Korean Journal of English Language and Linguistics, **Vol 25, May 2025, pp. 661-685**
DOI: 10.15738/kjell.25..202505.661

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Investigating the Impact of Interlocutor Type on English Oral Proficiency Interviews: A Comparative Analysis of Chatbot and Human Interlocutors

Yongkook Won · Sunhee Kim (Seoul National University)

Received: March 8, 2025
Revised: April 21, 2025
Accepted: May 13, 2025

Won, Yongkook (First author)
Visiting Researcher,
Center for Educational Research
Seoul National University
Email: purgatorio@snu.ac.kr

Kim, Sunhee
(Corresponding author)
Associate Professor,
Department of French Language Education, Adjunct Professor,
Learning Sciences Research Institute
Seoul National University
Email: sunhkim@snu.ac.kr

ABSTRACT

Won, Yongkook and Sunhee Kim. 2025. Investigating the impact of interlocutor type on English oral proficiency interviews: A comparative analysis of chatbot and human interlocutors. *Korean Journal of English Language and Linguistics* 25, 661-685.

Considering the recent emergence of voice chatbots as substitutes for human interlocutors in eliciting spoken responses during English oral proficiency interviews, this study examines how interlocutor type affects both fluency and holistic scores. Data were collected from 32 Korean college students, yielding 128 audio recordings across four distinct topics of varying complexity, with each topic administered via both chatbot and human interlocutors. Fluency features were analyzed using Praat software, while fluency and holistic scores were evaluated via many-facet Rasch measurement (MFRM) analyses by two raters. Results from Friedman and Wilcoxon tests indicate that both task complexity and interlocutor type influence temporal measures, although task complexity exerts a stronger effect on dysfluency measures. MFRM analyses further show that chatbot interlocutor difficulty significantly affects fluency but not holistic scoring, indicating distinct difficulty levels between interlocutors only in fluency scoring. Overall, these findings highlight both the potential and limitations of employing chatbot interlocutors in place of human interlocutors in oral proficiency interviews.

KEYWORDS

oral proficiency interview, artificial intelligence, English as a foreign language, speaking tests, chatbot

1. Introduction

Recent advancements in human–computer interaction technology have stimulated growing interest in the integration of artificial intelligence (AI) chatbots into foreign language education (Coniam 2014, Fryer et al. 2020, García Laborda et al. 2024, Huang et al. 2022, Kim et al. 2021, Kohnke et al. 2023, Yang 2022). Among these innovations, voice-based chatbots that simulate human conversation have demonstrated notable potential for supporting speaking practice and, more recently, for applications in language assessment contexts (Ockey and Chukharev-Hudilainen 2021, Timpe-Laughlin et al. 2022). While research has primarily focused on pedagogical applications and learner–chatbot interaction dynamics (Han 2020, Kim 2020, Yang et al. 2022), less is known about how interlocutor type (chatbot vs. human) may influence formal speaking assessments.

Prior studies have explored learners' speech production with general-purpose AI speakers and structured chatbot systems (Kim et al. 2019, Sung 2019, Yang et al. 2022), showing that AI interlocutors can elicit meaningful spoken output. Recent findings suggest that AI chatbot–based assessments may offer score outcomes comparable to human evaluations (Abida et al. 2023, Liu et al. 2025), emphasizing their potential for speaking evaluation. Nevertheless, significant challenges persist, particularly concerning the limited scoring accuracy of current AI systems and the threat to authenticity—defined as how the interlocutor's nature might alter examinee behavior and, consequently, the construct being assessed (Ross and Berwick 1992, Wu et al. 2025, Zhang et al. 2020). If examinees respond differently when interacting with a chatbot compared to a human, the validity of score interpretations may be compromised.

In speaking assessment research, interlocutor characteristics have long been recognized as a major source of construct-irrelevant variance (Chichon 2019, Fulcher 2003, McNamara 1996, Ockey and Li 2015). Early studies demonstrated that variability in interlocutor behavior could affect performance independently of test-takers' language ability. More recently, Ockey and Chukharev-Hudilainen (2021) and Ockey et al. (2023) developed a voice-based chatbot designed to assess interactional competence during paired speaking tasks. Their studies demonstrated that chatbots can provide standardized, ratable opportunities to observe interactional features. However, although they compared human- and chatbot-mediated interactions, their research primarily focused on interactional competence. They did not specifically focus on how interlocutor type might influence fluency measures or rater perceptions of fluency.

Fluency, especially as operationalized through temporal measures (e.g., speech rate, mean length of run) and dysfluency markers (e.g., filled pauses, hesitations), is closely tied to cognitive load and processing demands (Kormos and Dénes 2004, Lennon 1990). These cognitive demands may differ between AI-mediated and human-mediated interactions, yet few studies have directly compared the two modalities from a fluency-focused perspective. Given the growing educational applications of AI chatbots and the importance of fluency as a dimension of speaking performance (Hill et al. 2015, Wu et al. 2025), this gap warrants empirical attention.

To address this gap, the present study investigates how a standardized voice chatbot, built using Google Dialogflow, compares to a human interlocutor in eliciting fluency-related speech features and influencing rater evaluations during oral proficiency interviews. By combining automated fluency metrics with rater-based assessments, this study offers an integrated perspective on how chatbot-mediated interviews affect both speech production and its evaluation—an approach that remains underdeveloped in prior research (Jeon and Lee 2024, Wang et al. 2024).

Accordingly, this study addresses the following research questions (RQs):

1) How does interlocutor type (voice chatbot vs. human) affect fluency measures for Korean college

students during an oral proficiency interview?

2) How do raters evaluate an examinee's fluency and overall performance differently depending on whether the interlocutor is a voice chatbot or a human?

In doing so, this study contributes to the evolving field of AI-mediated speaking assessment by identifying conditions under which chatbot-led interviews might serve as valid substitutes for human-led interactions, particularly in fluency-focused evaluation contexts. It further provides empirical insights for test developers, policymakers, and educators aiming to balance assessment reliability, accessibility, and authenticity.

2. Literature Review

2.1 Interlocutors in the Oral Proficiency Interview

In an oral proficiency interview, an interlocutor is defined as a person "who talks with the candidates, and an assessor, who marks them" (British Council 2023). Interlocutors initiate and sustain interaction with test-takers, guiding conversations and eliciting language in a semi-structured format. According to the oral communication assessment model proposed by Ockey and Li (2015), scores in performance-based oral communication assessments are based on raters' evaluations of candidates' oral proficiency against predefined rating criteria. These assessments are influenced not only by examinees' language abilities but also by contextual variables such as task design, technological mediation, and interlocutor characteristics (McNamara 1996, Van Moere 2013). Given the numerous factors involved, it is plausible that variance unrelated to the intended construct may inadvertently be introduced. This construct-irrelevant variance can stem from any component of the assessment model, including rating criteria, task types, interlocutor attributes, technological environments, or their interactions (Ockey and Li 2015).

Within this framework, human variables—specifically the roles of raters and interlocutors—are important contributors to construct-irrelevant variance, yet their exact impact remains complex and sometimes elusive. Achieving consistency in evaluating examinee performance, regardless of who conducts or scores the interview, is critical for test validity (Brown 2012). However, substantial evidence documents that interlocutors can introduce construct-irrelevant score variance (Brown 2003, McNamara and Lumley 1997). Studies have found that interlocutor characteristics such as language proficiency (Davis 2009), ethnicity (Hou 2006), competence levels (Morton et al. 1997), interpersonal supportiveness (Chartrand and Bargh 1999, Clark 2002, Wilson and Wilson 2005), and specific areas of concern (May 2011) significantly affect test-taker performance. Notably, examinees tend to perform better when supported by empathetic interlocutors (McNamara and Lumley 1997, Morton et al. 1997), whereas in cases involving less supportive interlocutors, raters have sometimes adjusted scores upward to compensate (Lazaraton 1996). These findings underscore the intricate relationship between interlocutor behavior and assessment outcomes.

Given these challenges, researchers have increasingly explored whether standardized, technology-mediated interlocutors, particularly voice-based AI chatbots, could provide more consistent assessment conditions (Ockey and Chukharev-Hudilainen 2021, Yang et al. 2022). Ensuring consistency and minimizing construct-irrelevant variance introduced by interlocutors remains critical for the validity of speaking assessments. As AI chatbots are introduced as potential standardized interlocutors, it becomes essential to examine how they affect not only interactional competence but also fluency patterns and rater perceptions of test-taker speech.

2.2 Use of Chatbots in the Oral Proficiency Interview

Chatbots, originally referred to as "ChatterBot" (Mauldin 1994), are conversational agents that perform specific tasks by engaging in communication with humans via voice or text (Adamopoulou and Moussiades 2020). In the field of language education, chatbots have been increasingly integrated into both instructional and assessment contexts, particularly within computer-assisted language learning (CALL) frameworks. A growing body of research highlights their effectiveness in supporting second language (L2) oral skill development. Systematic reviews have consistently identified benefits such as enhanced learner motivation, reduced speaking anxiety, improved vocabulary and grammar acquisition, and overall gains in speaking ability (Ayedoun et al. 2015, Fryer and Carpenter 2006, Hsu et al. 2023, Tai and Chen 2022). Wollny et al. (2021) synthesized evidence showing that learners generally perceived chatbot-mediated interactions positively and experienced linguistic improvements across multiple domains. Similarly, Jeon and Lee (2024) found that speech-recognition chatbots fostered greater L2 exposure, self-regulated learning behaviors, and improved speaking confidence.

Despite these promising outcomes, research has also highlighted notable limitations. Challenges such as imperfect speech recognition for non-native speakers, diminished conversational spontaneity, and the reduced ability of chatbots to replicate the dynamic responsiveness of human interlocutors have been emphasized (Jeon and Lee 2024, Wollny et al. 2021). These limitations raise important concerns about the authenticity and validity of chatbot-mediated interactions, particularly in formal assessment settings (García Laborda et al. 2024, Huang et al. 2022).

