Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327

Arab World English Journal (AWEJ) Volume 17. Number 3, September 2026 Pp.-246-273
DOI: https://dx.doi.org/10.24093/awej/vol17no3.14

AI-Powered Automated Speaking Scoring and Human Evaluation: A Comparative Study of EFL Learners' Oral Proficiency

Sohaib Alam
Department of English Language and Literature, College of Sciences and Humanities
Prince Sattam Bin Abdulaziz University, Alkharj,
Kingdom of Saudi Arabia
Email: s.alam@psau.edu.sa

Received:06/28/2026 Accepted:09/02/2026 Published: 09/10/2026

Abstract
Recent advances in artificial intelligence have led to the development of automated speaking assessment systems capable of evaluating oral proficiency with increasing accuracy. This study compares Artificial Intelligence (AI) powered automated speaking scoring and human evaluation in assessing the oral proficiency of 74 Saudi English as Foreign Language (EFL) learners. Participants completed a picture-description speaking task, which was evaluated by both Claude and two trained human raters using identical holistic and analytic speaking rubrics. The study examined holistic speaking scores and five analytic dimensions: coherence, cohesion, content development, grammar, and vocabulary. Statistical analyses included descriptive statistics, comparative analyses, Intraclass Correlation Coefficients (ICC), Weighted Kappa coefficients, and Bland-Altman analysis. Results showed no significant difference between AI-generated and human-assigned holistic speaking scores. Similarly, cohesion and grammar demonstrated strong similarity between the two assessment approaches. However, significant differences were observed for content development, vocabulary, and coherence, with Claude consistently assigning slightly higher scores than the human raters. Agreement analyses revealed good-to-strong agreement across both holistic and analytic assessments. The findings suggest that AI-powered speaking assessment can produce evaluations broadly comparable to human judgments while demonstrating strong consistency across multiple dimensions of oral proficiency. The study supports the potential of AI-assisted speaking assessment as a reliable complement to human evaluation in EFL contexts.
Keywords: Artificial intelligence, automated speaking assessment, Claude, human evaluation, oral proficiency, self-efficacy, educational technology, SDG-4

How to cite: Alam, S. (2026). AI-Powered Automated Speaking Scoring and Human Evaluation: A Comparative Study of EFL Learners' Oral Proficiency. *Arab World English Journal, 17*(3): 246-273. https://dx.doi.org/10.24093/awej/vol17no3.14

Introduction

The current research aims to determine the degree of concordance between AI- automated, computer-generated scoring of students' oral proficiency and human judgment in different holistic and analytic approaches for assessing students' proficiency in English as a foreign language. Oral proficiency assessment is a multi-dimensional assessment process, which includes evaluating coherence, cohesion, content development, grammar, and vocabulary. Conventionally, oral language assessment has depended on trained human raters. It is believed that the judgments of human raters are the standard by which the oral language of others should be evaluated. However, this is problematic, as human raters tend to be subjective, vary in their task performance, get tired, and require many hours to assess a large number of oral productions (Davis & Papageorgiou, 2021; Metruk, 2018). In contrast to conventional speaking assessment, which relies primarily on human raters and established scoring criteria, recent technological developments have enabled AI, automated speech recognition (ASR), natural language processing (NLP), machine learning, and deep learning to be integrated into automated systems for evaluating oral performance (Zechner et al., 2014). Built on these are systems that advance from feature-based and regression models to neural networks, transformer architectures, multimodal learning systems, and large language models that can return a holistic and/or analytic score (Wang et al., 2021). A moderate to strong correlation was found between the AI-generated score and the human-assigned score in various speaking constructs, such as coherence, cohesion, content development, grammar, and vocabulary, in previous studies (Bamdev et al., 2021). More recently, the use of large language models (LLMs) such as Claude, ChatGPT, Gemini, and other generative AI models has revolutionized language assessment by offering the capacity to analyze fully transcribed, spoken or transcribed language in a more holistic way, provide context-appropriate interpretations, generate evaluative feedback, and score according to rubrics (Kasneci et al., 2023; Nikolovski et al., 2025). Due to this, studies have been conducted on the validity, reliability, and educational usefulness of such LLM-based assessment systems (Mendonça et al., 2025; Yavuz et al., 2025).

Several key questions remain unanswered, however, despite these developments. Prior research has mostly concentrated on the validation of automated scoring systems and the study of correlations between automated scores and human annotations, while a few have systematically compared the scores generated by AI and those assigned by humans in the context of either holistic or analytic assessment models of speaking (Metruk, 2018). Furthermore, it is not yet known to what degree the AI and human assessors provide the same answers regarding coherence, cohesion, content development, grammar, and vocabulary (Ockey & Chukharev-Hudilainen, 2021). Empirical research as an automated speaker evaluator is especially limited, and few studies have looked at both holistic and analytical scores of speeches and assessed them by various types of agreement in the EFL settings of Saudi Arabia. Assessing oral proficiency has become a pressing concern in language education and practice, not only high-stakes language testing but also instruction and evaluation, since the gap between automated evaluation and human evaluation on reliability and validity is of great concern (Zou et al., 2023). While previous studies have found high levels of correlation between automated and human ratings (Wang et al., 2021; Xu et al., 2021)., there is a disparity concerning each of the components of assessment components (AP) (coherence, cohesion, content development, grammar, and vocabulary) (Ogawa, 2022), and that automated systems are still far from processing higher-order discourse components and communicative aspects depending on different contexts (Ockey & Chukharev-Hudilainen, 2021). In addition, the evidence of LLM-based speaking assessment is still incomplete, especially with

respect to the extent to which generative AI models can capture the human evaluative judgments of various speaking parameters.

Against this background, the present study compares AI-powered automated speaking scoring and human evaluation of EFL learners' oral proficiency. Specifically, it compares AI-generated and human-assigned holistic speaking scores, and examines analytic scores across coherence, cohesion, content development, grammar, and vocabulary. Additionally, it investigates the level of agreement between the two assessment approaches. Through these objectives, the study seeks to provide further evidence regarding the validity and applicability of AI-powered speaking assessment in EFL contexts.

Traditionally, oral proficiency assessment in second or foreign language (SFL) situations is based on human evaluation, which refers to the use of trained raters to assess various aspects of oral proficiency, such as coherence, cohesion, content development, grammar, vocabulary, communicative effectiveness, and others, using either a holistic or an analytic rubric. Major language assessment systems still rely upon human raters because they can take into account the wider linguistic, pragmatic, discourse, and sociocultural context of the oral performance and evaluate the higher-order language constructs such as communicative appropriateness, coherence, creativity, and interactional competence (Metruk, 2018; Ockey & Chukharev-Hudilainen, 2021; Zechner et al., 2014). This has led to human judgment becoming the dominant benchmark for evaluating automated speech assessment systems (Davis & Papageorgiou, 2021). Human assessment may involve variation in rating flexibility, as well as differences in rubric interpretation, scoring strategies, evaluator fatigue, and expectations related to accent, cultural background, and topic familiarity, which may affect the reliability and validity of the scoring process (Clapham, 2000; Eckes, 2015; Isaacs & Thomson, 2013; Wang & Sun, 2018; Wind & Peterson, 2018). Researchers have pointed to the importance of rater training, rater calibration, and inter-rater reliability analyses (Brown & Abeywickrama, 2021; Davis & Papageorgiou, 2021) to help minimize these effects. Besides, in large-scale speaking tests, a significant amount of time, money, and trained staff is needed, as human scoring is more labour-intensive and expensive (Davis & Papageorgiou, 2021). While there is genuine promise for even more efficient, scalable, and consistent speaking assessment systems, the ability of such systems to match human judgments on coherence, cohesion, development of content, grammar, vocabulary, and overall communicative effectiveness is an important empirical question (Xu et al., 2021; Zechner & Bejar, 2006).

Literature Review

In the last twenty years, automated speaking assessment has matured significantly to meet the demand for a reliable, efficient, and unbiased method of assessing oral competence. Acoustic analysis (Zechner et al., 2009) speech-processing techniques; support vector machines; regression trees; and feature-based machine-learning models were used in early systems to analyze language use, discourse organization, content relevance, and linguistic accuracy (Zechner & Bejar, 2006). The development of Speech Rater, which combined ASR and NLP to assess the delivery, vocabulary, grammar, and content level within the context of pre-existing speaking rubrics developed by Xi et al., (2008) & Zechner et al., (2014), was a significant development. As a result of these developments, several concrete linguistic and acoustic markers of oral competency can be measured, as automated scoring systems are becoming increasingly sophisticated. These developments led to a more gradual instrumentalization of oral competence through measurable

linguistic and acoustic features, such as lexical sophistication, grammatical complexity, and discourse organization, which make it increasingly possible to create sophisticated automated scoring systems (Zechner et al., 2014). Later innovations added in neural networks, attention-based recurrent models, and other more advanced machine learning architectures, which showed significant correlations with the human ratings in both holistic and analytic scores (Xu et al., 2021). Recent advancements such as deep learning, transformer models, multimodal LLM, and speech-language foundation models have proven useful to the evaluation of coherence, cohesion, content development, grammar, and vocabulary (Fang et al., 2025; Wang et al., 2021). The appearance of LLMs is an important milestone since they are able to analyze speech in a more extensive linguistic, semantic, and discourse context, yielding more fine-grained evaluations of communicative effectiveness than the previous models (Kohnke et al., 2023; Li et al., 2024). In addition to generating scores, generative AI systems like Claude and GPT can offer explanatory feedback that mimics a human assessor's comments, thereby serving the purposes of formative assessment. Newer systems employ ASR, transformer-based language models, multimodal approaches, and adaptive scoring algorithms to produce scores and feedback for diagnostics (Polasa et al., 2025; Scaria et al., 2024; Zhou & Yu, 2025). Recent applications have also demonstrated benefits for speaking and broader English-language learning (Benaissa et al., 2025; Butarbutar et al., 2025; Grab, 2026; Lai, 2026). While some research has found moderate to high agreement between AI-generated and human-assigned scores, especially for coherence and cohesion, the validity of the scores relies on the accuracy of the ASR algorithm, the quality of training data, variations in accents, background noises, and non-native speech patterns (Mohammadkarimi et al., 2025; Wang & Min, 2026; Wang et al., 2021). In addition to persistent issues related to transparency, fairness, and algorithmic bias (Williamson et al., 2012; Xi, 2010), there are also difficulties in assessing interactional competence, creativity, pragmatic appropriateness, cultural interpretation, and context-sensitive communication (Hu & Hashim, 2026; Ockey & Chukharev-Hudilainen, 2021; Obera et al., 2025). The evidence has also raised concerns regarding assessment equity for non-native English speakers (Ndacyayisenga et al., 2026). In this light, newer studies have shown a growing preference for hybrid approaches that utilize both computer-based automated evaluations and human ones, which are more nuanced and context-sensitive. Given this, the possibility of using a combination of computer-based automated systems with human evaluations, referred to as hybrid models, is increasingly gaining traction in research, representing a collaboration between the efficiency and consistency of automated systems and the contextual sensitivity and nuance of human evaluation.

