DOI:10.59787/2413-5488-2025-50-2-18-31

[1]Zhibek Tleshova, [1]Zhanar Tusselbayeva, [1]Aelita Ichshanova, [1]Aigerim Urazbekova, [1]Meruyert Zhenisbayeva, [2]Ali Orymbayev

[1]*Astana IT University, Astana, Kazakhstan*
[2]*Astana International University (AIU), Astana, Kazakhstan*

RELIABILITY OF AI IN FOREIGN LANGUAGE SPEAKING ASSESSMENT: COMPARING AUTOMATED AND HUMAN SCORING AMONG UNDERGRADUATE IT STUDENTS IN KAZAKHSTAN

Abstract: The integration of Artificial Intelligence (AI) in language assessment, particularly in evaluating speaking skills, has introduced opportunities for greater consistency, efficiency, and scalability in educational contexts. This paper studies the reliability of AI-assisted speaking assessment compared to human-mediated evaluation, with a focus on inter-rater and intra-rater reliability in English as a Foreign Language (EFL) learning. This paper explores the strengths and limitations of AI in automated scoring, such as its capacity for standardization, alongside challenges related to validity, bias, and interpretability of results. This study reviews discrepancies between human and AI scoring due to subjective judgment and training limitations. The study emphasizes the need for standardized rubrics, rater training, and AI model calibration to enhance reliability. This paper concludes by proposing a hybrid assessment framework in which AI complements human raters, supported by methodological and technical improvements in speech recognition and natural language processing. This approach aims to optimize speaking proficiency evaluations while maintaining fairness and educational integrity.

Key words: AI in speaking evaluation, human-mediated assessments, inter-rater reliability, intra-rater reliability, AI-assisted assessment, rater severity.

Introduction

The use of artificial intelligence (AI) in foreign language speaking assessment has brought significant changes to education, particularly in the assessment of speaking skills. Since traditional language assessment methods often rely on human rater, AI-assisted assessment, on the other hand, offers greater standardization and scalability. However, reliability and validity of AI-generated ratings or scores continue to be the subject of debate compared to human assessments, particularly in the context of English as a Foreign Language (EFL) learners.

Artificial intelligence can improve the automated scoring process. Automated scoring involves three processes: feature extraction, feature evaluation and feature accumulation (Ercikan & McCafrey, 2022), i.e. the sum of growth by all additions. The first process presents the separation of specific elements of a response like words and sentences. The second analyses these elements and converts them into numeric values such as frequency and word length. The third feature combines these values to a single overall score. Artificial intelligence can enhance each of these steps by ensuring standardized and consistent evaluation. In contrast, individualized human raters assess responses by understanding the content and context and applying rubrics to assign a suitable score with some subjective judgment.

Automated assessment at large-scale language examinations can perform the teachers' role in grading and is termed as Computer-Assisted Language Learning (CALL) systems. The benefits of these systems are their accuracy in grading, these systems save human resources and improve efficiency (Wang, 2021). This is mostly relevant to the written forms of

examinations. However, oral examinations and speaking assignments are not yet automated or automated partly.

The present study is significant in that it attempts to research the opportunities of speaking assessment automation to enhance the reliability of the evaluation process. Here, this paper focuses on exploring the reliability of human-mediated speaking assessments compared to AI-assisted scoring, with a focus on inter-rater and intra-rater consistency. Additionally, our paper will examine key factors that influence rating reliability, such as rubric calibration, rater severity, grading and familiarity biases. Understanding these elements will contribute to improving the consistency and fairness of speaking assessments. Moreover, this research seeks to evaluate the opportunities and challenges that AI presents in human assessment of speaking skills. While AI has the potential to enhance efficiency and reduce grading time, concerns remain regarding bias, validity, and the ability of AI to provide meaningful feedback for language learners. By analyzing the role of AI in speaking proficiency evaluation, this study will offer insights into how AI can complement human raters and propose best practices for integrating AI into language assessment frameworks.

The current study targeted the following three research questions:

To what extent does human-mediated speaking assessment demonstrate inter-rater and intra-rater reliability compared to AI-assisted scoring in language proficiency evaluations?

What factors influence the inter-rater and intra-rater reliability of speaking proficiency evaluations?

What are the practical considerations for incorporating AI into human-led speaking assessment?

Literature Review

AI-assisted speaking assessment in TEFL

Artificial Intelligence (AI) is revolutionizing various educational environments by enabling personalized and interactive learning experiences. AI's use in language learning, specifically EFL contexts, is increasingly gaining attention. AI tools, such as Automatic Speech Recognition (ASR), allow students to practice speaking and receive feedback even when a native speaker is not present. AI-based systems can mimic human speech recognition, which has been shown to be beneficial for language learners in overcoming challenges related to fluency, pronunciation, and comprehension (Junaidi et al., 2020). Traditionally, EFL education emphasizes grammar, syntax, and written skills. However, studies have shown that this grammar-based approach has not been successful in improving fluency in spoken language. Over time, the focus in foreign language education has shifted towards achieving fluency and effective communication. Research points out that students often struggle with flow, fluency, pronunciation, and vocabulary in spoken English. Technologies like AI aim to bridge this gap by offering tools that replicate native speech environments. In the EAP context, AI tools such as Chivox, iFlytek, and Liulishuo assist university students in practicing speaking tasks necessary for academic success. These tasks often include presentations, group discussions, and answering questions related to subject-specific content (Zou et al., 2020).