Voice-based chatbots, which enable spoken exchanges, are particularly relevant for oral proficiency testing. Tai and Chen (2022) demonstrated that adolescent EFL learners interacting with AI assistants like Google Assistant produced more fluent and authentic speech and exhibited heightened engagement. Similarly, classroom-based studies by Kim et al. (2019) and Yang et al. (2022) reported that chatbot interactions encouraged longer speech turns and greater lexical diversity. In assessment-specific contexts, Ockey and Chukharev-Hudilainen (2021) developed the Interactional Competence Elicitor (ICE) to compare learners' paired discussions with either a human interlocutor or a chatbot. Their findings indicated that although chatbot partners could elicit coherent and ratable speech samples, learners generally received lower scores for interactional competence when interacting with chatbots, likely due to the scripted and less adaptive nature of AI responses. A subsequent study by Ockey et al. (2023) reinforced this observation, emphasizing the inherent trade-off between achieving standardization and preserving interactional authenticity in AI-mediated interviews.

Complementary studies further illustrate the pedagogical potential of conversational agents. Fryer and Carpenter (2006) found that chatbots promoted autonomous language practice through informal dialogues, while Ayedoun et al. (2015) showed that semantically driven chatbot interactions helped EFL learners simulate real-world conversations, reduce anxiety, and build communicative confidence. Xu et al. (2021) observed that although children interacting with conversational agents produced clearer responses, human interlocutors elicited more linguistically rich and varied output, suggesting that task type and learner characteristics may moderate chatbot effectiveness. Nonetheless, some research points to drawbacks: Fryer et al. (2017) reported declining engagement among university students when chatbots, rather than human partners, were used for language learning. Such mixed findings underscore the complex trade-offs involved in integrating chatbots into both educational and assessment settings.

Despite these advances, relatively few studies have rigorously investigated linguistic fluency in chatbot-mediated interviews or examined how rater perceptions vary depending on interlocutor type. While existing research demonstrates that chatbots can support language practice and elicit ratable samples (Abida et al. 2023,

Wu et al. 2025), important questions remain about how chatbot interaction influences critical fluency dimensions such as speech rate, pausing, and hesitation phenomena—features known to be highly sensitive to cognitive and affective conditions (Kormos and Dénes 2004, Lennon 1990). As voice-based AI technology becomes increasingly accessible and scalable, a deeper understanding of its effects on speech production and assessment validity is essential for the responsible integration of AI systems into high-stakes language testing.

2.3 Fluency Measures in Speaking

Fluency is a multidimensional construct that plays a central role in oral language proficiency. Fillmore (1979) approached the notion of fluency by defining it in terms of the quality of oral production and emphasized that fluency exists on a spectrum among native speakers—from temporal fluency (as exhibited by disc jockeys) to coherent discourse (as seen in scholars), socio-pragmatic aptitude (as demonstrated by broadcast journalists), and linguistic creativity (reflecting wit and imagination). In the context of L2 fluency, fluency is typically defined from the listener's perspective. Scholars working from this viewpoint have investigated the specific linguistic features of speech that correlate with the impression of fluency. Lennon (1990, 2000) distinguished between lower-order, or temporal, fluency and higher-order fluency; the latter involves the rapid, smooth, accurate, lucid, and efficient expression of thoughts in language. He also emphasized that fluency can vary among speakers depending on factors such as topic, speech context, interlocutor, mental state, and other variables. Moreover, conversational fluency encompasses not only verbal but also nonverbal communication (Bavelas et al. 2000).

In this study, fluency is operationalized through temporal and dysfluency measures that capture various aspects of speech production. Temporal features include speech rate (syllables per second), mean length of utterance (syllables per pause), phonation-time ratio, and articulation rate (syllables during phonation). These indicators provide insight into the flow and density of speech and are sensitive to both cognitive load and interactional conditions (de Jong et al. 2021, Kormos and Dénes 2004). Dysfluency features, such as repairs per AS-unit, filled pauses, and preparation time, reflect breakdowns in planning and delivery. They serve as important diagnostic markers of a speaker's real-time processing efficiency, especially under different task or interlocutor constraints. The analytic framework adopted in this study draws on well-established fluency models (Kormos and Dénes 2004, Towell et al. 1996) and is designed to detect subtle differences in how examinees respond to chatbot versus human interlocutors.

By combining automated fluency measurements with rater-based scoring, this study contributes to a more nuanced understanding of how AI-mediated interactions shape observable speech features and their evaluation in formal oral proficiency assessments.

3. Method

3.1 Participants

A total of 128 audio recordings were collected from an online oral proficiency interview, with 32 Korean university students each responding to four different prompts (see Table 1 for details). Participants were recruited from two large universities in South Korea through course announcements, email invitations, and voluntary sign-up forms. All participants were enrolled in English-related or linguistics courses at the time of the study and provided informed consent prior to participation.

Table 1. Interview Prompts for the Study

Interlocutor	Topic	Style	Task Complexity
Chatbot	Q1-1. Sport	Descriptive	Low
	Q1-2. Parenting	Hypothetical	High
Human	Q2-1. Gift	Descriptive	Low
	Q2-2. Immigration	Hypothetical	High

Note. See Appendix A for the complete set of interview prompts.

To ensure a consistent and interpretable performance range, participants were required to have a TOEIC score between 600 and 900, which corresponds to B1 to C1 proficiency levels on the Common European Framework of Reference for Languages (CEFR). This range was selected to include learners with sufficient spoken English ability to engage meaningfully in both human- and chatbot-mediated interviews, while still representing a broad spectrum of intermediate to advanced proficiency.

The sample size of 32 participants was deemed sufficient for the many-facet Rasch measurement (MFRM) analysis employed in this study. According to Linacre (1994), MFRM can produce stable and reliable estimates with participant samples as small as 30 when the design is fully crossed and raters evaluate all examinees across all conditions, as was the case in this study. Moreover, each participant completed four speaking tasks, resulting in 128 individual speaking samples, which provided adequate data density for evaluating interlocutor and rater effects in the MFRM model.

3.2 Oral Proficiency Interview

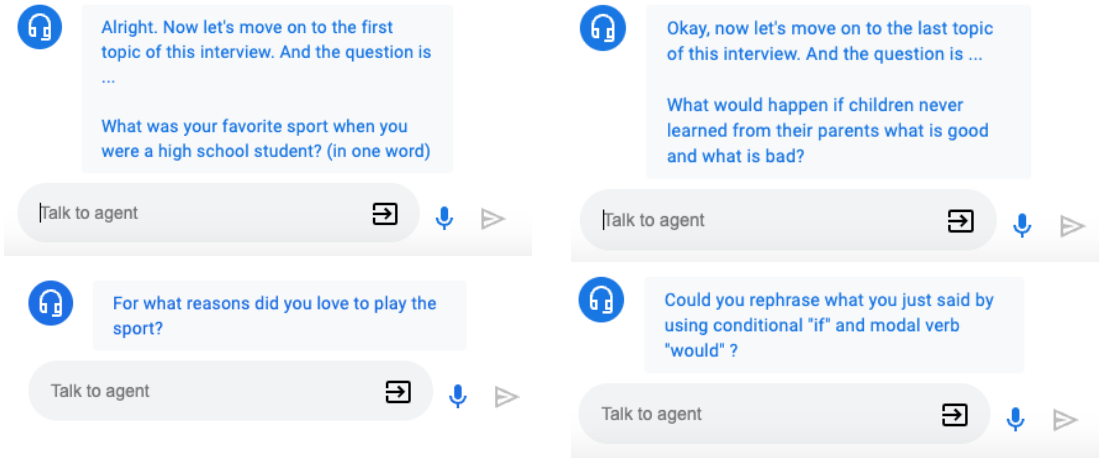
The oral proficiency interview in this study was designed to examine how interlocutor type affects learners' fluency and performance evaluation. Each participant completed two distinct interview segments: one administered by a human interlocutor and the other by a voice-based chatbot. Each segment featured two questions, yielding in a total of four tasks per participant.

While the primary focus of this study was to investigate the impact of interlocutor type (chatbot vs. human) on L2 learners' fluency and performance evaluation, task complexity was also included as a complementary factor to explore potential interaction effects. Task complexity was systematically manipulated following Robinson's (2001, 2011) Cognition Hypothesis and triadic componential framework, which classifies tasks based on cognitive demands, such as abstraction, reasoning, and hypothetical thinking. In this framework, tasks involving simple, personal descriptions (e.g., recounting past experiences) are categorized as low complexity, whereas those requiring the expression of hypothetical scenarios using conditional structures are considered high complexity due to their increased linguistic and cognitive demands. Including both task types enabled us to investigate whether the relationship between interlocutor type and fluency features remained consistent across varying levels of cognitive load, thereby providing a more nuanced understanding of how these two variables interact in shaping learners' spoken performance.

Interview topics were selected by two applied linguistics experts with extensive experience in language assessment and task design. Their selection goal was to ensure alignment with theoretical constructs of task complexity and the linguistic proficiency range of the target population. Questions Q1-1 (Sport) and Q2-1 (Gift), as described in Table 1, were designed to elicit descriptive, personal narratives and were categorized as low-complexity tasks. In contrast, Q1-2 (Parenting) and Q2-2 (Immigration) required participants to engage in hypothetical reasoning and use conditional structures, thereby meeting the criteria for high-complexity classification. The two tasks within each complexity level were assumed to be of comparable difficulty, based on

topic familiarity among Korean university students and consistency with CEFR B1–C1 speaking descriptors (e.g., B1: "Can give a prepared straightforward presentation on a familiar topic", and C1: "Can give a clear, well-structured presentation on a complex subject, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples") (Council of Europe 2020). A counterbalanced task design was employed so that each participant completed one low-complexity and one high-complexity task with both a chatbot and a human interlocutor, thereby minimizing potential order and content biases.