The similarities and differences between automated scoring with AI versus human scoring in the realm of language assessment have been a subject of great interest in research, with most studies comparing holistic and analytic speaking tasks. Initial studies showed positive results that showed that automated scoring methodologies can compare and match with manual scores with respect to coherence, content development, language use, and overall speaking proficiency. For instance, Zechner et al., (2009) showed that the SET-1 automatic speaking test was able to predict conventional human raters' scores with little error, and Zechner & Bejar (2006) reported score accuracies that approached human inter-rater agreement for machine-learning models based on support vector machines and regression trees. The same conclusion was reached by Xi et al., (2008), who showed that the ASR- and NLP-based feature extraction methods can deliver diagnostic information consistent with TOEFL speaking rubrics, and furnished evidence of the quality of the scoring provided by the Speech Rater system through validity analysis. More recent

classroom-based applications similarly report positive speaking outcomes following AI-supported instruction (Benaissa et al., 2025; Shikun et al., 2024).

All of these early investigations put together set a precedent for the feasibility of automatic speaking assessment, as well as gave early evidence that many of the linguistic features regarded by human speakers could be captured in computational models. The results indicated that automated systems could be used, especially in the evaluation of the grammatical competency, vernacular, context-related aspects, and language levels of discourse. Hence, automated speaking assessment became an alternative to the conventional human assessment procedures (Xi et al., 2008; Zechner et al., 2014).

Later research further validated the data on the agreement between AI and humans. Speaker-level correlation with human ratings as high as .73 (Zechner et al., 2014) across a variety of speaking dimensions, and an automated linguistic index predicting 52% - 61% of human speaking ratings (Crossley & McNamara, 2013) across the same dimensions were reported. As confirmed by Xu et al., (2021), who reported the correlations of .87 with human scoring, with scores reaching a high of 89% accuracy. Likewise, Wang & Min (2026) found moderate to high correlation on various speaking constructs ranging from .499-784, Mohammadkarimi et al., (2025) found high correlation between automated and human scoring for the holistic and analytic measurements of oral proficiency (0.67-68), Li et al., (2025) reported substantial agreement between the two methods for holistic and analytic scores (0.66-0.68), Scaria et al., (2024) showed moderate to good correlations between the two on holistic and analytic measures of oral proficiency (0.5-0.9), Fang et al., (2025) found the two methods to be substantially correlated on the holistic and analytic measures of oral proficiency (0.65-0.74), and Polasa et al., (2025) described high and statistically significant correlations, respectively, on holistic and analytic measures of oral proficiency (0.86 and 0.57-0.66).

The overall results across the different studies suggest that automated systems for speaking assessment can provide a solid evaluation of the speech that is equivalent to that of a human assessment. The level of agreement is, however, highly dependent on the constructs measured, the scoring system used, the learner population, and the technology used to run the assessment system. Therefore, the strong correlation between AI-generated scores and human-assigned scores does not necessarily indicate that the two score types are completely equivalent in terms of what they measure (Davis & Papageorgiou, 2021; Williamson et al., 2012).

Although these promising results have been found, there are studies that show that alignment is only moderate in different speaking constructs, scoring systems, assessment tasks, and model architectures. Davis & Papageorgiou (2021) found that no participant received exactly the same score from human and machine raters, and that the two scoring approaches demonstrated complementary strengths; Wang & Sun (2018) concluded that there were systematic differences between automated and human scoring classifications. While paired tests with computer have similar scores for grammar, discourse quality, and overall language performance as those with human partners, Ockey & Chukharev-Hudilainen (2021) have pointed out the lack of authenticity in computerized interaction. In addition, research involving only particular linguistic units has also shown differences in the level of consent. Ogawa (2022) found characteristics related to discourse to be important factors in predicting human ratings, while Hu et al., (2025) found that discourse complexity rather than grammatical accuracy was a more significant correlation with human ratings. Agreements are also found to differ across various assessment dimensions and across assessment conditions, as Bamdev et al., (2021), Li (2024), Matsuura & Suzuki (2023), and the

study on utterance-fluency refinement all showed beneficial results for automated systems in assessing vocabulary, grammar, coherence, and content development.

One of the significant trends that has become apparent in the literature is that learning outcome agreements are more likely to be supported by measurable linguistic constructs using measurable linguistic properties. Words, grammar, coherence, and content tend to have higher rates of agreement with AI due to the ability to accurately measure these components with linguistic and discourse markers. Recent evidence of AI-supported language learning further indicates positive effects across multiple English-language skills and speaking-related outcomes (Butarbutar et al., 2025; Grab, 2026; Lai, 2026). However, dimensions related to discourse organization, communicative effectiveness, interactional competence, and pragmatic appropriateness often do not produce a high degree of agreement as they are dependent on context and require a greater number of interpretations (Hu et al., 2025; Ockey & Chukharev-Hudilainen, 2021). Several studies have also revealed that integrated holistic and analytic assessment methods result in more holistic evaluations than one is able to achieve in using a single assessment method, thus supporting the integration of the automated and the human assessment (Metruk, 2018; Scaria et al., 2024; Wang & Min, 2026).

Nevertheless, important discrepancies between AI and human evaluation continue to be reported in the literature (Kobylarek & Birova, 2026). AI-based assessment systems have been shown to produce scores that differ significantly from those of human assessors, especially with regard to grammatical accuracy, as well as syntactic performance. For example, Obera et al., (2025) reported that AI-scored items were inconsistently marked with scores compared to teacher-scored items in grammar and syntax-related tasks; however, the scores were more consistent for the higher-order communicative traits and creative language uses. In this manner, Karatey & Xu (2025), Li et al., (2024), and Ockey and Chukharev-Hudilainen (2021), pointed out the problems that automated assessment systems have in evaluating a variety of aspects of speaking, such as interactional competence, discourse structure, semantic connections, and coherence. Moreover, a number of research studies have questioned the validity of the instruments designed for assessing students' oral language proficiency by using AI tools, as they only focus on linguistic aspects, and neglect the communicative effectiveness, pragmatic appropriateness, and cultural interpretation, which are all aspects that are usually the focus of human English for academic purposes (EAP) assessors in educational settings (Hu & Hashim, 2026; Obera et al., 2025). Researchers have also identified challenges associated with accent variation, spontaneous speech, ASR transcription errors, fairness, and contextual interpretation, all of which may affect the validity and reliability of automated scoring (Mohammadkarimi et al., 2025; Wang et al., 2021). Concerns about fairness have also been highlighted in recent evidence concerning the use of generative AI in assessment for non-native English-speaking learners (Ndacyayisenga et al., 2026). Another constraint in the previous studies is that most studies have concentrated on just correlation coefficients as a measure of scoring quality. While correlations can be helpful in terms of knowing the relationship between scores, they do not necessarily indicate a score equivalence or agreement. Two assessment systems can yield scores that are highly correlated, but systematically differ in rating learners. Therefore, several studies have indicated that agreement-based measures such as ICC, Kappa statistics, and Bland Altman analyses are recommended to gain a deeper understanding of AI-human consistency of scoring (Giavarina, 2015; Koo & Li, 2016).

Therefore, there is overall agreement in the literature on the validity and usefulness of AI-based automated speaking assessment, but the literature also consistently shows that the results of

AI and human assessment generally agree on the lower-order abilities, whereas significant differences can arise for the higher-order speaking competencies. Moreover, only a limited number of studies have compared holistic and analytic scores in speaking and investigated the extent to which there is agreement between holistic and other scores across individual dimensions (coherence, cohesion, content development, grammar, and vocabulary) with several agreement indexes. In addition, the recent rise of generative AI systems has also prompted a demand for more empirical research exploring the efficacy of these newer models compared to trained human assessors in real-world assessment settings for EFL learners. This discovery has inspired numerous scholars to explore the idea of blending automated and human assessment methods, integrating the benefits of automated scoring, such as its consistency, scalability, and efficiency, with human raters' contextual understanding and interpretive capabilities (Alam et al., 2025; Davis & Papageorgiou, 2021; Polasa et al., 2025; Wang & Min, 2026). The increasing role of GenAI in EFL communication also extends this discussion beyond assessment to learners' willingness to communicate in AI-mediated contexts (Zhong et al., 2026). Therefore, the present study aims to bridge these gaps by comparing holistic and analytical scores produced by AI with those produced by human teachers on both types of scoring methods, and by analyzing the degree of agreement between scores in the different dimensions of oral language (coherence, cohesion, content development, grammatical, and vocabulary components). The study uses several statistical methods for comparing the two types of scoring methods, including holistic and analytical, as well as the comparison of the agreement between them on the dimensions of oral language: coherence, cohesion, content development, grammar, and vocabulary.

