluation's technical capabilities, particularly for pronunciation practice and immediate feedback, they remain cautious about frequent implementation. The discrepancy between high personal innovativeness scores and low behavioral intention indicates that positive attitudes alone are insufficient to drive sustained technology use.

The practical implications of lower behavioural Intention scores suggest several implementation considerations. First, institutions may need to address students' technical self-efficacy through targeted training programs. Second, the integration of automated tools should emphasize their complementary rather than replacement role, addressing concerns about bias and reliability. Finally, the significant standard deviation in behavioural intention scores indicates that individualized approaches to technology integration may be more effective than universal implementation strategies.

Future correlation or regression analyses could examine which factors most strongly predict behavioral intention, potentially revealing whether self-efficacy, perceived usefulness, or attitude toward technology serves as the primary driver of intended usage among this population.

Overall, these results highlight both the potential and limitations of automated evaluation systems within the context of language learning.

3.2. Qualitative Findings

3.2.1 Theme Development Process

Qualitative data from open-ended questionnaire responses underwent thematic analysis following Miles and Huberman's (1994) framework. The analysis employed an inductive approach, allowing themes to emerge naturally from the data without predetermined categories. The coding process involved systematic identification of recurring patterns and meanings across participant responses.

The analysis process involved three concurrent activities: data reduction through systematic coding and categorization, data display through thematic matrices, and conclusion drawing through pattern identification and verification against original responses. This rigorous process yielded three primary themes that capture students' complex perceptions of both evaluation methods.

3.2.2 Summary of Qualitative Themes

Table 2. Summary of Qualitative Themes with Representative Quotes
Representative Quotes

Theme	Sub-themes	Representative Quotes	Frequency
Efficiency-Accuracy Tension	Speed vs. Reliability Transparency concerns	*"The AI gives fast feedback, but sometimes it doesn't recognize my pronunciation correctly" (P7)* *"I don't really understand how the system decides my score" (P12)*	78% of responses
Human Empathy and Contextual Understanding	Emotional connection Personalized feedback Intentionality recognition	*"When I speak with a human evaluator, I feel they understand my effort" (P19)* *"My teacher tells me how to fix it" (P5)*	82% of responses
Bias and Inclusivity Concerns	Accent discrimination Fairness issues Hybrid solution preference	*"AI judges me unfairly because of that [accent]" (P22)* *"AI is good for everyday practice, but for important exams, I would trust a human evaluator more" (P10)*	65% of responses

Note: P = Participant; Percentages indicate proportion of participants who mentioned theme

Theme 1: Efficiency-Accuracy Tension

The most prevalent theme revealed students' conflicted relationship with automated evaluation's core promise of efficient assessment. While participants consistently praised AI's speed and accessibility, this appreciation was tempered by significant concerns about accuracy and transparency. Students valued the immediacy of feedback, particularly for pronunciation practice, as one participant noted: "I can practice anytime and get instant results" (P15). However, this efficiency came with trade-offs that many found problematic.

The accuracy concerns centered on AI's inability to handle natural speech variations and conversational contexts. Students reported frustration when attempting more naturalistic pronunciation, feeling penalized for speech patterns they believed were correct. This tension reflects a fundamental challenge in automated speech recognition: balancing standardized assessment criteria with authentic communicative competence.

Transparency emerged as a critical factor influencing trust. The "black box" nature of AI scoring algorithms left students uncertain about evaluation criteria, undermining their confidence in the results. As P12 expressed, understanding the scoring logic could potentially increase trust, suggesting that algorithmic transparency might bridge the efficiency-accuracy gap.

Theme 2: Human Empathy and Contextual Understanding

In stark contrast to AI limitations, human evaluation was consistently framed through an empathy lens, with students emphasizing emotional connection and contextual sensitivity. Participants valued human raters' ability to recognize effort and intention beyond mere linguistic accuracy. This perception aligns with sociocultural learning theories that emphasize the importance of human interaction in language development.

The theme of personalized feedback emerged prominently, with students distinguishing between AI's "formulaic" responses and teachers' adaptive guidance. Human evaluators were perceived as capable of tailoring feedback to individual learning needs and emotional states. P5's distinction between AI telling "what's wrong" versus teachers explaining "how to fix it" illustrates this perceived difference in feedback depth and utility.