The chatbot interlocutor used in this study was developed using Google Dialogflow, a widely adopted natural language understanding platform. It was scripted to follow a fixed set of prompts and follow-up questions to maintain consistency across all participants. The chatbot employed a non-adaptive, rule-based structure to minimize variation in interlocutor behavior. Audio prompts were generated using NaturalReader (NaturalSoft Ltd. 2023), a text-to-speech engine that delivers clear, neutral speech. The chatbot interface was deployed via a web-based application and accessed through the online platform Zoom. A screenshot of a sample exchange between the chatbot and a participant is provided in Figure 1.



Low Complexity Topic "Sport" High Complexity Topic "Parenting"

Figure 1. Sample Interaction Between the Chatbot and a Participant

Each response was limited to two minutes, following established practices in oral assessment research that balance opportunities for elaboration with the need to maintain participant focus and minimize fatigue (Cotos 2014, Isbell and Winke 2019, Language Testing International 2018). This duration allows examinees to develop their ideas while keeping the sample concise and manageable for both participants and raters. If a participant's response was notably shorter than two minutes, both chatbot and human interlocutors provided follow-up prompts to elicit further elaboration—a strategy recommended in oral proficiency assessment guidelines to ensure sufficiently robust speech samples for analysis. Each session began with a one-minute warm-up conversation between the interlocutor and the examinee; this warm-up was not evaluated as part of the assessment.

3.3 Raters and Scoring Procedures

Two native English-speaking raters, both doctoral candidates in applied linguistics at a U.S. university, evaluated the speech samples. Each had formal training as an interlocutor and rater for Oral Proficiency Interviews

(OPIs), as well as over three years of experience teaching English as a Second Language (ESL) at the tertiary level in the United States. This combination of teaching and assessment expertise provided a strong foundation for evaluating L2 speech performance.

Examinees' responses were assessed solely from audio recordings using holistic and fluency criteria adapted from Cotos' (2014) rubric, originally developed to evaluate organization, fluency, lexico-grammar, and pronunciation. Both raters used a 13-point scale divided into four proficiency levels (see Appendix B). To ensure rating consistency, the audio clips were presented in a randomized order, and the sequence in which each rater evaluated the clips was also randomized to mitigate potential order effects. Prior to the main session, the raters completed a training phase using ten sample recordings to calibrate their use of the rating scales.

3.4 Fluency Measures

To analyze fluency features, each examinee's spoken responses were first transcribed using the Python *SpeechRecognition* library (Zhang 2017), then reviewed and corrected by trained human transcribers. Separate audio clips were created that included only the examinee's speech, excluding all interlocutor audio, to ensure clean data for acoustic analysis. As illustrated in Figure 2, each clip was cropped from the end of the examinee's response to the start of the next prompt or follow-up question. The cleaned audio files were then analyzed using a Praat script (de Jong et al. 2021) to extract core fluency features.

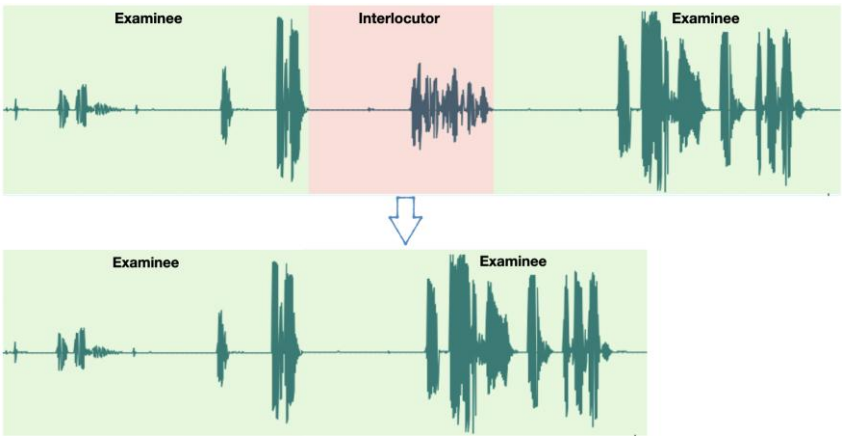


Figure 2. Editing to Extract Only the Examinees' Utterance

Table 2. Summary of Fluency Measures and Operational Definitions

Category	Fluency Measure	Operational Definition
Temporal Measures	Speech Rate	Syllables per second including pauses
	Articulation Rate	Syllables per second excluding pauses
	Phonation-Time Ratio	Proportion of speaking time to total duration
	Mean Length of Utterance	Average syllables between pauses ≥ 0.25 seconds
Dysfluency Markers	Repairs per AS-unit	Repetitions or reformulations divided by number of AS-units
	Filled Pauses per AS-unit	Number of "uh", "um", etc., per AS-unit
	Warm-up Preparation Time	Time from end of prompt to first filled pause (in seconds)
	Total Preparation Time	Time from end of prompt to first content-bearing response (in seconds)

Note. See Appendix C for a detailed summary of each fluency measure and its operational definition.

The fluency measures presented in Table 2 were selected to capture both temporal fluency (i.e., the speed and continuity of speech) and dysfluency features (i.e., breakdowns in speech production), which are widely recognized as core components of second language speaking ability (Kormos and Dénes 2004, Lennon 1990, Towell et al. 1996). These measures are commonly used in fluency research due to their relevance to listener judgments and their sensitivity to both individual and task-related variation (de Jong et al. 2021, Riggenbach 1991). Temporal measures—such as speech rate, articulation rate, phonation-time ratio, and mean length of utterance—provide insight into the rhythm, pacing, and density of spoken language. They reflect a speaker's ability to maintain a steady flow of speech under cognitive pressure (Jacewicz et al. 2009, Lennon 2000). In contrast, dysfluency markers—such as filled pauses, self-repairs, and preparation times—serve as indicators of planning difficulty and hesitation. These are particularly useful for assessing how fluency varies across interlocutor types and task complexities (Foster et al. 2000, Lambert et al. 2017). Preparation time was further divided into two components: warm-up preparation time (the interval before the first filled pause) and total preparation time (the interval before the first content-bearing response). Preparation time measures are illustrated in Figure 3. These metrics are commonly used in second language assessment research and were selected to support both automated and rater-based comparisons in the analysis of chatbot- versus human-mediated oral interviews.

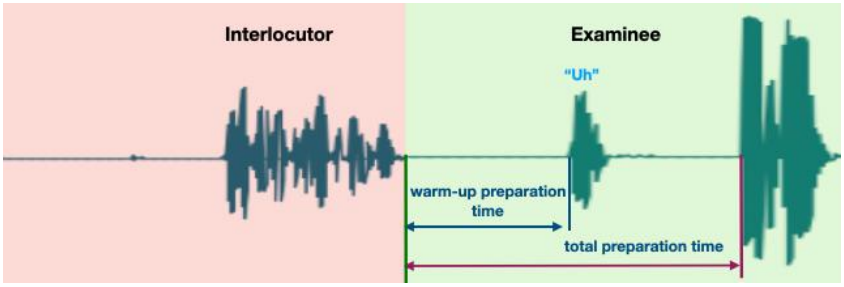


Figure 3. Warm-up Preparation Time and Total Preparation Time Measurements

3.5 Data Analysis

To investigate the effect of interlocutor type (chatbot vs. human) on examinees' fluency-related linguistic features (RQ1), Friedman tests were employed as omnibus non-parametric tests, followed by Wilcoxon signed-rank tests for post hoc pairwise comparisons. Non-parametric methods were selected not only because the fluency data violated normality assumptions, but also due to the specific distributional characteristics of the variables. For instance, speech rate and articulation rate are ratio-level measures bounded at zero and often positively skewed, while phonation-time ratio is a proportional variable constrained between 0 and 1, which introduces boundary effects that challenge parametric assumptions. Additionally, filled pauses per AS-unit and repairs per AS-unit are discrete, count-based indicators of dysfluency that are typically low in frequency and non-normally distributed. These characteristics make non-parametric methods more appropriate, as they do not require assumptions of normality, continuity, or homogeneity of variance. Given that the test involved four different topics, the significance level for the Wilcoxon signed-rank tests was adjusted using a Bonferroni correction ($\alpha = .05 / 6 = .008$) (Field 2009). To examine the effect of interlocutor type on examinees' fluency and holistic scores (RQ2), a rating scale model of MFRM was employed to compute fair scores. MFRM is an extension of the Rasch model that calibrates multiple facets—factors that influence assessment outcomes (Eckes 2015). By including raters, interview topics, and interlocutor type as facets, the difficulty levels of the chatbot and human interlocutor could be compared for both fluency and holistic scoring, while controlling for rater and topic effects. For both fluency

and holistic scores, a fully crossed (complete) MFRM design was used, in which each rater evaluated all examinees on all topics. This design was chosen to maximize parameter estimation accuracy and avoid missing data in the MFRM model (Eckes 2015). Fair average scores were estimated using FACETS version 3.80 (Linacre 2014).

4. Results

4.1 Impact of Interlocutor Type on Fluency Measures

This section examines how interlocutor type (chatbot vs. human) and task complexity (low vs. high) influenced fluency measures. The primary focus of the study was to investigate the effect of interlocutor type on fluency-related linguistic features in L2 speaking performance. In addition, task complexity was incorporated to determine whether the observed effects of interlocutor type remained consistent across varying cognitive demands or were moderated by task difficulty. The results revealed that interlocutor type significantly influenced certain aspects of speech initiation, particularly warm-up preparation time, while task complexity exerted a broader effect on both temporal fluency measures and dysfluency markers. Overall, these findings suggest that interlocutor type affected how quickly participants initiated speech, whereas the complexity of the speaking task had a greater influence on the fluency patterns sustained throughout the response.

4.1.1 Descriptive statistics

Table 3 presents descriptive statistics for fluency-related linguistic measures, organized by interlocutor type and task complexity. Among temporal measures, speech rate ranged from 0.49 to 2.93 syllables per second, with median values ranging from 1.32 (high-complexity *parenting* task) to 1.77 (low-complexity *gift* task). Mean length of utterance varied from 2.42 to 10.00 syllables, while the phonation-time ratio ranged from 0.45 to 0.57. Articulation rate spanned from 2.51 to 4.24 syllables per second across tasks. For dysfluency markers, repairs per AS-unit ranged from 0.00 to 4.00, and filled pauses per AS-unit ranged from 0.14 to 3.15. Warm-up preparation time varied from 0.52 to 30.55 seconds, and total preparation time ranged from 0.71 to 30.55 seconds.