The study by Junaidi et al. (2020) on the use of Lyra Virtual Assistant (LVA) demonstrates how AI can help secondary school students improve their speaking skills. Lyra, chosen for its affordability and functionality, allowed students to practice pronunciation and receive immediate feedback. The study compared students using LVA with a control group using traditional methods, showing significant improvement in pronunciation, grammar, flow, fluency, and vocabulary in the experimental group. Another study by Abdulhussein Dakhil (2025) investigated the impact of AI-mediated speaking assessment on the speaking performance and willingness to communicate (WTC) of intermediate Iraqi EFL learners. Forty participants were randomly divided into experimental and control groups, with the

experimental group receiving ten 60-minute sessions using the ELSA Speech Analyzer. Pre- and post-tests assessed speaking performance (grammar, vocabulary, pronunciation, intonation, fluency and flow), and the WTC scale measured communication willingness. Results showed significant improvements in grammar, vocabulary, intonation, and fluency for the experimental group, but no difference in pronunciation. Additionally, AI-mediated assessment enhanced WTC with both native and non-native speakers and in school contexts. Overall, AI-assisted speaking assessment proved effective in improving learners' speaking skills and communication willingness. In the same vein, the research by Zheng (2024) examined the use of an AI-assisted formative assessment platform in an English public speaking course. The platform utilized deep learning, automatic speech recognition, and writing evaluation to provide immediate feedback on speaking anxiety and competence. Fifty-two learners were randomly assigned to two groups: the control group (G1) used self-, peer, and teacher assessment, while the experimental group (G2) used self-, automated, and teacher assessment. Results showed that G1 reported higher social engagement, highlighting the importance of peer interaction in assessment. While G1 students were concerned about peer feedback quality, G2 students desired more detailed automated feedback. No significant differences were found in self-efficacy, engagement, or competence, suggesting that AI-assisted assessment can effectively supplement formative assessment and serve as a reliable learning aid.

Another study shows that students are generally receptive to the AI-powered presentation platform designed to provide students with more chances to practice their presentation skills without requiring faculty involvement. However, there are clear differences in the scoring abilities of AI and human raters. The results highlight limitations in both AI and human evaluation, suggesting that a collaborative approach combining AI and human intelligence could be beneficial (Chen et al., 2022). Furthermore, EAP Talk is the AI-powered platform aimed at improving the speaking abilities of English for Academic Purposes (EAP) learners. EAP Talk's impact on various speaking competencies, including fluency, grammar, vocabulary, pronunciation, and organization of ideas is found effective. EAP Talk is effective in enhancing EAP learners' speaking skills, with significant improvements observed in all evaluated areas. It can provide personalized feedback and the ability to tailor exercises to individual needs, which were highly valued by participants. However, some limitations were also identified, including the accuracy of speech recognition and automated scoring. Therefore, AI-assisted platforms like EAP Talk have the potential to complement traditional learning methods in EAP contexts, offering learners more personalized and adaptive learning opportunities (He et al., 2024).

Speaking assessment criteria

Educators argue that form and content in assessment are interconnected, requiring a balance between linguistic accuracy and structured arguments depending on the task (Moser, 2020). Speaking assessment commonly focuses on fluency, accuracy, pronunciation, grammar, and vocabulary. Fluency is often linked to speed, confidence, and minimal hesitations, that is described as the ability to use language naturally and effectively (Bailey, 2003; Makhlouf, 2021). While some researchers focus on fluency, others highlight the importance of content in effective communication (Harmer, 2015; Makhlouf, 2021). Webb, Newton, and Chang (2012) suggest that familiarity with words and expressions can help develop fluency. At the same time, accuracy is seen as a key indicator of proficiency, allowing speakers to communicate without errors and effective language control (Ellis, 2005; Makhlouf, 2021). Mispronunciation can lead to misunderstanding without practice, yet phonological training is sometimes overlooked in teaching (Vasbieva et al. 2016; Makhlouf, 2021). In standardized assessments like IELTS, accuracy is assessed through grammar, pronunciation,

and vocabulary (IELTS, 2007). Finally, vocabulary is vital in ensuring clear and meaningful communication (Schmitt, 2008; Zarei & Mahmoodzadeh, 2014; Ramezanali, 2017; Makhlouf, 2021). Each of these elements contributes to the overall learner's ability to communicate effectively, making them important for consideration in the language assessment.

These aspects included in predefined scaled assessment criteria enhance the overall objectivity and reliability of rating procedure. According to Dogan and Uluman (2017), this standardization minimizes subjective interpretation by different raters, leading to more consistent and fair evaluations.

Both traditional and AI assessments have limitations that must be addressed to ensure fair, accurate, and effective evaluation of student learning. The limitations of traditional assessment models mentioned by Yesilyurt (2023) are reliance on summative assessments, the difficulty in providing timely and personalized feedback, and the constraints of manual grading. Therefore, AI-driven innovations like automated scoring, speech recognition, multimodal analytics, and adaptive testing can transform language learning assessment. AI-powered assessments hold promise for improving efficiency and personalization, but they also have significant limitations. AI struggles to understand subtle nuances in language, creativity, and critical thinking that humans can easily grasp. Current AI assessment systems often rely on surface features of text (e.g., word count, sentence structure, grammar) rather than deeper understanding of content and argumentation. Many AI assessment systems are "black boxes," making it difficult to understand how they arrive at a particular score. So, it can be concluded that a responsible, human-centric integration of AI is needed to enhance pedagogy and the learner experience (Greene, Hoffman, & Stark, 2019; Selwyn, 2019; Yesilyurt, 2023).

