Exploring the Potential of a Video-Mediated Interactive Speaking Assessment

ETS RR–19-05

Gary J. Ockey
Veronika Timpe-Laughlin
Larry Davis
Lin Gu

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Exploring the Potential of a Video-Mediated Interactive Speaking Assessment

Gary J. Ockey,[1] Veronika Timpe-Laughlin,[2] Larry Davis,[2] & Lin Gu[2]

1 Iowa State University, Ames, IA

2 Educational Testing Service, Princeton, NJ

The construct of oral ability is multifaceted, but due to technological and practical constraints, the majority of computer-delivered speaking assessments are designed to measure only certain aspects of this ability. Most notably, interactional competence, which we define as the ability to actively structure dialogic speech in real time, is often not assessed. Innovations in technology, namely, computer-mediated video, make it possible for test takers in different locations to see and speak with others in real time and may make it achievable for computer-based tests to assess more aspects of oral communication, including interactional competence. This report gives the findings from a study that explored to what extent computer-mediated video, namely, Skype, could function in conjunction with a platform designed to present four innovative speaking tasks that could conceivably assess a broad construct of oral ability. The overarching goals of this project were twofold. First, we aimed to determine (a) the degree to which current computer video-mediated technology can be used effectively to deliver assessments remotely and (b) the extent to which participants felt that four specific tasks could assess speaking ability by means of this technology. The speaking tasks included giving short responses to an interlocutor's questions, summarizing a proposal, defending a position in a group discussion, and giving a prepared presentation and responding to questions from other participants. Two data collections were conducted: one with all 72 participants located in the United States and one with all 74 participants located in China. The findings provide preliminary evidence that the stability of computer video-mediated technology varies considerably, with technical disruptions being relatively few in the U.S. trial but very frequent in the China context. Moreover, the findings suggest that participants viewed the tasks as generally representing interactive speaking activities that they encounter in the oral language use domain, affording them the opportunity to demonstrate their oral abilities, and that these tasks can be effectively completed in a computer video-mediated environment when technology cooperates.

Keywords Interaction competence; speaking; computer video-mediated technology; assessment; Skype

doi:10.1002/ets2.12240

Oral ability is multifaceted, but due to various constraints, most computer-based oral assessments are designed to measure only certain aspects of this ability. Oral ability has been broadly defined as "the verbal use of language to communicate with others" (Fulcher, 2003, p. 23). Others have developed more detailed definitions of oral ability. For example, Ockey and Li (2015) asserted that the construct of second-language (L2) oral communication includes the following: (a) interactional competence, which refers to the ability to actively structure dialogic speech in real time and requires the ability to take turns appropriately, employ opening and closing gambits, effectively respond to others, and negotiate and develop topics with appropriate pragmatic use for a given context—abilities requiring both comprehension and oral production (Nakatsuhara, 2012; Ockey, 2018); (b) appropriate use of phonology, which relates to the effective use of both segmental and prosodic aspects of language; (c) appropriate and accurate use of vocabulary and grammar, which relates to vocabulary and grammatical breadth, that is, how many words and grammar structures an individual knows, and vocabulary and grammatical depth, or how well and effectively an individual can use these words and structures; and (d) appropriate fluency, which refers to naturalness of rate of speech, pausing, and repetition/language repair.

While appropriate and accurate use of phonology, vocabulary, and grammar, as well as appropriate fluency, may be assessed by many computer-delivered assessments, computer-delivered oral assessment tasks are generally asynchronous, and research has suggested that interactional competence may not be effectively assessed by such asynchronous tasks (Brooks & Swain, 2014; Ockey, Koyama, Setoguchi, & Sun, 2015; Theodoropulos, 2012).

Our overall goal was to create a computer-based assessment that would measure a broader construct of oral ability, one that includes interactional competence. Prior to the current study, we conducted preliminary research to identify

Corresponding author: V. Timpe-Laughlin, E-mail: vlaughlin@ets.org

a computer-mediated environment that would afford test takers an opportunity to demonstrate their oral abilities by directly conversing with each other. This effort included three such environments: Skype, a dedicated platform; WebEx, a Web-based environment; and Metamersive, a virtual environment that we developed (Ockey, Gu, & Keehner, 2017). Results from this preliminary study indicated that Skype provided a more stable connection than WebEx and was preferred by most test takers over the virtual environment because test takers wanted to see each other—not an avatar. We also learned from test takers' engagement with the virtual environment that they preferred to have the information necessary to take the test, for example, the prompts, presented by the technology rather than on paper, as we had with Skype and WebEx.

The purpose of the current study was to determine to what extent Skype could function in conjunction with a delivery platform designed to present four tasks that could conceivably assess our broad construct of oral ability. To be clear, this exploratory study was not designed to determine if the four tasks could actually assess the broad construct of oral ability; rather, it aimed to determine if Skype video chat technology could be used as part of the test delivery model (Mislevy, Almond, & Lukas, 2003) to aid in the effective presentation of these four tasks. Additionally, we aimed to determine the extent to which the test takers felt the test tasks were effectively delivered and provided them opportunities to demonstrate their oral abilities.

Technology

Limited research does exist on the feasibility of video-mediated technology for delivering oral assessments. For instance, face-to-face oral interactions have been compared to video-mediated interactions (Clark & Hooshmand, 1992; Cohen, 1982; Nakatsuhara, Inoue, Berry, & Galaczi, 2015; O'Conaill, Whittaker, & Wilbur, 1993; Rouhshad, Wigglesworth, & Storch, 2015). In general, these studies have shown that there are minor differences in these interactive media. For instance, Sellen (1995) found that the face-to-face mode "produced more interruptions and fewer formal handovers of the floor" (p. 401) than the video-mediated condition, while Rouhshad et al. (2015) found more incidences of negotiation of meaning in face-to-face mode than in computer video–mediated mode. Similarly, Nakatsuhara et al. (2015) observed minor differences in examinee output and examiner behavior in face-to-face and video-mediated versions of the IELTS speaking test. For example, in the video condition, test takers produced more clarification questions, while examiners reported speaking more slowly and articulating more carefully, although it should be noted that such differences had little impact on scores. Some of the early studies also reported various technological problems. For instance, while scores were not significantly different across media in Clark and Hooshmand's (1992) study, participants indicated that they preferred the face-to-face mode because of computer freezes and other technological glitches. In Nakatsuhara et al.'s (2015) study, the technology generally performed correctly, although test takers and examiners still reported issues with sound quality, which contributed to the differences in test-taker and examiner behavior.

Limited research comparing video-mediated and audio-only remote spoken interaction has been conducted. One exception is unreported research that was completed to import the task-delivery system in this study. Two speaking delivery modes were compared, one that was video mediated (Skype) and one virtual environment. In the virtual environment, test takers could see avatar representations of their speaking partners rather than actual video of their partners (Ockey et al., 2017). Results indicated that most test takers preferred to see their interlocutors because it was more natural and easier to understand what they said. However, the extent to which the different delivery modes actually led to different communication was not investigated.