Table 3. Descriptive Statistics of Fluency Measures across Topics (*N* = 32)

Fluency Measures		Interlocutor	Topic	Min	Max	Mdn	IQR
Temporal Measures	Speech Rate	Chatbot	Q1-1. Sport	0.49	2.43	1.62	0.57
			Q1-2. Parenting	0.58	2.04	1.32	0.57
		Human	Q2-1. Gift	0.91	2.93	1.77	0.67
			Q2-2. Immigration	0.56	2.68	1.58	1.06
	Mean Length of Utterance	Chatbot	Q1-1. Sport	2.42	10.00	4.23	1.99
			Q1-2. Parenting	2.48	5.63	4.08	1.53
		Human	Q2-1. Gift	2.42	8.70	4.25	1.53
			Q2-2. Immigration	2.52	8.29	4.17	2.37
	Phonation-Time Ratio	Chatbot	Q1-1. Sport	0.16	0.71	0.51	0.17
			Q1-2. Parenting	0.19	0.65	0.45	0.16
		Human	Q2-1. Gift	0.26	0.79	0.57	0.17
			Q2-2. Immigration	0.16	0.74	0.49	0.26
	Articulation Rate	Chatbot	Q1-1. Sport	2.64	3.86	3.21	0.52
			Q1-2. Parenting	2.51	2.92	3.19	0.51
		Human	Q2-1. Gift	2.58	3.89	3.23	0.40
			Q2-2. Immigration	2.53	4.24	3.30	0.61

Dysfluency Markers	Repairs per AS-unit	Chatbot	Q1-1. Sport	0.00	0.93	0.26	0.31
			Q1-2. Parenting	0.00	1.67	0.59	0.73
		Human	Q2-1. Gift	0.00	1.71	0.29	0.33
			Q2-2. Immigration	0.18	4.00	0.54	0.47
	Filled Pauses per AS-unit	Chatbot	Q1-1. Sport	0.19	1.57	0.64	0.51
			Q1-2. Parenting	0.44	2.18	1.06	0.58
		Human	Q2-1. Gift	0.14	1.53	0.86	0.59
			Q2-2. Immigration	0.35	3.15	1.11	0.52
	Warm-up Preparation Time (seconds)	Chatbot	Q1-1. Sport	0.90	12.89	2.93	2.33
			Q1-2. Parenting	1.10	30.55	5.96	6.93
		Human	Q2-1. Gift	0.52	7.63	1.42	0.82
			Q2-2. Immigration	0.57	29.77	1.61	0.85
	Total Preparation Time (seconds)	Chatbot	Q1-1. Sport	1.12	12.89	5.18	3.27
			Q1-2. Parenting	2.82	30.55	7.30	10.92
		Human	Q2-1. Gift	0.71	7.63	2.13	2.48
			Q2-2. Immigration	1.01	29.77	4.49	6.36

Note. Min: Minimum, Max: Maximum, Mdn: Median, IQR: Interquartile ranges

4.1.2 Statistical analysis of temporal fluency measures

To identify statistically significant differences, Friedman tests were conducted for each temporal measure across the four tasks. The results revealed significant omnibus effects for speech rate, phonation-time ratio, and articulation rate. Subsequently, pairwise Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = .05 / 6 = .008$) were performed to account for multiple comparisons (Field 2009). The results of these pairwise comparisons are summarized in Table 4.

Table 4. Wilcoxon Signed-rank Tests for Speech Rate, Phonation-Time Ratio, and Articulation Rate

Temporal Measures	Topic	Z [a]	p [b]	r [c]
Speech Rate	Q1-1. Sport vs. Q1-2. Parenting	–3.17	.002*	–0.56
	Q1-1. Sport vs. Q2-1. Gift	–3.19	.100	–0.56
	Q1-1. Sport vs. Q2-2. Immigration	–0.42	.674	–0.07
	Q2-1. Gift vs. Q1-2. Parenting	–4.94	< .001*	–0.87
	Q2-1. Gift vs. Q2-2. Immigration	–3.59	< .001*	–0.64
	Q1-2. Parenting vs. Q2-2. Immigration	–3.79	< .001*	–0.70
Phonation-Time Ratio	Q1-1. Sport vs. Q1-2. Parenting	–2.99	.003*	–0.53
	Q1-1. Sport vs. Q2-1. Gift	–3.35	.001*	–0.59
	Q1-1. Sport vs. Q2-2. Immigration	–0.13	.896	–0.02
	Q2-1. Gift vs. Q1-2. Parenting	–4.94	< .001*	–0.87
	Q2-1. Gift vs. Q2-2. Immigration	–4.62	< .001*	–0.82
	Q1-2. Parenting vs. Q2-2. Immigration	–3.48	.001*	–0.62
Articulation Rate	Q1-1. Sport vs. Q1-2. Parenting	–0.51	.610	–0.09
	Q1-1. Sport vs. Q2-1. Gift	–0.74	.460	–0.13
	Q1-1. Sport vs. Q2-2. Immigration	–1.98	.047	–0.35
	Q2-1. Gift vs. Q1-2. Parenting	–1.39	.164	–0.25
	Q2-1. Gift vs. Q2-2. Immigration	–1.54	.124	–0.27
	Q1-2. Parenting vs. Q2-2. Immigration	–2.70	.007*	–0.48

Note. [a] Wilcoxon signed-rank test; [b] Asympt. Sig (2-tailed): * $p < .008$; [c] Effect size: $r = .10$ is a small effect; $r = .30$ is a medium effect; and $r = .50$ is a large effect (Cohen 1992).

Pairwise comparisons for speech rate revealed significant differences between the low-complexity *Sport* task with a chatbot (Q1-1) and the high-complexity *Parenting* task with a chatbot (Q1-2); between the low-complexity *Gift* task with a human (Q2-1) and both the high-complexity *Parenting* task with a chatbot (Q1-2) and the *Immigration* task with a human (Q2-2); and between the two high-complexity tasks (Q1-1 and Q2-1). Notably, speech rate for the low-complexity tasks (Q1-1 and Q2-1) did not differ significantly by interlocutor type. However, the high-complexity *Parenting* task (Q1-2) showed significantly different speech rates compared to all other tasks, suggesting that task complexity, interlocutor type, or their interaction influenced this measure. For phonation-time ratio, significant differences were found between the low-complexity *Sport* task with a chatbot (Q1-1) and both high-complexity tasks (Q1-2 and Q2-2), and the low-complexity *Gift* task with a human (Q2-1) and both high-complexity tasks (Q1-2 and Q2-2). While the phonation-time ratio for the low-complexity *Sport* task with a chatbot (Q1-1) did not significantly differ from the high-complexity *Immigration* task with a human (Q2-2), it did differ from the other tasks. These results indicate that changes in phonation-time ratio cannot be solely attributed to interlocutor type. Articulation rate showed a significant difference only between the two high-complexity tasks: *Parenting* with a chatbot (Q1-2) and *Immigration* with a human (Q2-2). The articulation rate between the low-complexity *Sport* task with a chatbot (Q1-1) and the high-complexity *Immigration* task with a human (Q2-2) was not statistically significant after Bonferroni correction. Therefore, it is difficult to conclude that interlocutor type primarily influences articulation rate.

4.1.3 Statistical analysis of disfluency measures

The Friedman tests were conducted for all dysfluency measures, revealing significant differences across tasks for repairs per AS-unit, filled pauses per AS-unit, warm-up preparation time, and total preparation time. To further investigate these differences, post hoc Wilcoxon signed-rank tests were performed. The results of these comparisons are presented in Table 5.

Table 5. Wilcoxon Signed-rank Tests of Dysfluency Measures

Dysfluency Measures	Topic	Z [a]	p [b]	r [c]
Repairs per AS-unit	Q1-1. Sport vs. Q1-2. Parenting	–3.38	.001*	-0.60
	Q1-1. Sport vs. Q2-1. Gift	–1.15	.249	–0.20
	Q1-1. Sport vs. Q2-2. Immigration	–4.04	< .001*	–0.71
	Q2-1. Gift vs. Q1-2. Parenting	–2.46	.014	–0.44
	Q2-1. Gift vs. Q2-2. Immigration	–3.26	.001*	–0.58
	Q1-2. Parenting vs. Q2-2. Immigration	–0.89	.374	–0.16
Filled Pauses per AS-unit	Q1-1. Sport vs. Q1-2. Parenting	–4.34	< .001*	–0.77
	Q1-1. Sport vs. Q2-1. Gift	–1.87	.061	–0.33
	Q1-1. Sport vs. Q2-2. Immigration	–4.28	< .001*	–0.76
	Q2-1. Gift vs. Q1-2. Parenting	–3.48	.001*	–0.62
	Q2-1. Gift vs. Q2-2. Immigration	–3.76	< .001*	–0.66
	Q1-2. Parenting vs. Q2-2. Immigration	–0.60	.550	–0.11
Warm-up Preparation Time	Q1-1. Sport vs. Q1-2. Parenting	–4.75	< .001*	–0.84
	Q1-1. Sport vs. Q2-1. Gift	–4.47	< .001*	–0.79
	Q1-1. Sport vs. Q2-2. Immigration	–3.59	< .001*	–0.64
	Q2-1. Gift vs. Q1-2. Parenting	–4.92	< .001*	–0.87
	Q2-1. Gift vs. Q2-2. Immigration	–1.52	.130	–0.27
	Q1-2. Parenting vs. Q2-2. Immigration	–4.94	< .001*	–0.87

Total Preparation Time	Q1-1. Sport vs. Q1-2. Parenting	–4.32	< .001*	–0.76
	Q1-1. Sport vs. Q2-1. Gift	–4.32	< .001*	–0.76
	Q1-1. Sport vs. Q2-2. Immigration	–0.26	0.793	–0.05
	Q2-1. Gift vs. Q1-2. Parenting	–4.94	< .001*	–0.87
	Q2-1. Gift vs. Q2-2. Immigration	–4.13	< .001*	–0.73
	Q1-2. Parenting vs. Q2-2. Immigration	–3.40	.001*	–0.60

Note. [a] Wilcoxon signed-rank test; [b] Asympt. Sig (2-tailed): * $p < .008$; [c] Effect size: $r = .10$ is a small effect; $r = .30$ is a medium effect; and $r = .50$ is a large effect (Cohen 1992).