Score Reliability

Generalizability Theory (GT) built on Classical Test Theory and ANOVA, provides a unique conceptual framework for evaluating score reliability (Brennan, 2001, as cited in Wang & Luo, 2019). It views scores as samples from a broader perspective of testing conditions where higher reliability suggests better generalization to other contexts (Cronbach et al., 1972, as cited in Wang & Luo, 2019). The generalizability of scores depends not only on task-specific factors but also on external contexts that influence result interpretation and decision-making (Bachman, 1990, as cited in Wang & Luo, 2019). The level of inter-rater reliability can be assessed using several methods derived from Generalizability Theory. For determining agreement among raters on a specific item for an individual examinee Cohen's kappa coefficient is used. Additionally, statistical measurements such as Fleiss's kappa, Kendall's W and intra-class correlation coefficient (ICC) are commonly employed (Dogan & Uluman, 2017).

Performance task results can be influenced by various factors such as task design, interviewer, rating scales, and raters (Barkaoui, 2010; Eckes, 2005; McNamara, 1996, as cited in Wang & Luo, 2019), with raters playing a particularly significant role in score variability. Due to individual differences, raters may demonstrate inconsistent severity (Myford & Wolfe, 2003, as cited in Wang & Luo, 2019), interact with other facets (Kondo-Brown, 2002; Schaefer, 2008; Upshur & Turner, 1999, as cited in Wang & Luo, 2019), and deviate from standardized scoring practices (Eckes, 2005; Yan, 2014, as cited in Wang & Luo, 2019), potentially compromising fairness in test interpretation and use.

The consistency of marks given by various raters to the same performance or response is known as inter-rater reliability (IRR) (McHugh, 2012). In assessments where human raters are involved, for instance, grading an essay or rating a speech, it is crucial to analyze how much raters agree beyond mere coincidence. Some real-world assessments involve more than two raters, which necessitates an extended approach to IRR. Fleiss' kappa is an extension of Cohen's kappa that enables assessment of agreement among three or more raters (Zapf et al.,

2016). Conceptually, Fleiss' kappa also corrects for chance agreement, but it aggregates the ratings from multiple judges to produce a single coefficient of reliability. This statistic, like Cohen's, ranges from –1 to +1, where higher values indicate stronger reliability. A key advantage of Fleiss' kappa is its ability to handle any fixed number of raters, making it well-suited for panel evaluations or situations where several instructors or judges independently score the same set of performances (McHugh, 2012; Nichols, 2010). Interpreting Fleiss' kappa values involves categorizing the strength of agreement: values less than 0 indicate poor agreement; value 0.01–0.20 indicates slight agreement; value 0.21–0.40 means fair agreement; value 0.41–0.60 corresponds to moderate agreement; values 0.61–0.80 shows substantial agreement; and value 0.81–1.00 means almost perfect agreement (Nichols, 2010).

Rater effects often discussed in literature include severity, halo effect, central tendency effects, etc., and introduce systematic distortions in assessment outcomes as they come from the rater's judgment. These biases can threaten the validity of ratings by introducing extra factors that distort the evaluation process. The most common rater bias is severity effect when assessors consistently give overly harsh or lenient scores compared to other raters (Eckes, 2005). According to Eckes (2009), rater severity can be influenced by various factors like experience, personality, attitudes, demographics, workload and assessment purpose. While senior raters may be stricter to adhere to standards, less-experienced raters tend to be more lenient, but the author suggests that research on the stability and causes of these biases remains limited. The study by Eckes (2005) found that while women generally received higher scores than men in writing and speaking assessments, aligning with prior research, which is not considered to be a systematic gender bias, though some raters showed varying scoring tendencies. Based on the observed study results in rater severity, he prioritizes rater training and individual consistency rather than between raters, regular raters' monitoring for severity, leniency, consistency, and score adjustment to ensure fairness in examinee evaluation (Eckes, 2005).

According to Hardré (2014), grading bias means assigning different grades to student work of similar quality due to irrelevant factors, undermining the fairness of assessment. Even with the best intentions, teachers can unknowingly let bias influence their grading. The author argues that teachers' personal knowledge and perceptions of students can influence grading: they may grade more generously students who are positive and engaged, even if their performance is like others. When unsure about grading, teachers may rely on mental labels and grades based on perceived potential rather than objective performance.

The study by Park (2020) investigated how rater characteristics, especially familiarity with foreign accent influence oral assessments, focusing on the interrater reliability and rater severity among EFL raters. The findings showed that teachers with little familiarity with Korean accent demonstrated the highest consistency while heritage/native Korean speakers and teachers with some familiarity exhibited slightly lower but still high reliability in ratings. According to Bogorevich (2018), research on native and non-native raters in speaking and writing has shown conflicting results due to variations in rater populations, study designs and assessment conditions with no quantitative differences in scoring approaches while qualitative analyses reveal differences in rating approaches for specific speech features.