Oral Tasks

Many different tasks types have been used to assess oral ability. The four tasks used in the current study were chosen based on the desire to create an assessment that assesses all four of the aspects of oral ability in Ockey and Li's (2015) construct. These four tasks were (a) answering short questions about opinions and experiences posed by a moderator, (b) listening to a short video and retelling the content, (c) discussing the content of the video with other participants, and (d) presenting a brief prepared presentation and then answering questions from other participants. These tasks covered a variety of speaker configurations, including monolog, pair discussion with the moderator, and group discussion among participants. The aim of using group tasks was to provide test takers opportunities to demonstrate various aspects of their interactional competence, including the ability to appropriately respond to information (which would imply the ability to

comprehend content) as well as their fluency, pronunciation, and vocabulary/grammar. Alternatives to group discussion were also included because these types of oral tasks are commonly encountered in the target language use domain, and research has suggested that the personal characteristics of other test takers might impact the scores of test takers who are assessed in groups (e.g., O'Sullivan, 2002; Ockey, 2009). Thus, it may not be fair to test takers if their scores are based solely on group performance. Tasks were also chosen that were well documented in the language assessment literature and that were felt to closely represent tasks that test takers might encounter in the real world. Test tasks that closely simulate real-world tasks to which the inferences are meant to be drawn are more likely to be valid than ones that do not (Bachman & Palmer, 1996).

Responding to short questions, as done in Task 1, has been used for assessing L2 oral ability for decades. Such responses are almost always the major components of a one-on-one oral interview, which is the most commonly encountered direct test of oral ability (Luoma, 2004). In this format, the interviewer has control, making it possible to elicit targeted speech acts and maintain a rather high level of consistency in eliciting a language sample. Moreover, providing a short response to a question is a common task in the real world. For instance, many service encounters require employees to provide short answers to questions about products or services. Being able to respond appropriately to such questions would provide an opportunity to demonstrate certain aspects of interactional competence as well as the other aspects of oral ability defined by Ockey and Li (2015).

Presenting a summary of unprepared information, as in the second task, has become an increasingly popular method for assessing oral ability. It requires test takers to comprehend information and then paraphrase this information. Short summary tasks are commonly encountered in the real world, for example, in such contexts as relaying information about a product in a service encounter. Research on this type of task has suggested that it can be used to effectively assess test takers' oral ability. For instance, Frost, Elder, and Wigglesworth (2011) found that for oral summary tasks, the accuracy with which information was reproduced or reformulated could be used to effectively place test takers into ability levels.

In recent years, paired and group oral discussion, as in Task 3, has emerged as a common task for assessing a test taker's ability to communicate interactively (Bonk & Ockey, 2003; Ockey, 2014). In a typical group oral task, test takers are instructed to talk among themselves. The examiner has minimum, if any, involvement in carrying out the oral activities. This lack of control by the examiner makes it difficult to target specific speech acts, but it makes it possible to assess aspects of interactional competence, such as negotiation of meaning, opening and closing gambits, and turn taking (Brooks, 2009). Because of the relative lack of control by the examiner, group assessments have been shown to have lower reliability than one-on-one oral interviews (Van Moere, 2006), and it is recommended that group discussions be used in conjunction with other tasks when assessing oral ability (Bonk & Ockey, 2003).

The presentation of prepared information, as in the fourth task, is very common in many settings. For instance, it is a major type of oral task in university settings and often directly contributes to academic or professional success (Ferris & Tagg, 1996; Zappa-Hollman, 2007). Both graduate and undergraduate students engage in class presentations, and graduate students give conference presentations, course lectures, and dissertation or thesis defenses. Presentations are also commonly encountered in business settings where products, proposals, or other information may need to be delivered and discussed in formal or informal situations. The presentation task type provides opportunities for test takers to demonstrate their abilities in many aspects of the oral ability construct, including interactional competence in responding appropriately to questions by the other test takers. It also requires the other test takers to demonstrate certain aspects of interactional competence, including "interactive listening" (Ducasse & Brown, 2009).

To summarize, when used together, these four task types were selected with the aim of assessing a broad construct of oral ability, one that includes interactional competence, pronunciation, grammar/vocabulary, and fluency. Moreover, the tasks are representative of ones that test takers are likely to encounter in real-world communication.

Project

The overarching goals of the project were twofold. First, we aimed to determine (a) the degree to which current computer video-mediated technology can be used effectively to deliver assessments remotely and (b) the extent to which participants felt that the four tasks mentioned in the literature review could assess oral ability by means of this technology to indicate this important group of stakeholders' satisfaction with the assessment system.

Research Questions

To achieve these goals, the following, more specific research questions were investigated.

Technology Oriented

1. To what extent is computer-mediated video technology feasible (in terms of usability and stability) for conducting interactive remote oral assessments?
2. What are participant perceptions of the effectiveness of the use of computer video–mediated technologies on oral tests?

Task Oriented

3. To what extent do participants believe that the following test tasks, when delivered in a computer-video environment, represent tasks they encounter in the real world and provide them opportunities to demonstrate their oral abilities?
    a. Giving short responses to questions asked by an interlocutor
    b. Summarizing a proposal
    c. Defending a position in a group discussion
    d. Giving a prepared presentation and responding to questions from other test takers

Methods

Overall, two trials were conducted in 2014 and 2015. In late 2014 and early 2015, a trial was conducted with participants located in the United States at three U.S. academic institutions: a community college in the West, a university in the Midwest, and a university located in the eastern United States. A second trial was conducted in China in June 2015 involving schools and universities in three different Chinese cities. During both data collections, the moderator was located at Educational Testing Service in Princeton, New Jersey, while each participant was located in a geographically separate location.

U.S. Trial Participants

Seventy-two university-level English-language learners ($N = 72$) participated in the U.S. study. Participants were recruited by local site coordinators, who were staff of each local institution. A total of 38 men and 34 women participated in the study and were between 22 and 45 years of age. English language proficiency of the participants varied considerably, although based on their performance in the study, individuals recruited from the community college appeared to have relatively low proficiency (roughly low intermediate), while many individuals recruited in the Midwest were undergraduate and graduate students with relatively high proficiency (approximately high intermediate to advanced). A mix of students and community members were recruited in the East and showed a range of proficiency. First languages of the participants included Arabic ($n = 5$), Chinese ($n = 21$), French ($n = 2$), languages indigenous to India ($n = 10$), Japanese ($n = 13$), Korean ($n = 9$), Malay ($n = 1$), Portuguese ($n = 2$), Russian ($n = 2$), Spanish ($n = 4$), Turkish ($n = 1$), and Vietnamese ($n = 1$).[1] The largest group of participants had been in the United States less than 6 months ($n = 31$), but a fair number had more experience in the United States: 6 months to 1 year ($n = 11$), 1–2 years ($n = 18$), and more than 3 years ($n = 11$). Roughly half were graduate students ($n = 37$), and the others were undergraduates ($n = 34$). Participants majored in a variety of areas: natural sciences ($n = 20$), business ($n = 23$), social sciences ($n = 16$), and humanities ($n = 12$). All students indicated that they spent more than an hour using a computer every day, the majority of whom indicated that they spent 4 or more hours per day on the computer ($n = 62$). Most participants reported using video chat regularly: once a month ($n = 21$), weekly ($n = 18$), multiple times per week ($n = 16$), and daily ($n = 9$), suggesting that most were familiar with the technology used in the study.

