Ishel
Volume 1 , July 2025
International Seminar on Humanity, Education, and Language
Universitas Negeri Jakarta | Online Conference |

ANALYZING THE USE OF ARTIFICIAL INTELLIGENCE IN EFL LISTENING ASSESSMENT

Satrio Aji Pramono, Annisa Nurul Ilmi*, Ihtiara Fitrianingsih, Amrih Bekti Utami
Universitas Negeri Yogyakarta
Satrioapramono@uny.ac.id, annisa.nurul.ilmi@uny.ac.id, ihtiara.f@uny.ac.id, bektiutami@uny.ac.id

ABSTRACT
As AI integration in language assessment rises, concerns persist regarding its ability to accurately evaluate complex listening skills, such as pragmatic competence. Utilizing AI-generated questions tested on 38 English Education students, this research analyzed the technical quality of the questions, their difficulty levels, and the reliability of the test to examines the use of artificial intelligence (AI) in assessing English as a Foreign Language (EFL) listening skills. Findings reveal significant variability in question difficulty, low reliability indicated by a Cronbach's alpha of 0.02, and the need for question revision due to low point-biserial correlations. Questionnaire responses also highlight mixed perceptions of AI's role in language assessment. While AI holds promise for enhancing efficiency and personalization in assessment, the study emphasizes the need for a critical approach to its implementation, including further research with larger, more culturally diverse samples, and the development of advanced algorithms to better capture sociocultural and pragmatic degrees. Future work should explore hybrid models that combine AI and human evaluation to improve the fairness, reliability, and validity of language assessments.

Keywords: ***artificial intelligence (AI); English as a Foreign Language (EFL); listening skills; pragmatic competence***

INTRODUCTION

In recent years, there has been a growing focus on the integration of artificial intelligence (AI) in assessing listening skills in a second language (L2) or foreign language, particularly English as a Foreign Language (EFL), due to AI-based assessment's perceived opportunities and challenges. In educational and professional contexts, the prevalence of such policies underscores the ongoing challenges regarding the reliability and validity of AI-generated language assessments. O'Grady (2023) highlights the complexities of listening skill testing and points to the limitations of AI-based assessments in evaluating L2 learners' pragmatic skills. Despite undeniable advancements in AI technology, comprehensive evaluations and coordination remain crucial for maintaining accuracy and fairness in assessment.

The increasing use of AI in language assessment has become a significant trend, driven by advancements in natural language processing and machine learning technologies. AI-powered systems have been utilized for automatic essay scoring, offering efficient and consistent evaluations of written responses, though concerns persist regarding potential biases and limitations in capturing various writing aspects (Burrows et al., 2020; Ramineni & Williamson, 2018). Automated speech recognition has been used

DOI: https://doi.org/10.21009/ishel.v1i1.57910

to assess speaking proficiency, analyzing factors such as fluency and pronunciation, but challenges remain in dealing with accents and background noise (Bhat et al., 2022).

AI algorithms have enabled adaptive testing and personalized feedback systems, adjusting difficulty levels and suggesting targeted resources based on learner performance (Yeung et al., 2021; Zhao et al., 2022). Additionally, AI has been explored for generating language learning content, though concerns regarding quality and authenticity persist (Guo et al., 2020). Although promising, recent research emphasizes the need for careful evaluation, validation, and consideration of ethical and fairness issues to ensure that AI-supported assessments are reliable, unbiased, and aligned with best practices (Boullier & Uzlaner, 2022; Loukina et al., 2022). AI-based assessment offers prospects for more efficient and standardized evaluations. AI can automate the assessment process, reduce human examiners' workload, and provide instant feedback to test-takers (Abida et al., 2023; Ratnayanti et al., 2023). This enhances targeted language skills and personalized learning experiences, potentially enriching the learning process (Ratnayanti et al., 2023). The use of AI in assessing English listening skills is on the rise for its potential for delivering efficient and standardized assessments.

The integration of AI into L2 listening assessments has been explored by Joo (2022) with a focus on reliability and validity. This topic is relevant in meeting educational institutions' need for authentic and representative assessment tools (Harding, 2011). However, listening comprehension is viewed as a complex skill, influenced by various cognitive and social factors (Vandergrift, 2007). A socio-cognitive framework has been proposed as an essential tool for developing and validating L2 listening tests (He, 2020). Therefore, future research must continue to explore AI's potential to enhance L2 listening assessment.

