

Exploring the Potential of Conversational AI for Assessing Second Language Oral Proficiency

YASIN KARATAY  AND JING XU 

*Cambridge University Press & Assessment
Cambridge, UK*

Abstract

Interactional Competence (IC) is an important subcomponent of oral proficiency, but many computer-mediated oral English assessments fall short in assessing this construct mainly due to technological limitations. Spoken Dialogue Systems (SDSs) have shown promise in assessing L2 oral communication, yet further investigation is needed on their effectiveness in eliciting IC features in high-stakes assessment contexts. This study is unique in that it analyzed both test takers' and an SDS's spoken discourse in human–computer interactions. Using an SDS to simulate an examiner in an IELTS Speaking task, the study explored how well the system mimicked human–human interaction and elicited IC features from test takers, focusing on the IC features documented in prior research. Thirty participants completed the SDS-mediated test, with their performances rated by two trained raters. Semi-structured interviews with 10 test takers were conducted following the assessment. The findings revealed that the SDS successfully elicited key IC features, which helped distinguish test takers at different proficiency levels, with reliable scoring across raters. Most test takers found the SDS competent, though some noted its limitations in nonverbal communication and conversational flow. These results suggest that SDSs have potential in oral proficiency assessments and provide valuable insights for refining SDS design to ensure reliable and valid assessments.

doi: 10.1002/tesq.70003

INTRODUCTION

Interactional competence is concerned with the ability to deal with a dynamic process of oral communication and is considered a key element in models of communicative language ability (Galaczi & Taylor, 2018). Although interactional competence (IC) is an integral part of oral proficiency in various contexts of language use, it is missed by a number of large-scale, high-stakes, computer-mediated English language assessments, such as Test of English as a Foreign Language Internet-based Test (TOEFL iBT), Pearson Versant, and Duolingo English Test, which predominantly use semi-direct, monologic speaking tasks to circumvent technological constraints and increase speaking assessment efficiency and accessibility (Isaacs, 2016; Xu, Jones, Laxton, & Galaczi, 2021). These well recognized English language assessments have unanimously been criticized for construct underrepresentation and lack of authenticity (Chun, 2006; Wagner, 2020; Xu, 2015).

Direct speaking assessment facilitated by an interlocutor, either in a face-to-face or video-call format, can simulate real-life social interactions (Lee, Mullooly, Devine, & Galaczi, 2024). However, human-mediated oral interviews are resource-intensive, costly, and difficult to scale up (Galaczi, 2010; Timpe-Laughlin, Sydorenko, & Dombi, 2024). Additionally, they may suffer from interlocutor idiosyncrasies. That is, while conversations in real life are impromptu and unstructured, previous research has indicated that varied interlocutor behavior and proficiency in a direct speaking test may impact test-taker performance and hence undermine test validity and fairness (e.g., Davis, 2009; Nakatsuhara, 2018). Thus, it remains a challenge to balance between interlocutor standardization and authenticity, ensuring that candidates are examined under similar conditions while enabling natural and free-flowing interactions (Roever & Kasper, 2018; Seedhouse & Nakatsuhara, 2018).

Spoken dialog systems (SDSs) can be powered by conversational artificial intelligence (AI) technology designed to mimic the role of an interlocutor in real-time oral communication. They can extract the semantics of speech input, including L2 speech, and output prompt and natural responses in synthetic speech (Litman, Strik, & Lim, 2018). This cutting-edge AI technology looks promising in addressing the aforementioned challenge that direct speaking assessment has been facing. That is, the AI-powered SDSs, assumingly consistent in its interactional behaviors, can potentially elicit some aspects of IC when “conversing” with a language learner (Karatay, 2022; Ockey & Chukharev-Hudilainen, 2021).

Within this context, the present study aimed to investigate the potential of using a self-developed SDS, powered by Open AI's GPT-4o, to moderate an interactive discussion task that is similar to the third part of the International English Language Testing System (IELTS) Speaking test. Specifically, we are interested in finding out how well the SDS-mediated speaking task can elicit interactive oral performances that contain IC features similar to those identified in human-mediated assessment contexts in previous research (e.g., Seedhouse, 2019; Seedhouse & Nakatsuhara, 2018). The study focused on evidence collection from two main aspects including (1) the language functions observed in human-machine interactions, and (2) test takers' perceptions of the SDS.

LITERATURE REVIEW

The Construct of Interactional Competence

Theoretically informed by conversation analysis, the construct of IC describes a language learner's ability to participate in social interaction and display appropriate behaviors in the social contexts of language use (Galaczi & Taylor, 2018; Roever & Ikeda, 2022).

Evidence about mastery of IC may include but is not limited to:

- management of turn-taking;
- repair of mishearings, misunderstandings, and unintended overlaps;
- openings and closings of conversations;
- management of possible misalignment by means of politeness;
- design of speech to address the recipient's social status and shared background knowledge;
- backchannelling, using both verbal ("right," "yeah," "wow," "that's great") and/or nonverbal expressions (eye contact, facial expressions, head movements), to indicate active reciprocity of information;
- topic management;
- displaying behaviours in accordance with one's social role (Roever & Ikeda, 2022, p. 8).

The construct of IC is embraced either explicitly or implicitly by a few influential psycholinguistic models of speaking. Bygate (1987), for

example, describes turn-taking as “international routines” in his model of speaking as a process (p. 26). Based on the model, a speaker needs to assess the situation of communication (e.g., telephone conversations, interviews, service encounters, or casual encounters) and relies on his or her knowledge of the discourse structures to handle turn-taking in a “conventional way” (Bygate, 1987, p. 23). Additionally, based on who the interlocutor is, the speaker needs to deal with meaning negotiation, determining the level of explicitness of the speech content and controlling the level of specificity of the language. Levelt (1999) discussed assessment of the interactional contexts in “conceptual preparation” of his Modular Model of Speech Production (pp. 89–94). In agreement with Clark (1996), Levelt considers oral communication as “a form of joint action” (Levelt, 1999, p. 89). That is, participants must be aware of some goal of the communication as well as their common ground; they also exercise their Theory of Mind (knowledge of others’ mental states) to monitor the mutually shared state of information with the interlocutor. Thus, conceptual preparation involves guiding the intended focus shift of the interlocutor and monitoring whether the interlocutor is following the speech act. In Kormos’s (2006) Model of Bilingual Speech Production, which is an adaptation of Levelt’s model, a speaker’s language choice is largely dependent on sociolinguistic factors such as the communication situation, and his or her relationship with the interlocutor (p. 169). In addition, Kormos reckons that L2 speakers may struggle with monitoring the accessibility of their speech to interlocutors (Kormos, 2006, p. 176).

Following the IC definition in Ockey and Chukharev-Hudilainen (2021), this study conceptualizes IC as the ability of an individual required for effective interaction, rather than as a competence co-constructed between participants (Mehan, 1979). This approach allows for the separation of individual test-taker performance from the collaborative nature of interactions (Lam, Galaczi, Nakatsuhara, & May, 2023; Young, 2011) in an assessment context, which is critical for defining the construct of conversational speaking assessment for individual test takers. To incorporate IC into L2 oral assessments, researchers have expanded on this conceptualization by identifying the key components of IC. For example, Galaczi and Taylor’s (2018) tree model divides IC into macro features, such as turn management, topic management, interactive listening, breakdown repair, and nonverbal behavior. In this model, macro features are depicted as branches, micro features as leaves, and broader contexts, such as speech acts and events, as roots. Topic management refers to the ability to manage topics in interaction, interactive listening involves actively listening to an interlocutor, turn management

pertains to alternating between speaker and listener roles, and repair management refers to strategies or techniques used in communication to rectify misunderstandings. While nonverbal behavior is a critical component of IC, it was excluded from this study due to technical limitations.