China Trial Participants

Adult L1 Chinese English language learners were recruited from each of three cities in China: Guangzhou, Nanjing, and Shanghai. A total of 74 learners, 13 men and 61 women, participated in the study. In Guangzhou and Shanghai, the participants were undergraduate students who majored in English at public 4-year universities. In Nanjing, the participants were recruited from students who attended business English classes at a private language learning institution. Except for one, all participants were between 22 and 45 years old. Except for two students, all participants reported that they were studying at a university, being enrolled either in an undergraduate program ($n = 54$) or in a master's degree program ($n = 18$). Among the university participants, most students ($n = 66$) majored in areas that were related to English language, such as English literature, English translation, and medical English. English proficiency of participants in all locations was typically intermediate to upper intermediate, based on their performance in the study tasks.

In terms of English learning experience, about 87% reported that they had studied English for 9 years or more, while approximately 77% reported that they attended English classes more than 5 hours a week. At the time of data collection, the majority of participants (about 80%) had never visited an English-speaking country. Those who had experience abroad reported that their stays in an English-speaking country ranged from 2 weeks or less to more than 2 years. When being asked to self-rate their oral English-speaking ability, about 25% said that they could easily communicate complex ideas in English. Approximately half of the participants reported that they could communicate complex ideas but needed to work hard to do it, while the rest (i.e., 25%) indicated that they needed to work hard to communicate even basic ideas in English.

With regard to experience with computer technologies, especially computer video-mediated technologies, the majority of the participants (90%) indicated that they spent more than 4 hours per week using a computer. Moreover, two-thirds of the participants reported using video chat regularly — once a month ($n = 10$), weekly ($n = 11$), multiple times per week ($n = 11$), and daily ($n = 17$) — suggesting that most of them were familiar with the technology used in the study. The remaining one third reported that they very rarely or never used a computer to video chat.

When asked about the number of people they usually chatted with via video, about 35% indicated that they chatted regularly with two or more people at the same time. The most frequently cited purposes were to chat with friends, reported by 53 participants, and to chat with family, reported by 44 participants. Only six participants reported that they used video chat for work or schoolwork. Most participants (about 75%) reported that they were either very comfortable or comfortable with learning new computer programs and technologies. Furthermore, all participants expressed an interest in learning new computer programs and technology.

Oral Tasks

As shown in Figure 1, each session was delivered through an online interface that featured the Skype window on the left and the Educational Testing Service (ETS) delivery platform on the right. Participants could see each other as well as the moderator via Skype and were guided through each task by verbal input from the moderator and instructions provided on the ETS platform. A local coordinator was also present to supervise the session and help with any technical problems.

Each session consisted of four tasks, as mentioned earlier, presented in the following order: (a) responding to a set of short questions, (b) summarizing a proposal, (c) defending a position in a small-group discussion, and (d) presenting information and responding to questions. Participants were not allowed to take notes during any of the four tasks.

Task 1: Responding to Short Questions

The moderator asked each participant three questions. First, participants were given 10 seconds each to briefly introduce themselves. Following the introduction question, each participant was given 30 seconds to respond to each of the next two questions (i.e., Questions 2 and 3), which were different for each participant. In an attempt to maximize the interactive nature of this task, Questions 2 and 3 were designed as follow-up questions, intended to elicit participants' opinions about a rather accessible topic in a business context. The following questions showcase an example:

Q2: Do you prefer to shop online or to go to a store? Why?

Q3: Are customer reviews on Web sites important for deciding whether to buy a product? Why or why not?

After the moderator read each question, a countdown timer appeared on the guidance screen to help the participant gauge the amount of time to answer the question (see Figure 2).

Figure 1 Interface for the interactive oral tasks.

Task 2: Summarizing a Proposal

First, the moderator very briefly described a scenario where a company faced a problem, such as the need to reduce stressful working conditions or to increase recruitment of new employees. Participants were asked to imagine that they were employees of the company, and then three proposals for dealing with the problem were presented by coworkers, with each appearing in a short video. Participants were told that they would hear three proposals and would be asked to summarize one of them. They were also told not to take notes while listening to the proposals. Although everyone watched all three video proposals, each participant, prior to listening, was assigned by the moderator to summarize the next proposal.

Proposals were scripted and written to be similar in terms of word count (approximately 200 words), word frequencies, and grammatical difficulty. Actors featured in the video proposals were L1 speakers of U.S. English instructed to speak clearly and naturally and, to the extent possible, pace and pauses were controlled across speakers. Videos of the proposals were presented on the right side of the screen, as shown in Figure 3. Immediately following the conclusion of the video, the participants were given 45 seconds to summarize the proposal content. An example of a proposal script is provided in Appendix A.

Task 3: Defending a Position in a Small-Group Discussion

Participants were asked to have an informal discussion on the workplace problem introduced in Task 2, continuing in their roles as employees of the given company. The purpose of the task was to demonstrate the ability to engage in a discussion by interacting with professional peers. Before engaging in the discussion, the moderator informed the participants that they should each argue in favor of the proposal they summarized in Task 2, using information from the proposal as well as their own experiences and logic to convince their partners that their option was the best. In addition, although participants were encouraged to refer to proposal content, they were also asked to avoid "resummarizing" the proposal during the discussion. The following text is an example of the specific instructions given to participants:

> The three of you will now have an informal discussion about this topic. The purpose of this activity is to demonstrate that you can engage in a discussion with professional peers. Now that each of you've just summarized one option on how to cut costs, try to convince your partners that your option is the best and the other two options are not as good. You do not have to come to an agreement. You have also already summarized your position, so another

Figure 2 Task 1: Responding to short questions.

summary isn't needed. But feel free to use the information from the input in the summary tasks as well as draw on your own experiences and logic to convince your partners that your opinion is the best.

Key words to remind the participants of the three proposals were displayed in the platform during the "Discussion" section. Six minutes were allotted for the discussion, and a countdown timer was displayed on the screen (see Figure 4)

Task 4: Presenting Information and Responding to Questions

For Task 4, all participants gave a short prepared presentation on a business-related topic of their choice. At least 2 days prior to the date of the data collection session, they were provided with guidelines for preparing and giving a 2-minute presentation (see Appendix B). The presentation needed to be visually supported by a one-page presentation slide, produced by the participant according to the guidelines, which was uploaded to the platform prior to the beginning of the session (see Figure 5). While the presenter spoke, this visual was displayed and was intended to help provide context for the presentation. Following the presentation, there was a 2-minute question-and-answer session in which the other participants asked the presenter questions. The moderator did not participate in this discussion.

Questionnaires

Background Questionnaire

A survey was designed to collect participants' demographic information, their English-language-learning background, and their use of computers and computer-mediated video technology (see Appendix C). The information about test takers' backgrounds provided in the participants' section was gleaned from this questionnaire.