Pairwise comparisons for repairs per AS-unit revealed significant differences between the low-complexity tasks (*Sport* with a chatbot [Q1-1] and *Gift* with a human [Q2-1]) and the high-complexity tasks (*Parenting* with a chatbot [Q1-2] and *Immigration* with a human [Q2-2]). Specifically, repair frequency was significantly lower in the low-complexity tasks, suggesting that task complexity, rather than interlocutor type, was the primary factor influencing this dysfluency marker. Similarly, filled pauses per AS-unit were significantly more frequent in the high-complexity tasks than the low-complexity ones, again indicating a stronger influence of task complexity over interlocutor type. However, warm-up preparation time displayed a different pattern. Participants spent significantly less time preparing to speak when interacting with a human interlocutor (*Gift* [Q2-1] and *Immigration* [Q2-2]) compared to a chatbot (*Sport* [Q1-1] and *Parenting* [Q1-2]), suggesting that interlocutor type had a meaningful impact on the initial planning phase of speech production. Total preparation time also differed significantly across most task pairs. The high-complexity *Parenting* task with a chatbot (Q1-2) required significantly more preparation time than both the low-complexity *Sport* task with a chatbot (Q1-1) and the low-complexity *Gift* task with a human (Q2-1). Similarly, the high-complexity *Immigration* task with a human interlocutor (Q2-2) required more preparation time than the *Gift* task, again pointing to task complexity as the main determinant of overall preparation demands.

In summary, the analysis of temporal measures revealed that speech rate, phonation-time ratio, and articulation rate were influenced by task complexity and, to some extent, interlocutor type. Speech rate was generally slower for high-complexity tasks, regardless of interlocutor. Phonation-time ratio exhibited a more complex interaction between task and interlocutor variables. Articulation rate was primarily affected by the topic of the high-complexity tasks. Regarding dysfluency measures, repairs per AS-unit and filled pauses per AS-unit were more strongly associated with task complexity, with higher rates observed in more cognitively demanding tasks. In contrast, warm-up preparation time was significantly shorter with human interlocutors, indicating that interlocutor type influenced the initiation of speech. Total preparation time was mainly driven by the cognitive load of the task. Overall, these findings highlight the distinct—yet occasionally interacting—roles of interlocutor type and task complexity in shaping different dimensions of L2 speaking fluency.

4.2 Impact of Interlocutor Type on Fluency and Holistic Scores

This section examines the effect of interlocutor type (chatbot vs. human) on examinees' fluency and holistic performance scores using MFRM analysis. Overall, the results reveal that interlocutor type significantly influenced fluency scores, with human interlocutors associated with lower fluency performance. However, no substantial effect was found for holistic scores, suggesting that overall performance ratings remained stable across interaction modes.

4.2.1 Descriptive statistics

Table 6 presents descriptive statistics for examinees' fluency and holistic scores across task topics. For fluency scores, the mean ranged from 5.68 (*Parenting*) to 6.71 (*Gift*), with standard deviations ranging from 2.28 (*Gift*) to 2.65 (*Immigration*). For holistic scores, the mean ranged from 3.89 (*Parenting*) to 5.06 (*Gift*), with standard deviations between 2.46 (*Sport*) and 2.62 (*Immigration*). Skewness and kurtosis values ranged from –0.52 (kurtosis for fluency scores on *Immigration*) to 1.33 (skewness for holistic scores on *Sport*), all within the acceptable range of ±2.00 (Bachman 2004). These results suggest that the ratings were approximately normally distributed.

Table 6. Descriptive Statistics of Fluency and Holistic Scores across Topics (*N* = 32)

Scores	Interlocutor	Tasks	Min	Max	Mean	SD	Skewness	Kurtosis
Fluency	Chatbot	Q1-1. Sport	1.46	11.67	6.40	2.42	0.37	0.43
		Q1-2. Parenting	0.47	10.18	5.68	2.42	0.16	–0.38
	Human	Q2-1. Gift	3.04	12.62	6.71	2.28	0.63	0.39
		Q2-2. Immigration	1.43	11.77	6.01	2.65	0.14	–0.52
Holistic	Chatbot	Q1-1. Sport	1.40	10.59	4.08	2.46	1.33	0.97
		Q1-2. Parenting	1.40	9.88	3.89	2.51	1.32	0.48
	Human	Q2-1. Gift	1.82	11.59	5.06	2.48	1.02	0.71
		Q2-2. Immigration	1.43	10.58	4.45	2.62	0.76	–0.46

Note. Min: Minimum, Max: Maximum, SD: Standard Deviation.

4.2.2 Main effects of fluency scores by interlocutor type

Table 7 presents summary Rasch statistics from the rating scale model applied to fluency scores. The mean fluency score for examinees was 0.03 logits, with rater severity, task difficulty, and interlocutor difficulty centered at zero. The infit and outfit mean-square values for examinees, raters, tasks, and interlocutor type ranged from 0.93 to 0.96. These values fall within the acceptable range of 0.5 to 1.5 (Linacre 2014), indicating a good fit of the model to the data. For examinee measures, the separation ratio (G = 4.54) and strata index (H = 6.38) suggest that the assessment could statistically distinguish between more than five proficiency levels (Linacre 2014). In contrast, the separation index (G = 1.96) and strata (H = 2.95) for interlocutor type indicate a meaningful difference in difficulty between the chatbot and human interlocutor conditions. Within the MFRM framework, *difficulty* refers to the extent to which a given condition makes achieving a particular score more challenging. A higher difficulty level means that—after accounting for rater severity and task complexity—test-takers tend to receive lower scores under that condition.

Table 7. Rasch Measurement Summary Statistics of Fluency Scores

Statistic	Examinees	Raters	Topics	Interlocutor Type
Measures				
Mean (SE)	0.03 (0.27)	0.00 (0.07)	0.00 (0.09)	0.00 (0.07)
RMSE	0.27	0.07	0.07	0.07
Adjusted (True) SD	1.21	0.35	0.08	0.08
Infit MS				
Mean	0.93	0.95	0.95	0.95
Outfit MS				
Mean	0.96	0.96	0.96	0.96
Separation (G)	4.54	5.25	2.07	1.96
Strata (H)	6.38	7.34	3.10	2.95
Separation reliability	.95	.97	.81	.79

Note. Raters, tasks, and interlocutor type were centered at zero.

Table 8 presents the rating scale difficulty estimates (and standard errors) in centered logits for the two interlocutor types: a voice chatbot and a human interlocutor. The infit and outfit mean-square values, ranging from 0.84 (infit for the chatbot) to 1.07 (infit/outfit for the human), indicate a good fit to the measurement model. A fixed chi-square test (χ^2 = 4.8, *df* = 1, *p* = .03), significant at α = .05, suggests that the difficulty levels associated with the two interlocutor types are statistically different. The sample separation index (G = 1.96) and strata (H = 2.95), with a reliability of .79 (see Table 7), further support the presence of two distinct difficulty levels. Given that this reliability approaches the conventional .80 threshold, it can be concluded that—after controlling for rater severity and task complexity—interacting with a human interlocutor presents a significantly higher level of difficulty for achieving fluency scores than interacting with a chatbot.

Table 8. Interlocutor Type and Infit and Outfit Mean-squares in Fluency Scores

Interlocutor Type	Measures (logits)	SE	Infit Mean-square	Outfit Mean-square
Chatbot	–0.10	0.07	0.84	0.84
Human	0.10	0.07	1.07	1.07

4.2.3 Main effects of holistic scores by interlocutor type

Table 9 presents summary Rasch statistics from the rating scale model applied to holistic scores. The mean proficiency score for examinees was 1.11 logits, while rater severity, task difficulty, and interlocutor difficulty were each centered at zero. The infit and outfit mean-square values for examinees, raters, tasks, and interlocutor type ranged from 0.95 to 1.05, all within the acceptable range of 0.5 to 1.5 (Linacre 2014), indicating successful estimation. The examinee separation ratio (G = 4.43) and strata index (H = 6.24) suggest that the assessment could reliably distinguish among more than five proficiency levels (Linacre 2014). In contrast, the interlocutor type separation ratio (G = 0.14) and strata index (H = 0.52) indicate no meaningful difference in difficulty between the two interlocutor conditions.

Table 9. Rasch Measurement Summary Statistics of Holistic Scores

Statistic	Examinees	Raters	Topics	Interlocutor Type
Measures				
Mean (SE)	1.11 (0.29)	0.00 (0.07)	0.00 (0.10)	0.00 (0.07)
RMSE	0.29	0.07	0.10	0.07
Adjusted (True) SD	1.28	0.84	0.11	0.00
Infit MS				
Mean	0.95	1.05	1.00	1.00
Outfit MS				
Mean	0.96	0.96	0.96	0.96
Separation (G)	4.43	11.62	1.13	0.14
Strata (H)	6.24	15.83	1.83	0.52
Separation reliability	.95	.99	.56	.02

Note. Raters, tasks, and interlocutor type were centered at zero.

Table 10 presents the rating scale difficulty estimates (and standard errors) in centered logits for holistic scores across the two interlocutor types: a voice chatbot and a human interlocutor. The infit and outfit mean-square values, ranging from 0.94 (outfit for the chatbot) to 1.05 (infit for the human), fall within the acceptable range and align with expectations of the measurement model. A non-significant fixed chi-square test (χ^2 = 1.00, *df* = 1, *p* = .31), failed to reject the null hypothesis, indicating no statistically significant difference in difficulty between the two

interlocutor types. Additionally, the low separation index (G = 0.14) and strata (H = 0.52), with a reliability of .02 (see Table 9), indicate that the chatbot and human interlocutor did not differ meaningfully in difficulty once rater severity and task complexity were accounted for.