Analytic Dimensions of Speaking Assessment

Analytic speaking assessment does not rely on a summative performance-based grade but breaks down speaking skills into various components to gain deeper insight into students' strengths and weaknesses. Human evaluators as well as automated speaking assessment systems typically evaluate dimensions that include coherence, cohesion, content development, grammar, and vocabulary, all of which are part of oral proficiency (Zechner et al., 2014). The method provides diagnostic feedback on specific parts of a speaker's utterances and makes more accurate comparisons between human and AI evaluations of speaking performance (Brown & Abeywickrama, 2021). In this study, five analytic dimensions were analyzed. A key aspect for evaluating the coherence of spoken discourse is its logical composition and evolution, evaluated through discourse structure, topic development, and sequencing, of which automated systems increasingly monitor the discourse-level features, applying NLP and language-model technologies to analyze these aspects (Li et al., 2024; Ockey & Chukharev-Hudilainen, 2021). Cohesion is related to the use of cohesive devices, which can be recognized with the help of discourse analysis techniques (Crossley & McNamara, 201). These devices are capable of making ideas coherent or maintaining that coherence within the text. In content development, speakers demonstrate the level of elaboration of ideas, supply appropriate details, and demonstrate sufficient content to meet task requirements (Brown & Abeywickrama, 2021). It is generally considered that grammar represents the correct use of the syntactic and morphological structures of the language and is a reliable marker of the linguistic accuracy of the language level as well as linguistic complexity (Bamdev et al., 2021). Vocabulary is defined as the quantity, suitability, and utility of words in use within communication, and automated systems using NLP techniques to examine lexical features and vocabulary diversity (Crossley & McNamara, 2013; Wang et al., 2021). While automated systems

can be useful for many of these factors, there may be challenges in assessing contextual appropriateness and communicative effectiveness that led to discrepancies between AI predictions and human evaluations (Crossley & McNamara, 2013).

Theoretical Framework

The present study is grounded in Construct Validity Theory, Multi-Faceted Rasch Measurement Theory, and sociocultural perspectives on assessment technology. Construct Validity Theory posits that assessment scores should accurately represent the construct being measured and that sufficient evidence should support score interpretation and use (Kane, 2013; Messick, 1989). As coherence and cohesion, content development, grammar, and vocabulary are considered to be multidimensional aspects of oral competence, comparing the scores of AI-generated and human-assigned oral scores can provide evidence to see if the two approaches measure the same construct of oral competence. Another aspect of Multi-Faceted Rasch Measurement Theory is that there are other conditions that affect the outcome of an assessment, including the person doing the assessment (the rater), how the assessment is scored, and the assessment conditions themselves (Linacre, 1989). While the present study does not use Rasch analysis, the theory offers a framework of how various aspects of the scoring process, such as scoring severity and consistency, and scoring rubric interpretation, may differ between AI systems and human evaluators (Eckes, 2015; Linacre, 1989). In addition, sociocultural perspectives have been taken into account, which emphasize that language assessment is not just about accuracy in language, but also about meaning in the context, communicative appropriateness, interactional effectiveness, and sociocultural factors (Xi, 2010). These are especially important for automated assessment, where a major concern is the linguistic and computational aspects, while people take the broader communicative and contextual aspects into account. They also focus on the fairness, bias, context-dependent interpretation, and effects of training data in technology-mediated evaluation (TME) (Williamson et al., 2012). These theoretical explanations can then be used to understand the score differences, construct representation, and the level of agreement between the AI-powered automated speaking score and human scores in terms of holistic and analytic components of EFL oral proficiency.

The Current Study

The current study compares AI-powered automated speaking scoring and human evaluation of EFL learners' oral proficiency. Specifically, the first objective of the study is to compare the holistic speaking scores assigned by an AI-powered automated speaking scoring system and trained human evaluators. The second objective is to compare the analytic speaking scores assigned by the two evaluation approaches across the five major dimensions of oral proficiency, namely coherence, cohesion, content development, grammar, and vocabulary. The third objective is to examine the level of agreement between AI-generated and human-assigned scores in both holistic and analytic assessments of EFL learners' oral proficiency. By addressing these objectives, the study seeks to determine whether automated speaking assessment can produce scoring outcomes comparable to those of trained human raters and to identify the dimensions in which convergence or divergence between the two assessment methods may occur. Accordingly, the study addressed the following research questions and hypotheses:

RQ1: To what extent do AI-generated and human-assigned holistic speaking scores differ in evaluating EFL learners' oral proficiency?

RQ2: To what extent do AI-generated and human-assigned analytic speaking scores differ across coherence, cohesion, content development, grammar, and vocabulary?
RQ3: What is the level of agreement between AI-powered automated scoring and human evaluation in holistic and analytic assessments of EFL learners' oral proficiency?
Based on the objectives and research questions of the study, the following hypotheses were formulated:
Ho1: No significant difference exists between AI-generated and human-assigned holistic speaking scores in evaluating EFL learners' oral proficiency.
Ho2: No significant differences exist between AI-generated and human-assigned analytic speaking scores across coherence, cohesion, content development, grammar, and vocabulary.
Ho3: AI-generated and human-assigned speaking scores demonstrate a high level of agreement in both holistic and analytic assessments of EFL learners' oral proficiency.

Methodology

Participants

The subjects were 74 male undergraduates EFL learners from the freshmen level of an English language program at a public university in Saudi Arabia. The participants were all native speakers of Arabic, aged between 19 and 23 years, and at the B1 (bachelor) level in the Common European Framework for Reference for Languages (CEFR). Convenience sampling was used to recruit participants in the 2025-2026 year, which included freshmen English classes. The study was carried out at a male campus, thus restricting the study to male learners only. Given the similar sample size required for comparative and agreement analyses, and to be comparable to the previous studies on automated and human speaking assessment (Xu et al., 2021; Zechner et al., 2014), this sample size was determined to be appropriate. In order to reduce potential sources of extraneous variability and to ensure a valid comparison between AI-generated and human-assigned speaking scores, learners with comparable linguistic and proficiency backgrounds were selected. The B1 level learners were chosen as they were able to carry out lengthy speaking tasks without losing coherence, cohesion, content, grammar, or vocabulary. Voluntary participation, ethical approval, and informed consent were gained before data were collected, and recordings and assessment data were anonymous and kept for research only.

Human Raters

Two independent male instructors in English Language Teaching (ELT) with more than five years of experience in conducting English language assessment were evaluated for their speaking performances. Neither evaluator was involved in the teaching of the participants, so an independent assessment would be conducted. Before scoring, the raters attended a calibration session to discuss the holistic and analytic rubrics, scoring guidelines, and assessment procedures, followed by pilot recordings that were not part of the final data set. In the event of any differences in scores, the rubric descriptors and scoring criteria were discussed to create a common understanding. All recordings were then independently reviewed with the study's holistic and analytic rubrics. The ICC was used to assess the inter-rater reliability, which was satisfactory between the two raters (Koo & Li, 2016). Based on known assessment practices, human-evaluation scores for each assessment dimension were computed as the average of the two rater scores (average rater effects) and stability of scoring (Brown & Abeywickrama, 2021; Clapham, 2000).

Speaking Tasks

To assess oral performance, a picture-description task was employed, and it was scored both by the AI system and by the human raters. These tasks are commonly used in speaking evaluations due to their ability to encourage spontaneous production of language and provide consistency in task items across evaluators (Fulcher, 2015). Participants were asked to look at a known-coloured picture for one minute, which was chosen to elicit longer oral production in each participant, and to describe the picture for two-three minutes. A person was shown in the picture, a thing, an action, a setting, and an event were described by participants without any prompts or corrective feedback given. The task was designed to generate connected speech and ask the learners to plan and coordinate their ideas as well as use a variety of linguistic resources, making it possible to assess the coherence, cohesion, content development, grammar, vocabulary, and communicative effectiveness (Brown & Abeywickrama, 2021). The AI-generated and human-made samples of the evaluation were compared, correlated, and agreed upon using the resulting speech samples. All of the participants had to perform the same task under the same administration, so that the automated scoring using AI for speaking is reliable when compared to human evaluation (Fulcher, 2015).

Assessment Instruments

Holistic Speaking Rubric

Overall oral proficiency (Brown & Abeywickrama, 2021) was assessed using a holistic speaking rubric. The rubric used a 10-point scale that ran from very limited (1) to excellent (10) and offered only one score: Communicative Effectiveness. Assessment was based on: clarity of the message, maintainability of clarity, organization of thoughts, linguistic appropriateness, and comprehensibility. The criteria used were the same for both the AI-driven algorithm and the human raters, promoting consistency and comparability. The rubric was piloted for content validity by three specialists in English Language Teaching and language assessment, and the final version was consistently used in this study.

Analytic Speaking Rubric

Five aspects of oral proficiency - coherence, cohesion, content development, grammar, and vocabulary were assessed by means of an analytic speaking rubric. Analytic assessment helps to give detailed diagnostic information about specific aspects of speaking performance (Fulcher, 2015). All dimensions were rated on a scale from 1 (very weak performance) to 5 (excellent performance). The following areas were assessed: coherence (logical organization and progression of ideas), cohesion (use of the linguistic devices to connect ideas and create continuity within discourse), content development (adequacy, relevance, and elaboration of ideas in relation to the speaking task), grammar (accuracy, syntactic control), and vocabulary (lexical range, appropriateness, and effectiveness of expression). The analytic criteria used by both the AI system and human raters were the same. Subject-matter experts reviewed the rubric descriptors, and during calibration, they had them pilot tested for clarity and applicability for the assessment procedures, leading to systematic comparisons of AI-assigned scores and human-assigned scores across the five dimensions to support content validity and consistency of the assessments.

AI-Powered Automated Speaking Scoring System

Claude (Anthropic AI) was employed as the AI-powered automated speaking scoring system because of its advanced natural language processing capabilities and ability to perform

rubric-based language assessment (Kohnke et al., 2023). Participants completed the picture-description task by interacting directly with Claude, and their answers were transcribed by the automatic speech-recognition (ASR) technology embedded in the assessment platform. The transcripts were then judged using a prompt that was consistent among the human raters and was created using the same holistic and analytic rubric. Overall, the AI assessment prompt provided by Claude has resulted in an overall score for the speaking and the following sub scores for coherence, cohesion, content development, grammar, and vocabulary. Claude has generated an overall score and sub scores for coherence, cohesion, content development, grammar and vocabulary. The complete AI assessment prompt used for the automated scoring is included in Appendix A. All participants got exactly the same instructions, criteria for scoring, and the same procedures for assessment; the same prompt was applied for the entire study to ensure consistency and comparability in scoring across all participants' responses. Statistical analysis of all AI-generated scores was automatically recorded. The standardised procedure provided uniform assessment conditions and reduced the human scoring variability (Kane, 2013; Kohnke et al., 2023).