This study adopts Galaczi and Taylor’s framework to define and operationalize the construct of IC. Based on the framework, the micro-level features of IC that were observed by prior research (e.g., Galaczi, 2014; Lam, 2021; Saeki et al., 2024) in discourses produced by test takers in human-mediated interactive oral tasks were analyzed. In face-to-face oral proficiency interviews such as the IELTS Speaking Test, candidate performance is evaluated based on their ability to engage in structured interactions that integrate turn-taking, adjacency pairs, and topic development (Seedhouse, 2019; Seedhouse & Nakatsuhara, 2018). Part 3 of the IELTS Speaking Test, often referred to as a “two-way discussion,” requires candidates to respond to a series of topic-based questions, which consist of both an adjacency pair component (eliciting a response) and a topic component (eliciting development of the topic). While the examiner initiates the topic using scripted questions, candidates are expected to extend the topic by adding details, expressing opinions, and maintaining coherence, which are essential for achieving high scores (Seedhouse, 2019).

Research has shown that high-scoring candidates tend to produce extended turns, utilize varied and sophisticated lexical choices, and construct identities aligned with their academic and professional aspirations (Seedhouse, 2019; Seedhouse & Nakatsuhara, 2018). In contrast, lower-scoring candidates often struggle to develop topics as fully, with responses that may be tangential, limited in scope, or lacking coherence. The structured yet co-constructed nature of Part 3 allows differentiation in candidate performance while maintaining standardization. These features of topic management and development informed the design of the current study, which aimed to simulate similar structured interactional patterns; however, using an SDS as an interlocutor.

Criticisms on Computer-Mediated Speaking Assessment

While examiner-delivered speaking tests like IELTS, Occupational English Test (OET), and Cambridge English Qualifications still use direct, interactive tasks, many test providers have computerized speaking assessments using semi-direct tasks over the past two decades. PhonePass, the predecessor of the Pearson Versant Speaking test, pioneered the use of constrained tasks such as reading sentences

aloud, repeating sentences aloud, saying opposite words, and building a sentence from phrases, to cater for the limitations of automated scoring technology (Bernstein, De Jong, Pisoni, & Townshend, 2000). Launched in 2005, TOEFL iBT includes six monologic speaking tasks and was designed to accommodate the use of the automated scoring engine, SpeechRater (Xi, 2008). Following the example of TOEFL and Pearson, other programs like Aptis, Linguaskill, Duolingo English Test, and some local language tests adopted computer-mediated, semi-direct speaking assessments, which have become the norm due to their efficiency in large-scale testing.

In light of the growing role of the communicative approach to language teaching, learning, and assessment (Canale & Swain, 1980), criticisms of semi-direct speaking assessments have never stopped. The main argument is that constrained or monologic speaking tasks are unauthentic and underrepresent or even misrepresent the construct of oral proficiency for real-world communications. Chun (2006) criticized PhonePass for altering the nature of speaking and failing to ask test takers to create discourse that would be expected in real-life language use situations. Alderson (2009) noted the construct underrepresentation of the TOEFL iBT Speaking test. According to his appraisal, the ability to engage in interactive speaking such as clarifying one's understanding, questioning what has been heard, and challenging others' opinions is a necessary skill in academic study but is completely missed by the test (Alderson, 2009, p. 628). Xu (2015) argues that spontaneous, interactive speaking involves a different and more demanding underlying cognitive process than monologic speaking. That is, the former requires extra attentional resources devoted to active listening, including assessment of the listener's understanding and adapting the speech content based on the listener's reaction. Wagner (2020) explicitly calls out the deficiency of the Duolingo English Test in assessing IC, which is essential for the target language use domain of the test. That is, test takers will be expected to engage in interactive conversations and participate in class discussions in an English-medium university setting.

Roever and Ikeda (2022) examined how test takers' performances on TOEFL iBT monologic speaking tasks reflected their IC. Their study involved 150 non-native English speakers across three cohorts: pre-entry students, recent-entry students, and longer-term university students. In addition to the TOEFL iBT, participants completed a role-play task where their IC was rated. The overall correlation between TOEFL iBT speaking scores and IC was medium to high ($r = .76$), but this correlation was lower for pre-entry and recent-entry students ($r = .57$). The researchers concluded that TOEFL iBT scores only partially reflect interactional abilities. Additionally, Roever and Ikeda found that IC scores varied widely within each TOEFL iBT score

level, often exceeding one or two standard deviations. This variation indicated a dissociation between the constructs measured by monologic tasks and the role-play task. Discourse analysis of two participants with the same TOEFL iBT speaking score showed that one demonstrated strong turn-taking and social awareness, while the other showed limited recipient orientation and failed to manage turn-taking effectively. It is noteworthy that a high correlation between the scores of two tests does not necessarily mean the constructs of the two tests overlap. For example, Bernstein, Van Moere, and Cheng (2010) reported a strong correlation ($r = .92$) between the Versant speaking test scores and the oral proficiency interview (OPI) scores. However, Versant, which mainly uses constrained speaking tasks such as reading aloud, repetition, and short responses, has been criticized for construct underrepresentation, that is, failing to assess a full spectrum of oral communication skills (Chun, 2006, 2008).

Building upon these criticisms and in addition to its focus on IC, this study also considers two critical aspects of language assessment: authenticity and effectiveness. Authenticity, often defined as the extent to which tasks reflect real-world language use situations (Bachman & Palmer, 1996), is a key consideration in the validation of speaking assessments. Authenticity can be further divided into two types, that is, situational authenticity and interfractional authenticity. The former is concerned with the correspondence between the characteristics of the test tasks and those of the specific target language use (TLU) situations; the latter is about the involvement of test takers' language ability in accomplishing a test task (Bachman, 1991). SDS-mediated tasks can potentially enhance both situational and interfractional authenticity by enabling interactive and dialogic exchanges, which closely resemble real-life communication situations and involving test takers' interactional competence in task completion, while maintaining consistency in test delivery across test takers (Karatay, 2022; Ockey, Chukharev-Hudilainen, & Hirsch, 2023). Effectiveness, on the other hand, refers to the efficacy of an SDS in moderating a conversational task as an interlocutor. Previous studies (e.g., Gokturk & Chukharev, 2024; Ockey & Chukharev-Hudilainen, 2021) have demonstrated that SDSs can potentially provide these two benefits by minimizing variability in interlocutor behaviors and encouraging meaningful, extended responses.

Assessing International Competence Using SDSs

The inadequacy of monologic tasks in assessing oral interactional abilities has recently stimulated a small body of research to explore the use of SDSs to deliver interactive speaking assessments. Chukharev-

Hudilainen and Ockey (2021) developed a prototype SDS, called Interaction Competence Elicitor (ICE), for a paired oral discussion task. Based on data from 42 participants' oral interactions with the SDS and a human interlocutor, questionnaires, and semi-structured interviews, they concluded that ICE was able to generate task-appropriate utterances useful for eliciting test takers' interactional behaviors in 90% of cases. They also found that even some of the system failures that originated from automated speech recognition (ASR) or natural language understanding (NLU) components prompted test takers to repair the "mishearing" or "misunderstanding" displayed by the SDS.

In a follow-up study, Ockey and Chukharev-Hudilainen (2021) compared human ratings on discourses from human-SDS and human-human interactions in the same paired oral task. Their Generalizability (G) study found similar scores for fluency, pronunciation, grammar, and vocabulary across both conditions, but IC scores were more reliable in the human-computer interactions. While raters viewed human-human interactions as more natural and authentic, they acknowledged that SDS provided a slightly more standardized assessment of oral proficiency, including IC. Interviews indicated raters prioritized discourse naturalness over standardization. Despite favoring human-human interactions, the researchers saw potential for SDSs in assessing IC in high-stakes contexts.

Karatay (2022) developed a specialized SDS for three role-play scenarios in a Tourism English oral communication test, gathering validity evidence for its use in speaking assessment. The SDS acted as a hotel guest making complaints, while the test taker acted as a receptionist. The SDS used keyword spotting for responses, with 30 Turkish learners assessed by four raters. Post-task surveys and interviews indicated that the participants found the task authentic and effective, and raters were confident in its consistency across proficiency levels. Discourse analysis identified five construct-relevant strategies, and G-theory analysis showed high inter-rater reliability (.94), with 67% of score variability attributed to oral communication ability. Karatay (2022) called for further research on SDS applications in ESP speaking assessments to expand construct coverage.