Tasks Questionnaire

A questionnaire was designed to obtain participants' perceptions and feedback on each task regarding (a) the extent to which they found the tasks to be engaging, (b) the extent to which the tasks represented real-world activities in which the participants engage on a regular basis, and (c) the extent to which these tasks functioned effectively in assessing their

Figure 3 Task 2: Summarizing a proposal.

oral skills, given the technology used to present them. For each question, participants were asked to indicate to what extent they agreed with a statement on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). This questionnaire is provided in Appendix D.

Technology Questionnaire

A technology questionnaire was designed to determine (a) the extent to which the Skype video technology was effective in delivering the tasks, (b) the extent to which the ETS-built guidance platform was effective in delivering the tasks and task instructions, and (c) the extent to which communication via Skype resembled face-to-face interaction. As with the task questionnaire, Likert-scale items were used to indicate participants' agreement with various statements. The questionnaire is provided in Appendix E.

Focus Group

While the questionnaires were administered for both the U.S. and Chinese trials, an additional focus group was conducted as part of the China data collection to obtain information from participants about their perceptions regarding tasks, technology, and interactivity (see Appendix F for the focus group protocol).

Equipment

In the U.S. trial, each of the three data collection sites was provided with identical laptop computers: Lenovo Enhanced Experience 3 with a Core i5 Intel processer in a Lenovo ThinkPad laptop featuring Windows 7. The laptop computers at each of the three sites were hard-wired to an Internet connection during data collection. The moderator at ETS also used a similar Lenovo ThinkPad. In the China trial, it was not possible to provide computers to each site, and so equipment that was already in place was used for the study, once again using a wired Internet connection. In both trials, participants and moderator wore headphones during the session.

Figure 4 Task 3: Group discussion.

Procedures

Prior to each session, participants were asked to prepare a 2-minute presentation and an accompanying slide and were provided with the guidelines shown in Appendix B. They were asked to bring their presentation slide on a memory stick to the data collection session for which they were scheduled. After a participant arrived, the site coordinator at each of the three sites welcomed the participant, asked him or her to complete a consent form, and uploaded the presentation slide to the testing platform. A name selected by the participant by which he or she would be called during the assessment was also uploaded to the testing platform by the site coordinator. This name was displayed to the right of Skype window so the other participants could refer to each other throughout the activities. After the technology was ready, the participant was given a set of instructions. The participant was told that (a) the focus of this research study was to test the technology; (b) he or she should not be concerned about technical problems or questions that seemed harder or easier than others in their group; (c) the participant would be speaking to either one or two other participants, who were at one of the other institutions involved in the study; and (d) he or she would be guided by a moderator at ETS. Then, the participant was placed in front of a computer and handed a set of USB headphones. The moderator, a researcher at ETS in Princeton, NJ, then guided the participant through each of the four tasks. Participants completed the sessions in groups of two or three. Data collection sessions were conducted in U.S. locations from November 2014 through February 2015, while data collection in China took place in June 2015.

Two particular differences between the U.S. and China trials need to be noted. First, during the U.S. data collection, participants were instructed to stand up when giving their presentation. This approach was taken to mimic the real-life situation of giving a professional presentation. However, owing to technological issues when rearranging the computer screen and camera, this approach was dropped for the China trial, and participants remained seated throughout data collection sessions. Second, the final step also differed slightly between trials. In the U.S. trial, the participants were logged out of the system after completing all tasks. Then the site coordinator asked them to complete the task, technology, and biographical information questionnaires (Appendices C, D, and E) before paying and thanking them for their participation. In the China trial, participants remained online after completing the session, and an L1 Chinese-speaking professor from one of the participating universities logged on to lead a 15-minute focus group in Chinese, during which the participants were asked about their experience with the tasks and the technology. After completing the focus group, respondents in

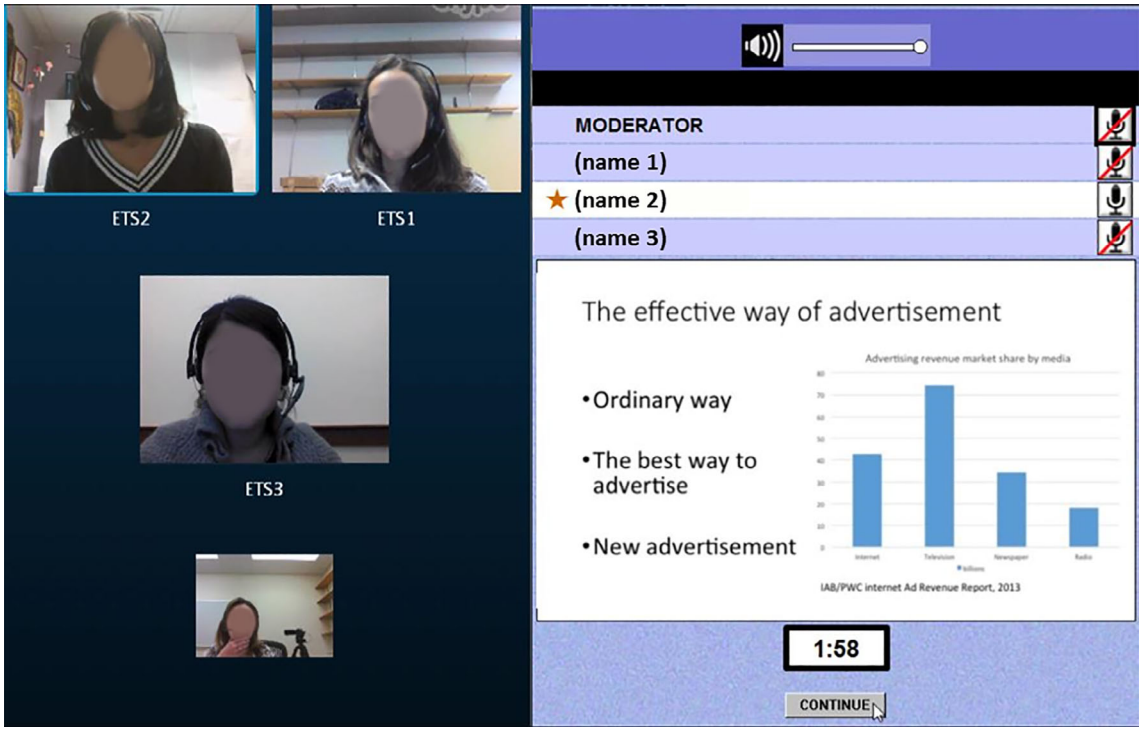


Figure 5 Task 4: Prepared oral presentation and question and answers.

China also completed the same three questionnaires and were paid and thanked for participating. In both data collections, the sessions and — in the case of China — the subsequent focus group were video recorded using Camtasia screen capture software. All 28 U.S. sessions were recorded, while of the 25 China sessions, the recording for one session was lost due to a system error. Each data collection session took approximately 45 – 60 minutes per group of participants.