Table 10. Interlocutor Type and Infit and Outfit Mean-squares in Holistic Scores

Interlocutor Type	Measures (logits)	SE	Infit Mean-square	Outfit Mean-square
Chatbot	–0.05	0.07	0.95	0.94
Human	0.05	0.07	1.05	0.98

In summary, the MFRM analysis of fluency scores reveals two distinct difficulty levels associated with the interlocutor types: human and chatbot. Conversely, the MFRM analysis of holistic scores indicates that the two interlocutor types posed comparable levels of difficulty. These findings suggest that examinees' fluency scores were influenced by interlocutor type, whereas their holistic scores remained largely stable across interaction modes.

5. Discussion

5.1 Effects of Interlocutor Type on Fluency Measures

Ockey and Li (2015) proposed an assessment framework for oral communication in which interlocutors' personal characteristics are considered one of several key elements—alongside raters, rating criteria, technological aspects, and examinees' oral communication abilities—that jointly influence test scores. According to this framework, these components can affect both the scores assigned and examinee performance, which should be reflected in the linguistic features of examinees' output. In the present study, among the temporal measures examined, speech rate and articulation rate were influenced primarily by task complexity, whereas phonation-time ratio was affected by both task complexity and interlocutor type. Variations in speech rate and articulation rate, both of which depend on the quantity of syllables produced within a given time frame, may arise from differences in cognitive load (Mora et al. 2024), thereby impacting examinees' temporal speech patterns during the limited speaking period. However, this finding contrasts with findings from Won (2020), who reported that task complexity influenced phonation-time ratio but not speech rate or articulation rate among international students. This discrepancy may stem from differences in examinee populations, task design, or interlocutor behavior across studies.

The observed variation in phonation-time ratio by task complexity likely reflects examinees' additional planning time in response to cognitively demanding tasks. More complex tasks appear to elicit longer silent or preparatory periods before initiating speech, aligning with the view that higher cognitive load increases pre-speech planning. Moreover, the influence of interlocutor type on phonation-time ratio suggests that the nature of the conversational partner (human vs. chatbot) shapes not only content production but also temporal fluency patterns. This supports prior findings that AI-mediated conversations evoke different cognitive and affective responses compared to human interaction (Azizimajd 2023, Han 2020, Wang et al. 2024).

Regarding dysfluency measures, repairs per AS-unit, filled pauses per AS-unit, and total preparation time were significantly influenced by task complexity, while warm-up preparation time was affected by interlocutor type. Although task complexity had a stronger overall effect, the sensitivity of warm-up preparation time to interlocutor type deserves special attention. Examinees tended to produce more initial filled pauses (e.g., "uh," "um") before

substantive responses when interacting with human interlocutors—a phenomenon less prominent during chatbot interactions. This pattern reflects the communicative role of filled pauses in natural conversations, especially when visual social cues such as eye contact are present (Kosmala 2022). Thus, the shorter warm-up preparation times observed with human interlocutors may reflect examinees' readiness to engage in socially contingent interaction, where visual cues and turn-taking expectations prompt earlier speech initiation. In contrast, the longer preparation times with chatbots likely stem from the absence of dynamic social feedback, reducing pressure to respond promptly and allowing more time for internal planning. Additionally, because filled pauses are typically counted as a single syllable, they may inflate the phonation-time ratio more than other temporal measures such as speech rate or articulation rate—potentially explaining the observed interaction between phonation-time ratio and interlocutor type.

These results support the view that spoken fluency is co-constructed by both participants in a conversation (Lazaraton 1996, Peltonen 2022, Sato 2014, Wilson and Wilson 2005). Human interlocutors often provide subtle social cues, such as backchannels and gaze, that facilitate smoother speech onset and turn-taking. The "chameleon effect" in social psychology illustrates how individuals naturally mimic their interlocutor's gestures and speech patterns, fostering smoother conversational exchanges (Chartrand and Bargh 1999). Empirical studies have demonstrated that interpersonal mimicry between interlocutors facilitates the smoothness of interactions and enhances the establishment of rapport (Chartrand and Bargh 1999, Kennedy et al. 2024, Lakin et al. 2003). In contrast, the static, less adaptive nature of AI chatbots, characterized by limited paralinguistic feedback, may introduce interactional uncertainty at speech initiation, resulting in longer pauses and fragmented speech. This interpretation is supported by earlier studies indicating that while chatbots can elicit ratable speech samples, they often lack the interactional naturalness that characterizes human-to-human communication (García Laborda et al. 2024, Hill et al. 2015, Huang et al. 2022, Wu et al. 2025).

Overall, the results suggest that speaking fluency is shaped not only by cognitive load and task design but also by the interactive affordances and social dynamics created by different types of interlocutors. As such, AI-mediated speaking assessments must account for potential interlocutor effects on both temporal fluency and dysfluency dimensions to ensure valid interpretation of examinee performance.

5.2 Effects of Interlocutor Type on Fluency and Holistic Scoring

This study found that interlocutor type significantly influenced fluency scores, with examinees producing higher fluency scores when interacting with a chatbot compared to a human interlocutor. A plausible explanation, supported by previous research, is that interacting with a chatbot reduces social pressure, cognitive load, and anxiety, thereby facilitating smoother and more fluent speech production (Fryer and Carpenter 2006, Hsu et al. 2023, Jeon and Lee 2024). This finding aligns with studies suggesting that chatbot-mediated interactions can create a lower-stress environment, encouraging more fluid and rapid language production (Azizimajd 2023, Han 2020, Wang et al. 2024).

Additionally, as Ockey and Chukharev-Hudilainen (2021) pointed out, the use of a standardized computer interlocutor may provide a more uniform testing environment, reducing variability caused by interlocutor differences. In contrast, face-to-face interactions often impose additional demands on speakers, such as managing social cues, turn-taking, and interpreting non-verbal signals (Clark 2002, McNamara and Lumley 1997), which may inhibit spontaneous language production. These dynamic interactional demands may increase cognitive burden, thereby reducing fluency during human-mediated interviews. Furthermore, the predictable and structured nature of chatbot prompts, without unexpected conversational moves, may make it easier for test-takers to

formulate responses quickly compared to the more varied demands typical of human interlocutors (Ockey and Chukharev-Hudilainen 2021). Such task predictability likely reduces planning demands and allows examinees to maintain speech continuity without needing to adjust to shifting conversational dynamics (García Laborda et al. 2024, Hill et al. 2015).

The findings of this study partially support the notion that the nature of the interlocutor can influence L2 speakers' performance, as evidenced by statistically significant differences in fluency measures across varying task complexities and interlocutor types. Specifically, fluency scores suggest that examinees performed more favorably when interacting with a chatbot than with a human interlocutor. This pattern aligns with recent research indicating that technological mediation can buffer speakers against real-time anxiety triggers that often disrupt fluency (Huang et al. 2022, Wu et al. 2025). However, these results contrast with those of Ockey and Chukharev-Hudilainen (2021), who found higher ratings for interactional competence in interviews with human interlocutors—suggesting that while human interlocutors may better elicit complex interactive behaviors, they do not necessarily promote more fluent speech. Notably, their studies emphasized interactional competence rather than temporal fluency, which may account for the differing outcomes. Taken together, this divergence highlights that distinct aspects of speaking proficiency (e.g., fluency vs. interactional competence) may be differentially affected by interlocutor type.

In contrast, when considering holistic scores, the results indicate no statistically significant effect of interlocutor type. This absence of statistical significance may reflect minimal or negligible differences in overall linguistic output attributable to interlocutor type, and/or the possibility that raters subconsciously adjust their evaluations to compensate for fluency variations depending on the interlocutor. Such rater adaptations are plausible given the complex, integrative nature of holistic scoring, which often balances multiple linguistic features beyond surface fluency alone. In this study, the latter explanation appears more plausible, given the observed significant differences in phonation-time ratio and warm-up preparation time across interlocutor types. Thus, it appears that interlocutor type directly affects specific fluency features and fluency scores, while raters may make adjustments during holistic scoring to account for these variations. This interpretation aligns with previous findings (e.g., Won and Kim 2023) on lexico-grammatical rating adjustments, where raters were found to integrate multiple linguistic signals and modify their severity to maintain perceived fairness. Such rater flexibility suggests that human raters are sensitive to contextual factors and strive to preserve score validity across different interactional conditions. Because fluency scoring focuses narrowly on measurable features (e.g., speech rate, preparation time), whereas holistic scoring considers broader communicative competence, raters may be more likely to consciously or unconsciously compensate for interlocutor-driven differences in the holistic rating process.

In sum, the findings of this study indicate that for holistic scores, there is no significant difference in perceived task difficulty between chatbot and human interlocutors. Consequently, oral proficiency interviews using chatbot interlocutors could serve as a viable alternative to human-led interviews for holistic proficiency assessment. Nevertheless, this does not imply that interlocutor type has no influence on examinee performance overall, as evidenced by differences in fluency features and fluency scores. The evidence also cautions that chatbot use may selectively favor certain speaking dimensions (e.g., fluency) while underserving others (e.g., interactional competence), depending on assessment goals (Ockey et al. 2023). Moreover, the differences observed suggest important considerations for test fairness and validity. While chatbots offer advantages in standardization and practical scalability, they may also limit opportunities for natural co-constructed interaction, which is critical if interactional competence is an intended target of assessment. Factors such as assessment criteria and rater training could also influence holistic scoring outcomes, particularly when varying interlocutor characteristics are involved, as emphasized in prior research (Chartrand and Bargh 1999, Clark 2002, Lazaraton 1996, McNamara and Lumley

1997, Morton et al. 1997, Song 2017, Wilson and Wilson 2005). Therefore, careful design of assessment frameworks and rater calibration procedures is essential to ensure that chatbot-mediated interviews uphold standards of validity, reliability, and fairness.