Timpe-Laughlin et al. (2024) explored the use of an SDS in role-play tasks for learning and low-stakes assessment. Forty-seven tertiary-level English learners completed the same role-play task in two modes: via a computer SDS and face-to-face with a human interlocutor. The researchers compared the two modes on linguistic complexity (syntactic complexity, lexical variety, fluency) and pragmatic functions (number/type of requests), and surveyed participants' perceptions. They found similar linguistic complexity across modes, but human interactions had shorter speaking time, more turns, and more rapport

building, clarification, and backchanneling moves. Participants made more direct requests in SDS than with human interlocutors. Post-task questionnaires indicated a slight preference for human interactions, attributed to nonverbal cues. Despite differences, the researchers supported SDS for teaching and assessing L2 oral interaction.

RESEARCH QUESTIONS

Given the great promises of SDSs in addressing the limitations of the current mainstream computer-mediated speaking assessment, the present study explored the use of this technology for delivering an interactive speaking task adapted from a high-stakes oral English test. The study posed four research questions as follows:

1. What types of interactive language functions are elicited by SDS, and how frequently do these functions occur in test-taker responses?
2. What are test takers' perceptions of the SDS-mediated speaking test in terms of authenticity and effectiveness?

METHODOLOGY

Research Design

This study employed a convergent, parallel-mixed methods design (Creswell & Plano Clark, 2011) to explore the use of SDS in assessing L2 oral proficiency. Both quantitative and qualitative data were used to provide a holistic understanding of the efficacy of a hybrid approach to SDS design that combines a rule-based keyword spotting technique and a large language model, that is, GPT-4o (OpenAI, 2024) for the dialogue management. Quantitative methods were used to assess the types and frequency of interactive language functions elicited by the SDS. Qualitative methods were used to explore the perceptions of test takers regarding the authenticity and effectiveness of the SDS-mediated speaking test.

Participants

Thirty non-native English speakers (15 females and 15 males) with a diverse range of first language (L1) backgrounds, proficiency levels,

and educational or professional contexts sat the SDS-mediated speaking test. Their native languages included Turkish ($n = 14$), Persian ($n = 5$), Mandarin Chinese ($n = 3$), Indian ($n = 3$), Korean ($n = 2$), Arabic ($n = 2$), Pashto ($n = 1$), and Indonesian ($n = 1$). Their self-reported English proficiency levels were varied, including three beginners, five pre-intermediate, nine intermediate, eight upper-intermediate, and five advanced learners. The test takers had a mix of educational and professional experiences, including postgraduate students, engineers, language instructors, and finance specialists.

Two senior raters with MA and PhD degrees in Applied Linguistics and substantial experience of rating high-stakes-speaking exams evaluated the test takers' oral performances in the SDS-mediated speaking test. They were in late 30s and both non-native speakers of English with postgraduate degrees in applied linguistics. Both had English as a Second/Foreign Language teaching experiences at various universities. Although they were considered as proficient users of English, none of them had had any experience of assessing human-computer interaction. Thus, they underwent an online rater training and calibration session. Both the test takers and raters agreed to take part in the study and received incentives for their participation.

Instruments

AI interlocutor SDS. The AI interlocutor SDS used in this study simulates human-like interactions in high-stakes speaking tests, such as IELTS, through a hybrid approach combining rule-based mechanisms and a large language model (LLM). It includes five core components: automated speech recognition (ASR), natural language understanding (NLU), a dialogue manager (DM), natural language generation (NLG), and text-to-speech (TTS). Throughout this manuscript, we consistently use the term 'SDS' when referring to the interactive spoken assessment system as a whole. The term 'AI interlocutor' specifically refers to the dialogue management component powered by GPT-4o.

The SDS utilized Google's cloud-based Speech-To-Text API for speech recognition for the sake of affordability, fast processing times, and robust accuracy in transcribing spoken language (Filippidou & Moussiades, 2020; McCrocklin & Edalatshams, 2020). The system also utilized Google's Text-to-Speech API, specifically employing the English (US) Standard voice (en-US-Standard-E), for generating synthesized speech responses.

The interaction begins when the test taker clicks "Start conversation" and speaks into the system. The ASR transcribes speech into text, processed by the NLU, which integrates GPT-4o for enhanced

interpretation. Unlike standard SDS systems, our NLU integrates GPT-4o to improve the interpretation of linguistic elements and enhance understanding. The DM controls the interaction, applying predefined rules to ensure topic coherence and guide the test taker, adjusting question complexity in real time based on CEFR proficiency. NLG generates responses, which TTS converts into speech (see Appendix S1 for a sample dialogue).

While large language models such as GPT-4o are effective in accommodating a wide range of test-taker inputs, previous research (e.g., Sydorenko et al., 2024) has revealed their tendency to produce overly verbose responses, which could reduce opportunities for test takers to produce extended speech. To mitigate this issue, our system was carefully designed with prompts that minimize verbosity and prioritize eliciting meaningful, test-taker-generated responses.

This study aimed to simulate Part 3 of the IELTS Speaking Test using the SDS. IELTS measures English proficiency for study or work in English-speaking environments, with a 9-band grading system from 1 (Non-User) to 9 (Expert) across Listening, Reading, Writing, and Speaking. The 11–14-minute Speaking Test is conducted face-to-face with an examiner who assigns a score based on four criteria: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation (25% each; see Appendix S2 for the rating scale). The Part 3 of the IELTS Speaking Test, referred to as a “two-way discussion,” involves an examiner and candidate engaging in an interactive discussion on complex issues related to the previous Part 2 topic. The interaction in Part 3 is structured as a series of topic-based question-answer adjacency pairs (Seedhouse, 2019), where the examiner poses open-ended questions, and the candidate responds with extended answers, offering opinions, justifications, and elaborations. This section is designed to assess the candidate’s ability to discuss abstract ideas, engage in reasoning, and demonstrate flexibility in managing the flow of a discussion. In this study, the SDS simulated the examiner’s role in Part 3 in eliciting interactive language and measuring oral proficiency.

Test taker survey and interview protocol. The survey used a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). The test taker survey (Appendix S3), informed by previous research (i.e., Chukharev-Hudilainen & Ockey, 2021; Timpe-Laughlin et al., 2024), gathered test takers’ perceptions of the SDS-mediated task, focusing on comfort, interaction with the AI, and perceived effectiveness in assessing speaking skills. It also examined their preferences between AI and human interlocutors and the naturalness of the SDS responses. Post-test interviews with 10 participants provided further insights into

the perceived authenticity, effectiveness, and overall experience of the SDS-mediated speaking test (see Appendix S4 for the interview protocol).

Rater training protocol. Following the guidelines of Knoch, Fairbairn, and Jin (2021), two raters completed a 2-hour online training session to ensure consistent application of the IELTS Speaking mark scheme. The session began with an introduction to the test task, task prompt, and rating scale, followed by a review of benchmark performances that illustrated clear examples of various proficiency levels. Raters then practiced scoring a randomized set of test taker performances, covering all dimensions of the rating scale. After each rating, the raters discussed their decisions, justifying their scores based on specific criteria in the rubric, and any incorrect ratings were discussed by the first author by explaining how the criteria were applied. The sample performances used during training were excluded from the analysis.

Procedures

Data collection occurred between June and July 2024. The participants were recruited via convenience sampling from individuals known to the researchers, with a focus on ensuring diversity in L2 English proficiency and L1 backgrounds. Each participant received a unique URL to access the SDS-mediated speaking test, which they completed remotely on their laptops. Before starting the task, participants completed a consent form to ensure ethical compliance. The test task took an average of 10 minutes to complete, followed by another 10 minutes to complete a post-test survey. The survey included questions addressing major topics such as participants' comfort with the SDS, perceptions of pacing, response naturalness, and the impact of features such as the computer's accent and avatar.

The SDS automatically recorded the audio files of the human-computer interactions. The oral interactions in the audio files were automatically transcribed by the ASR of the SDS and these transcriptions were subsequently manually checked and corrected at the word level by the first author to ensure accuracy. Throughout the data collection period, the first author and the system developer monitored participants' test-taking processes remotely to address any technical issues, such as login errors. All recordings were of acceptable quality and were successfully analyzed as part of the study. After the test, all 30 participants completed the post-test survey, and 10 took part in semi-structured interviews.