Methodology

The current study sample consists of the participants learning English as a Foreign language in their first year of studies at B1 proficiency level from the Computer Science, Cybersecurity, Software Engineering, Media Technologies and Smart Technologies Departments of a higher education institution in Astana (Astana IT University). A systematic random sampling method was employed selecting every third student from a total number of 274 students, resulting in 91 participants. Of these, 70 students agreed to participate and signed

informed consent forms. Random sampling is where every individual has an equal chance of being selected from the population. Simple random sampling guarantees that each person has the same probability of being included. In this approach, the researcher compiles a numeric list of the entire sample size and employs a computer program to generate random numbers (Acharya, 2013). The participants answered predetermined speaking cards. The cards included topics related to the use of technology in the healthcare system. Students' oral responses were audio recorded for further assessment. All the recorded responses were initially assessed in three subgroups by raters, who were instructing these students, using the scaled assessment criteria rubric which included the following sections: content and organization, fluency, vocabulary accuracy, grammar accuracy, pronunciation and clarity, and time management.

The next step included cross-checking of subgroup recordings by independent raters to enhance reliability and consistency in assessment. Following human evaluation, the recordings were graded by the AI model which included two interconnected scripts: speech recognition, designed to convert audio files into text, and the second, which performs the analysis and evaluation of the resulting transcription (Figure 1). Out of 70 results 6 were excluded due to some technical errors. The scores given for fluency, pronunciation and clarity, and time management by human raters were also removed from the analysis due to AI model's constraints or limitations (AI model failed to provide scores for these abovementioned criteria). Hence, excluding these criteria from analysis is made strategically rather than intentionally to balance strengths and constraints of the AI model.

Natural language processing and speech recognition technologies allow for automating the process of audio transcription and text analysis which are relevant for educational and research purposes. The interaction of these components allows not only to automate the process of oral speech processing but also to perform a detailed analysis of speech quality considering the predefined assessment criteria. The application of such methods is especially valuable for evaluating students' oral speech, preparing data for machine learning, and automated analysis of audio files in various professional fields.

A speech recognition system was utilized. A script with the "Whisper AI" performs key tasks in processing audio files: converting them into a unified format, splitting long audio fragments into smaller parts, and then transcribing speech into text. This tool is based on the Whisper library developed by OpenAI (Radford et al., 2022), which uses neural network algorithms for highly accurate speech recognition. In addition, the script applies pydub (Pydub, 2023) to process audio files and ffmpeg (FFmpeg, 2023) to convert files to WAV format with the required parameters.

The script execution process starts with checking the audio file passed in the command line arguments. If the file format is other than WAV, it is automatically converted by ffmpeg to a 16 kHz monaural format. This step is necessary to ensure optimal performance of the Whisper model, since it is trained on data with similar characteristics. Once the audio file has been reduced to the desired format, it is analyzed for length. If the audio is longer than 30 seconds, the file is automatically split into parts of the specified length using the pydub library. This splitting is necessary because transcribing long files can be difficult both in terms of computational resources and model accuracy.

Next, each of the received audio parts is passed to the Whisper neural network model, which performs speech-to-text conversion. The received text is aggregated into a single string and written to a file with the extension _transcription.txt. Additionally, the function of deleting temporary files containing intermediate parts of the audio recording is implemented, which allows for optimizing the use of disk space. As a result of the script operation, the user receives a text transcription of the audio file, which allows for use in further analysis, machine learning, or other areas of natural language processing.

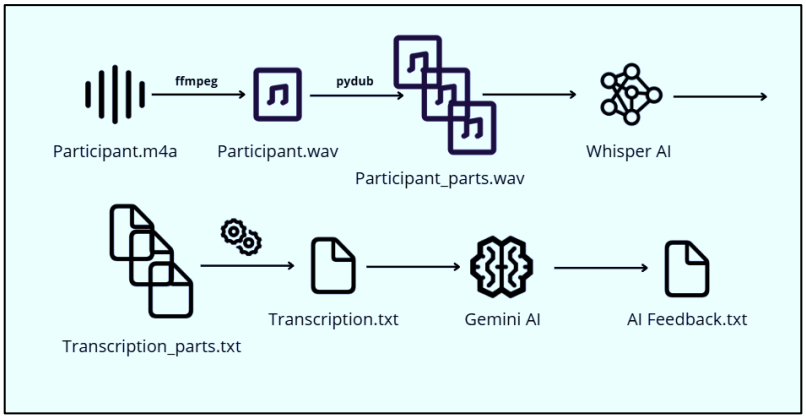
The generative AI script is designed to analyze transcribed text using the capabilities of the Gemini language model provided by Google. The main goal of this script is an automated evaluation of speech quality according to the predefined assessment criteria. For this purpose, the Google Generative AI API (Google AI, 2024) is used, which allows integrating the model into text processing and using it as an expert evaluation tool.

The script starts by loading the transcribed text from the file passed in the command line arguments. Then the evaluation model is applied to the text, working based on a specially prepared prompt. The prompt specifies evaluation criteria that include parameters such as speech organization and content, fluency, adherence to time frames, vocabulary proficiency, grammatical correctness, and pronunciation. To interact with the Gemini model, a chat session is created in which the transcription is sent as input. The model analyzes the text and generates a response containing a numerical score for each of the given criteria.