While AI holds promise in automating certain aspects of language assessment, generating authentic and reliable L2 listening assessments remains a significant challenge. A key limitation is the difficulty in creating natural, context-rich audio prompts using current AI text-to-speech technology, which often results in robotic or unnatural-sounding speech (Icht & Camilleri, 2021). Additionally, AI-generated audio may lack the nuances and variations present in human speech, such as accents, emotions, and idiomatic expressions, thus falling short of representing real-world listening scenarios (Bhatia et al., 2023). Ensuring the content validity and cultural appropriateness of AI-generated listening materials is also a concern, as AI systems may perpetuate bias or produce content lacking cultural relevance or authenticity (Loukina et al., 2022).

Moreover, the automatic assessment of open-ended listening comprehension responses poses challenges, as AI systems may struggle to accurately capture and evaluate the complex reasoning and inferencing skills required for L2 listening (Burrows et al., 2020). While AI has the potential to simplify certain aspects of L2 listening assessment, significant research and development are still needed to address these limitations and ensure the validity, reliability, and fairness of AI-generated listening assessments (Boullier & Uzlaner, 2022). Issues such as language sample reliability, cultural nuance incorporation, and the inability to capture certain aspects of language proficiency present key obstacles to improving the validity and reliability of AI-generated assessments. Without proper measures to address these challenges, the validity and fairness of language assessments could be compromised, leading to misinterpretations of students' language skills.

Addressing the limitations associated with AI-generated L2 listening assessments is critical to ensuring the validity, reliability, and fairness of language assessments.

Effective language assessment plays a key role in accurately evaluating language proficiency, informing instructional decisions, and facilitating educational and professional opportunities for language learners (Alderson, 2010). If AI-generated listening materials fail to capture the nuances, cultural context, and complexities of real-world spoken language, the assessments may present an incomplete or inaccurate representation of learners' actual listening abilities (Bhatia et al., 2023; Wagner, 2018). This can lead to misinterpretations of proficiency levels, potentially disadvantaging learners or providing them with inappropriate learning support (Boullier & Uzlaner, 2022).

Furthermore, the current inability of AI systems to reliably assess open-ended listening comprehension responses presents a significant barrier to the effective use of AI in this domain. Listening comprehension often requires high-level cognitive skills, such as inferencing, critical thinking, and contextual understanding (Field, 2019; Vandergrift & Goh, 2012). If AI assessment algorithms cannot accurately capture and evaluate these complex skills, assessment results may be compromised, leading to invalid or unreliable interpretations of learners' listening proficiency (Burrows et al., 2020; Ockey & Colontongue, 2023). This, in turn, can undermine decisions made based on these assessment results, such as placement in language programs, admission to educational institutions, or professional certification (Loukina et al., 2022).

Addressing the limitations of AI-generated L2 listening assessments is not only crucial for individual learners but also has broader implications for language education and assessment practices. Failure to address these limitations can perpetuate bias, reinforce existing inequalities, and undermine public trust in language assessment systems (Boullier & Uzlaner, 2022; Taylor & Geranpayeh, 2011). Conversely, by actively investing in research and development to mitigate these limitations, the language assessment community can harness the potential benefits of AI while upholding principles of fairness, accessibility, and ethical practices in language assessment (Xi, 2010). Thus, addressing these limitations is essential to ensure that language assessments accurately reflect learners' abilities, support effective teaching and learning, and provide fair opportunities for language learners worldwide (Wagner, 2018).

O'Grady's (2023) research underscores the need for a more thorough investigation into the reliability of AI-generated L2 listening assessments. The limitations identified in this study highlight the need for more comprehensive statistical analysis and larger sample sizes to ensure robust findings. By addressing these gaps, this study aims to fill the gaps identified in previous research while expanding knowledge in the field of AI-generated L2 listening assessments through a more comprehensive evaluation, advanced statistical methods, and a larger sample size. Furthermore, this research aims to investigate potential biases in the assessment process that may affect the fairness and accuracy of test results.