Students also emphasized human capacity for interpreting pragmatic competence—elements like tone, humor, and cultural context that automated systems struggle to process. This finding suggests that despite technological advances, human judgment remains crucial for assessing complex communicative competencies that extend beyond mechanical accuracy.

Theme 3: Bias and Inclusivity Concerns Leading to Hybrid Preferences

Perhaps the most pedagogically significant theme involved students' awareness of algorithmic bias, particularly regarding accent and dialect variations. Participants with non-standard accents reported feeling marginalized by AI systems that appeared to privilege native-speaker norms. P22's concern about being "judged unfairly" because of accent reveals awareness of how AI training data limitations can perpetuate linguistic hierarchies.

This bias concern directly influenced students' preferences for hybrid assessment models. Rather than viewing automated and human evaluation as competing approaches, participants conceptualized them as complementary systems serving different purposes. The hybrid model emerged as a pragmatic solution: utilizing AI for routine practice and preliminary feedback while reserving human evaluation for high-stakes assessment and nuanced judgment.

Students' hybrid preferences reflect sophisticated understanding of each system's strengths and limitations. They recognized AI's value for consistent, immediate feedback on technical aspects while appreciating human evaluators' capacity for holistic assessment and motivational support. P10's distinction between "everyday practice" and "important exams" illustrates this nuanced approach to technology integration in assessment.

3.2.3 Implications for Assessment Design

These qualitative findings reveal that student perceptions extend beyond simple preference statements to encompass complex understanding of assessment validity, fairness, and pedagogical effectiveness. The themes suggest that successful implementation of automated evaluation requires addressing transparency, inclusivity, and complementarity rather than replacement of human judgment. Students' sophisticated articulation of hybrid models indicates readiness for integrated assessment approaches that leverage both technological efficiency and human expertise.

Discussion

Student Perceptions Reveal Complex Technology Acceptance Patterns

As summarized in Table 2, three primary themes emerged from the qualitative analysis: efficiency-accuracy tensions, human empathy and contextual understanding, and bias and inclusivity concerns leading to hybrid preferences. This study found that English Education students demonstrate nuanced perceptions toward automated and human evaluation that extend beyond simple preference

dichotomies to encompass sophisticated understanding of assessment validity and pedagogical effectiveness.

Automated Evaluation Acceptance Through Technology Acceptance Model Lens

The positive reception of automated evaluation tools aligns closely with the Technology Acceptance Model (TAM), particularly regarding Perceived Usefulness and Perceived Ease of Use dimensions. This study found that students highly valued AI's speed and accessibility (M = 3.9 for feedback effectiveness), supporting Davis's (1989) proposition that perceived usefulness strongly influences technology acceptance.

However, this study found a critical divergence from traditional TAM predictions in the gap between positive attitudes and behavioral intentions. Despite high Personal Innovativeness scores (M = 4.0), students showed significantly lower Behavioral Intention to use automated evaluation frequently (M = 2.7). This contradiction suggests that in educational assessment contexts, trust and transparency serve as crucial mediating factors beyond TAM's core constructs. While TAM explains much of AI's initial appeal, students' continued preference for human evaluators is better explained through Self-Efficacy Theory.

Self-Efficacy Theory Explains Human Evaluator Preferences

Students' strong preference for human evaluation can be understood through Bandura's (1997) Self-Efficacy Theory, which emphasizes social persuasion and emotional support in building learner confidence. This study found that human evaluators provided personalized feedback that enhanced students' self-efficacy beliefs, offering "how to fix it" guidance rather than AI's standardized responses.

The emotional connection students reported with human evaluators reflects self-efficacy theory's emphasis on encouragement and specific performance feedback. This study found that students valued human evaluators' ability to recognize "effort" and "intention," indicating that motivational feedback is integral to effective language assessment. This preference for human empathy, however, intersects with critical concerns about fairness that are best understood through assessment validity frameworks.

Assessment Validity Framework Illuminates Bias Concerns

Through Messick's (1989) assessment validity framework, students' bias concerns reveal sophisticated understanding of fairness issues. This study found that participants with non-standard accents perceived AI systems as privileging native-speaker norms, demonstrating intuitive awareness of construct-irrelevant variance—factors that influence scores unrelated to the intended construct.

Students' expressions of feeling "marginalized" by AI evaluation reflect understanding that validity extends beyond statistical properties to encompass social consequences. This study found that algorithmic bias concerns were not merely technical complaints but reflected awareness of how assessment practices can perpetuate linguistic hierarchies. These validity concerns directly informed students' preference for hybrid assessment approaches.