6. Conclusion and Implications

This study investigated oral communication assessments featuring both a chatbot and a human interlocutor, with particular emphasis on temporal and dysfluency measures. The findings revealed notable differences in fluency measures across various interview scenarios. In terms of fluency scoring, distinct proficiency levels were observed among examinees when interacting with different interlocutor types, indicating varying levels of perceived difficulty depending on whether the interlocutor was human or AI-based. However, no significant differences emerged in holistic scores, suggesting that while interlocutor type significantly affected specific fluency-related features, it did not broadly influence raters' overall judgments of communicative competence. These results highlight the differing impacts of interlocutor type on fluency versus holistic scoring processes.

From a theoretical perspective, this study provides evidence that interlocutor type can influence raters' interpretations of rating scales in oral proficiency interviews. This outcome supports Ockey and Li's (2015) oral communication assessment model, which posits that interlocutor characteristics may interact with raters, potentially altering their rating severity for speaking performances. Such an interaction suggests that raters may apply rating scales differently depending on the nature of the interlocutor, particularly when assessing narrowly defined linguistic features such as fluency.

From a practical standpoint, the findings underscore both the potential and limitations of substituting human interlocutors with chatbot interlocutors—particularly in specific assessment contexts where fluency is the primary focus, such as low-stakes speaking practice tests, automated placement tests, or formative assessments emphasizing speed and fluency. In these contexts, chatbot-mediated interviews may offer reliable, scalable alternatives to human-led interviews without significantly compromising assessment validity. However, in high-stakes holistic speaking tests—where broader dimensions of communicative competence, including interactional competence and pragmatic appropriateness, are critical—greater caution is warranted when relying solely on chatbots.

Based on these findings, several practical recommendations emerge: chatbot interactions should be designed to more closely approximate human conversational dynamics (e.g., through backchanneling or varied prompting strategies) if the test construct assumes human interaction; raters should receive targeted training to recognize and adjust for performance differences stemming from interlocutor type, especially if human raters are involved while AI-chatbots serve as interlocutors; and hybrid assessment models, incorporating both chatbot- and human-mediated components, may be appropriate for high-stakes speaking tests to balance efficiency with authenticity.

To deepen understanding of these issues, future research should examine how chatbot and human interlocutors differentially influence a broader range of linguistic constructs beyond fluency, including interactional and pragmatic skills. Future studies should also include more diverse participant populations and employ advanced statistical models capable of controlling for a wide variety of extraneous variables, thereby producing more generalizable insights into the effects of interlocutor type on oral communication performance. As automated speaking tests continue to expand, careful attention to interlocutor effects will be essential to ensuring valid, fair, and effective evaluations of learners' communicative abilities.

References

Abida, F. I. N., R. Kuswardani, O. Purwati, A. Rosyid and E. Minarti. 2023. Assessing language proficiency through AI chatbot-based evaluations. *Proceedings of International Conference on Islamic Civilization and Humanities* 1, 138-145.

Adamopoulou, E. and L. Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications* 2, 100006.

Ayedoun, E., Y. Hayashi and K. Seta. 2015. A conversational agent to encourage willingness to communicate in the context of English as a foreign language. *Procedia Computer Science* 60, 1433-1442.

Azizimajd, H. 2023. Investigating the impacts of voice-based student-chatbot interactions in the classroom on EFL learners' oral fluency and foreign language speaking anxiety. *Technology Assisted Language Education* 1(2), 61-83.

Bachman, L. F. 2004. *Statistical Analyses for Language Assessment*, Cambridge University Press.

Bavelas, J. B., L. Coates and T. Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology* 79(6), 941-952.

British Council. 2023. *Interlocutor*. Available online at https://www.teachingenglish.org.uk/professional-development/teachers/knowing-subject/d-h/interlocutor.

Brown, A. 2003. Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20(1), 1-25.

Brown, A. 2012. Interlocutor and rater training. In G. Fulcher and F. Davidson, eds., *The Routledge Handbook of Language Testing*, 413-425. Routledge.

Chartrand, T. L. and J. A. Bargh. 1999. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76(6), 893-910.

Chichon, J. 2019. Factors influencing overseas learners' Willingness to Communicate (WTC) on a pre-sessional programme at a UK university. *Journal of English for Academic Purposes* 39, 87-96.

Clark, H. H. 2002. Speaking in time. *Speech Communication* 36(1-2), 5-13.

Cohen, J. 1992. A power primer. *Psychological Bulletin* 112(1), 155-159.

Coniam, D. 2014. The linguistic accuracy of chatbots: Usability from an ESL perspective. *Text & Talk* 34(5), 545-567.

Cotos, E. 2014. *Oral English Certification Test (OECT): Rater Manual,* Iowa State University.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Companion Volume)*. Available online at https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Davis, L. 2009. The influence of interlocutor proficiency in a paired oral assessment. *Language Testing* 26(3), 367-396.

de Jong, N. H., J. Pacilly and W. Heeren. 2021. PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice* 28(4), 456-476.

Eckes, T. 2015. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments,* Peter Lang GmbH.

Field, A. 2009. *Discovering Statistics using SPSS,* Sage Publications.

Fillmore, C. J. 1979. On fluency. In C. J. Fillmore, D. Kempler and W. S. Y. Wang, eds., *Individual Differences in Language Ability and Language Behavior*, 85-101. Academic Press.

Foster, P., A. Tonkyn and G. Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied*

Linguistics 21(3), 354-375.

Fryer, L. and R. Carpenter. 2006. Bots as language learning tools. *Language Learning & Technology* 10(3), 8-14.

Fryer, L. K., M. Ainley, A. Thompson, A. Gibson and Z. Sherlock. 2017. Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. *Computers in Human Behavior* 75, 461-468.

Fryer, L., D. Coniam, R. Carpenter and D. Lăpușneanu. 2020. Bots for language learning now: Current and future directions. *Language Learning & Technology* 24(2), 8-22.

Fulcher, G. 2003. *Testing Second Language Speaking,* Pearson.

García Laborda, J., S. Madarova and T. Magal Royo. 2024. Issues in the design and implementation of chatbots for oral language assessment. *Journal of Research in Applied Linguistics* 15(2), 43-54.

Han, D.-E. 2020. The effects of voice-based AI chatbots on Korean EFL middle school students' speaking competence and affective domains. *Asia-pacific Journal of Convergent Research Interchange* 6(7), 71-80.

Hill, J., W. R. Ford and I. G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49, 245-250.

Hou, Y.-C. 2006. A cross-cultural study of the perception of apology: Effect of contextual factors, exposure to the target language, interlocutor ethnicity and task language. Unpublished master's thesis, National Sun Yat-sen University, Taiwan.

Hsu, M.-H., C. Pei-Shih, and C.-S. Yu. 2023. Proposing a task-oriented chatbot system for EFL learners speaking practice. *Interactive Learning Environments* 31(7), 4297-4308.

Huang, W., K. F. Hew and L. K. Fryer. 2022. Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning* 38(1), 237-257.

Isbell, D. and P. Winke. 2019. ACTFL Oral Proficiency Interview–computer (OPIc). *Language Testing* 36(3), 467-477.

Jacewicz, E., R. A. Fox, C. O'Neill and J. Salmons. 2009. Articulation rate across dialect, age, and gender. *Language Variation and Change* 21(2), 233-256.

Jeon, J. and S. Lee. 2024. The impact of a chatbot-assisted flipped approach on EFL learner interaction. *Educational Technology & Society* 27(4), 218-234.

Kennedy, O., N. Kuwahara, T. Noble and C. Fukada. 2024. The effects of teacher nodding: Exploring mimicry, engagement, and wellbeing in the EFL classroom. *Frontiers in Education*, 9.

Kim, H.-S., Y. Cha and N. Y. Kim. 2021. Effects of AI chatbots on EFL students' communication skills. *Korean Journal of English Language and Linguistics* 21, 712-734.

Kim, H., D. K. Shin, H. Yang and J. H. Lee. 2019. A study of AI chatbot as an assistant tool for school English curriculum. *Journal of Learner-Centered Curriculum and Instruction* 19(1), 89-110.

Kim, Y. 2020. Analysis of chatbots and chatbot builders for English language learning. *Multimedia-Assisted Language Learning* 23(4), 161-182.

Kohnke, L., B. L. Moorhouse and D. Zou. 2023. ChatGPT for language teaching and learning. *RELC Journal* 54(2), 537-550.

Kormos, J. and M. Dénes. 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System* 32(2), 145-164.

Kosmala, L. 2022. Exploring the status of filled pauses as pragmatic markers: The role of gaze and gesture. *Pragmatics & Cognition* 29(2), 272-296.

Lakin, J. L., V. E. Jefferis, C. M. Cheng and T. L. Chartrand. 2003. The chameleon effect as social glue: Evidence

for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior* 27(3), 145-162.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition* 39(1), 167–196.

Language Testing International. 2018. *ACTFL OPIc Examinee Handbook.* Available online at https://www.languagetesting.com/pub/media/wysiwyg/PDF/opic-examinee-handbook.pdf

Lazaraton, A. 1996. Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing* 13(2), 151-172.

Lennon, P. 1990. Investigating fluency in EFL: A quantitative approach. *Language Learning* 40(3), 387-417.

Lennon, P. 2000. The lexical element in spoken second language fluency. In H. Riggenbach, ed., *Perspectives on Fluency*, 25-42. University of Michigan Press.

Linacre, J. M. 1994. Sample size and item calibration stability. *Rasch Measurement Transactions* 7(4), 328.

Linacre, J. M. 2014. *A User's Guide to FACETS (Version 3.80)* [Computer software]. Available online at https://www.winsteps.com/a/Facets-ManualPDF.zip

Liu, X. J., J. Wang and B. Zou. 2025. Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback as feedback enhancement. *Journal of English for Academic Purposes* 75, 101505.

Mauldin, M. L. 1994. ChatterBots, TinyMuds, and the Turing Test: Entering the Loebner Prize competition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 16-21.

May, L. 2011. Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly* 8(2), 127-145.

McNamara, T. F. 1996. *Measuring Second Language Performance,* Longman.