The assessment results are saved in a file with the _result.txt extension, which allows researchers to analyze the dynamics of students' performance, automatic verification of speech quality in educational and research settings. This tool provides the possibility of automated verification of oral speech, integration with distance learning systems, the assessment of speaking skills, and linguistic analysis.

Figure 1

Model of speaking assessment by AI



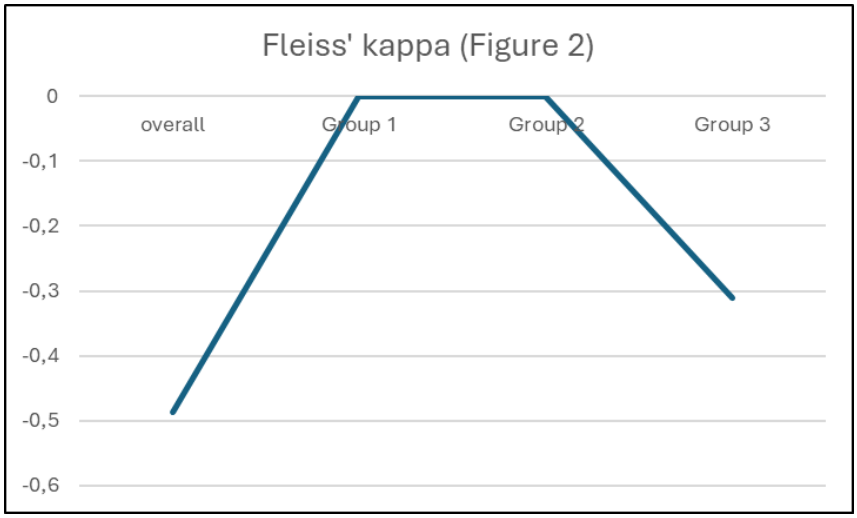
The collected data was processed in the following way: to analyze the reliability of four rater assessments (rater 1, rater 2, rater 3, and AI), Microsoft Excel as a tool for statistical calculations was used. The inter-rater reliability was examined using Fleiss' kappa to measure the level of agreement across four different raters, providing insight into the consistency of their grades. Intra-rater reliability was assessed using descriptive statistics for rater score analysis (mean, standard deviation, and standard error mean) to determine the consistency of a single rater. Additionally, rater severity was calculated to evaluate the extent to which individual raters differed in their scores, identifying the leniency or severity degree.

Findings

To assess inter-rater reliability, Fleiss' kappa was calculated for the entire dataset (all three groups of raters) and separately for each group (group 1, group 2, and group 3). Overall Fleiss' kappa equals to -0.487, indicating poor agreement across all rater assessments. Fleiss' kappa of group 1 (rater 1, rater 3, and AI) is 0.09 which corresponds to slight agreement. Group 2 (rater 2, rater 1, and AI) shows 0.02 which indicates slight agreement. Group 3 (rater 3, rater 2, and AI) has Fleiss' kappa of -0.31 which demonstrates poor agreement among raters (Figure

2). A negative kappa value suggests disagreement beyond chance, meaning that AI and human raters do not follow a consistent scoring pattern.

Figure 2
Fleiss' Kappa Inter-rater Reliability



To evaluate intra-rater reliability, descriptive statistics were computed separately for each group. As shown in Table 1, there are notable inconsistencies in mean scores among different assessors. In Group 1, assessor 1 assigned the highest mean score (39.00), whereas AI provided the lowest (26.05), with assessor 3 falling in between (28.33). In Group 2, assessor 1 and assessor 2 demonstrated a high level of agreement, with mean scores of 38.68 and 38.05, respectively. However, AI's mean score was significantly lower (31.89), suggesting a different evaluation approach. In Group 3, assessor 3 assigned the highest mean score (41.33), while AI once again provided the lowest (24.95), and assessor 2's score (31.58) was closer to AI than to assessor 3.

Table 1 also illustrates the variability of scores through standard deviation (SD) and standard error (SE). A lower SD indicates that the rater gives scores that do not vary much demonstrating high consistency, while a high SD means the raters' scores range widely with lower consistency in assessment. AI exhibits the highest SD in Groups 2 and 3 (10.3 and 10.08, respectively), highlighting greater variability in its scoring patterns compared to rater 1 and rater 2 who have significantly lower SD (4.03 and 6.27 respectively). In Group 1, assessor 1 has the highest SD (9.85), suggesting less consistency in scoring compared to other human assessors. Regarding SE, AI's values range from 1.9 to 2.4 across groups, implying a lower degree of confidence in its mean scores compared to some human assessors. Notably, assessor 1 in Group 2 has the lowest SE (0.9), indicating a high level of scoring precision.

Table 1

Assessment Scores across Three Groups

	Group	Mean Score	Standard Deviation	Standard error
Assessor 1	1	39	9.85	2.15
Assessor 3	1	28.33	8.69	1.9
AI	1	26.05	8.8	1.9
Assessor 2	2	38.05	6.27	1.4
Assessor 1	2	38.68	4.03	0.9
AI	2	31.89	10.3	2.4
Assessor 3	3	41.33	7.03	1.4
Assessor 2	3	31.58	5.24	1.07
AI	3	24.95	10.08	2.06

The analysis of rater severity further confirms these disagreements, as it measures how much each assessor's mean score deviates from the overall mean score (Figure 3). AI consistently displays negative severity values across all groups, indicating that it is systematically stricter in scoring compared to human raters. Conversely, assessor 3 exhibits the highest positive severity, particularly in Group 3, showing a more lenient evaluation approach. Meanwhile, assessors 1 and assessor 2 generally align more closely with the overall mean, showing their relative consistency. Overall, AI scores were systematically lower than human scores across most categories, highlighting the AI's greater severity.