Results and Discussion

In the following, we present the questionnaire responses and focus group comments for the two main areas of interest: technology and tasks. We first focus on the technology, including both the computer video-mediated *Skype* technology and the ETS-developed task-delivery platform. Then we present the participants' perceptions of the four tasks used to assess oral ability by means of video-mediated technology. Throughout, we compare findings from both the U.S. and China trials.

RQ1: Feasibility of Computer Video-Mediated Technology

To determine the extent to which computer-mediated video technology is feasible in terms of usability and stability for conducting interactive oral assessments across geographically distant locations, we examined video from each of the 28 U.S. and 24 China sessions captured and compared this information to responses from the technology questionnaire. The results are summarized in Table 1.

U.S. Trial

Although the large majority of the sessions were technological successes, we encountered some problems associated with technology and/or human error. Most (18 of the 28) sessions had no technology problems, excluding human error, but some had minor problems, and a few had major technological problems. The majority of problems overall occurred during the first few days of data collection. In seven of the sessions, problems resulted from human error, such as having the headset muted, uploading the wrong slide, accidentally disconnecting the Internet cable when the student stood up

Table 1 Frequency of Technology and Human Error Issues

Issue	United States,[a] *n* (%)	China,[b] *n* (%)
Sessions with issues due to technology	10 (36)	24 (100)
Slides not visible to all members	5 (18)	12 (50)
Video of proposals not visible	4 (14)	Common[c]
Skype call dropped (audio and video)	3 (11)	5 (21)
Dropped video (but audio functioned appropriately)	3 (11)	22 (92)
Sessions with issues due to human error[d]	7 (25)	–

Note. The number of specific problems listed is greater than 10, given that sometimes more than one problem occurred in a session. [a]$N = 28$. [b]$N = 24$. [c]Time lag in the delivery of the video occurred with differing frequency and severity across sites, and it was not possible to establish an exact frequency. [d]For example, mute button on, wrong slides uploaded, Internet cord disconnected during session, not ready for session.

to give a presentation, and not having the system ready at the start of the session. In eight of the sessions, participants were not always able to see the information delivered by the platform. The presentation slides were not visible to all participants in six sessions, and the video of the proposal (Task 2) was not visible in four sessions. The most serious problems occurred in two sessions when Skype calls were dropped; both were probably caused by human error when moving the laptop from standing to sitting, or vice versa. In two other sessions, computer screens appeared to freeze temporarily (17 s in one session and 1 min 19 s in the other) while the call was automatically reconnected, but these disruptions were quite short in duration and may not have had a critical impact on the participants and their oral communication. Three sessions were conducted that included visual problems with Skype, but the problems were not associated with disruption to audio communication. A summary of these results is provided in Table 1.

China Trial

In the China trial, technology issues were much more frequent, oftentimes impacting the progress of the sessions. While the majority of sessions during the U.S.-based trial encountered no technology problems, each China session was challenged by difficulty of some kind, sometimes with instances of several different kinds of problems. For example, dropped video in the Skype feed was observed in most sessions (92%) and affected all three of the sites in China but was especially common in one location where there seemed to be less Internet bandwidth. In most cases, the video would return after a period of time; however, at one site, video capability was sometimes unavailable for an entire task. Participants whose video feed was dropped also reported not being able to see the video feeds of other individuals. In contrast to the U.S.-based data collection, there were also more instances of completely dropped calls (21%). Challenges in content delivery were also common. Approximately 50% of the time a presentation slide could not be viewed by one or more participants, and particularly common were delays in the start of the video presentation in Task 2 and the display of slides in Task 4. That is, participants in China sometimes reported a delay in the start of the video, although the video was already playing for the moderator. The issues in displaying participants' presentation slides in Task 4 were partly due to operator error in uploading slides at the beginning of the session and in some cases were apparently due to connectivity problems at the local test site. Overall, technology issues were much more common in the China trial than in the U.S. trial, which appeared to be largely due to problems in Internet connectivity.

RQ2: User Perceptions of the Effectiveness of Computer Video-Mediated Technology

Responses to the technology questionnaire were analyzed to understand test takers' perceptions of the effectiveness of the use of computer video-mediated technologies for oral assessment tasks. The results for both data collections are presented in the following tables; as an aid to summarizing patterns in the data, responses of "agree" and "strongly agree" were aggregated into a category labeled "high" and represent an approval rating, while responses of "disagree" and "strongly disagree" were combined into a category labeled "low." Means above 4 on the 5-point scale are discussed as indicating general agreement with the statement presented; however, this cutoff is somewhat arbitrary and is not meant to indicate statistical significance.

Table 2 Test-Taker Perceptions of the Effectiveness of Using Skype Technology

	User friendly		Engaging		Visual clarity		Clear speaker ID		Seeing others helped communication		Good sound quality		Easy to express self		Easy to respond		Tech. interruptions		Tech. pauses	
Statistic	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN
N	72	74	72	74	72	74	72	74	72	74	72	74	72	74	72	74	72	74	72	74
Mean	4.2	3.9	4.3	4.3	4.4	3.4	4.5	4.1	4.3	4.1	4.4	4.0	3.9	3.8	4.2	3.8	2.1	4.0	2.1	3.8
SD	0.8	0.9	0.7	0.7	0.8	1.3	0.6	0.9	0.7	0.8	0.8	1.0	1.0	0.8	0.8	0.8	1.1	1.0	1.0	1.1
Frequency																				
High	63	57	64	66	69	38	69	63	61	60	67	59	47	46	57	51	9	61	6	49
Low	2	6	2	0	3	22	3	6	1	4	4	8	4	2	2	5	49	8	49	11
Strongly agree	27	17	28	30	36	19	40	25	30	23	41	22	24	14	30	15	3	26	1	22
Agree	36	40	36	36	33	19	29	38	31	37	26	37	23	32	27	36	6	35	5	27
Neutral	7	11	6	8	0	14	3	5	10	10	1	7	21	26	13	17	14	5	17	14
Disagree	0	5	2	0	2	18	0	6	1	4	4	7	3	2	2	5	23	5	23	8
Strongly disagree	2	1	0	0	1	4	0	0	0	0	0	1	1	0	0	0	26	3	26	3

Note. CN = China; US = United States; 5 = strongly agree; 4 = agree; 3 = neither agree nor disagree; 2 = disagree; 1 = strongly disagree.

Overall, participants' reactions to the Skype technology used to deliver the tasks were generally positive (Table 2). The large majority of participants in both trials indicated that the computer video-mediated environment and platform were engaging (U.S. $M = 4.3$, China $M = 4.3$) and that it was easy to identify who was speaking (U.S. $M = 4.5$, China $M = 4.1$). Participants also felt that seeing others helped with communication (U.S. $M = 4.3$, China $M = 4.1$) and that the sound quality was good (U.S. $M = 4.4$, China $M = 4.0$). Smaller majorities of participants in both trials agreed that use of Skype made it easy to express themselves (U.S. $M = 3.9$, China $M = 3.8$).