Recording Instruments, Procedures and Tasks

The data were collected by individually describing pictures under controlled classroom conditions. The participants were given standard instructions and were asked to perform the task in direct interaction with Claude, with the help of a digital voice recorder and an external microphone which was connected to the recorder to capture each of the performances. This procedure made possible to get the same speech samples for the evaluation of the AI and human. All recordings were checked for quality, anonymously coded and secured in a database for analysis. The human raters were given only the anonymized recordings and not the identities of the participants or the scores generated by the AI, which reduced the risk of any potential bias and allowed for an independent evaluation. To enhance the credibility of comparisons over AI-powered and human scoring, the same speech samples were used for both scoring types (Brown & Abeywickrama, 2021; Fulcher, 2015).

Scoring Procedures

The scores were obtained by both AI and human judging of the same speeches. After completing the picture description task, Claude was scored on coherence, cohesion, content development, grammar, and vocabulary, following a standardized assessment prompt, to generate a holistic score as well as individual scores on each of the five components. Every student had a score recorded for analysis, that was produced by AI. All AI-generated scores were generated with the same standardized prompt and assessment process for all participants, to ensure the scores were produced in a stable manner. At the same time speaking was recorded, anonymized and identified with a code number that was sent to two human judges who were blind to the identities of the speakers as well as the scores generated by the AI (Brown & Abeywickrama, 2021; Clapham, 2000). After a rubric calibration session with pilot recordings to assure consistency of scoring, the raters independently scanned each recording, applying the same holistic and analytic rubrics. The ICC method was used to test interrater reliability (Koo & Li, 2016), and the mean of the two ratings was calculated for each participant and assessment dimension (Clapham, 2000). Finally, the two sets of human scores (AI-generated and averaged) were combined and analyzed for comparisons and agreements. A valid comparison of AI-powered and human evaluation of oral proficiency was

carried out with the use of the same speech sample, the same scoring criteria, and standardized procedures.

Procedures for Analysis

Descriptive, comparative, and agreement-based statistical procedures were used to analyze the data. The descriptive statistics of means, standard deviations, minimum values, and maximum values were obtained for the holistic and analytic scores obtained from both Claude and the human raters. The normality of the scores was determined by the Shapiro Wilk test. The results were used in the independent-samples t-tests to compare holistic, coherence, cohesion, content development, grammar, and vocabulary scores. A repeated-measures ANOVA was used to analyze the patterns of scores on the five analytic dimensions. The items were assessed for agreement with the human raters by Intraclass Correlation Coefficients (ICC), Weighted Kappa analysis, and Bland-Altman analysis. Effect sizes were also computed to quantify the differences found and to help interpret the results.

Results

Preliminary Analysis and Normality Testing

Prior to inferential analyses, score distributions were examined using the Shapiro-Wilk test. As shown in Table 1, holistic (W = .982, p = .214), coherence (W = .976, p = .119), cohesion (W = .981, p = .193), content development (W = .985, p = .307), and grammar scores (W = .971, p = .058) were normally distributed, as their p-values exceeded .05. In contrast, vocabulary scores showed a significant departure from normality (W = .968, p = .042). Accordingly, independent-samples t-tests were used for holistic, coherence, cohesion, content development, and grammar scores, whereas the Mann Whitney U test was applied to vocabulary scores. These results confirmed the suitability of the selected statistical procedures for subsequent analyses.

Table 1. *Shapiro–Wilk normality test results*

Variable	W	p	Decision
Holistic	.982	.214	Normal
Cohesion	.976	.119	Normal
Content Development	.981	.193	Normal
Vocabulary	.968	.042	Non-normal
Grammar	.985	.307	Normal
Coherence	.971	.058	Normal

Analysis of Holistic Speaking Scores

To address the first research question, the holistic speaking scores generated by Claude and the human raters were compared. As shown in Table 2, the human raters assigned a mean score of 7.12 (SD = 1.08), whereas Claude assigned a slightly higher mean score of 7.34 (SD = 1.12). The observed mean difference was 0.22 points.

An independent-samples t-test indicated that this difference was not statistically significant, t(146) = 1.21, p = .228. The effect size was small (d = 0.20), suggesting a negligible practical difference between the two assessment approaches. Table 2 further illustrates the close similarity between the mean scores assigned by Claude and the human raters, indicating that both approaches produced comparable evaluations of overall oral proficiency.

Table 2. *Descriptive statistics and comparison of holistic speaking scores*

Assessment Method	Mean	SD	Min	Max	t	p	Cohen's *d*
Human Raters	7.12	1.08	4.50	9.20			
Claude (AI)	7.34	1.12	4.70	9.50	1.21	.228	0.20

The results indicate that Claude and the human raters evaluated overall speaking performance in a similar manner. Therefore, the results support *Ho1*, which proposed that no significant difference would exist between AI-generated and human-assigned holistic speaking scores.

Analysis of Analytic Speaking Scores

The second research question examined differences between AI-generated and human-assigned scores across the five analytic dimensions of speaking assessment: coherence, cohesion, content development, grammar, and vocabulary. As shown in Table 3, Claude assigned slightly higher mean scores than the human raters across all dimensions. The smallest differences were observed for cohesion (0.08) and grammar (0.12), whereas larger differences were found for content development (0.26), vocabulary (0.33), and coherence (0.34). Human-rater mean scores were 3.63, 3.86, 3.72, 3.69, and 3.58 for coherence, cohesion, content development, grammar, and vocabulary, respectively, compared with 3.97, 3.94, 3.98, 3.81, and 3.91 assigned by Claude.

A repeated-measures ANOVA revealed a significant effect of assessment dimension, $F(4, 292) = 4.83$, $p = .002$, $\eta^2p = .12$, indicating that the degree of similarity between AI-generated and human-assigned scores varied across speaking constructs. As illustrated in Table 3, cohesion and grammar demonstrated the closest correspondence between the two assessment approaches, whereas vocabulary and coherence showed the largest discrepancies.

Table 3. *Comparison of analytic speaking scores*

Dimension	Human Mean	Human SD	Claude Mean	Claude SD	Mean Difference
Cohesion	3.86	0.64	3.94	0.61	0.08
Content Development	3.72	0.67	3.98	0.59	0.26
Vocabulary	3.58	0.71	3.91	0.65	0.33
Grammar	3.69	0.68	3.81	0.66	0.12
Coherence	3.63	0.73	3.97	0.70	0.34

The results suggest that although Claude and the human raters produced broadly similar analytic evaluations, the extent of agreement varied across individual speaking dimensions. The strongest similarity was observed for cohesion and grammar, whereas vocabulary and coherence demonstrated comparatively greater divergence. The following sections provide detailed analyses of each analytic dimension separately.

Analysis of Cohesion Scores

The overall comparison of analytic speaking scores was followed by a breakdown of these scores for each dimension. Analytic dimension was the first dimension analyzed. As revealed by the human raters, the mean cohesion score for this segment of the questionnaire was 3.86 (SD = 0.64) and for the Claude segment, it was 3.94 (SD = 0.61). The mean difference between the scores of human raters and Claude was only 0.08 points, with the scores ranging from 2.10 to 4.90 for the human raters, and 2.20 to 5.00 for Claude. However, this difference was not statistically significant $t(146) = 0.79$, $p = .431$; $d = 0.13$). Representative similarity between the two assessment methods

is also demonstrated in Table 4 where the human raters gave a mean of 3.86, while Claude gave a mean of 3.94. The results show a high similarity between the AI-generated and the human-assigned cohesion scores, signifying consistency between the two evaluation approaches in the evaluation of whether linguistic devices are used effectively to link ideas and make them continuous in spoken discourse.

Table 4. *Descriptive statistics and comparison of fluency scores*

Assessment Method	Mean	SD	Min	Max	t	p	Cohen's *d*
Human Raters	3.86	0.64	2.10	4.90			
Claude (AI)	3.94	0.61	2.20	5.00	0.79	.431	0.13

Overall, the results indicate substantial similarity between AI-generated and human-assigned cohesion scores and suggest that cohesion was one of the dimensions in which automated speaking assessment most closely approximated human evaluation.

Analysis of Content Development Scores

Content development was the second analytic dimension examined. As shown in Table 5, the human raters assigned a mean content development score of 3.72 (SD = 0.67), whereas Claude assigned a higher mean score of 3.98 (SD = 0.59). Scores ranged from 2.00 to 4.80 for the human raters and from 2.30 to 5.00 for Claude, resulting in a mean difference of 0.26 points. An independent-samples t-test revealed a statistically significant difference between the two assessment approaches, $t(146) = 2.18$, $p = .031$. The effect size was small to moderate ($d = 0.42$), indicating a meaningful difference in content development evaluation. Table 5 also shows that Claude consistently assigned higher content development scores than the human raters, supporting the statistical findings reported there.

Table 5. *Descriptive statistics and comparison of pronunciation scores (N = 74)*

Assessment Method	Mean	SD	Min	Max	t	p	Cohen's *d*
Human Raters	3.72	0.67	2.00	4.80			
Claude (AI)	3.98	0.59	2.30	5.00	2.18	.031	0.42

These results indicate that Claude evaluated content development slightly more favorably than the human raters. Nevertheless, the moderate difference suggests that both assessment approaches yielded broadly comparable evaluations of content development performance.

Analysis of Vocabulary Scores

The third analytic dimension analyzed was vocabulary. The vocabulary score data did not meet the assumption of normality and the Mann-Whitney U test was used. Table 6 shows that the human raters gave a mean score of 3.58 (SD = 0.71) to the vocabulary section of the test, while Claude gave a higher mean = 3.91 (SD = 0.65). The difference in scores for the human raters and for Claude was 0.33 mean points, with scores ranging from 1.90 to 4.90 for the human raters and from 2.20 to 5.00 for Claude. The Mann-Whitney U test showed a small-to-moderate effect size (r = .24), and there was a statistically significant difference between the 2 assessment approaches (U = 2206.50, p = .009). As shown in Table 6, in terms of vocabulary, Claude consistently received higher ratings than the human raters, suggesting that there was a greater difference between the two.