Figure 3

Rater Severity across Three Rater Assessments

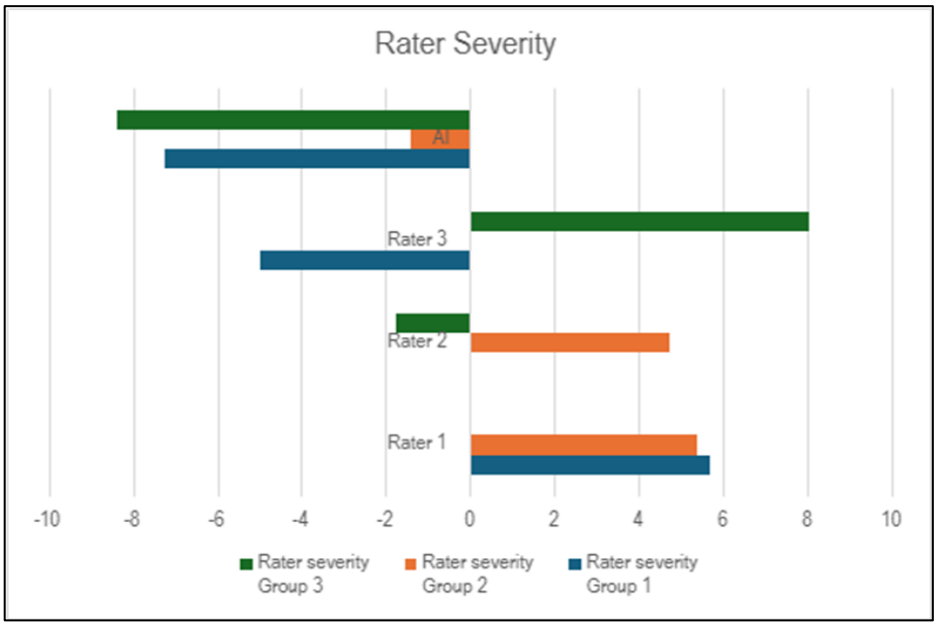


Table 2 demonstrates that AI seems to be tougher than human raters when assessing the content and organization and grammar categories. It shows bigger differences in scoring this category. It means that assessing how ideas are organized and accuracy in grammar can be subjective, and both AI and human raters see it differently in terms of foreign language norms and error detection precision. Vocabulary is the most consistent category among all raters

compared to the content and organization category. As it is agreed well on whether students have sufficient language to express themselves, suggesting that vocabulary is easier to judge objectively.

Table 2

Descriptive Statistics for Human Raters and AI Scores by Assessment Criteria

		Rater 1		Rater 3		AI	
	Criteria	M	SD	M	SD	M	SD
Group 1	Content & Organization	25.9	7.34	16.57	7.42	13.8	7.05
	Vocabulary	6.42	1.5	6.38	1.16	5.57	0.92
	Grammar	6.04	1.24	5.9	1.04	4.66	1.19
	Total	39	9.85	28.33	8.69	26.04	8.8
		Rater 2		Rater 1		AI	
	Criteria	M	SD	M	SD	M	SD
Group 2	Content & Organization	24.73	5.08	26.52	3.65	21.05	7.56
	Vocabulary	6.42	1.07	6.26	0.8	6.31	1.33
	Grammar	6.89	0.65	6.42	0.96	4.57	1.98
	Total	38.05	6.26	38.68	4.02	31.89	10.29
		Rater 3		Rater 2		AI	
	Criteria	M	SD	M	SD	M	SD
Group 3	Content & Organization	27	5.27	19.75	4.52	15	7.51
	Vocabulary	6.95	0.95	5.66	0.76	5.45	1.64
	Grammar	7.08	1.17	6.08	0.71	4.2	1.88
	Total	41.33	7.02	31.58	5.24	24.95	10.08

Discussion

These findings suggest a significant inconsistency in scoring patterns among assessors and AI. Particularly, the negative kappa value in Group 3 implies that raters were inconsistent due to disagreement. Such disagreement may stem from differences in individual assessors' interpretation of the rubric or variability in their scoring criteria. This confirms that inconsistencies remain constant aligning with previous research (Eckes, 2005; Wang & Luo, 2019). AI, in particular, may apply a stricter or fundamentally different evaluation method compared to human raters, leading to persistent discrepancies across all groups. This raises concerns about the alignment of AI-based scoring with human judgment, particularly when used as an objective assessment tool. Moreover, the slight agreement observed in groups 1 and 2 suggests that there is room for improvement in rubric calibration and training to enhance scoring consistency among human raters too.

The variations in mean scores and standard deviations indicate discrepancies in rating patterns, which have implications for intra-rater reliability. It varies notably across assessors and AI, and discrepancies among human assessors, especially in Group 1, highlight potential differences in how they interpret assessment rubric, suggesting the need for further calibration or training to improve scoring consistency. The findings by Limgomolvilas and Sukserm (2025) indicate that while multiple raters can improve reliability, a single well-trained assessor can still provide consistent evaluations in a resource-limited setting. A detailed analytical

rubric, its calibration, and proper rater training can make assessments efficient and reliable, even with just one rater.