In other areas, reactions to the Skype technology differed across trials, with differences likely related to the much higher frequency of technology problems in the China trial. For example, only a small number of U.S.-based participants agreed with statements that there were inappropriate interruptions ($n = 9$) or unnatural pauses ($n = 6$) because of technology, suggesting that generally, technological problems were not a common concern. By contrast, the majority of China-based participants agreed with statements that they experienced interruptions ($n = 61$) and unnatural pauses ($n = 49$) due to technology issues. Similarly, U.S. participants were more likely to report good visual clarity (they could identify who was speaking; U.S. $M = 4.4$, China $M = 3.4$), although most participants in both trials reported that they could easily identify who was speaking, suggesting that visual input was not critical for speaker identification. A higher frequency of U.S. participants also regarded the video-mediated technology as more user friendly (U.S. $M = 4.2$, China $M = 3.9$) and found it easy to respond to others through the technology (U.S. $M = 4.2$, China $M = 3.8$). Overall, these results are consistent with the frequency of technical problems in the U.S. and China trials reported earlier in Table 1.

Participant reactions to the platform used to deliver task content were also generally positive (Table 3). Large majorities in both trials agreed that the onscreen instructions were useful (U.S. $M = 4.3$, China $M = 4.5$) and that it was helpful to have a timer to monitor time remaining to respond (U.S. $M = 4.5$, China $M = 4.3$). In the U.S. trial, 15 participants also made open-ended comments in response to the question regarding the timer, but comments focused on the time allowed to make a response rather than the visual affordance of a countdown timer on the screen. Reactions were mixed on this point: Some participants felt that the time given was too long, others thought that it was still short, and several felt that there should be no timer at all.

In addition to overall clarity, clarity of the slides used in the presentation (Task 4) was reported slightly differently between the two data collections. Participants in the U.S. trial were less likely to agree that the slides were clear than were China-based participants (U.S. $M = 3.7$, China $M = 4.0$), a difference that may be related to the fact that in the U.S. data collection, speakers were asked to stand while giving their presentations. One U.S. participant remarked,

> It is almost impossible to read the slides (esp. the data in the inserted image) while I was standing during the presentation. I think the windows for participants should be minimized, so that the presenter can see what they are presenting.

Table 3 Technology Questionnaire: ETS-Built Delivery Platform

	Instructions useful		Timer useful		Slides clear		Easy to upload		Screen buttons	
Statistic	US	CN	US	CN	US	CN	US	CN	US	CN
N	72	74	72	74	72	74	70	74	70	74
Mean	4.3	4.5	4.5	4.3	3.7	4.0	4.0	4.0	3.9	3.9
SD	0.7	0.7	0.8	0.9	1.4	1.0	0.9	0.9	0.8	1.0
Frequency										
High	65	70	62	63	45	59	49	60	41	52
Low	2	1	2	5	17	7	3	7	0	7
Strongly agree	34	45	44	38	29	27	24	21	21	23
Agree	31	25	18	25	16	32	25	39	20	29
Neutral	5	3	7	6	10	8	18	7	29	15
Disagree	2	1	2	4	9	5	2	6	0	6
Strongly disagree	0	0	0	1	8	2	1	1	0	1

Note. CN = China; US = United States.

Table 4 Technology Questionnaire: Comparison of Tasks to Face-to-Face Experience

	Effective interaction		Easy-to-complete task		Like real-world speaking		Like technology-mediated communication		Different from face-to-face communication	
Statistic	US	CN	US	CN	US	CN	US	CN	US	CN
N	70	74	70	74	70	74	61	74	69	74
Mean	4.2	3.9	4.4	4.0	4.1	3.6	4.2	3.9	3.1	4.1
SD	0.7	0.8	0.6	0.77	0.8	1.0	0.7	0.9	1.1	0.7
Frequency										
High	61	55	67	60	57	43	55	54	33	64
Low	1	4	0	4	4	11	1	7	22	4
Strongly agree	24	15	30	17	23	10	22	17	2	19
Agree	37	40	37	43	34	33	33	37	31	45
Neutral	8	15	3	10	9	20	5	10	14	5
Disagree	1	3	0	4	4	10	1	7	13	4
Strongly disagree	0	1	0	0	0	1	0	0	9	0

Note. CN = China; US = United States.

Another suggested using dual-screen monitors to make everything more visible for participants. These comments were taken into account during the second, China data collection, where participants sat while delivering their presentations. No open-ended comments were made by China-based participants regarding the visibility of the presentation slides.

Additionally, five questions were included in the questionnaire that asked participants to compare the video-mediated environment to their daily face-to-face interactions (Table 4). Participants in the U.S. trial were generally more positive than their China-based counterparts; large majorities in the U.S. trial indicated that the computer-mediated environment allowed them to effectively interact ($M = 4.2$) and complete the tasks ($M = 4.4$). Moreover, U.S. participants felt that the tasks were similar to their real-life technology-mediated communication, such as FaceTime ($M = 4.2$). Participants had more mixed views on the extent to which the computer video-mediated environment was like their daily face-to-face communications, with about half of the U.S. group agreeing with the statement that communication was different between the two modes. Although participants mentioned a variety of reasons for this difference, a common comment was that during computer-mediated communication, it was difficult to make direct eye contact or maintain a social presence through gestures or physical contact.

Large majorities of China-based participants similarly expressed general agreement with statements related to the usability of the technology, but the proportion of individuals indicating strong agreement with various statements was noticeably lower than for U.S. respondents (Table 4). For example, only 17 China participants (22%) "strongly agreed" that the computer-mediated environment allowed them to easily complete the task compared to 30 (43%) of U.S. participants. This difference is perhaps not surprising given the much higher frequency of technical problems observed in the

Table 5 Task 1: Short Responses to Questions Asked by a Moderator

	Engaging		Low anxiety		Like real life		Opportunity to show ability to answer questions		Opportunity to get accustomed with others' speech varieties		Opportunity to show language ability	
Statistic	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN
N	72	74	72	74	72	74	72	74	72	74	71	74
Mean	4.1	3.8	4.1	4.0	4.2	3.4	4.2	3.9	4.2	4.4	4.0	3.7
SD	0.8	0.8	0.9	1.0	0.8	1.1	0.9	1.0	0.8	0.8	0.9	0.9
Frequency												
High	61	53	54	57	60	35	59	53	63	66	51	47
Low	5	6	4	9	2	19	5	8	3	2	4	7
Strongly agree	22	15	29	26	30	12	31	22	27	36	24	15
Agree	39	38	25	31	30	23	28	31	36	30	27	32
Neutral	6	15	14	8	10	20	8	13	6	6	16	20
Disagree	5	6	4	8	1	18	4	6	3	2	3	6
Strongly disagree	0	0	0	1	1	1	1	2	0	0	1	1

Note. CN = China; US = United States.

China trial. Difference between the trials was especially evident when comparing interaction via the technology to face-to-face interaction, where a large majority (86%) indicated that the two modalities were different, compared to less than half of U.S. participants. Among China participants who indicated the two modalities were different, the most commonly cited reasons were that the technology provided relatively limited visual input (typically seen as a disadvantage) and that interacting online was less stressful than face-to-face communication with strangers. In their written comments, participants did not generally elaborate on why they felt online communication was less stressful, but oral comments made in the focus group session suggested that the lack of physical presence of the interlocutor was an important factor.