Data Analysis

To examine the types of interactive language functions (RQ1), the transcribed interactions were analyzed using discourse analysis (Galaczi, 2014; Lazaraton, 2002). The SDS-mediated interactions were explicitly standardized, with all conversations strictly limited to exactly 4 minutes. Therefore, frequency counts of interactive language functions inherently represent occurrences per standardized 4-minute interaction. For this analysis, the transcripts were divided into distinct turns, where a turn was defined as a segment of speech that begins and ends when another speaker (typically the interlocutor) contributes to the interaction (Crookes, 1990). Each turn was examined for the presence or absence of IC features, using the principles of the next-turn-proof procedure (Hutchby & Wooffitt, 2008), which was applied to determine the speech act of each utterance based on how it was interpreted in the subsequent turn. This approach relies on the principle that each turn reflects the speaker's understanding of the prior turn, making interaction a co-constructed process.

The segmented transcripts were coded based on an interactional data coding scheme adapted from Galaczi (2014), using NVivo 12. Based on the data collected and the construct of IC of interest in this study, some modifications were made to the coding scheme (e.g., adding interruptions as a macro feature based on Bilmes (1997)). The final version of the coding scheme consists of five main categories (i.e., topic development, turn-taking, listener support, repairs, and interruptions) and 21 specific codes for both AI interlocutor and test taker utterances (see Appendix S5). The first author and a second coder coded the data. The second coder received an approximately 2-hour training on applying the coding scheme to the conversational data using two randomly chosen transcripts. During the training, both coders discussed the codes together to ensure that they had a consistent understanding of the coding scheme. To establish the reliability of the coding, both coders first coded randomly chosen four out of the remaining 28 transcripts (around 14% of the data) independently, which yielded high inter-coder reliability (Krippendorff's $\alpha = .87$) (Stemler & Tsai, 2008). Thus, the second coder coded the rest of the data independently.

Test taker perceptions (RQ2) were analyzed through discourse synthesis (Spivey & King, 1989) of interview data, complemented by descriptive survey statistics. The qualitative findings were organized around emerging themes, which were supported by representative excerpts from the interview responses of the test takers. Additionally, to ensure the qualitative analysis was grounded on reliably scored

performance data, two independent, experienced raters assessed the test-takers' SDS-elicited oral responses using the IELTS rubric. The inter-rater reliability for SDS-elicited oral responses was found to be excellent (Koo & Li, 2016), with an intraclass correlation coefficient of 0.948 for average measures (95% CI [0.874, 0.978]).

FINDINGS

Types and Frequencies of Interactive Language Functions in SDS-Mediated Speech (RQ1)

The first research question focused on the types and frequency of interactive language functions in human–computer interactions. Discourse analysis of the transcribed interactions revealed five major interactional functions: topic development, turn-taking, listener support, interruptions, and repairs. Table 1 presents the frequency of these patterns. As seen by the total counts and *n* values in the table, nearly all interactive patterns were consistently observed in task responses from both the SDS and test takers across different score levels.

As shown in the table, the distribution of interactive patterns varied across categories. In the *Topic Development* category, the SDS contributed more to certain aspects, such as extending topics, particularly other-initiated topics (*n* = 189, 69.5% of its contributions). Test takers showed a more balanced approach, with 39.6% of their contribution (*n* = 90) focused on extending self-initiated topics and 31.7% (*n* = 72) on developing other-initiated topics. Topic initiation was nearly evenly split between the SDS (*n* = 74, 27.2%) and test takers (*n* = 65, 28.6%). *Turn-taking* was also balanced, with the SDS taking 282 turns and test takers 258. However, test takers always took turns after a gap, while the SDS initiated 85.1% of its turns (*n* = 240) after a 3-second gap, with 14.9% (*n* = 42) occurring as unintended interruptions due to a technical glitch in the ASR component.

In the *Listener Support* category, the SDS contributed 27 instances of *confirmations of comprehension*, making up 100% of its listener support actions, while the test takers contributed only seven instances, of which four were *backchannelling* (57.1%), and three were confirmations of comprehension (42.9%). Notably, the SDS did not employ any backchannelling because of existing technical constraints. *Interruptions* were primarily employed by the SDS, with the most common type being interruptions due to ASR problems (*n* = 42, 53.8%). Interruptions also arose from silences that were longer than 3 seconds (*n* = 36, 46.2%). Finally, *Repairs* were handled almost equally between the SDS and the

TABLE 1
Overall Use of Interactive Patterns in Task Responses

Category		Feature	n^1	Total count ²	% in relation to interactive category
Topic development	Test taker	Initiation	30	65	28.6
		Extension: 'Own' Topic	30	90	39.6
		Extension: 'Other' Topic	30	72	31.7
		Total		227	100.0
	AI	Initiation	30	74	27.2
		Extension: 'Own' Topic	5	9	3.3
Extension: 'Other' Topic		30	189	69.5	
Total			272	100.0	
Turn-taking	Test taker	After a gap	30	258	100.0
		No-gap-no-overlap	0	0	0.0
		Total		258	100.0
	AI	After a 3 seconds gap	30	240	85.1
		No-gap-no-overlap	30	42	14.9
		Total		282	100.0
Listener support	Test taker	Backchannelling	2	4	57.1
		Confirmations of Comprehension	2	3	42.9
		Total		7	100.0
	AI	Backchannelling	0	0	0.0
		Confirmations of Comprehension	25	27	100.0
		Total		27	100.0
Interruption	AI	Interruption_Silence	17	36	46.2
		Interruption_STT	29	42	53.8
		Total		72	100.0
Repair	Test taker	Clarification Request	17	35	49.3
		Rephrasing	22	36	50.7
		Total		71	100.0
	AI	Clarification Request	3	3	4.0
		Rephrasing	27	72	96.0
		Total		75	100.0

¹Number of test takers who used a particular interactive pattern.

²Total frequency of occurrences across all participants.

test takers, although the SDS engaged more frequently in *rephrasing* ($n = 72$, 96% of its repairs) compared to *clarification requests* ($n = 3$, 4%). The test takers split their repairs more evenly, with 49.3% consisting of *clarification requests* ($n = 35$) and 50.7% involving *rephrasing* ($n = 36$).

We also examined whether the distribution of interactive language functions varied by test takers' proficiency levels. Table 2 presents descriptive statistics on interactive patterns by score level (see Appendix S6 for the full dataset). Considerable differences in medians

TABLE 2
Interactive Patterns Across Score Bands

	8.5 (<i>n</i> = 2)			4 (<i>n</i> = 6)		
	Mdn	Min	Max	Mdn	Min	Max
Initiation	6.50	5	8	5.50	0	11
Extension: 'Own' Topic	8.50	8	9	2.00	2	2
Extension: 'Other' Topic	8.00	8	8	3.00	3	3
Initiation (AI)	6.00	6	6	3.00	2	4
Extension: 'Own' Topic (AI)	0.00	0	0	1.00	1	1
Extension: 'Other' Topic (AI)	18.50	16	21	6.00	4	8
After a gap	9.00	7	11	22.50	19	26
No-gap-no-overlap	0.00	0	0	0.00	0	0
After a 3 seconds gap (AI)	10.00	9	11	21.50	18	25
No-gap-no-overlap	1.00	1	1	3.00	2	4
Backchannelling	0.00	0	0	0.00	0	0
Confirmations of Comprehension	0.00	0	0	0.00	0	0
Backchannelling (AI)	0.00	0	0	0.00	0	0
Confirmations of Comprehension (AI)	2.00	2	2	1.00	0	2
Interruption_Silence (AI)	0.00	0	0	6.00	4	8
Interruption_STT Problem (AI)	2.50	2	3	2.50	2	3
Interruption_End-of-Test (AI)	2.00	2	2	2.00	2	2
Clarification Request	4.00	4	4	2.00	0	4
Rephrasing	2.50	2	3	2.50	2	3
Clarification Request (AI)	0.00	0	0	0.00	0	0
Rephrasing (AI)	3.00	3	3	8.50	7	10

Note. The proficiency levels represent scores on the IELTS Speaking scale, ranging from 1 (Non-User) to 9 (Expert). Participants with Band 8.5 demonstrated near-native fluency and interactional competence, while those with Band 4 showed limited ability to maintain conversations and produce coherent responses.