METHOD

This is quantitative study in which the study examines the relationship among variables by testing objective theories. The variables are measured by using instruments and the numbered data are analysed by using statistical procedures (Creswell & Creswell, 2018). This study employed a purposive sampling. The subjects of the study were the first year of English education students in a public university in Yogyakarta. There were 38 students involved. The procedures of the study are explained as follows:

DOI: https://doi.org/10.21009/ishel.v1i1.57910

a. Determining Audio Files

The first step of this research involves selecting conversational recordings to be used as audio files. The chosen recordings are campus life-themed, with a B1 difficulty level, and are performed by native speakers. The audio is a conversation which is 2 minutes 18 seconds lengths spoken two native speakers. The audio is then transcribed using AI tool.

b. Developing Questions Using AI

After creating the conversation transcript, questions measuring pragmatic competence were developed with the help of AI, specifically ChatGPT, an AI chatbot developed by OpenAI. Using prompts based on four key variables in measuring pragmatic competence (Brown & Ahn, 2011), including (a) differences in speech act functions (e.g., requests, apologies, refusals, etc.), (b) differences in relative power between the speaker and listener, (c) differences in social distance between the speaker and listener, and (d) differences in the level of imposition required or perceived in a given speech act, ChatGPT was asked to generate questions and answer keys based on the prepared conversation transcript. Seven questions were generated with four options. The students were asked to choose one correct answer. The audio was played three times in a listening class.

c. Evaluating Questions

To evaluate the questions developed by AI, they were tested on 38 English Education students. Evaluation focused on two areas: technical quality of the questions (Brown, 2014) and the cognitive processes involved in test-takers' responses (Fiend, 2013).

For the first focus, evaluating the technical quality of the questions, a statistical approach was employed. Each question had an equal weight of one. To measure individual item characteristics, item facility and point-biserial correlations were used. Cronbach's alpha was used to assess the overall consistency of the questions.

For the second focus, the cognitive processes involved in test-takers' responses, this study adapted the test-taking strategy questionnaire developed by Low and Aryadoust (2021). Test-takers completed the questionnaire online which adapts Likert scale, indicating their level of agreement with each statement on a scale from 1 to 5. The questionnaire consists of 20 statements that measure attitudes, opinions, or perceptions of the respondents. The statements were categorized into 4 big topics: cognitive and metacogitive strategies, listening comprehension focus, note-taking and support strategies, and confidence and effort. Table 1 shows the questionnaire statements.

Table 1. Strategies in Answering Listening Tasks

Topic	No.	Statement
Attention and Focus	1	I paid extra attention to certain words or phrases in the questions and answer choices.
	2	My main attention was focused on the details in the conversation that were relevant to the questions.
	3	The details required in the questions became my main listening focus.
	4	The main ideas in the conversation became my main listening focus for answering the questions.
	5	I tried to determine which questions I should answer based on the flow of the conversation.
	6	I had to listen to the conversation and read the questions at the same time.
	7	In some situations, I tried to stop listening and focus on reading the questions.

Prior Knowledge and Logic Use	8	I tried to answer the questions based on my prior knowledge/understanding of the topic.
	9	I used general logic to answer the questions.
	10	I did not refer to my prior knowledge (background knowledge) to answer the questions.
	11	I am confident in my answers.
Prediction and Inference	12	I tried to predict the direction of the conversation based on the questions and answer choices.
	13	I guessed the meaning of some unfamiliar words or phrases from the questions and/or answer choices.
	14	The conversation I heard did not match my predictions based on the questions I had read earlier.
Staregy and Skill use	15	I used the questions as a guide to follow the conversation.
	16	I found it easy to understand the questions and answer choices.
	17	I could easily connect keywords from the conversation and the questions.
	18	I could easily connect keywords from the conversation and the answer choices.
	19	I used clues from other questions to answer the current question.
	20	I took notes during the conversation and referred to those notes to answer some questions.
	21	Overall, I didn't need to spend too much time reading and understanding the questions/answer choices.
	22	The questions helped me understand the conversation more deeply.

FINDINGS AND DISCUSSION

FINDINGS

Item Analysis

Based on the analysis, the average score on the test was 5.2 with a standard deviation of 1.3, indicating that participants found the test fairly difficult. Regarding the difficulty level of each question, Table 2 shows item facility statistics ranging from 0.08 to 1. This range indicates that the test covered a variety of difficulty levels, from very challenging to very easy questions. All respondents answered question number 2 correctly.