Table 6. *Descriptive statistics and comparison of vocabulary scores (N = 74)*

Assessment Method	Mean	SD	Min	Max	U	p	r
Human Raters	3.58	0.71	1.90	4.90			
Claude (AI)	3.91	0.65	2.20	5.00	2206.50	.009	0.24

Overall, the results indicate that Claude consistently assigned higher vocabulary scores than the human raters, suggesting differences in how lexical range, word choice, and vocabulary sophistication are evaluated.

Analysis of Grammar Scores

The fourth analytic dimension studied was grammar. In Table 7, the human raters gave a mean grammar score of 3.69 (SD = 0.68) while Claude gave a higher mean grammar score of 3.81 (SD = 0.66). The mean difference between the scores of the human raters and Claude was 0.12 points, with scores ranging from 2.10 to 4.90 for the human raters and from 2.20 to 5.00 for Claude. This difference was not statistically significant, t(146) = 1.02, p = .309, and small effect size (d = 0.18) by an independent-samples t-test. As shown in Table 7, the mean grammar scores given by Claude were quite comparable to those given by the human raters, indicating that both assess grammatical performance in a similar way.

Table 7. *Descriptive statistics and comparison of grammar scores (N = 74)*

Assessment Method	Mean	SD	Min	Max	t	p	Cohen's *d*
Human Raters	3.69	0.68	2.10	4.90			
Claude (AI)	3.81	0.66	2.20	5.00	1.02	.309	0.18

The results indicate substantial agreement between Claude and the human raters in evaluating grammatical accuracy and language structure. The absence of a statistically significant difference suggests that automated speaking assessment can assess grammar with a level of consistency comparable to that of trained human evaluators.

Analysis of Coherence Scores

The last dimension considered was coherence. As per Table 8, the coherence score given by human raters was an average of 3.63 (SD = 0.73), while the average coherence score given by Claude was 3.97 (SD = 0.70). The maximum score obtained by the human raters was 1.80, and the maximum score obtained by Claude was 5.00, with a difference of 4.90 points between the two at the extreme, and a mean difference of 0.34 points for the five analytic dimensions. There was a significant difference between the two assessment approaches t(146) = 2.73, p = .007, d = 0.47; a moderate effect size). The number of coherence scores assigned by Claude was also higher than the human raters' scores, as shown in Table 8, making the statistical result that coherence showed the largest differences in scores between AI-generated and human-assigned scores.

Table 8. *Descriptive statistics and comparison of coherence scores* (N = 74)

Assessment Method	Mean	SD	Min	Max	t	p	Cohen's *d*
Human Raters	3.63	0.73	1.80	4.90			
Claude (AI)	3.97	0.70	2.10	5.00	2.73	.007	0.47

The results suggest that discourse organization, logical progression of ideas, and overall communicative coherence represented the dimensions in which the two assessment approaches diverged most noticeably. Nevertheless, the observed difference remained moderate in magnitude, indicating that both assessment approaches still produced broadly comparable evaluations.

Agreement Between AI and Human Evaluation

To address the third research question, agreement analyses were conducted using ICC, Weighted Kappa statistics, and Bland-Altman analysis. Table 9 presents the ICC values obtained for holistic and analytic scores. As shown in Table 9, the ICC coefficients ranged from .73 to .86, indicating good-to-strong agreement across all assessment dimensions. The highest agreement was

observed for cohesion (ICC = .86), followed by holistic assessment (ICC = .84) and grammar (ICC = .82). Content development (ICC = .78), vocabulary (ICC = .74), and coherence (ICC = .73) also demonstrated acceptable levels of agreement.

Table 9. *Intraclass correlation coefficients between AI and human scores*

Dimension	ICC	Interpretation
Holistic	.84	Strong Agreement
Cohesion	.86	Strong Agreement
Content Development	.78	Good Agreement
Vocabulary	.74	Good Agreement
Grammar	.82	Strong Agreement
Coherence	.73	Good Agreement

Table 10 presents the Weighted Kappa coefficients for the six assessment categories. The Kappa values ranged from .61 to .78, indicating substantial agreement between Claude and the human raters. The highest agreement was observed for cohesion (κ = .78) and holistic assessment (κ = .76), whereas coherence (κ = .61) and vocabulary (κ = .63) produced comparatively lower values. Content development (κ = .69) and grammar (κ = .72) demonstrated moderate-to-high levels of agreement. Nevertheless, all coefficients remained within the range generally interpreted as substantial agreement.

Table 10. *Weighted kappa statistics between AI and human scores*

Dimension	Weighted Kappa	Interpretation
Holistic	.76	Substantial Agreement
cohesion	.78	Substantial Agreement
Content Development	.69	Substantial Agreement
Vocabulary	.63	Substantial Agreement
Grammar	.72	Substantial Agreement
Coherence	.61	Substantial Agreement

Figure 1 presents the agreement profile across the analytic dimensions based on ICC values. As shown in Figure 8, fluency (ICC = .86) and grammar (ICC = .82) demonstrated the highest levels of agreement, whereas vocabulary (ICC = .74) and coherence (ICC = .73) produced comparatively lower agreement values. However, all dimensions exceeded the threshold generally considered acceptable for agreement studies.

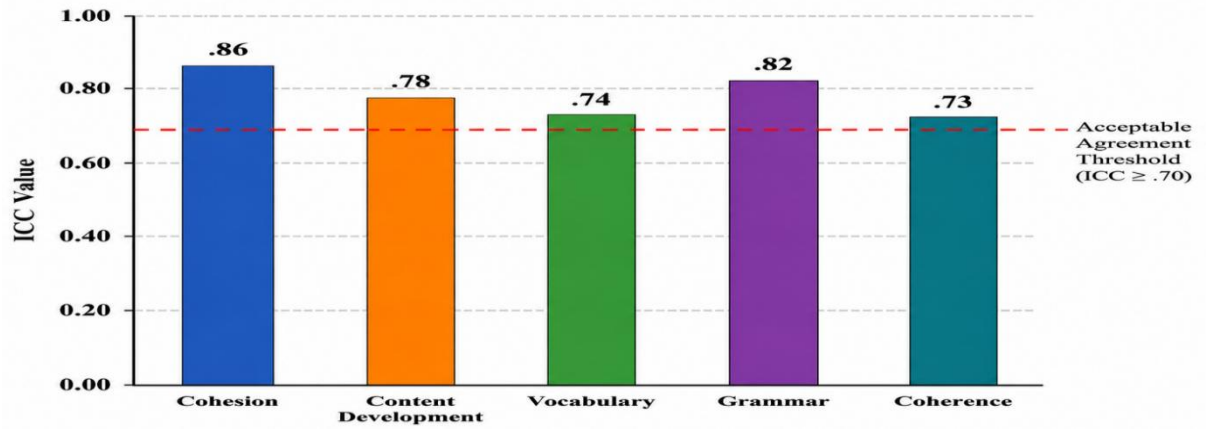
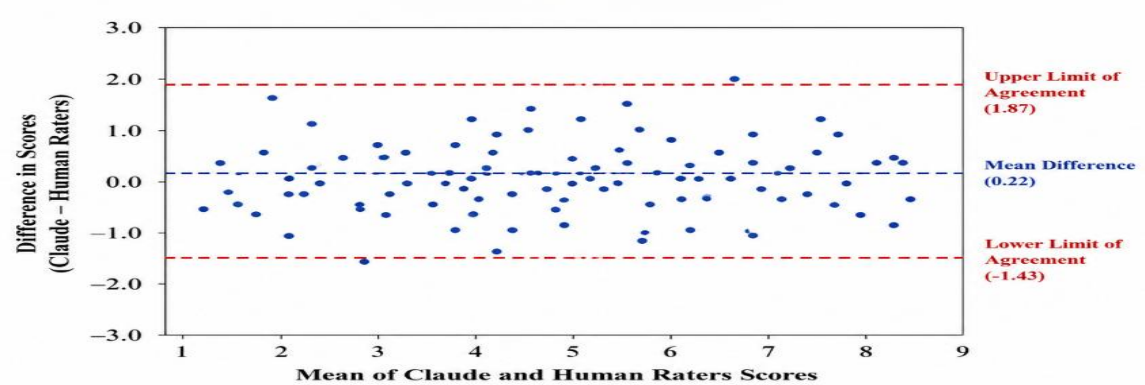


Figure 1. Agreement plot data across analytic dimensions

To further examine consistency between the two assessment methods, a Bland-Altman analysis was conducted for holistic speaking scores. Figure 2 presents the agreement pattern

between Claude and the human raters. The mean difference between the two assessment approaches was 0.22 points, with upper and lower limits of agreement of 1.87 and - 1.43, respectively. Most observations fell within these limits, indicating acceptable agreement and the absence of substantial systematic bias.



Statistic	Value
Mean Difference	0.22
Upper Limit of Agreement	1.87
Lower Limit of Agreement	-1.43

Note. The plot shows the difference between Claude and human ratings against their mean. The dashed lines represent the mean difference and the 95% limits of agreement.

Figure 2. Bland-Altman plot data for holistic scores

The agreement analyses demonstrate that Claude and the human raters produced highly comparable evaluations of EFL learners' oral proficiency across both holistic and analytic assessment frameworks.

Comparison of AI and Human Evaluation

The results of the comparative analysis and agreement analysis give a full picture of how much similarity there was between the evaluation of speaking performance by the human Claude and the trained human raters. The findings provided answers to the three study research questions, which were concerned with (1) Differences in holistic scores, (2) Differences in analytic scores on five speaking dimensions, and (3) Agreement between analytic and holistic scores.

Regarding holistic assessment, Claude and the human raters were very similar. Claude's mean holistic score (7.34) did not differ significantly from that of human raters (7.12), which yielded a mean difference of 0.22 points as reported in Table 2. This difference was not statistically significant $t(146) = 1.21$, $p = .228$), and the effect size was small ($d = 0.20$). Table 2 also showed there is close similarity between the two assessment approaches. The results suggest that the AI system used in the assessment system performed on the general oral proficiency at a level similar to the human assessors.