were observed between the lowest (Band 4) and highest (Band 8.5) levels. For instance, Band 8.5 test takers showed higher use of self- and other-initiated topic extensions (8.5 and 8) compared to Band 4 (2.0 and 3.0, respectively). This suggests that higher-proficiency test takers were more adept at developing both their own and SDS-initiated topics. As proficiency decreased, so did the frequency of topic development. For example, Participant 18 (Band 8.5) extended the topic effectively as in the following example:

AI: *Using a calendar is a great way to stay organized. Do you prefer a **digital calendar or a paper one**? Why?—Extension: "Other" Topic (SDS)*

TT: *I generally prefer **the digital calendar** because like it is accessible to any other like electronic devices ... (continues)—Extension: "Other" Topic (Test taker)*

AI: *That's interesting! So, digital calendars make it easier for you to stay organized. Can you give me an example of how a digital calendar has helped you manage your time more effectively?—Extension: "Other" Topic (SDS)*

TT: *Definitely like uh yeah you know maybe people use like a paper-based calendar but like ... (continues)—Extension: "Other" Topic (Test taker)*

This interaction highlights how the SDS facilitated deeper exploration of the topic while the test taker engaged in elaboration and example-giving. A similar attempt by the SDS was observed with Participant 13 (Band 4), who initially faced difficulty in the interaction, with several pauses and instances of hesitation. The SDS made multiple efforts to elicit further details from the test taker; however, the test taker was unable to provide this information, ultimately prompting the SDS to change the topic:

AI: *People use various tools and methods to organize their time. Can you give me an example of what you use or what you've seen others use?*

TT: *Can you please repeat the question?*

AI: ***For example, some people use calendars. How about you?***

TT: *Okay I have a schedule and I prefer that.*

AI: *That's a good start! So, you mentioned you have a schedule. **Can you tell me a bit more** about how you use it to organize your time?*

TT: *First of all, I wake up early and I have breakfast. After the breakfast I study.*

AI: *That's great! So, you have a routine that begins early. **Why do you think** waking up early is important for organizing your time?*

TT: *Because if I if I don't wake up early, I'll be late for everything.*

AI: *That's interesting. **Can you tell me more** about what happens if you don't wake up early?*

TT: *If I don't work wake up early, my day is over quickly so I don't I don't like it.*

AI: ***That's okay. Let's move on.** Now let's talk about gift giving. In your country, when do people usually give gifts?*

In the *turn-taking* category, the turns were relatively balanced between the SDS (282 turns) and test takers (258 turns). Lower-proficiency test takers relied more on the SDS to manage

conversational gaps, with a higher frequency of SDS-initiated turns after long pauses compared to higher-proficiency test takers. Figure 1 illustrates the number of turns taken across proficiency bands. Due to system design, test takers could only initiate turns *after a gap* ($n = 258$), while the SDS predominantly initiated turns *after a 3-second gap* ($n = 240$). Additionally, the SDS exhibited 42 instances of *no-gap-no-overlap* turn-taking, caused by a technical glitch in the speech-to-text component, leading to unintended interruptions.

Listener support through backchannelling and confirmations of comprehension was another feature observed in the responses. Although the ASR component did not transcribe test takers’ speech when the text-to-speech component was active, the audio recordings captured both parties’ speech. This allowed the raters to hear and analyze test takers’ backchannelling directly from the audio files. As a result, the reported frequency of backchannelling ($n = 4$) reflects the manual analysis of the recorded audio files, not solely the ASR-generated transcripts. However, the SDS did not produce backchannelling due to design limitations, as it was not programmed to include this feature.

As for the *confirmations of comprehension*, while only two test takers employed this feature three times in total, the SDS was able to employ it 27 times for the majority of the test takers ($n = 25$). For example, after presenting a prompt, the SDS responded to Participant 18 (Band 8) with supportive feedback: “That’s interesting! So, digital calendars make it easier for you to stay organized. Can you give me an example of how a digital calendar has helped you manage your time more effectively?”. This

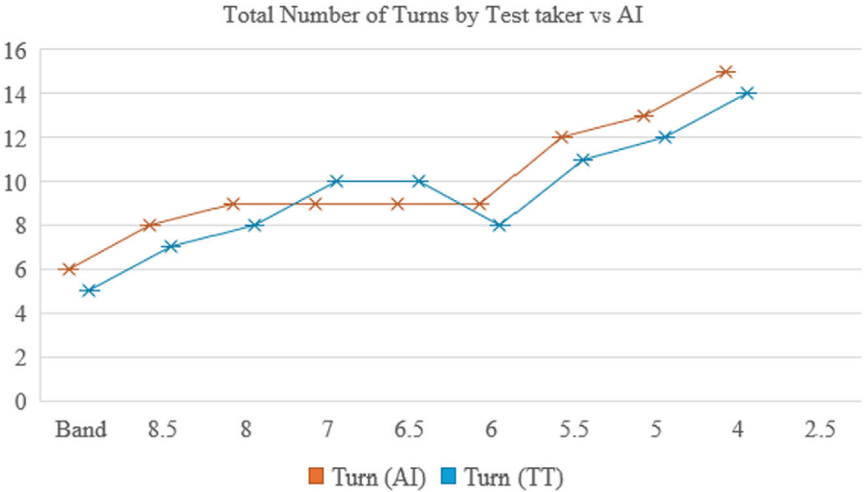


FIGURE 1. Total number of turns by test takers and AI across score bands.

encouraged the participant to further elaborate, extending the interaction with detailed examples.

Interruptions in the SDS-mediated speaking test stemmed from pauses and technical issues with the ASR system. Two main types of interruptions were identified as follows: those caused by silence (linked to hesitations or fluency issues) and technical glitches in the ASR component. Band 4 test takers experienced more interruptions due to longer silences, with a median of six per response, while Band 8.5 had no such interruptions. ASR-related interruptions were consistent across bands (*Mdn* = 2.5 for each).

Finally, in the *Repairs* category, *Clarification Requests* had a median of 4.0 for Band 8.5 and 2.0 for Band 4, showing higher usage among higher-proficiency learners. *Rephrasing* was used similarly across both bands, with a median of 2.5. These figures suggest that while rephrasing was employed at comparable levels across both proficiency groups, clarification requests were more frequently used by higher-proficiency learners. The excerpt below shows a typical clarification request from a test taker:

Participant 20 (Band 5)

AI: *Can you give me an example of a time when staying organized was particularly challenging for you?*

TT: *I don't know what you mean. **Please repeat your question?***

AI: *Of course. Can you tell me about a time when it was very difficult for you to stay organized?*

Test Takers' Perceptions of the SDS-Mediated Speaking Test (RQ2)

Test takers provided both positive feedback and constructive criticisms regarding their experience with the SDS-mediated speaking test. The findings from the questionnaire and open-ended responses highlighted several key themes related to communication and comprehension, interaction quality and pacing, enjoyment and preference for AI, and challenges related to technical issues and interruptions. Table 3 displays test takers' perceptions of the SDS-mediated test based on the questionnaire items.

Overall communication. Test takers generally felt comfortable during the test ($M = 4.0$), with 21 participants (70%) agreeing. Similarly, they were satisfied with the clarity of computer speech ($M = 4.0$). They