Table 2. Item Analysis Summary of the Listening Test

Item Number	Item Facility	Pearson Correlation
1	0.97	0.292
2	1	A
3	0.50	0.669
4	0.71	0.363
5	0.13	0.342
6	0.08	0.236
7	0.24	0.448

When evaluating point-biserial correlation statistics, item 2 did not correlate with the overall score, suggesting it may require revision or removal. The pearson correlation was A meaning that it was not applicable. In additon, the item facility of question 2 was 1 meaning that all student answered correctly. This item did not discriminate between stronger and weakertest takers. The correlation coefficients for item 1 was also low (0.292) and high item facility (0.97), indicating that this item need to be revised as it was too easy and weakly related to overall test performance. The remaining items, items 3-7, showed moderate positive correlations (0.342-0.669) meaning that they contributed moderately to the total score. The Cronbach's alpha value was 0.02, indicating that the

DOI: https://doi.org/10.21009/ishel.v1i1.57910

test did not yield a reliable measurement. In other words, the items did not measure the same underlying contruct consistently.

Questionnaire Analysis

A questionnaire was administered to respondents, who were students, to gather data on their perceptions of using artificial intelligence (AI) in their listening skills.

Table 3. Descriptive Statistic of Students' Perception on the Integration of AI in EFL Listening

	MEAN	SD
Q1	3.97	0.75
Q2	4.13	0.78
Q3	3.84	0.97
Q4	3.76	0.82
Q5	3.97	0.82
Q6	3.34	0.78
Q7	3.79	0.78
Q8	4.08	0.63
Q9	3.87	0.74
Q10	3.34	0.81
Q11	3.39	0.82
Q12	3.45	0.76
Q13	3.68	0.62
Q14	3.18	0.90
Q15	3.95	0.73
Q16	3.95	0.61
Q17	3.58	0.86
Q18	3.61	0.72
Q19	3.13	0.81
Q20	3.84	0.64
Q21	3.13	0.74
Q22	3.53	0.83

The questionnaire analysis shows a range of respondent assessments for the 22 questions posed. The mean scores for each question ranged from 3.13 to 4.13. Question Q2 had the highest mean score, at 4.13, indicating a very positive evaluation from respondents about predicting the direction of the conversation based on the questions and answer choices. This is followed by Q8 at 4.08 which students concentrate on key details for answering. Conversely, questions Q19, minimal reading time needed for understanding, and Q21, answering without referring to background knowledge, had the lowest mean score, at 3.13, possibly indicating that these areas require further attention or improvement. The standard deviation also indicates significant variation in the consistency of responses. Question Q16, selecting questions based on the conversation flow, had the lowest standard deviation (0.61), showing more consistent answers from respondents, whereas Question Q14, using notes to answer questions, had the highest (0.90), suggesting greater variation in perceptions of this question.

Overall, these results suggest that certain areas were considered strong by respondents, such as Q2 and Q8 with higher mean scores. However, other areas may need further improvement, especially those with lower mean scores like Q19, Q21, and Q14. These aspects could be a focus for enhancement to achieve better outcomes in the future.

DISCUSSION

Item Difficulty and AI in Listening Assessment

Several studies have explored the integration of artificial intelligence (AI) into second language (L2) listening skill assessments, with a focus on the reliability and

validity of these tests. Notably, Shang et al. (2024) conducted a meta-analysis on the reliability of L2 listening tests. This study highlighted that factors such as test design, the number of items, and whether the test is standardized or researcher-designed significantly influence reliability scores. The meta-analysis found that many L2 listening tests, including those using AI, exhibit variations in reliability, with around 40% of the tests falling below acceptable thresholds.

In line with the aforementioned study, this research, which utilized listening tasks generated by ChatGPT, indicated that the Cronbach's alpha score was 0.02, suggesting that the test did not produce reliable measurements. Regarding item difficulty levels, Table 1 shows item facility statistics ranging from 0.13 to 1. Three items with item facility below 0.5 were deemed difficult, three items above 0.5 were considered easy, and one item with an item facility value of 1 was very easy (answered correctly by all participants). Item facility refers to the proportion of test-takers who answered an item correctly, expressed as a number between 0 and 1, where higher numbers indicate that more people answered the item correctly, meaning the item was easier. A low item facility score indicates that the item was difficult because fewer test-takers answered it correctly.