Hybrid Assessment Models as Theoretical Integration

Perhaps most significantly, this study found that students conceptualized hybrid assessment models that theoretically integrate multiple validity frameworks. Rather than viewing automated and human evaluation as competing approaches, participants demonstrated sophisticated understanding of complementary validity strengths: AI for consistency and efficiency (addressing reliability concerns), human evaluation for contextual sensitivity and motivational support (addressing consequential validity).

Students' distinction between "everyday practice" and "important exams" reflects nuanced understanding of assessment stakes and purposes. This study found that students intuitively recognized that different assessment contexts require different validity priorities formative assessment benefiting from AI's immediate feedback capabilities, while summative assessment requiring human judgment for holistic evaluation and fairness considerations (Aeni et al, 2025)

Implications for Technology Integration Theory

These findings extend existing technology acceptance theories by highlighting the unique considerations that emerge in assessment contexts. This study found that traditional TAM constructs require supplementation with trust, transparency, and fairness dimensions when applied to educational evaluation tools. The gap between positive attitudes and behavioral intentions suggests that assessment technologies face higher adoption barriers than general educational tools due to their consequential nature for learner outcomes and identity.

Furthermore, this study found evidence supporting sociocultural learning theories that emphasize human interaction's irreplaceable role in language development. Students' preference for human empathy and contextual understanding suggests that effective technology integration in language assessment must preserve rather than replace the social dimensions of learning and evaluation.

Contributions to Assessment Practice

This study found that successful implementation of automated evaluation requires addressing three critical factors: algorithmic transparency to build trust, inclusive training data to ensure fairness, and complementary rather than replacement integration with human evaluation. The findings suggest that the future of language assessment lies not in choosing between automated and human evaluation but in thoughtfully orchestrating their respective strengths within coherent assessment ecosystems that serve both efficiency and pedagogical depth..

4. CONCLUSION

This study offers a novel contribution to the literature by highlighting how English Education students in an Indonesian higher education context conceptualize automated and human assessment as complementary rather than competing, challenging the dominant binary framing in previous research. A key finding reveals a significant implementation gap: despite students reporting high levels of personal innovativeness and generally positive attitudes toward AI-based assessment tools, their behavioral intention to use such systems regularly remains low. This divergence from traditional Technology Acceptance Model (TAM) predictions underscores the need for expanded theoretical frameworks that incorporate trust, transparency, and fairness as central mediators of technology acceptance in educational assessment. Qualitative findings further illustrate students' nuanced understanding of AI's limitations, particularly regarding efficiency-accuracy tensions, the absence of human empathy, and concerns over algorithmic bias—especially in relation to non-standard accents. Notably, students proposed hybrid assessment models that allocate formative feedback tasks to AI while reserving high-stakes evaluative roles for human assessors, reflecting a sophisticated learner-driven vision of integrative assessment. However, this study is limited by its cross-sectional design and reliance on self-reported data, which may not fully capture long-term usage patterns or actual performance outcomes. Future research should adopt longitudinal and experimental methodologies to track changes in perception and effectiveness over time, explore individual differences through advanced statistical modeling, and expand TAM frameworks to include socio-cultural and pedagogical dimensions. Further investigation into optimal hybrid implementation strategies, as well as transparency interventions that enhance trust in AI systems, is also recommended. These findings

underscore the importance of culturally responsive, pedagogically informed assessment designs that preserve the value of human judgment while strategically leveraging the efficiencies of AI.

REFERENCES

Aeni, N., Khang, A., Al Yakin, A., Yunus, M., & Cardoso, L. (2024). Revolutionized teaching by incorporating artificial intelligence chatbot for higher education ecosystem. In *AI-centric modeling and analytics* (pp. 43–76). CRC Press.