TABLE 3
Test Takers' Ratings of the SDS-Mediated Test

Question item	Mean (SD)	Agree ¹	Neutral ¹	Disagree ¹
1. I would prefer to have the option to choose my accent in future tests.	4.3 (0.9)	24 (80.0%) ²	5 (16.7%)	1 (3.3%)
2. I felt that the accent of the computer was easy to understand.	4.2 (0.9)	23 (76.7%)	6 (20.0%)	1 (3.3%)
3. I felt the computer spoke at a pace that was easy to understand.	4.0 (1.1)	20 (66.7%)	7 (23.3%)	3 (10.0%)
4. I understood what the computer said to me.	4.0 (1)	21 (70.0%)	7 (23.3%)	2 (6.7%)
5. I enjoyed talking to the computer examiner.	4.0 (1)	21 (70.0%)	7 (23.3%)	2 (6.7%)
6. I felt comfortable communicating during the computer-mediated test.	4.0 (0.9)	21 (70.0%)	7 (23.3%)	1 (3.3%)
7. I felt that the response of the computer was appropriate.	4.0 (0.9)	21 (70.0%)	7 (23.3%)	1 (3.3%)
8. I felt that the computer understood my responses well.	3.9 (0.9)	20 (66.7%)	8 (26.7%)	2 (6.7%)
9. I felt that the accent of the computer had a positive impact on my performance in the test.	3.9 (0.9)	20 (66.7%)	8 (26.7%)	2 (6.7%)
10. I felt that the choice of accent made the test experience more enjoyable.	3.9 (0.9)	20 (66.7%)	8 (26.7%)	2 (6.7%)
11. I felt that the response of the computer was natural.	3.9 (0.7)	21 (70.0%)	8 (26.7%)	1 (3.3%)
12. I felt that the pace of the conversation was appropriate.	3.7 (1.1)	17 (56.7%)	9 (30.0%)	4 (13.3%)
13. I would prefer the computer examiner for future speaking tests.	3.7 (1.1)	17 (56.7%)	9 (30.0%)	4 (13.3%)
14. I felt that the test allowed me to fully demonstrate my oral communication skills.	3.7 (1)	17 (56.7%)	9 (30.0%)	3 (10.0%)
15. I had sufficient time to think about my responses during the test.	3.2 (1.6)	13 (43.3%)	7 (23.3%)	10 (33.3%)
16. I felt that the choice of avatar made the test experience more enjoyable.	3.2 (1.5)	13 (43.3%)	8 (26.7%)	10 (33.3%)
17. I felt that the avatar made me comfortable.	3.1 (1.4)	12 (40.0%)	8 (26.7%)	10 (33.3%)
18. I felt that the avatar's appearance had a positive impact on my performance in the test.	3.0 (1.4)	11 (36.7%)	8 (26.7%)	11 (36.7%)
19. I felt that the computer interrupted me too often.	3.0 (1.3)	11 (36.7%)	9 (30.0%)	11 (36.7%)

¹The responses were categorized into three groups based on their Likert scale ratings. "Agree" includes responses rated 4 or 5, "Neutral" includes responses rated 3, and "Disagree" includes responses rated 1 or 2. Frequencies and percentages were calculated for each category to provide a clearer understanding of participants' perceptions.

²Percentages and counts are based on a total of 30 participants.

reported being adequately understood by the computer ($M = 3.9$, 70%, $n = 21$) and found its responses natural ($M = 3.9$, 70%, $n = 21$) and appropriate ($M = 4.0$, 70%, $n = 21$). This feedback suggests that

test takers viewed the SDS as a satisfactory interlocutor, with a positive communication experience overall. However, some reported communication breakdowns and technical issues that they felt negatively impacted their performance, as seen in Excerpt 1.

Excerpt 1: (Participant 16). And as I said, you start explaining the question and someone or something interrupts you and **you have to start over**. And all the plans you've made in your mind just collapsed. So, of course, this affects your overall performance a little bit.

Some test takers (43.3%, $n = 13$) also noted that they did not feel it necessary to change their normal speaking behaviors when interacting with the SDS, such as modifying their language, adapting their communication style, simplifying their speech, or slowing down. This view can be observed in the Excerpt 2 below:

Excerpt 2: (Participant 33). You just need to answer the question, **doesn't matter who asks it**. So, computer or human, you just need to explain.

Interaction pace. The pacing of the interaction received a mean score of 3.7, with 17 participants (56.7%) agreeing that the pace was appropriate. However, some participants (43.3%, $n = 13$) expressed concerns about the time available to think and plan their responses ($M = 3.2$), suggesting that they felt rushed during parts of the conversation. Excerpt 3 highlights a test taker's observation on pacing.

Excerpt 3: (Participant 15). It took me a couple of turns until I found out that you need to wait a little bit once you're done with your talking. The more you talk the longer it listens. So, you need to wait so that it can speak, but that waiting time was a little bit long and weird for me.

Additionally, a notable proportion of participants (40%, $n = 12$) felt that the computer interrupted them too often ($M = 3.0$), which could be an indication that the timing of the SDS was perceived to be disruptive to the flow of their responses.

Enjoyment and preference. Test takers generally enjoyed the interaction with the SDS ($M = 4.0$, 70%, $n = 21$). However, when asked about their preferences for AI versus human interlocutors, they offered mixed responses, with a mean score of 3.7. A majority (56.7%, $n = 17$) preferred the AI interlocutor because it made them feel more relaxed, as shown in Excerpt 4:

Excerpt 4: (Participant 2). Before I experienced this AI, I'd have said a human, but after I experienced it, I would say AI because you feel more relaxed.

In contrast, 30% ($n = 9$) of the participants expressed a clear preference for a human interlocutor, emphasizing the importance of non-verbal communication and building a personal rapport with the interlocutor. These mixed preferences likely reflect participants' prior experiences with human interlocutors in everyday or conversational settings, rather than a direct comparison to a human examiner, as participants were not specifically asked about their experience on operational IELTS Speaking test. While the SDS-mediated speaking assessment offers some advantages, including creating a generally comfortable interaction for many test takers, the absence of nonverbal cues and relationship building, a social aspect of human communication, may limit its appeal for some test takers.

Response naturalness and appropriateness. The responses of the SDS were generally perceived as appropriate ($M = 4.0$, 70%, $n = 21$) and natural ($M = 4.0$, 70%, $n = 21$). However, as discussed in Excerpt 3, some test takers felt that the delayed responses occasionally undermined the natural flow of the conversation, as illustrated in Excerpt 5:

Excerpt 5: (Participant 15). I can say it was meaningful, but being natural? I have a doubt there because **its responses were a little bit late**, so you understand that there is a computer which analyses your speech and tries to give a response.

DISCUSSION

The present study aimed to explore the potential of using an SDS to assess L2 English learners' oral proficiency, including interactional competence, in high-stakes assessment contexts. Our investigation into the efficacy of the SDS focused on three areas: the interactive language functions observed in the human-computer interactions, test takers' perceptions of the SDS, and the reliability of human rating on learner speech elicited by the SDS. The findings revealed important implications for using an SDS in L2 speaking assessment, particularly in addressing the shortcomings of the existing large-scale, computer-mediated oral English assessments.

Regarding the interactive language functions, the SDS demonstrated a capacity to elicit a wide range of communicative functions, particularly in topic development and turn-taking. Regardless of their

proficiency levels, test takers were able to produce most interactive language functions in their responses, suggesting that the SDS could enhance the standardization of dialogic speaking assessments by providing equal opportunities for test takers to demonstrate their abilities (Gokturk & Chukharev, 2024; Ockey et al., 2023; Ockey & Chukharev-Hudilainen, 2021). The general consistency in SDS management of interactions across proficiency levels may mitigate concerns about interlocutor variability in human-mediated tests (Davis, 2009). Previous research has noted the challenge of balancing standardization and authenticity in human-delivered assessments (Roever & Kasper, 2018; Seedhouse & Nakatsuhara, 2018). Similar to prior studies (, 2021; Karatay, 2022), this research found that SDS, with its consistent interactional behaviors, has potential to address this issue.

Similar to Seedhouse and Nakatsuhara's (2018) study on spoken discourse produced in the face-to-face IELTS Speaking test, this study found that higher-level test takers demonstrated stronger abilities in topic development than their lower-level counterparts, with fewer turns and a greater capacity to expand on topics (Seedhouse, 2019). The same finding from both face-to-face and SDS-mediated speaking assessments may suggest that topic development is a reliable indicator of oral proficiency across different testing modalities. However, some interactive features typically seen in face-to-face tests, such as back-channelling, varied intonation use, and turn management in a no-gap-no-overlap manner, were notably absent in the SDS-mediated spoken discourse produced by test takers. These features, important for managing flow and demonstrating active listening, were also missing in the SDS's responses. This finding confirms the challenge raised by Seedhouse and Nakatsuhara (2018) that some interactive features essential for human-human interactions are very difficult to replicate in a computer-mediated speaking context. The absence of these elements suggests the current limitations of an SDS in simulating the nuances of human-human interactions and shed light on directions for further improvement of the SDS.

Unlike earlier SDS studies that solely focused on test takers' oral performances (e.g., Chukharev-Hudilainen & Ockey, 2021; Gokturk & Chukharev, 2024; Karatay, 2022), this study also examined the spoken discourse produced by the SDS. By doing so, we were able to assess SDS's capacity in mimicking not only surface-level features of interaction, such as turn-taking, but also more complex conversational behaviors typically exhibited by human examiners. Our analysis of AI-generated utterances revealed that while the system had limitations in handling more fluid conversation, such as turn-taking in a no-gap-no-overlap manner and signaling active listening through

backchannelling, it could effectively mimic a structured interview, as seen in the IELTS Speaking test, by asking appropriate follow-up questions.