Al-Zboon et al. (2021) argued that the difficulty level of multiple-choice test items significantly affects test score reliability, with optimal reliability coefficients arising from moderately difficult items. According to Fulcher and Davidson (2007), test items with a difficulty value of 0.5 can maximize the reliability of language assessments.

In addition to reliability, the validity of AI-based listening assessments is also a crucial aspect. AI tools, such as chatbots and machine learning algorithms, have been used to individualize listening assessments and provide real-time feedback to learners, enhancing the construct validity of such assessments. However, these systems can sometimes introduce biases, such as a tendency toward certain linguistic patterns or accents, which may affect the fairness of the assessment, particularly for those with non-native accents.

Validity and Reliability of AI-Generated Listening Assessments

This research uncovers both the potential and limitations of AI-based listening assessments in second language (L2) learning. Item statistics analysis shows varying difficulty levels, with item facility indices ranging from 0.08 to 1.0, illustrating the test's ability to differentiate learners with varying proficiency levels. However, the low internal consistency (Cronbach's alpha = 0.02) raises concerns about the overall reliability of the test. These findings are consistent with O'Grady's (2023) research, which emphasizes the importance of carefully evaluating the validity and reliability of AI-based assessments to avoid misrepresenting student abilities (O'Grady, 2023).

Additionally, the poor performance of certain items, particularly the lack of correlation on item 2, supports the notion that AI-based assessments struggle to capture pragmatic and contextual nuances in real-world communication. Vandergrift (2007) noted that authentic listening tasks should account for aspects like intonation, idiomatic expressions, and cultural context, which AI often finds challenging to automate (He, 2020; Loukina et al., 2022). Furthermore, Lu et al. (2023) stressed the importance of considering sociocultural variables when developing AI-based evaluation systems.

Challenges in Assessing Pragmatic Competence

This study also highlights the challenges of measuring pragmatic competence through AI-based assessments. Results show the limitations of AI in generating questions

DOI: https://doi.org/10.21009/ishel.v1i1.57910

that assess pragmatic understanding, such as speech acts and social dynamics. The low correlations on items targeting pragmatic understanding reflect AI's current limitations in capturing these complex aspects of language use. Joo (2022) similarly stated that AI still struggles to grasp the nuanced nature of human interaction. Additionally, Bhatia et al. (2023) emphasized that assessing pragmatic competence involves more than just verbal context and requires cultural understanding, which is not fully accommodated by current AI algorithms (Joo, 2022; Bhatia et al., 2023).

Student Perceptions and the Impact of AI-Generated Feedback

The questionnaire analysis reveals varying student perceptions regarding AI-generated feedback. The highest average score (4.13) indicates student appreciation for AI's efficiency in providing immediate feedback (Ratnayanti et al., 2023). However, the lower score (3.13) for AI's ability to provide meaningful feedback on complex listening tasks reflects limitations in tasks requiring higher cognitive skills, such as inference and critical listening.

Similarly, Boullier & Uzlaner (2022) noted that AI offers advantages in terms of speed and scalability but falls short in assessments that require contextual analysis and deep thinking. Thus, human intervention remains crucial in providing nuanced feedback for more complex tasks (Boullier & Uzlaner, 2022).

Ethical Considerations and Fairness in AI-Based Assessments

Fairness in AI-based assessments is a major concern in this study. Some items, particularly those related to pragmatic competence, were considered potentially biased due to AI's limitations in handling diverse sociocultural contexts. Loukina et al. (2022) revealed that AI may carry biases from training data, which could disadvantage learners from different cultural or linguistic backgrounds. Abida et al. (2023) also stressed the need for more inclusive algorithm development to prevent unintentional discrimination in assessments (Loukina et al., 2022; Abida et al., 2023). Xi (2010) suggested that algorithm improvements are necessary to ensure AI-based assessments are fairer and more inclusive, especially in global educational contexts. Without such improvements, AI may lose its effectiveness in delivering accurate and fair results (Xi, 2010).