Aeni, N., Muthmainnah, M., Nurfadhilah, A. S., & Inayah, F. (2025). AI-Driven Classroom Conversations: Revolutionizing Education 5.0 for Enhanced Student Engagement in Speaking Class. In *Innovations in Educational Robotics: Advancing AI for Sustainable Development* (pp. 173-192). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-6165-8.ch009

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burston, J. (2021). Intelligent computer-assisted language learning: A review of the field. *Computer Assisted Language Learning*, *34*(5–6), 429–449. https://doi.org/10.1080/09588221.2021.1901745

Chapelle, C. A., & Chung, Y.-R. (2021). The promise of NLP and speech processing in language assessment. *Language Testing*, *38*(2), 189–200. https://doi.org/10.1177/0265532220925731

Dai, H., Ai, H., & Lin, C. (2023). Generative AI as a feedback provider in second language writing: Comparing ChatGPT and human instructors. *Computers & Education*, *201*, 104785. https://doi.org/10.1016/j.compedu.2023.104785

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Fryer, L. K., & Carpenter, R. (2022). Chatbot language learning: Moving beyond the hype. *Computer Assisted Language Learning*, *35*(3), 203–218. https://doi.org/10.1080/09588221.2022.2032184

Godwin-Jones, R. (2020). Dealing with complexity in language learning: Language learning analytics and AI. *Language Learning & Technology*, *24*(1), 1–9. https://doi.org/10125/44707

He, Y., Chen, J., & Liu, Z. (2020). Learner perceptions of automated essay scoring and feedback: A mixed-methods study. *System*, *92*, 102279. https://doi.org/10.1016/j.system.2020.102279

Hummel, S., & Donner, M.-T. (Eds.). (2023). *Student assessment in digital and hybrid learning environments*. Springer.

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2021). Artificial intelligence applications in education. *International Journal of Emerging Technologies in Learning*, *16*(11), 196–213. https://doi.org/10.3991/ijet.v16i11.19657

Johnson, K., & Valente, M. (2023). Artificial intelligence in second language acquisition: Enhancing self-regulated learning through adaptive scaffolding. *Language Learning & Technology*, *27*(1), 1–18. https://doi.org/10.1017/LLT.2023.105

Kim, G.-m., & Lee, S.-j. (2016). Korean students' intentions to use mobile-assisted language learning: Applying the technology acceptance model. *International Journal of Contents*, *12*(3), 47–53. https://doi.org/10.5392/IJoC.2016.12.3.047

Knoch, U., & Macqueen, S. (2020). *Assessing English proficiency in the age of automation: A critical perspective on AI-driven testing*. Routledge.

LeMay, D. J., Morin, M. M., Bazelais, P., & Doleck, T. (2018). Modeling students' perceptions of simulation-based learning using the technology acceptance model. *Clinical Simulation in Nursing*, *20*, 28–37. https://doi.org/10.1016/j.ecns.2018.04.004

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

Lu, X., & Zhang, W. (2023). AI and human teachers in language education: Bridging efficiency and

adaptability. *Journal of Educational Technology*, *40*(3), 215–230. https://doi.org/10.1080/09588221.2023.2256789

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications.

Pereira, F. D., Oliveira, E., Rodrigues, L., Cabral, L., Oliveira, D., Carvalho, L., Silva, I., Brandão, A., Isotani, S., & Mello, R. F. (2023). Evaluation of a hybrid AI-human recommender for CS1 programming assignments. In *European Conference on Technology Enhanced Learning* (pp. 289–303). Springer.

Rahayu, T. (2021). AI-assisted language learning for non-native English speakers. *Journal of English Educators Society*, *6*(1), 33–49. https://doi.org/10.21070/jees.v6i1.849

Selwyn, N. (2022). *Education and technology: Key issues and debates* (3rd ed.). Bloomsbury Academic.

Sun, Y., Wang, J., & Liu, C. (2022). The impact of artificial intelligence on second language acquisition: A systematic review. *Frontiers in Psychology*, *13*, 1049139. https://doi.org/10.3389/fpsyg.2022.1049139

Wang, T., & Liu, Y. (2023). Advancing pragmatic competence assessment through AI-driven discourse analysis. *Computer-Assisted Language Learning*, *36*(2), 189–210. https://doi.org/10.1080/09588221.2023.2184967

Wang, Y., & Lin, C. (2023). Bridging AI and human interaction in language learning: Challenges and future directions. *Computer Assisted Language Learning*, *36*(2), 145–163. https://doi.org/10.1080/09588221.2023.1874563

Yastibas, A. E., & Yastibas, G. C. (2021). Learners' opinions on the use of AI-based tools in EFL classrooms. *Education and Information Technologies*, *26*, 3587–3606. https://doi.org/10.1007/s10639-021-10439-1

Zou, D., Huang, Y., & Xie, H. (2022). A review of research on AI-supported language learning and teaching. *Computer Assisted Language Learning*, *35*(1–2), 1–25. https://doi.org/10.1080/09588221.2022.2054844