The current study also found a strong inter-rater reliability ($ICC = .95$) on L2 speech elicited by the SDS. This dovetails with Ockey and Chukharev-Hudilainen's (2021) finding that the SDS-mediated speaking test resulted in more consistent ratings of interactional competence than the face-to-face counterpart moderated by an examiner. A plausible explanation for this finding could be that raters' cognitive load was greatly reduced when the job of managing a conversation with the test taker was taken by the SDS, thus allowing them to concentrate on rating. For example, Brown, Iwashita, and McNamara (2005) found that well-trained speaking raters were only able to attend to four conceptual categories in a rating scale. For this reason, an interlocutor in face-to-face oral interviews such as IELTS and Cambridge English Qualifications only assigns holistic scores rather than analytical scores. The strong inter-rater reliability could also be attributed to the high consistency of SDS behavior across test takers. The variability of interlocutor behaviors undermines the reliability of human-mediated speaking assessments, as noted by a number of studies (e.g., Davis, 2009; Nakatsuhara, 2018).

Test taker perceptions of the SDS were generally positive, with many participants expressing comfort and satisfaction with AI's ability to understand their responses and ask appropriate follow-up questions. However, some test takers reported challenges in conversing with the SDS due to the lack of nonverbal cues and occasional unexpected interruptions and shifts of topics initiated by AI. These drawbacks of the SDS reduce the naturalness and authenticity of human-computer interactions and may lead to test takers' preference for human-human interactions (Timpe-Laughlin et al., 2024). Nonetheless, the standardization of interlocutor behaviors introduced by the SDS was appreciated by many participants, which aligns with earlier findings reported by Karatay (2022), suggesting that SDS-mediated tasks can contribute to consistency and reliability in delivering oral proficiency interviews.

CONCLUSION

This study explored the potential of using an SDS, instead of a trained examiner, to moderate an interactive discussion task from a high-stakes English-speaking test. It was found that the SDS consistently elicited some key interactional competence features, as seen in face-to-face oral proficiency interviews, and that such features were

useful in distinguishing between higher- and lower-proficiency test takers. While most participants, including both the test takers and the raters, perceived the SDS as a competent and reliable interlocutor, certain limitations, such as the absence of nonverbal cues and occasional unnatural interruptions in human–computer interactions, highlight the areas for improvement. Despite these challenges, the findings suggest that the SDS technology holds promise for expanding the construct coverage of computer-mediated speaking assessments, which predominantly use monologic speaking tasks at the moment, reducing speaking examiners’ cognitive load in test administration, and enhancing the standardization and fairness of speaking assessments.

The implications of these findings may also extend to TESOL practitioners. In EFL contexts, where logistical challenges and resource constraints often limit opportunities for speaking practice, SDS-mediated tasks could seem a promising solution for self-regulated learning outside the classroom. Aside from being used in language assessment, these tasks may be equally helpful for learners to practice interactive speaking skills independently when conversational partners are not accessible. The SDS may be programmed to adapt its speech content, pace, and accent based on learners’ proficiency levels and learning needs. However, further research is needed on using an SDS in the learning contexts. Furthermore, SDS-mediated assessments, when combined with automated scoring and feedback technologies (Xi, 2023), could offer teachers valuable insights into learner performance patterns and enable more targeted instructional strategies.

Despite these promising findings, this study had several limitations that should be acknowledged. First, the sample size was relatively small, which limits the generalizability of the results to a larger test-taker population with diverse proficiency levels and linguistic backgrounds. Additionally, participants were recruited via convenience sampling and were not operational IELTS test takers. While the study was not specifically designed for IELTS but used it as a context to explore SDS-mediated speaking tasks, the lack of operational test takers may limit the generalizability of the findings to actual high-stakes assessment contexts. Future research should include broader samples, including operational test takers, to enhance external validity.

Second, the SDS investigated in this study was unable to either produce or identify nonverbal communication features such as facial expressions, gestures, and gaze, which are considered crucial elements in human–human interactions and may affect human ratings (see Burton, 2024). For this reason, we recommend further research that incorporates multimodal features into the SDS design, for example, using human-like avatars to provide nonverbal cues to the test takers.

Third, although interactions with the SDS aimed to replicate natural conversational dynamics, the SDS's occasional interruptions or misunderstandings of test-taker speech may have affected the authenticity and naturalness of interactions. Such issues may disrupt test-takers' pace in speech communication and hinder them from fully demonstrating their oral proficiency, potentially differentiating these interactions from typical human-mediated assessments.

While this study focused on interactional features, further research is also needed to explore the linguistic features and other aspects of the oral construct elicited by SDS-mediated tasks. A repeated-measures comparative study examining face-to-face and SDS-mediated interactions could provide deeper insights into how the two formats elicit both interactional and linguistic features. Such research would help evaluate the extent to which SDS-mediated tasks align with traditional assessment methods in capturing a comprehensive range of oral proficiency constructs.

Finally, while the LLM in this study was instructed to generate follow-up questions based on test testers' oral proficiency, further research is needed to assess how well the SDS tailor its responses in real-time to varying oral performances. This could help refine the SDS's ability to adjust question complexity and manage conversational flow more effectively, ensuring a more individualized assessment experience for each test taker.

AUTHOR CONTRIBUTIONS

Yasin Karatay: Conceptualization, methodology, formal analysis, writing original draft, revision. Jing Xu: Conceptualization, methodology, writing original draft, revision.

ACKNOWLEDGEMENTS

We are grateful to Ms. Jodie Watson for her assistance in deploying the spoken dialogue system into the test interface and Dr Leyla Karatay for helping with the data analysis.

CONFLICT OF INTEREST

None.

FUNDING INFORMATION

This research project was funded by Cambridge University Press & Assessment who covered the incentives for the thirty participants and the fees for using the GPT-4o model.

PARTICIPANT CONSENTS

The study was conducted in accordance with ethical guidelines, and all participants provided informed consent prior to their involvement. Participants were fully informed about the purpose and procedures of the research and voluntarily agreed to participate by completing a consent form.

THE AUTHORS

Yasin Karatay is a Senior Research Manager at Cambridge University Press & Assessment, where he leads research projects on conversational AI and spoken dialogue systems in L2 speaking assessment and learning contexts. His research interests include computer-based language testing, spoken dialogue systems, automated writing evaluation, and test validation. He holds a PhD in Applied Linguistics and Technology from Iowa State University. Before joining Cambridge, he was a lecturer at Düzce University (Türkiye) and a teaching and research assistant at Iowa State University (USA). Following the completion of his PhD, he worked as a Postdoctoral Research Associate at Iowa State University, contributing to the revision of a Global Online Course project sponsored by the U.S. Department of State. His current work focuses on leveraging cutting-edge technologies to enhance L2 speaking assessment and learning materials. He also teaches at the MSt in English Language Assessment program at the University of Cambridge.

Jing Xu is Head of Propositions Research-English at Cambridge University Press & Assessment where he oversees innovative research on technology-mediated English language assessment and learning. He received his PhD in Applied Linguistics & Technology from Iowa State University. His research interests are centered on the application of cutting-edge AI technologies to English language education and the related validity issues. In collaboration with his colleagues at Cambridge, his current work focuses on automarking of constructed responses in high-stakes L2 English assessment and AI-mediated English-speaking tasks. He is Co-Chair of the Automated Language Assessment SIG at the International Language Testing Association (ILTA) and a Subject Matter Lead for the Cambridge Institute for Automated Language Teaching and Assessment (ALTA). He also teaches at the MSt in English Language Assessment program at the University of Cambridge.