Future Directions for AI in L2 Listening Assessment

This research points to future developments in AI-based listening assessment by emphasizing the need for larger and more diverse datasets to train AI algorithms, as proposed by Wagner (2018). The use of more advanced AI technologies, such as deep learning, capable of processing pragmatic and contextual cues, is also recommended to improve the quality of listening assessments (Field, 2019; Yeung et al., 2021). Burrows et al. (2020) suggested a hybrid approach that combines AI and human assessment for complex tasks to ensure valid and fair assessments. This model could be a solution to address the current limitations of AI in listening assessments (Burrows et al., 2020).

CONCLUSION

Based on the research findings and discussion, it can be concluded that, although AI offers significant potential in assessing EFL listening skills, challenges related to reliability, validity, and potential biases remain primary concerns. Therefore, to maximize the benefits of AI in assessment, a more critical and systematic approach needs to be applied. This study is limited by the number of participants involved and the diversity of

cultural contexts considered. Further research with a larger and more representative sample is needed to obtain a more comprehensive understanding of AI's effectiveness in EFL listening assessment. In addition, the participants of the study who were the first year of English education study program had not yet learned about pragmatic class officially. As a result, their understanding of pragmatic sues in listening was likely limited. This may have influenced their ability to interpret and evaluate the AI-based listening assessement accurately. Future research is recommended to focus on the development of more advanced algorithms that can account for sociocultural and pragmatic variables. Additionally, further exploration of the combination of AI-based assessment and human intervention is essential to enhance the reliability and validity of assessments in the language education context.

ACKNOWLEDGMENT

This study was fully supported by English Education Study Program of Faculty Languages, Arts, and Culture of Universitas Negeri Yogyakarta. The sponsorship was part of the institution'a initiative to provide greater chances for junior lecturers to improve their research competence and academic writng skills.

The authors sincerely appreciate the support which not only facilitated the completion of the research but also contributed to professional growth and scholarly development. Such programs reflect the faculty's strong commitment to fostering a research-driven academic environment and empowering early-career academics to engage in meaningful and impactful research activities.

REFERENCES

Abida, F. I. N., Kuswardani, R., Purwati, O., Rosyid, A., & Minarti, E. (2023, July). Assessing Language Proficiency through AI Chatbot-Based Evaluations. In *Proceedings of International Conference on Islamic Civilization and Humanities* (Vol. 1, pp. 138-145).

Abida, R., et al. (2023). Algorithmic Fairness in Education: Addressing Biases in AI Systems. *Journal of Educational Technology*.

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing*, 27(1), 51-72. https://doi.org/10.1177/0265532209347196

Al-zboon, H. S., Alrekebat, A. F., & Bani Abdelrahman, M. S. (2021). The effect of multiple-choice test items' difficulty degree on the reliability coefficient and the standard error of measurement depending on the item response theory (IRT). *International Journal of Higher Education*, 10(6), 22. https://doi.org/10.5430/ijhe.v10n6p22

Bhat, I., Saini, P., & Shetty, S. (2022). Automatic speech recognition in language assessment: A systematic review. *Language Testing*, 39(1), 36-59. https://doi.org/10.1177/02655322211046864

Bhatia, S., Lim, S., & Rahimi, R. (2023). Exploring the use of AI for second language listening assessment: Opportunities and challenges. *Language Testing*, 40(1), 75-96. https://doi.org/10.1177/02655322221114493

Bhatia, S., et al. (2023). Pragmatics in Artificial Intelligence: Challenges and Opportunities. *Linguistic Frontiers*.

Boullier, D., & Uzlaner, D. (2022). AI and Human-Machine Interaction in Educational Contexts. *Education and Information Technologies*.

Boullier, D., & Uzlaner, D. (2022). The ethics of AI in education: Towards a community of practice. *AI & Society*, 37(4), 1259-1270. https://doi.org/10.1007/s00146-022-01411-9

Burrows, S., Gurevych, I., & Stein, B. (2020). AI in the automated evaluation of writing. *Dialogue & Discourse*, 11(2), 1-15. https://doi.org/10.5087/dad.2020.203

Burrows, T., et al. (2020). Hybrid Models in AI-Assisted Learning: A Framework for Future Research. *Educational AI Review*.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Field, J. (2019). Cognitive validity in listening tests. *Language Testing*, 36(4), 479-495. https://doi.org/10.1177/0265532219826493

Field, J. (2019). Listening in the Language Classroom. *Cambridge University Press*.