When the five analytic dimensions were used, a more differentiated pattern of findings emerged from the analysis. Overall, Claude scored marginally higher than the human raters on each of the dimensions (see Table 3). The mean differences of cohesion (0.08) and grammar (0.12) were the smallest, while content development (0.26), vocabulary (0.33), and coherence (0.34) were the largest. The repeated-measures ANOVA was significant, $F(4, 292) = 4.83$, $p = .002$, $\eta^2p = .12$, showing that the similarity between the AI-generated and human-assigned scores was significantly different across the five speaking constructs. Table 3 showed that the two methods of assessment

had the greatest difference in scores for vocabulary and coherence, but the least difference in scores for cohesion and grammar.

The individual analyses also revealed that there were no significant differences found in cohesion and grammar. More specifically, the cohesion scores that Claude rated (M = 3.94) were not significantly different from the human raters (M = 3.86), t(146) = 0.79, p = .431. Similarly, the grammar scores assigned by Claude (M = 3.81) and the human raters (M = 3.69) did not differ significantly, t(146) = 1.02, p = .309. However, notable variations were found with respect to the content, word choice, and organization of the materials. Content development scores also varied significantly t(146) = 2.18, p = .031; with Claude giving a higher mean score (3.98) than the human raters (3.72). For vocabulary, Claude also assigned a higher mean score (3.91) than the human raters (3.58). A Mann–Whitney U test confirmed that this difference was statistically significant, $U = 2206.50$, $p = .009$. Here, $U = 2206.50$ represents the test statistic obtained from the Mann–Whitney U test. Similarly, coherence scores varied significantly, t(146) = 2.73, p = .007: Claude rated the coherence of the writing at a mean score of 3.97, while the human raters rated the coherence of the writing at a mean score of 3.63. These differences were statistically significant, but their effect sizes were small to moderate, which indicates minimal practical differences between the two ways of assessing.

The results of the analyses of the agreement provided further evidence of consistency between Claude and the human raters. The ICC values for this ranged from .73 to .86, suggesting good to strong reliability within all assessment's dimensions (Table 9). The highest ICC was found for cohesion (ICC = .86), followed by holistic assessment (ICC = .84) and grammar (ICC = .82). Agreement for content development (ICC = .78), vocabulary (ICC = .74), and coherence (ICC = .73) also was acceptable. In like manner, Table 10 revealed that there was a good amount of agreement between the two across all the dimensions as evidenced by the Weighted Kappa values of .61 to .78. Further, it was illustrated that the level of agreement was the highest for cohesion and grammar, and the lowest for vocabulary and coherence (Figure 1). The Bland-Altman analysis presented in Figure 2 found an average difference of 0.22 points between the two methods of assessment and limits of agreement of upper 1.87 and lower -1.43 points. The majority of observations were within these bounds, signifying that there were good agreement and no large systematic error.

The results showed that the Claude and human raters gave similar oral proficiency assessment for the EFL learners in general. AI models of automated speaking scoring hold promise for measuring overall speaking ability as well as several dimensions of oral proficiency. For the first study hypotheses (*Ho1*), it was found that there is no statistically significant difference between the holistic speaking scores assigned by Claude and the human raters; therefore, *Ho1* was accepted. *Ho2* was eliminated due to the significant differences found in content development, vocabulary, and coherence. Lastly (*Ho3*), the consistency of AI-generated with human assigned scores in the agreement analyses was satisfactory between the holistic and analytic assessment framework. In general, the study results show that automated speaking scoring with AI is a reliable and promising way of measuring oral proficiency in the assessment of EFL speaking skills, and can be used as one of the important tools in the assessment of oral proficiency in EFL.

Discussion

The present study investigated the validity of automated scoring using AI for the assessment of EFL learners' oral proficiency in comparison with human scoring. The results

showed no significant differences between the holistic rating scores given by Claude and the human raters. The mean score was slightly higher for Claude, but the difference was small in magnitude with a negligible effect size between the two approaches, suggesting that there is a significant similarity between the two approaches. The finding indicates that modern-day evaluation systems based on AI are capable of assessing foreign language skills in a similar way to trained human assessors. The same conclusions have been drawn in other research (Bamdev et al., 2021; Obera et al., 2025) which demonstrates the possibility of achieving holistic scoring via automatic speech recognition (ASR), natural language processing (NLP), and machine-learning (ML) systems, in order to closely simulate human evaluations (Wang et al., 2021). Considering the similarity of the holistic scores in both AI-generated and human-assigned categories, from a construct theoretic perspective, the two methods were measuring the same aspects of oral proficiency, thus demonstrating the construct validity of AI-based speaking evaluation (Kane, 2013; Messick, 1989).

When the dimensions of the analyses are compared, differing degrees of similarity in the analyses of the Claude and human raters are noted. A joint assessment from cohesion provided the highest level of agreement and there was no statistically significant difference between the two assessment approaches. Cohesion is affected by the effective use of linking devices, referential expressions, and logical links between ideas, making it an appropriate content for automated assessment using discourse- and text-analysis methods. The near congruence noticed in the present study indicates that AI systems can be used to correctly detect linguistic link elements that help to achieve the connectedness and continuity of text (Alam et al., 2026a). This finding suggests that automated assessments can effectively measure cohesion, in a way very similar to human assessment.

The development of the content was also fairly comparable, although Claude frequently gave higher scores than the human ratings. Previous research suggests that differences between automated and human assessment may arise from variations in how relevant, sufficiently developed, and adequate oral responses are interpreted and evaluated (Davis & Papageorgiou, 2021). For automated systems, the focus is on semantic analysis, topic relevance, and information density while human raters might consider various other measures, such as overall communicative effectiveness or appropriateness in context. Thus, Claude might have been more likely to award answers that included more semantically similar content; the human raters more likely to scrutinize supporting details for quality and relevance. The results indicate that there could be differences in the interpretation of discourse elaboration in content-related features, but AI systems are effective at evaluating content-related features.

There were mixed responses for vocabulary and grammar. The vocabulary scores were statistically significantly different, with Claude consistently assigning higher scores. Large language models may focus more on lexical diversity, complexity and semantic richness while human evaluators may emphasize contextual appropriateness, semantic precision and communicative suitability. As a result, some of the word selection was imprecise in context, but Claude might have been aggressive about rewarding the broader lexical use of such words. This is what can make the vocabulary scores achieved by the AI system appear higher than in reality. Previous studies have reached similar conclusions, finding less agreement between AI and human assessments of vocabulary, as effectiveness depends on factors other than computational analysis of the text (Crossley & McNamara, 2013). In contrast, there was no statistically significant difference for grammar, and relatively high agreement coefficients. According to this finding,

grammatical assessment systems can be reliably performed by automatic systems based on natural language processing of speech transcription, as has been revealed in other studies on grammatical assessment that showed similar performance by AI systems and trained human raters (Bamdev et al., 2021).

The most considerable discrepancy was observed for coherence. Although agreement coefficients remained acceptable, coherence produced the highest mean difference and comparatively lower agreement values. While coherence is commonly evaluated by human raters by means of logical organization, topic development, discourse progression, semantic connectedness, and communicative effectiveness, automated systems usually attempt to measure coherence by transcribing these elements in terms of lexical cohesion and semantic relationships found in the transcripts. As a result, Claude was able to infer meaning and links between thoughts, even though from the human standpoint, there was a less clear discourse organization. This finding is consistent with previous studies showing that coherence and discourse organization remain among the most difficult dimensions for automated speaking assessment systems to evaluate accurately because they require interpretation of higher-order discourse relationships and contextual meaning (Li et al., 2024; Ockey & Chukharev-Hudilainen, 2021; Obera et al., 2025). The findings also support sociocultural perspectives of assessment, which emphasize that communicative competence extends beyond measurable linguistic forms to include contextual meaning and socially situated communication (Xi, 2010).

Overall, the agreement analyses also revealed that the judgments made by both Claude and the human subject raters were very similar, both in the holistic and analytic rubrics. The ICCs showed good to strong agreement, the Weighted Kappa values showed substantial agreement, and the Bland-Altman analysis showed a mean difference of only 0.0–0.1 cm, with no significant systematic bias. The findings demonstrate a high degree of correspondence between Claude and human raters across the holistic and analytic assessment frameworks, supporting the proposed expectation of high agreement between the two assessment approaches. The highest agreement was found for cohesion and grammar while comparatively lower agreement was found for vocabulary and coherence. The results indicated that AI-enhanced speaking evaluation platforms can provide accurate and consistent judgements on various aspects of speaking. These findings are consistent with previous research that found high inter-agreement between automated and human scoring (Davis & Papageorgiou, 2021; Zechner et al., 2014). However, the differences in the degree of consensus suggest that the performance of AI systems is greater when assessing features of the language and discourse than those that are specific to the context of the communication.

One major trend that was clear from the results is that Claude consistently scored a little higher than the human raters on all of the assessment components. While these differences tended to be rather minimal, the direction of the differences was consistent, suggesting a possible AI leniency effect. Large language models tend to focus on efficacy and semantic appropriateness, as opposed to human evaluators who might be more conservative in their approach to scoring. This means that AI systems could potentially rate students higher for conveying meaning rather than for linguistic correctness. AI systems might thus give a higher score to students who demonstrate ways of expressing meaning even if they make linguistic mistakes. Such a tendency is a general pattern of LLM performance in the scoring of native LLM learners' texts. This finding substantiates previous research, where it was found that AI-based assessment systems might offer more favorable ratings than human raters (Wang et al., 2021).

The results can also be analyzed using Multi-Faceted Rasch Measurement Theory. Linacre (1989) noted that outcomes of assessment are affected by several dimensions, such as how the rater scores the task, the nature of the task, or the ability of the learner. The ICC and Weighted Kappa scores are strong, suggesting that the scoring method added little measurement variability and that Claude was more or less a stable facet of assessment, similar to trained human scorers. However, the variance in content development, vocabulary, and coherence among the scores suggests that such variation is not completely ruled out by automated systems, because differences in what it is that AI systems and human evaluators are scoring communicative performance still exist (Wang & Sun, 2018).