Rater severity results indicate that AI systematically assigns lower scores compared to human assessors, particularly in Groups 1 and 3 which suggests that AI may be using a stricter or fundamentally different evaluation method compared to human raters. Variations in severity suggest that strict scoring tendencies of AI may need recalibration to align better with human evaluators, while slight disagreements among human raters indicate the necessity for comprehensive assessment criteria to enhance reliability.

In terms of criterion-based assessment, inter-rater discrepancies are notable in content evaluation, vocabulary assessment seems to show higher inter-rater reliability, and the grammar category shows more alignment with human raters in several groups while AI suggests more disagreement with human raters. These differences suggest that human raters may account for nuances in speech that AI does not, leading to greater variation. The findings propose that slight inconsistencies among human raters can result from familiarity bias or grading bias. Group 1 has moderate disparity where rater 1 (group 1 instructor) appears to be the most lenient and in Group 3, rater 3 (group 3 instructor) appears to have a tendency to give higher scores. This points to a strong leniency or familiarity bias in rater 3 for Group 3 and moderate leniency in rater 1 for group 1. Whereas the trend in Group 2 implies that it has higher agreement in scoring among human raters, suggesting reliable and consistent assessment. This is consistent with study conducted by Hardré (2014) where grading bias can be observed among human raters caused by various factors.

Conclusion

This study emphasizes that differences in assessment are conditioned not only by subjective perception of the assessment process but also by inconsistencies in how assessment criteria are understood and applied both by human raters and the AI model. However, AI ameliorates this process by standardization of variables to be considered. At the same time, we must constantly monitor programs and revise protocols as they need to be. Hence, continuous monitoring of rating quality is essential. It also conforms with Eckes (2005), who recommends regular revision of assessment protocols to raise the rater's consistency in terms of criteria and task design.

Reliability among human raters and AI has not been observed. This might correlate with the fact that the AI model was trained to assess native speakers mainly. Moreover, even though data training was administered (study materials on the content, assessment rubric, and prompts were introduced to the AI model), the results obtained were not complete. Furthermore, the strict requirements can justify the severity of the AI model; it can penalize small mistakes and observe minor inaccuracies in language usage. Human raters may allow nuanced judgment (familiarity bias) and leniency in the language usage being themselves non-native speakers. Addressing inter-rater inconsistencies through standardized training and recalibrating AI models to align more closely with human raters could improve inter-rater reliability in future assessments.

Developing comprehensive analytical rubrics can help reduce differences in rater's evaluations. Regular training sessions and calibration activities should be conducted to minimize inconsistencies among raters and a single rater. Ongoing evaluation of rater's performance can be conducted. Utilizing technology from AI models will ensure efficiency and optimization in large-scale assessments. This aligns with Limgomolvilas and Sukserm (2025) who advocate for detailed rubrics to enhance rating consistency. However, the need to evaluate multiple factors simultaneously within a limited time presents a challenge for raters. To ensure fair grading, educational institutions and instructors should use strategies that identify, reduce, and prevent bias in their assessment practices, including professional

development on recognizing bias, sharing assessment tools, and grading assignments together (Hardré, 2014).

Implementing AI in human-led conversation evaluation requires methodological and technical considerations to ensure objectivity, and consistency with human experts. From a methodological perspective, AI models must be trained to evaluate conversational speech based on the established CEFR frames of reference to maintain consistency with human experts. The AI should complement by acting as a co-assessor depending on the needs of the educational institution. AI-generated assessments and feedback should be clear so that humans can understand and validate the AI's decisions. From a technical perspective, high-quality Automatic Speech Recognition (ASR) models such as Whisper are needed for accurate speech transcription, especially for non-native speakers. AI should also use Natural Language Processing (NLP) to analyze cohesion and lexical diversity, providing detailed feedback on spoken responses. Real-time feedback tools are necessary for assessing pronunciation and grammar while remaining easy for users. By tackling these methodological and technological challenges, AI can substantially increase the effectiveness of human-led spoken language assessments, while preserving the reliability and objectivity of language evaluations.

Conflict of Interest Statement

The authors declare no potential conflicts of interest regarding the research, authorship, or publication of this article.

Author Contributions

Tleshova Zhibek: Editing, Annotation, Administration, Reviewing; Zhanar Tusselbayeva: Literature Review, Data Collection And Preparation, Grading Students' Works, Methodology; Aelita Ichshanova: Literature Review, Data Collection And Preparation, Grading Students' Works, Methodology; Aigerim Urazbekova: Literature Review, Data Collection And Preparation, Grading Students' Works, Methodology; Meruyert Zhenisbayeva: Literature Review, Methodology, Discussion; Ali Orynbayev: Software Development, Tools And Scripts Building And Processing.

References

Acharya, A. S., Prakash, A., Saxena, P., & Nigam, A. (2013). Sampling: Why and how of it. Indian Journal of Medical Specialties, 4(2), 330-333. DOI: 10.7713/ijms.2013.0032

Bogorevich, V. (2018). Native and Non-Native Raters of L2 Speaking Performance: Accent Familiarity and Cognitive Processes. *Northern Arizona University ProQuest Dissertations & Theses, 2018. 10821820.*

Chen, J., Lai, P., Chan, A., Man, V., & Chan, C. H. (2022). AIdakhil-assisted enhancement of student presentation skills: Challenges and opportunities. *Sustainability*, *15*(1), 196.