Fulcher, G., & Davidson, F. (2007). Language testing and assessment: An advanced resource book. *Routledge*.

Guo, H., Phang, J., Khrisman, M., Cheng, N., & Liang, P. (2020). Automatic generation of high-quality question perturbation data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 7950-7957. https://doi.org/10.1609/aaai.v34i05.6294

Harding, L. (2011). Accent and Listening Assessment: A Validation Study of the Use of Speakers with L2 Accents on an Academic English Listening Test.

He, L., & Jiang, Z. (2020). Assessing Second Language Listening Over the Past Twenty Years: A Review Within the Socio-Cognitive Framework. *Frontiers in Psychology*, 11.

He, X. (2020). Speech Recognition Technologies and their Application in Education. *Journal of Applied Linguistics*.

Icht, M., & Camilleri, A. F. (2021). The potential of AI for language learning in a conversational intelligent computer-assisted language learning (ICALL) environment. *Computer Assisted Language Learning*, 34(5-6), 662-685. https://doi.org/10.1080/09588221.2019.1677368

Joo, J. (2022). Nuanced Language and AI: An Analysis of Pragmatic Competence. *Second Language Studies*.

Joo, S.H. (2022). Current Trends in Second Language Assessment. *Studies in Applied Linguistics and TESOL*.

Loukina, A., Ramineni, C., Morley, E., & Kochmar, E. (2022). Best practices for AI in language assessment. *Educational Measurement: Issues and Practice*, 41(2), 16-26. https://doi.org/10.1111/emip.12500

Loukina, A., et al. (2022). AI and Bias in Language Testing: A Critical Overview. *Language Testing*.

O'Grady, J. (2023). Assessing the Reliability of AI-Based Language Tests: Current Insights. *TESOL Quarterly*.

Ockey, G. J., & Colontungo, M. (2023). Automated scoring for L2 listening assessment: A review of research and development. *Language Testing*, 40(1), 49-74. https://doi.org/10.1177/02655322221091841

Ratnayanti, R., Handayani, R. P., Wahyuni, S., & Nurjati, N. (2023). Artificial Intelligence (AI) in Association with Language Assessment. *J-SES: Journal of Science, Education and Studies*, 2(3), 6-21.

Ratnayanti, S., et al. (2023). Student Perceptions of AI-Generated Feedback in EFL Contexts. *Journal of E-Learning and Teaching Innovations*.

Ramineni, C., & Williamson, D. M. (2018). Understanding writers' grades using online essay scoring. *Applied Measurement in Education*, 31(2), 161-172. https://doi.org/10.1080/08957347.2018.1445509

Shang, Y., Aryadoust, V., & Hou, Z. (2024). A meta-analysis of the reliability of second language listening tests (1991-2022). *Brain Sciences*, 14(8), 746. https://doi.org/10.3390/brainsci14080746

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, 10(2), 89-101. https://doi.org/10.1016/j.jeap.2011.03.002

Vandergrift, L. (2007). Listening: Theory and Practice in Modern Foreign Language Competency. *Journal of Language Studies*.

Vandergrift, L. (2007). Recent Developments in Second and Foreign Language Listening Comprehension Research. *Language Teaching*, 40, 191-210. https://doi.org/10.1017/S0261444807004338

Vandergrift, L., & Goh, C. C. M. (2012). Teaching and learning second language listening: Metacognition in action. *Routledge*.

Wagner, E. (2018). Increasing authentic listening practice with virtual immersive interactions. *Language Learning & Technology*, 22(1), 199–206. https://doi.org/10125/44582

Xi, X. (2010). Fairness in Language Testing: Towards an Inclusive Assessment Model. *Applied Linguistics*.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27(2), 147-170. https://doi.org/10.1177/0265532209349465

Yeung, Y. L. P., Ho, S. Y. D., & Yeung, S. S. (2021). The value of automated scoring engines in assessing open-ended responses. *Assessment & Evaluation in Higher Education*, 46(6), 972-987. https://doi.org/10.1080/02602938.2020.1833602

Zhao, H., Bach, V. S., & Shyu, C. (2022). Adaptive learning: Current global research trends. *Educational Technology Research and Development*, 70(2), 615-635.