RQ3: User Perceptions of the Tasks

Responses to the task questionnaire and focus group comments were used to explore participants' perceptions of the tasks (i.e., giving a short response to a question asked by an interlocutor, summarizing a proposal, defending a position in a group/paired discussion, and giving a prepared presentation and responding to questions from other test takers). Participants were asked to judge if the tasks (a) were similar to those they might encounter in their oral language use in real life, (b) would afford them the opportunity to demonstrate their oral abilities, and (c) could be effectively completed in a computer video–mediated environment.

As can be seen in Table 5, participants in the U.S. trial and respondents in the China trial generally had positive opinions of Task 1, giving a short response to questions asked by an interlocutor. Large majorities of participants in both trials reported that they felt comfortable speaking in front of the other test takers (low anxiety). The U.S. participants were somewhat more likely to agree that the task was engaging (U.S. $M = 4.1$, China $M = 3.8$) and that the task gave them a good chance to show their ability to answer questions without preparation time (U.S. $M = 4.2$, China $M = 3.9$) or to demonstrate their speaking ability generally (U.S. $M = 4.0$, China $M = 3.7$). A greater difference was seen in response to the question "In real life, I use English to answer brief questions" (U.S. $M = 4.2$, China $M = 3.4$), which is perhaps not surprising given that U.S. participants were living in an English-speaking country at the time of data collection. However, Chinese participants were more positive than their U.S. counterparts in their reaction to the question "The activity gave me the chance to get accustomed to the other test takers' language varieties and accents" (U.S. $M = 4.2$, China $M = 4.4$), a finding that confirms our original, intended conceptualization of this first task. For example, regarding this aspect, one Chinese student commented that "the first part can help [him] get a chance to be more familiar with moderator and other group members."[2]

Table 6 displays the user perceptions concerning the second task, in which participants were asked to summarize a proposal delivered via video embedded in the guidance platform. Participants from both trials found the task quite engaging (U.S. $M = 4.1$, China $M = 4.3$). They were able to understand the video-delivered proposals (U.S. $M = 4.0$, China $M = 4.5$), suggesting that the task was not too difficult for the large majority of them. Moreover, they found the visuals quite useful

Table 6 Task 2: Summarizing a Proposal

	Engaging		Able to understand proposal		Taking notes would have helped		Visuals were useful		Had enough time to summarize		Like real life		Opportunity to show language ability	
Statistic	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN
N	72	74	72	74	72	74	71	74	72	74	72	74	72	74
Mean	4.1	4.3	4.0	4.5	3.9	3.8	4.1	4.1	3.6	3.8	3.8	3.0	3.9	4.1
SD	0.8	0.7	1.1	0.7	1.0	1.1	1.0	0.9	1.0	0.9	1.0	1.2	0.9	0.7
Frequency														
High	59	70	56	70	47	50	56	58	46	53	52	27	56	61
Low	3	2	8	1	7	14	7	5	13	9	9	30	5	1
Strongly agree	24	27	29	39	24	23	28	31	12	17	20	8	18	18
Agree	35	43	27	31	23	27	28	27	34	36	32	19	38	43
Neutral	10	2	8	3	18	10	8	11	13	12	11	17	11	12
Disagree	3	2	6	1	7	12	6	5	12	9	7	25	2	1
Strongly disagree	0	0	2	0	0	2	1	0	1	0	2	5	3	0

Note. CN = China; US = United States.

(U.S. $M = 4.1$, China $M = 4.1$), indicating that they felt video was an effective medium to deliver the input. Like for Task 1, participants in China were less positive than U.S. participants regarding the authenticity of the task, but compared to Task 1, both groups appeared to view the retell task as less authentic (Task 1, U.S. $M = 4.2$, China $M = 3.4$; Task 2, U.S. $M = 3.8$, China $M = 3.0$). However, individuals in the China trial viewed Task 2 as a better opportunity to demonstrate their English-language abilities than the previous task (Task 1, $M = 3.7$ vs. Task 2, $M = 4.1$).

The majority of respondents would have preferred to be able to take notes while listening to the proposal (U.S. $M = 3.9$, China $M = 3.8$), although opinions were mixed, especially for the Chinese participants. While 50 participants in the China trial agreed that taking notes would have helped them to complete the task more effectively, some noted in open-ended comments that such an approach would be less realistic. For example, one Chinese participant commented that "[t]aking notes will help student summarise, but it's take time to write down, normally, when people speaking, they will not take note." Perceptions also varied regarding whether the amount of time to complete the task was adequate (U.S. $M = 3.6$, China $M = 3.8$), where relatively narrow majorities agreed that they had enough time to summarize the proposal content. In open-ended comments, participants expressed mixed perceptions regarding preparation time. For example, one U.S. participant wrote, "It could have been better if at least 10 seconds are provided for preparation," while another felt that "[i]t was great to make a quick answer, of a proposal. It shows how better is my thinking and English."

Table 7 presents the results from the participant questionnaire for Task 3, defending a position in a group discussion. On average, Task 3 was considered the most engaging of all four activities (U.S. $M = 4.3$, China $M = 4.5$), with only a single candidate in the China trial indicating disagreement. Participants felt that it gave them an opportunity to demonstrate their ability to interact with others (U.S. $M = 4.1$, China $M = 4.2$) and show their language ability (U.S. $M = 4.1$, China $M = 4.2$). They were less positive about the authenticity of this task, particularly the Chinese group (U.S. $M = 3.7$, China $M = 2.8$), who rated this activity as the least authentic among all four tasks. We suspect that small-group discussion activities may not be as common in Chinese learning environments as in the U.S. context. A sizable minority of China participants (17 individuals, 23%) felt that the task did not provide equal speaking time. On the other hand, most disagreed with the statement that "defending my position made me feel uncomfortable" (U.S. $M = 2.2$, China $M = 2.4$).

While overall perceptions seemed to be rather positive regarding Task 3, both positive and critical perspectives were seen in open-ended comments. A number of participants appreciated the interactive nature of the group discussion, highlighting its value for the workplace domain. For example, one participant in the China trial noted that this discussion task "tests our critical thinking, which is really important for productive conversations. Many Chinese people tend to follow others which is not good for business." Others emphasized the opportunity to practice conversational, turn-taking skills, as in the case of one China participant who said, "I think it's good that it was not orderly because this pushed us to take our chance to talk." However, some students also raised concerns about this task format, specifically the potential impact of interlocutor on performance. For instance, one China participant questioned, "What if I'm a person who is not that aggressive at talking while the other two are quite different from me and I don't get many chances to defend myself?"