Dogan, C. D., & Uluman, M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational Sciences: Theory and Practice, 17*(2), 631–651. https://doi.org/10.12738/estp.2017.2.0321

Eckes, Thomas (2005). Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly, 2(3), 197–221.* doi:10.1207/s15434311laq0203_2

Eckes, Thomas. (2009). Many-facet Rasch measurement.

Ercikan & McCaffrey (2022). Optimizing Implementation of Artificial-Intelligence-Based Automated Scoring: An Evidence Centered Design Approach for Designing Assessments for AI-based Scoring. Validity Arguments Meet Artificial Intelligence in Innovative Educational Assessment https://doi.org/10.1111/jedm.12332

Hardré, P. L. (2014). Checked Your Bias Lately? Reasons and Strategies for Rural Teachers to Self-Assess for Grading Bias. *Rural Educator*, *35*(2), n2.

He, H., Zou, B., & Du, Y. (2024, May 13). Bridging the Gap: Linking AI Technology Acceptance to Actual Improvements in EAP Learners' Speaking Skills. https://doi.org/10.31219/osf.io/syb62

International English Language Testing System. (2007). *IELTS handbook 2007*. Retrieved from https://www.ielts-writing.info/EXAM/docs/IELTS_Handbook_2007.pdf

Junaidi, J. (2020). Artificial intelligence in EFL context: rising students' speaking performance with Lyra virtual assistance. *International Journal of Advanced Science and Technology Rehabilitation*, *29*(5), 6735-6741.

Limgomolvilas, S., & Sukserm, P. (2025). Examining rater reliability when using an analytical rubric for oral presentation assessments. *LEARN Journal: Language Education and Acquisition Research Network, 18*(1), 110–134. https://doi.org/10.70730/JQGY9980

Makhlouf, M. K. I. (2021). Effect of artificial intelligence-based application on Saudi preparatory-year students' EFL speaking skills at Albaha University. *International Journal of English Language Education, 9*(2), 1–25. https://doi.org/10.5296/ijele.v9i2.18782

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(*3*), 276–282. DOI: 10.11613/BM.2012.031

Moser, A. (2020). *Written corrective feedback: The role of learner engagement: A practical approach*. Springer Cham. https://doi.org/10.1007/978-3-030-63994-5

Nichols, T. R., Wisner, P. M., Cripe, G., & Gulabchand, L. (2010). Putting the kappa statistic to use. *The Quality Assurance Journal*, 13(*3–4*), 57–61. https://doi.org/10.1002/qaj.481

Park, M. S. (2020). Rater Effects on L2 Oral Assessment: Focusing on Accent Familiarity of L2 Teachers. *Language Assessment Quarterly, 17(3), 231-243.* doi:10.1080/15434303.2020.1731752

Wang, J., & Luo, K. (2019). Evaluating rater judgments on ETIC Advanced writing tasks: An application of generalizability theory and many-facets Rasch model. *Papers in Language Testing and Assessment, 8*(2), 91–116.

Webb, S., Newton, J., & Chang, A. (2012). Incidental learning of collocation. *Language Learning, 62*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x Yesilyurt, Y. E. (2023). AI-Enabled Assessment and Feedback Mechanisms for Language Learning: Transforming Pedagogy and Learner Experience. In G. Kartal (Ed.), *Transforming the Language Teaching Experience in the Age of AI* (pp. 25-43). IGI Global. https://doi.org/10.4018/978-1-6684-9893-4.ch002

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16(*1*). https://doi.org/10.1186/s12874-016-0200-9

Zheng, C., Chen, X., Zhang, H., & Chai, C. S. (2024). Automated versus peer assessment: Effects of learners' English public speaking.

Zou, B., Liviero, S., Hao, M., & Wei, C. (2020). Artificial intelligence technology for EAP speaking skills: Student perceptions of opportunities and challenges. *Technology and the psychology of second language learners and users*, 433-463.

Information about authors

Zhibek Tleshova - Candidate of Pedagogical Sciences, Associate professor, Astana IT University, e-mail: zhibek.tleshova@astanait.edu.kz, ORCID 0000-0001-5095-5436 ***(corresponding author)***

Zhanar Tusselbayeva – Candidate of Pedagogical Sciences, Associate professor, Astana IT University, e-mail: zhanar.tusselbayeva@astanait.edu.kz, ORCID 0000-0002-0832-7898

Aelita Ichshanova – Master of Arts, Senior-lecturer, Astana IT University, e-mail: aelita.ichshanova@astanait.edu.kz, ORCID 0000-0003-4099-855X

Aigerim Urazbekova – MSc in TESOL, Senior-lecturer, Astana IT University, e-mail: aigerim.urazbekova@astanait.edu.kz, ORCID 0000-0002-5641-0303

Meruyert Zhenisbayeva – MA in Foreign Philology Sciences, Senior-lecturer, Astana IT University, e-mail: meruyert.zhenisbayeva@astanait.edu.kz, ORCID 0000-0002-4858-3394

Ali Orymbayev – Master's student in Computer Engineering and Software, Astana International University (AIU), e-mail: phigadamer@proton.me, ORCID 0009-0003-0166-5653