The results of this study also have implications in Saudi Arabian EFL contexts. With a massive number of students who are studying English, AI-based assessment systems could be an effective solution to offer a streamlined approach to evaluating oral proficiency that includes quick, on-the-spot scoring, feedback, and thorough information about the student's learning needs. The capabilities match the growing incorporation of Artificial Intelligence (AI) in the realm of education in Saudi Arabia. The major discrepancies found in the content construction, vocabulary and coherence, however, suggest that AI systems should be used in conjunction, not in replacement of human evaluators (Alam et al., 2026b). The results thus validate hybrid models of assessments that blend the ease and uniformity of AI-based scoring with the contextual understanding and interpretative judgement of human raters (Davis & Papageorgiou, 2021; Polasa et al., 2025).

Based on the results, the automatic oral assessment tool for EFL by using AI can be considered a reliable tool for measuring oral proficiency in EFL. The results of the holistic assessment, cohesion and grammar did not show significant differences and had high agreement levels across each of these areas, suggesting that modern AI technologies can to some extent analyze many aspects of a speaker's performance with a fair degree of accuracy. However, there are significant differences in terms of content, vocabulary and coherence which suggest that there is still some room for the contribution of human knowledge and interpretation to oral skills. The findings are in line with the ongoing debate about the best assessment methods for the future, which considers hybrid approaches that integrate AI technology with assessment by human judges as preferred methods, leading to reliable assessment procedures (Polasa et al., 2025).

Key implications of the findings for language assessment and incorporating AI into oral proficiency assessment are discussed. Claude's high degree of correspondence with the human raters indicates that AI-based assessment systems can deliver quick, scalable, and consistent speaking evaluation, thus, relieving the burden of evaluation and speeding up the feedback process. This potential is of particular importance in EFL settings in Saudi Arabia where teachers have to evaluate high numbers of students. The marked differences, however, in content development, vocabulary acquisition, and coherence suggest that the need for human expertise will remain to assess discourse-level, contextual and communicative aspects of oral proficiency. For EFL settings, therefore, a hybrid assessment model may be the most useful and valid strategy for speaking assessment, as it can be done using AI-generated assessment with human-trained interpreters, while maintaining the benefits of efficiency and consistency provided by the AI system.

There are a number of limitations to be mentioned when considering the results. Firstly, the study sample was limited (N = 74) and included male Arabic-speaking EFL learners from one Saudi university, so the results may not be generalized to other educational contexts, other Arabic-speaking learners, and female learners. Second, it is uncertain how much agreement between AI

and humans might exist at proficiency levels other than intermediate (B1); and third, all participants were at the bachelor (B1) proficiency level. Third, a picture description task was used which was not a comprehensive representation of all language activities used in language assessment. Fourth, the study only examined Claude, whereas various AI scoring systems could yield varying patterns. Finally, the study investigated the agreement and reliability between AI and human raters without exploring the decision-making processes used. Further studies are recommended to involve larger and more diverse samples, different levels of achievement, different speaking tasks, comparative studies of differences among various AI-based assessment systems.

Conclusion and Recommendations

Results show that there is a high degree of similarity between the assessment of oral competence of EFL learners carried out using AI-assisted automated speaking assessment and the assessment performed by a human. Results point to no statistically significant difference between holistic scores generated by the AI and holistic scores assigned by humans, and statistically good to strong agreement was present in both holistic and analytic assessment scores. Cohesion and grammar were the most similar and there was no significant difference between them while with content development, vocabulary and coherence, their similarities were acceptable but differences significant. These findings indicate that AI systems show promise of assessing linguistic & discourse features that are easily analyzed systematically and computationally which are more problematic at assessing higher order discourse organization, contextual interpretation and communicative effectiveness. In addition, the high Intraclass Correlation Coefficients, high Weighted Kappa and satisfactory Bland-Altman agreement limits suggest that assessment using artificial intelligence can yield reliable and consistent judgements of oral proficiency.

Future studies should investigate alignment between AI and human scores with larger and more diverse groups of learners, with learners of varying abilities, in a variety of speaking tasks, and with a wider linguistic background. Further research is needed to explore how task difficulty, discourse-type and language level affect scoring agreement, and whether gender differences emerge when considering the best ways to teach communities to be accurate. Further investigation is needed to examine the impact of various factors (language level, task complexity and discourse type) on scoring agreement, as well as gender disparities in scoring on the task of training communities to be accurate. Comparative studies of different system of speaking assessment using Artificial Intelligence would also help to understand the strengths and weaknesses of the automated speaking assessment and development of validity, reliability and equity practices in speaking assessment.

Funding
The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2026/02/45289).
Acknowledgement
The authors extend their appreciation to Deanship of Research and Graduate Studies for constant support and thankful to each participant for the present study.
Conflict of Interest: The author declares no conflict of interest.
Authenticity: This manuscript is an original work.

Artificial Intelligence Statement: The author has reviewed and edited the content and took full responsibility for the content of the publication. There is no use of AI in this manuscript.

About the Author:

Dr. Sohaib Alam is an Associate Professor of Applied Linguistics in the Department of English Language and Literature at Prince Sattam Bin Abdulaziz University (Alkharj, Saudi Arabia). He holds a Ph.D. in Applied Linguistics and brings over a decade of international university-level teaching and research experience across India and Saudi Arabia. His primary research interests with investigations into AI-assisted language teaching and learning. https://orcid.org/0000-0002-9972-9357

References

Alam, S., Amir, Saudi, T. T., Ahmad, F., Kralik, R., Rasheed, T. (2026a). Impact of Meta AI generated corrective feedback in the writing classroom: effects on L2 writing linguistic errors. *World Journal of English Language, 16(*3), 119-130. https://doi.org/10.5430/wjel.v16n3p119.

Alam, S., Ahmad, F., Mohamed, A. O., Dawood, J. A., Obloberdiyevna, D. S., & Rasheed, T. (2026b). Artificial intelligence in Saudi English language classrooms: A narrativereview. *World Journal of English Language, 16*(5), 254. https://doi.org/10.5430/wjel.v16n5p254

Alam, S., Hameed, A., Balogh, Z., & Haque, S. A. (2025). Digitalisation of education: challenges and issues in virtual pedagogical competence. *Journal of Education Culture and Society*, *16*(1), 165-177. https://doi.org/10.15503/jecs2025.2.165.177

Bamdev, P., Grover, M. S., Singla, Y. K., Vafaee, P., Hama, M., & Shah, R. R. (2023). Automated speech scoring system under the lens: evaluating and interpreting the linguistic cues for language proficiency. *International Journal of Artificial Intelligence in Education*, *33*(1), 119–154. https://doi.org/10.1007/s40593-022-00291-5

Benaissa, M., Attou, Y., Seddik, M., & Nfissi, A. (2025). Can AI-based teacher-bots (T-bots) enhance students' speaking skills? A case study of "My AI" on Snapchat in Moroccan higher education. *Arab World English Journal (AWEJ), Special Issue on CALL*(11), 144–159. https://doi.org/10.24093/awej/call11.9

Brown, H. D., Abeywickrama, P. (2021). *Language assessment: principles and classroom practices*. Pearson.

Butarbutar, R., Sujarwo, Gunu Pure, E. A., Sulviana, & Rizal, A. (2025). Unlocking the impact of AI-assisted online interactive learning environment in English language learning. *The Journal of Asia TEFL, 22*(4), 646–662. https://doi.org/10.18823/asiatefl.2025.22.4.2.646

Clapham, C. (2000). Measuring second language performance.: Tim McNamara. *TESOL Quarterly, 34*(2), 376-377. https://doi.org/10.2307/3587967

Crossley, S. & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology, 17*(2), 171–192. https://doi.org/10.64152/10125/44329

Davis, L., & Papageorgiou, S. (2021). Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 437–455. https://doi.org/10.1080/0969594X.2021.1979466

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. *Analyzing and Evaluating Rater-Mediated Assessments.* Berlin, Germany: Peter Lang Verlag. Retrieved on Aug 20, 2026. https://doi.org/10.3726/978-3-653-04844-5

Fang, Y. H., Lo, T. H., Sung, Y. T., & Chen, B. (2025, December). Beyond Modality Limitations: A Unified MLLM Approach to Automated Speaking Assessment with Effective Curriculum Learning. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1-7). IEEE.

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge.

Giavarina, D. (2015). Understanding Bland-Altman analysis. *Biochemia medica*, *25*(2), 141-151. https://doi.org/10.11613/BM.2015.015

Grab, M. O. (2026). Embedding design thinking and artificial intelligence in EFL classrooms: Effects on secondary learners' oral proficiency. *Social Sciences & Humanities Open, 14*, 103465. https://doi.org/10.1016/j.ssaho.2026.103465

Hu, H., & Hashim, H. (2026). What shapes automated ratings in computer-based Englishspeaking tests? Perspectives from analytic complexity, accuracy, fluency, andpronunciation indices. *VIAL-Vigo International Journal of Applied Linguistics*. *23*(1), 51-82. https://doi.org/10.35869/vial.v0i23.6467

Hu, H., Said, N. E. M., & Hashim, H. (2025). Human ratings and complexity, accuracy, and fluency (CAF) indices: A correlational study of a standardised monologic English speaking test in China. *SAGE Open, 15*(2), 1-13. https://doi.org/10.1177/21582440251343944

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Karatay, Y. and Xu, J. (2025), Exploring the potential of conversational AI for assessing second language oral proficiency. *TESOL Quarterly, 59*(1), S220-S250. https://doi.org/10.1002/tesq.70003

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kobylarek, A., & Birova, J. (2026). AI as the philosophy of a matrix society. educational implications. *Journal of Education Culture and Society*, *17*(1), 9-18. https://doi.org/10.15503/jecs2026.1.9.18

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *Relc Journal*, *54*(2), 537-550. https://doi.org/10.1177/00336882231162868

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Lai, Z. C.-C. (2026). AI in EFL blended learning: Impact on speaking performance and learning resilience. *Asian-Pacific Journal of Second and Foreign Language Education, 11*, 11. https://doi.org/10.1186/s40862-025-00367-4

Li, H. (2024). Design and development of an oral English evaluation system based on speech recognition technology and a scoring algorithm. *Journal of Electrical Systems, 20*(6), 75-84. https://doi.org/10.52783/jes.2616

Li, J., Huang, J., & Sheeran, T. J. (2025). ChatGPT-4o as an AI peer assessor in EFL speaking classrooms: Examining scoring reliability and feedback effectiveness. *SAGE Open, 15*(3), 1-15. https://doi.org/10.1177/21582440251369938

Li, J., Yan, B., Lo, T., Wang, Y., Hsu, Y., & Chen, K. (2024). *Automated speaking assessment of conversation tests with novel graph-based modeling on spoken response coherence*. *IEEE Spoken Language Technology Workshop (SLT)*, Macao, pp. 825-832. https://doi.org/10.1109/SLT61566.2024.10832136.