McNamara, T. F. and T. Lumley. 1997. The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing* 14(2), 140-156.

Mora, J. C., I. Mora-Plaza and G. Bermejo Miranda. 2024. Speaking anxiety and task complexity effects on second language speech. *International Journal of Applied Linguistics* 34(1), 292-315.

Morton, J., G. Wigglesworth and D. Wiliams. 1997. Approaches to the evaluation of the interviewer performance in oral interaction tests. In G. Brindley and G. Wigglesworth, eds., *Access: Issues in English Language Test Design and Delivery*, 175-196. National Centre for English Language Teaching and Research.

NaturalSoft Ltd. 2023. *NaturalReader* [Computer software]. Available online at https://www.naturalreaders.com/index.html

Norris, J. M. and L. Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4), 555-578.

Ockey, G. J. and E. Chukharev-Hudilainen. 2021. Human versus computer partner in the paired oral discussion test. *Applied Linguistics* 42(5), 924-944.

Ockey, G. J. and Z. Li. 2015. New and not so new methods for assessing oral communication. *Language Value* 7(1), 1-21.

Ockey, G. J., E. Chukharev-Hudilainen and R. R. Hirch. 2023. Assessing interactional competence: ICE versus a human partner. *Language Assessment Quarterly* 20(4-5), 377-398.

Peltonen, P. 2022. Connections between measured and assessed fluency in L2 peer interaction: A problem-solving perspective. *International Review of Applied Linguistics in Language Teaching* 60(4), 983-1011.

Plough, I. C., S. L. Briggs and S. Van Bonn. 2010. A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing* 27(2), 235-260.

Riggenbach, H. 1991. Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes* 14(4), 423-441.

Robinson, P. 2001. Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson, ed., *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*, 287-318. Cambridge University Press.

Robinson, P. 2011. Task-based language learning: A review of issues. *Language Learning* 61(s1), 1-36.

Ross, S. and R. Berwick. 1992. The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14(2), 159-176.

Sato, M. 2014. Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System* 45, 79-91.

Song, M.-Y. 2017. Nonnative raters' perceptions and judgments of Korean English learners' fluency and pronunciation level. *Korean Journal of English Language and Linguistics* 17(4), 787-815.

Sung, M.-C. 2019. Development of a flowchart-based English-speaking chatbot for Korean primary students' negotiation of meaning. *Primary English Education* 25(4), 101-122.

Tai, T.-Y. and H. H.-J. Chen. 2022. The impact of intelligent personal assistants on adolescent EFL learners' listening comprehension. *Computer Assisted Language Learning* 37(3), 1-28.

Timpe-Laughlin, V., T. Sydorenko. and P. Daurio. 2022. Using spoken dialogue technology for L2 speaking practice: What do teachers think? *Computer Assisted Language Learning* 35(5-6), 1194-1217.

Towell, R. 2002. Relative degrees of fluency: A comparative case study of advanced learners of French. *International Review of Applied Linguistics in Language Teaching* 40(2), 117-150.

Towell, R., R. Hawkins and N. Bazergui. 1996. The development of fluency in advanced learners of French. *Applied Linguistics* 17(1), 84-119.

Van Moere, A. 2013. Raters and ratings. In A. J. Kunnan, ed., *The Companion to Language Assessment*, 1358-1374. John Wiley & Sons, Inc.

Wang, C., B. Zou, Y. Du and Z. Wang. 2024. The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. *System* 127, 103533.

Wilson, M. and T. P. Wilson. 2005. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review* 12(6), 957-968.

Wollny, S., J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger and H. Drachsler. 2021. Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence* 4, 654924.

Won, Y. 2020. The effect of task complexity on test-takers' performance in a performance-based L2 oral communication test for international teaching assistants. *Journal of the Korea English Education Society* 19(1), 27-52.

Won, Y., and S. Kim. 2023. The impact of topic selection on lexico-grammatical errors and scores in English oral proficiency interviews of Korean college students. *Education Sciences* 13(7), 695.

Wu, T.-T., I. P. Hapsari and Y.-M. Huang. 2025. Effects of incorporating AI chatbots into think–pair–share activities on EFL speaking anxiety, language enjoyment, and speaking performance. *Computer Assisted Language Learning*, 1-39.

Xu, Y., D. Wang, P. Collins, H. Lee and M. Warschauer. 2021. Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. *Computers & Education* 161, 104059.

Yang, H., H. Kim, J. H. Lee and D. Shin. 2022. Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL* 34(3), 327-343.

Yang, J. 2022. Perceptions of preservice teachers on AI chatbots in English education. *International Journal of*

Internet, Broadcasting and Communication 14(1), 44-52.

Zhang, A. 2017. *Speech Recognition (Version 3.8)* [Computer software]. Available online at https://github.com/Uberi/speech_recognition#readme.

Zhang, M., L. Yao, S. J. Haberman and N. J. Dorans. 2020. Assessing scoring accuracy and assessment accuracy for spoken responses using human and machine scores. In K. Zechner and K. Evanini, eds., *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*, 32-58. Routledge.

Appendix A. Interview Prompts for the Study

Interlocutors	Topics (Style; Complexity)	Prompts
Chatbot	Q1-1. Sport (Descriptive; Low Complexity)	What was your favorite sport when you were a high school student? For what reasons did you love to play the sport? Please rephrase what you just said by using past tense verbs, such as "loved", "played", and "went".
	Q1-2. Parenting (Hypothetical; High Complexity)	What would happen if children never learned from their parents what is good and what is bad? Could you rephrase what you just said by using conditional "if" and modal verb "would"? Please rephrase what you have said by using one of the "if conditionals" below: A: if the weather improved, we could go for a walk. (It is not likely that the weather will improve.) B: If the weather had improved, we could have gone for a walk. (The weather did not improve—fine weather is therefore an impossible condition)
Human	Q2-1. Gift (Descriptive; Low Complexity)	Describe the best gift you have ever received.
	Q2-2. Immigration (Hypothetical; High Complexity)	What would happen if people were allowed to immigrate freely to any country?

Appendix B. Scoring Rubric

Table B.1 Holistic Scoring Rubric

Level	Score	Description
Excellent	13	Communication is like that of an educated N-American native speaker; always fluent & effective; no effort needed to understand.
Very Strong	11-12	Communication is fairly close to that of an educated N-American native speaker; always fluent & effective; little effort needed to understand.
Strong	9-10	Communication is generally effective; fluent most of the time; performance slightly weakens with more complicated topics and tasks; little effort needed to understand.
Adequate	7-8	Communication is fairly effective; fluent most of the time; cannot sustain performance with more complicated topics and tasks; is able to compensate for the limited aspects of communication; some effort needed to understand.
Limited	5-6	Communication somewhat effective; speaker expresses ideas freely, but has problems that impede communication; more effort needed to understand.
Very Limited	3-4	Some communication takes place, but speaker struggles to express ideas and/or has significant communication problems that make it difficult for listener to understand.
Poor	1-2	Communication generally ineffective; multiple significant problems exhibited; much effort needed to understand.
Not competent	0	Totally ineffective communication; listener can only catch a few words.

Table B.2 Fluency Scoring Rubric

Level	Score	Description
Excellent	13	Native-like delivery
Very Strong	11-12	Effective pace; smooth delivery & fluency, with good use of English rhythm & focal stress to highlight meaning.
Strong	9-10	Delivery at a fair pace; fluency generally smooth & with good rhythm.
Adequate	7-8	Delivery at a fair pace; fluency often smooth & with good rhythm, but occasionally choppy or too even.
Limited	5-6	Delivery may be overly slow or fast; fluency with choppy flow, or very even rhythm, or word by word at times, or inappropriate intonation patterns.
Very Limited	3-4	Pace is often slow because of halting fluency with hesitations and/or repetitions.
Poor	1-2	Inappropriate pace may significantly interfere; broken fluency because of pauses, hesitations, & false starts.
Not competent	0	Not fluent; pace of delivery severely interferes.

Appendix C. Summary of Temporal and Dysfluency Measures Used in the Study

Category	Fluency Measure	Operational Definition
Temporal Measures	Speech Rate	• Speech rate is calculated by dividing the total number of syllables by the total duration of the utterance (Riggenbach 1991). *Number of Syllables / Total Utterance Duration (in seconds)*
	Articulation Rate	• Articulation rate is calculated by dividing the number of syllables by the actual speaking time, excluding pauses. It reflects the speed of producing speech segments and excludes hesitations and emotional expressions, unlike speech rate, which includes them (Jacewicz et al. 2009). *Number of Syllables / Speaking Time (in seconds)*
	Phonation-Time Ratio	• Phonation-time ratio represents the proportion of actual speaking time relative to the total time allotted for the utterance (Towell 2002). *Speaking Time / Total Utterance Duration (in seconds)*
	Mean Length of Utterance	• Mean length of utterance is determined by averaging the number of syllables between pauses of at least 0.25 seconds, which are considered reliable indicators of speech run boundaries (Towell et al. 1996). *Number of Syllables / Number of Utterances*
Dysfluency Markers	Repairs per AS-unit	• Repairs per AS-unit is calculated by dividing the number of self-repairs, such as repetitions or reformulations, by the number of AS-units. *Number of Repairs / Number of AS-unit*
	Filled Pauses per AS-unit	• Filled pauses per AS-unit is measured by dividing the number of filled pauses (e.g., "uh," "um") by the number of AS-units. *Number of Filled Pauses / Number of AS-unit*
	Warm-up Preparation Time	• Warm-up preparation time refers to the duration (in seconds) from the end of the interlocutor's prompt to the examinee's first filled pause.
	Total Preparation Time	• Total preparation time refers to the duration (in seconds) from the end of the interlocutor's prompt to the examinee's first content-bearing response.

Note. An AS-unit is a grammatical construct that includes an independent clause and any subordinate clauses (Foster et al. 2000). The AS-unit is more suitable than the T-unit or c-unit for analyzing spoken data (Foster et al. 2000, Norris and Ortega 2009, Plough et al. 2010).

Examples in: English
Applicable Languages: English
Applicable Level: Elementary