REFERENCES

- Alderson, J. C. (2009). Test review: Test of English as a foreign language: Internet-based test (TOEFL iBT®). *Language Testing*, 26(4), 621–631. <https://doi.org/10.1177/0265532209346371>
- Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, 25(4), 671–704. <https://doi.org/10.2307/3587082>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.
- Bernstein, J., De Jong, J., Pisoni, D., & Townshend, B. (2000). Two experiments on automatic scoring of spoken language proficiency. In P. Delcloque (Ed.), *Proceedings of InSTIL2000: Integrating speech technology in learning* (pp. 57–61). Dundee, UK: University of Abertay.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355–377. <https://doi.org/10.1177/0265532210364404>
- Bilmes, J. (1997). Being interrupted. *Language in Society*, 26(4), 507–531. <https://doi.org/10.1017/S0047404500021035>
- Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series, 2005*, i–157.
- Burton, J. D. (2024). Evaluating the impact of nonverbal behavior on language ability ratings. *Language Testing*, 41(4), 729–758. <https://doi.org/10.1177/02655322241255709>
- Bygate, M. (1987). *Speaking*. Oxford, UK: Oxford University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. <https://doi.org/10.1093/applin/I.1.1>
- Chukharev-Hudilainen, E., & Ockey, G. J. (2021). *The development and evaluation of interactional competence elicitor (ICE) for oral language assessments*. Princeton, NJ: Educational Testing Service.
- Chun, C. W. (2006). COMMENTARY: An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly*, 3(3), 295–306. https://doi.org/10.1207/s15434311laq0303_4
- Chun, C. W. (2008). Comments on “Evaluation of the usefulness of the versant for english test: A response”: The author responds. *Language Assessment Quarterly*, 5(2), 168–172. <https://doi.org/10.1080/15434300801934751>
- Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.
- Crookes, G. (1990). The utterance and other basic units for second language discourse analysis. *Applied Linguistics*, 11, 183–199. <https://doi.org/10.1093/applin/11.2.183>
- Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26(3), 367–396. <https://doi.org/10.1177/0265532209104667>
- Filippidou, F., & Moussiades, L. (2020). A benchmarking of IBM, Google and wit automatic speech recognition systems. In *Artificial intelligence applications and innovations: 16th IFIP WG 12.5 international conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, proceedings, part I 16* (pp. 73–82). Cham, Switzerland: Springer International Publishing.

- Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, 35(5), 553–574. <https://doi.org/10.1093/applin/amt017>
- Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Proceedings of the computer-based assessment (CBA) of foreign language speaking skills* (pp. 29–51). Brussels, Belgium: European Union.
- Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219–236. <https://doi.org/10.1080/15434303.2018.1453816>
- Gokturk, N., & Chukharev, E. (2024). Exploring the potential of a spoken dialog system-delivered paired discussion task for assessing interactional competence. *Language Assessment Quarterly*, 21(1), 60–99. <https://doi.org/10.1080/15434303.2023.2289173>
- Hutchby, I., & Wooffitt, R. (2008). *Conversation analysis* (2nd ed.). Cambridge, UK: Polity Press.
- Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 131–146). Berlin, Germany: Walter de Gruyter.
- Karatay, Y. (2022). *Development and validation of spoken dialog system-based oral communication tasks in an ESP context* (publication No. 29165842). Doctoral dissertation, Iowa State University ProQuest Dissertations Publishing.
- Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Sheffield, UK: Equinox Publishing.
- 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>
- Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Lam, D. (2021). Don't turn a deaf ear: A case for assessing interactive listening. *Applied Linguistics*, 42(4), 740–764. <https://doi.org/10.1093/applin/amaa064>
- Lam, D., Galaczi, E., Nakatsuhara, F., & May, L. (2023). Assessing interactional competence: Exploring ratibility challenges. *Applied Pragmatics*, 5(2), 208–233. <https://doi.org/10.1075/ap.00014.lam>
- Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge, UK: UCLES/Cambridge University Press.
- Lee, H.-w., Mullooly, A., Devine, A., & Galaczi, E. (2024). Exploring interaction in video-call paired speaking tests: A look at scores, language, and perceptions. *Applied Linguistics*, 45(3), 533–552. <https://doi.org/10.1093/applin/amad046>
- Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford, UK: Oxford University Press.
- Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3), 294–309. <https://doi.org/10.1080/15434303.2018.1472265>
- McCrocklin, S., & Edalatshams, I. (2020). Revisiting popular speech recognition software for ESL speech. *TESOL Quarterly*, 54(4), 1,086–1,097. <https://doi.org/10.1002/tesq.3006>
- Mehan, H. (1979). *Learning lessons: Social organization in the classroom*. Cambridge, MA: Harvard University Press.
- Nakatsuhara, F. (2018). Investigating examiner interventions in relation to the listening demands they make on candidates in oral interview tests. In G. J. Ockey

- & E. Wagner (Eds.), *Assessing listening: Moving towards authenticity* (pp. 206–225). Philadelphia, PA: John Benjamins. <https://doi.org/10.1075/llt.50.14nak>
- 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>
- Ockey, G. J., Chukharev-Hudilainen, E., & Hirch, R. R. (2023). Assessing interactional competence: ICE versus a human partner. *Language Assessment Quarterly*, 20(4–5), 377–398. <https://doi.org/10.1080/15434303.2023.2237486>
- OpenAI. (2024). ChatGPT (May 2024 version) [Large language model]. <https://chat.openai.com/>
- Roever, C., & Ikeda, N. (2022). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing*, 39(1), 7–29. <https://doi.org/10.1177/02655322211003332>
- Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, 35(3), 331–355. <https://doi.org/10.1177/0265532218758128>
- Saeki, M., Takatsu, H., Kurata, F., Suzuki, S., Eguchi, M., Matsuura, R., . . . Matsuyama, Y. (2024). Intella: Intelligent language learning assistant for assessing language proficiency through interviews and roleplays. In V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, & K. Komatani (Eds.), *Proceedings of the 25th annual meeting of the special interest group on discourse and dialogue* (pp. 385–399). Kyoto, Japan: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.sigdial-1.34>
- Seedhouse, P. (2019). The dual personality of ‘topic’ in the IELTS speaking test. *ELT Journal*, 73(3), 247–256. <https://doi.org/10.1093/elt/ccz009>
- Seedhouse, P., & Nakatsuhara, F. (2018). *The discourse of the IELTS speaking test: Interactional design and practice*. Cambridge, UK: UCLES/Cambridge University Press.
- Spivey, N. N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Research Quarterly*, 24, 7–26. <http://www.jstor.org/stable/748008>
- Stemler, S., & Tsai, J. (2008). Best practices in interrater reliability three common approaches. In J. Osborne (Ed.), *Best practices in interrater reliability three common approaches* (pp. 29–49). Thousand Oaks, CA: SAGE Publications, Inc. <https://doi.org/10.4135/9781412995627>
- Sydorenko, T., Dombi, J., Agrawal, A., Thorne, S. L., Lee, J. I., & Tao, Y. (2024). Spoken dialogue systems and ChatGPT for second language pragmatics research. In K. Sadeghi (Ed.), *Routledge handbook of technological advances in researching language learning* (pp. 378–391). London, UK: Routledge.
- Timpe-Laughlin, V., Sydorenko, T., & Dombi, J. (2024). Human versus machine: Investigating L2 learner output in face-to-face versus fully automated role-plays. *Computer Assisted Language Learning*, 37(1–2), 149–178. <https://doi.org/10.1080/09588221.2022.2032184>
- Wagner, E. (2020). Duolingo English test, revised version July 2019. *Language Assessment Quarterly*, 17(3), 300–315. <https://doi.org/10.1080/15434303.2020.1771343>
- Xi, X. (2008). What and how much evidence do we need? Critical considerations in validating an automated scoring system. In C. Chapelle, Y. R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 102–114). Ames, IA: Iowa State University.
- Xi, X. (2023). Advancing language assessment with AI and ML—leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, 20(4–5), 357–376. <https://doi.org/10.1080/15434303.2023.2291488>

- Xu, J. (2015). *Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech*. Doctoral dissertation, Iowa State University, Iowa State University Digital Repository. <https://doi.org/10.31274/etd-180810-4474>
- 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>
- Young, R. (2011). IC in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426–443). London, UK: Routledge.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. The transcript of a sample dialog between a test taker and the SDS.

Appendix S2. IELTS rating scale.

Appendix S3. Test taker survey.

Appendix S4. Test taker semi-structured interview protocol.

Appendix S5. Coding scheme for interactional competence features.

Appendix S6. Findings of interactional features by score level.