Linacre, J. M. (1989). Rasch Models from Objectivity: A Generalization. Retrieved from https://eric.ed.gov/?id=ED308211

Matsuura, R., Suzuki, S. (2023). Prompt-Independent Automated Scoring of L2 Oral Fluency by Capturing Prompt Effects. *In: Wang, N., Rebolledo-Mendez, G., Matsuda, N.,Santos, O.C., Dimitrova, V. (eds) Artificial Intelligence in Education. AIED 2023*.Lecture Notes in Computer Science, vol 13916. Springer, Cham.https://doi.org/10.1007/978-3-031-36272-9_62

Mendonça, P. C., Quintal, F., & Mendonça, F. (2025). Evaluating LLMs for automated scoring in formative assessments. *Applied Sciences*, *15*(5), 2787-2805. https://doi.org/10.3390/app15052787

Messick, S. (1989). *Validity. In R. linn (org.), Educational Measurement*.(13-103). *New York, NY: American Council on Education and Macmillan Publishing Company*.

Metruk, R. (2018). Comparing holistic and analytic ways of scoring in the assessment of speaking skills. *Journal of Teaching English for Specific and Academic Purposes, 6*(1), 179–189. https://doi.org/10.22190/JTESAP1801179M

Mohammadkarimi, E., Omar, J. A., & Rashid, A. S. (2025). Advancements and Challenges in Automated Evaluation of Spoken Language Proficiency Using AI. In N. Jomaa (Ed.), *Using AI Tools in Text Analysis, Simplification, Classification, and Synthesis* (pp. 225-256). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-9511 0.ch008

Ndacyayisenga, J. C. Y., Kidega, C., & Can, L. A. (2026). The impact of generative artificial intelligence tools on assessment equity for non-native English-speaking medical students: A systematic review. *BMC Medical Education, 26*, 989. https://doi.org/10.1186/s12909-026-09303-7

Nikolovski, V., Trajanov, D., & Chorbev, I. (2025). Advancing AI in higher education: a comparative study of large language model-based agents for exam question generation, improvement, and evaluation. *Algorithms*, *18*(3), 1-34. https://doi.org/10.3390/a18030144

Obera, J. A., Ojochegbe, A. T., Obera, J. A., & Ojochegbe, A. T. (2025). Automated scoring systems and teacher assessments in English language education. *Journal of World Science, 4*(10), 1352–1358. https://doi.org/10.58344/jws.v4i10.1517

Ockey, G. J., & Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics, 42*(5), 924–944. https://doi.org/10.1093/applin/amaa067

Ogawa, C. (2022). CAF indices and human ratings of oral performances in an opinion-based monologue task. *Language Testing in Asia, 12*(1), 1-18. https://doi.org/10.1186/s40468-022 00154-9

Polasa, P., Laoaree, S., Thanadunpremdet, T., Rodjananant, N., & Kritsuthikul, N. (2025). "Chai Calibrated Hybrid Assessment for IELTS Speaking with Human-Referenced Validation," *2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Phuket, Thailand, 2025, pp. 1-6, doi: 10.1109/iSAI-NLP66160.2025.11320542.

Scaria, N., Kennedy, S. J. J., Latinovich, T., & Subramani, D. N. (2024). *Evalyaks: Instruction tuning datasets and LoRA fine-tuned models for automated scoring of CEFR B2 speaking assessment transcripts*. *Computers and Education: Artificial Intelligence*, *10*(2026), 1-20. https://doi.org/10.1016/j.caeai.2025.100539

Shikun, S., Grigoryan, G., Huichun, N., & Harutyunyan, H. (2024). AI chatbots: Developing English language proficiency in EFL classroom. *Arab World English Journal (AWEJ), Special Issue on ChatGPT*, 292–305. https://doi.org/10.24093/awej/ChatGPT.20

Wang, X., Evanini, K., Qian, Y., & Mulholland, M. (2021). *Automated scoring of spontaneous speech from young learners of English using transformers*. *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, 2021, pp. 705-712. https://doi.org/10.1109/SLT48900.2021.9383553

Wang, Y., & Min, S. (2026). The human–machine correlation in automated speech evaluation: A three-level meta-analysis. *Language Testing*, *43*(2), 165-196. https://doi.org/10.1177/02655322251387816

Wang, Z., & Sun, Y. (2018). Comparison of human rater and automated scoring of test takers' speaking ability and classification using item response theory. *Psychological Test and Assessment Modeling, 60*(1), 81-100.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, *31*(1), 2-13.

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161-192. https://doi.org/10.1177/0265532216686999

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading?. *Language Testing*, *27*(3), 291-300.https://doi.org/10.1177/0265532210364643

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech using SpeechRaterSM v1.0. *ETS Research Report Series, 2008*(2), 1-113. https://doi.org/10.1002/J.2333-8504.2008.TB02148.X

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: examining automarker reliability. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 411–436. https://doi.org/10.1080/0969594X.2021.1979467

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*(1), 150-166. https://doi.org/10.1111/bjet.13494

Zechner, K., & Bejar, I. I. (2006). *Towards automatic scoring of non-native spontaneous speech*. Human Language Technology Conference of the North American Chapter of the

Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA. https://doi.org/10.3115/1220835.1220863

Zechner, K., Evanini, K., Yoon, S., Davis, L., Wang, X., Chen, L., Lee, C. M., & Leong, C. W. (2014). *Automated scoring of speaking items in an assessment for teachers of English as a foreign language*. Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications. Baltimore, Maryland USA. https://doi.org/10.3115/V1/W14-1816

Zhong, W., Zhou, Z., Wang, L., & Yang, Y. (2026). Developing and validating multi-scenario EFL willingness to communicate scale across classroom, extracurricular, and GenAI mediated scenarios. *BMC Psychology, 14*, 819. https://doi.org/10.1186/s40359-026 04607-1

Zhou, T., & Yu, H. (2025). *Research on deep learning-based English speaking automatic scoring system*. CISAI 2025: 2025 8th International Conference on Computer Information Science and Artificial Intelligence, China. https://doi.org/10.1145/3773365.3773381

Zou, B., Du, Y., Wang, Z., Chen, J., & Zhang, W. (2023). An investigation into artificial intelligence speech evaluation programs with automatic feedback for developing EFL learners' speaking skills. *SAGE Open*. 13(3), 1-9. https://doi.org/10.1177/21582440231193818

Appendix A. AI Assessment Prompt Used for Automated Speaking Scoring

AI Assessment Prompt for Claude-Based Evaluation of Oral Proficiency

You are an experienced English language assessor evaluating the oral proficiency of an English as a Foreign Language (EFL) learner. Carefully analyze the following speech transcript and assess the learner's performance according to the rubric criteria provided below. Evaluate the transcript objectively, consistently, and strictly according to the scoring descriptors.

Assessment Criteria

Holistic Speaking Score (1–10)

Assign one overall score representing the learner's general oral proficiency and communicative effectiveness.

10–9 (Excellent): Communication is clear, coherent, and highly effective. Ideas are well organized and fully developed. Language use is accurate, appropriate, and varied, with only minimal errors.

8–7 (Good): Communication is generally clear and effective. Ideas are organized and adequately developed. Minor language errors may occur but do not interfere with understanding.

6–5 (Satisfactory): Communication is understandable but may contain noticeable weaknesses in organization, development of ideas, or language use. Errors occasionally affect clarity.

4–3 (Limited): Communication is often unclear. Language problems and organizational weaknesses frequently interfere with understanding.

2–1 (Very Limited): Communication is severely restricted and difficult to understand because of major linguistic and organizational problems.

Analytic Speaking Scores (1–5)

Coherence

Evaluate the logical organization, sequencing, and progression of ideas throughout the response.

5: Ideas are logically organized and consistently developed.
4: Ideas are generally well organized with minor lapses.
3: Organization is adequate but occasionally inconsistent.
2: Ideas are poorly organized and difficult to follow.
1: Little or no logical organization is evident.

Cohesion

Evaluate the effective use of linking devices, transitions, references, and other cohesive features.
5: Cohesive devices are used effectively and naturally.
4: Generally effective use of cohesive devices.
3: Some cohesive devices are used appropriately.
2: Limited use of cohesive devices.
1: Minimal or ineffective cohesion.

Content Development

Evaluate the relevance, adequacy, elaboration, and completeness of ideas in relation to the speaking task.
5: Ideas are fully developed with relevant details and supporting information.
4: Ideas are adequately developed with appropriate support.
3: Ideas are partially developed.
2: Limited development of ideas.
1: Very limited content related to the task.

Grammar

Evaluate grammatical accuracy, syntactic control, and appropriateness of sentence structures.
5: Consistently accurate and varied grammatical structures.
4: Generally accurate grammar with minor errors.
3: Noticeable grammatical errors that do not seriously impede understanding.
2: Frequent grammatical errors that affect communication.
1: Persistent grammatical errors that severely hinder communication.

Vocabulary

Evaluate lexical range, appropriateness, precision, and effectiveness of word choice.
5: Wide and appropriate vocabulary used effectively.
4: Good range of vocabulary with occasional inaccuracies.
3: Adequate vocabulary for the task.
2: Limited vocabulary that restricts expression.
1: Very limited vocabulary.

Output Format

Provide the evaluation in the following format only:
Holistic Score: [1–10]
Coherence: [1–5]
Cohesion: [1–5]
Content Development: [1–5]
Grammar: [1–5]
Vocabulary: [1–5]
Brief Justification:
Provide a concise explanation (2–3 sentences) justifying the assigned scores based on the learner's performance according to the rubric criteria.