Table 7 Task 3: Defending a Position in a Group/Paired Discussion

	Engaging		Equal speaking time		Uncomfortable task		Like real life		Opportunity to show ability to interact		Opportunity to show language ability	
Statistic	US	CN	US	CN	US	CN	US	CN	US	CN	US	CN
N	72	74	72	74	72	74	72	74	72	74	72	74
Mean	4.3	4.5	3.9	3.2	2.2	2.4	3.7	2.8	4.1	4.2	4.1	4.2
SD	0.7	0.7	0.9	0.9	1.0	1.1	1.1	1.2	0.8	0.6	1.0	0.7
Frequency												
High	63	68	55	48	9	10	47	24	61	70	60	66
Low	0	1	7	17	51	50	12	33	3	1	5	1
Strongly agree	30	41	16	12	0	6	17	4	24	22	25	24
Agree	33	27	39	36	9	4	30	20	37	48	35	42
Neutral	9	5	10	9	12	14	13	17	8	3	7	7
Disagree	0	1	6	16	32	40	10	22	2	1	2	1
Strongly disagree	0	0	1	1	19	10	2	11	1	0	3	0

Note. CN = China; US = United States.

Table 8 Task 4a: Giving a Prepared Presentation

	Engaging		Skype was appropriate		Like real life		Opportunity to show ability to give a presentation		Opportunity to show language ability	
Statistic	US	CN	US	CN	US	CN	US	CN	US	CN
N	70	74	71	74	71	74	71	74	71	74
Mean	4.3	4.2	4.1	3.9	4.2	4.3	4.3	4.4	4.1	4.2
SD	0.7	0.7	0.9	0.8	1.0	0.7	0.7	0.6	0.9	0.7
Frequency										
High	64	62	58	57	59	63	64	72	61	67
Low	2	0	5	5	4	0	1	1	3	1
Strongly agree	29	24	26	17	33	31	32	29	25	26
Agree	35	38	32	40	26	32	32	43	36	41
Neutral	4	12	8	12	8	11	6	1	7	6
Disagree	2	0	5	5	2	0	1	1	1	1
Strongly disagree	0	0	0	0	2	0	0	0	2	0

Note. CN = China; US = United States.

Given that Task 4 consisted of two parts, giving a prepared presentation and responding to questions from other test takers, the results for these two subtasks are presented separately. We will begin with the findings for the presentation and follow them with the results obtained for the Q&A part.

As shown in Table 8, giving a prepared presentation was generally perceived quite positively. Participants felt that the task was engaging (U.S. $M = 4.3$, China $M = 4.2$), that Skype was quite appropriate for this task (U.S. $M = 4.1$, China $M = 3.9$), that the task was authentic (U.S. $M = 4.2$, China $M = 4.3$), and that the task provided them with a good opportunity to demonstrate their ability to give a presentation (U.S. $M = 4.3$, China $M = 4.4$) and show their language ability (U.S. $M = 4.1$, China $M = 4.2$). Despite the overall positive evaluation of this task, participants also raised particular concerns. For example, one China participant pointed out that "[g]iving a prepared speech may fail to get the real speaking standard of the candidates," while another wrote, "As far as this part, I think it is better to give presentation without the preparations. In that way, students would be allowed to show their abilities and thinking procedures." Still another China participant cautioned that "[i]t is difficult to avoid cheating in this task," perhaps referring to the possibility of memorizing a text written by a third party. In the U.S. trial, several participants commented that it was difficult to see their slides while standing up to make their presentations, and as mentioned, earlier presenters were allowed to sit in the China trial.

Table 9 summarizes perceptions related to the second part of Task 4, where the presenter responded to questions from other participants. Participants were generally positive about this activity and felt that the task was similar to their own

Table 9 Task 4b: Responding to Questions From Other Participants

	Like real life		Comfortable responding to questions		Comfortable asking questions		Opportunity to show ability by responding to questions		Opportunity to show ability by asking questions	
Statistic	US	CN	US	CN	US	CN	US	CN	US	CN
N	69	74	70	74	70	74	70	74	70	74
Mean	3.9	4.1	4.1	4.0	4.0	4.1	4.1	4.2	3.9	4.0
SD	0.9	0.9	0.8	0.9	0.9	0.7	0.8	0.7	1.0	0.7
Frequency										
High	53	63	59	59	55	63	57	67	51	57
Low	9	6	3	6	4	1	2	1	8	2
Strongly agree	18	25	21	20	20	21	25	24	20	16
Agree	35	38	38	39	35	42	32	43	31	41
Neutral	7	5	8	9	11	10	11	6	11	15
Disagree	9	6	3	6	3	1	1	1	7	2
Strongly disagree	0	0	0	0	1	0	1	0	1	0

Note. CN = China; US = United States.

experiences with presentations (U.S. $M = 3.9$, China $M = 4.1$). They felt comfortable answering questions about their presentations (U.S. $M = 4.1$, China $M = 4.0$) and asking questions of other presenters (U.S. $M = 4.0$, China $M = 4.1$). However, participants felt that responding to questions about their presentations provided them with slightly better opportunities to demonstrate their language ability than asking questions about others' presentations.

Although the question-and-answer session was generally perceived positively, some candidates commented that they found it difficult to relate to some topics chosen by other test takers. For example, one China trial participant maintained that "[t]he topic choosing is quite important. If the participant choose something away from our life, it would be a little bit hard for us to understand him or her." Another China participant argued that "Task 4 can show what student want to talk during the [?], however, some student prepare different talk, different level, we cannot join the talk if they really not familiar with these topic. It's hard to discuss with others." Similarly, a few people in the U.S. trial mentioned the potential impact of working with a lower-level speaker, such as one person who commented, "Sometime, I feel it is hard to ask question. The ability of speaking English another test taker will easily affect me." On the other hand, several China participants commented that they appreciated the opportunity to select and prepare a topic in advance. However, concerns raised about the accessibility of the topics selected by test takers highlight a significant challenge for employing this type of task in an operational assessment.

In the China focus group, comments addressed a variety of issues related to task design and topics (Table 10). A number of individuals commented that they felt the tasks and topics were realistic, but the most common task-related comment was that more preparation and/or response time was needed. Participants were not allowed time to prepare their responses, and nine individuals specifically commented on this fact. Five individuals commented on the technology problems experienced in the sessions, but this represented only 8% of participants. The low frequency of technology-related comments was somewhat surprising given the prevalence of problems during the China trial, and we speculate that participants may have been used to such disruptions or that they understood the focus group questions to address the tasks rather than the user experience.

Those commenting on interactivity within the test generally expressed positive attitudes, although two individuals felt that participants should be assigned turns to speak, and one expressed nervousness toward speaking with unfamiliar interlocutors. Other comments made by the participants were that the experience was enjoyable, relaxing, and good practice for speaking. Two individuals liked the opportunity to prepare the presentation before the session, while another two questioned the feasibility of this approach. When asked to state a preference for an online or face-to-face assessment format, a total of 69% of the participants who commented (43 individuals out of 62 in total) said that they would prefer an online version (Davis, Timpe-Laughlin, Gu, & Ockey, 2018). Common reasons for preferring an online format included a belief that it was less stressful than speaking with someone in the same room (20 participants) and that it was convenient (8 participants). However, the possibility of technology problems was mentioned as a caveat by 15 individuals who

Table 10 Focus Group Responses to the Question "What Do You Think of the Activities Overall?"

Subject	Individuals,[a] *n* (%)	Groups,[b] *n* (%)
Tasks and topic		
Positive		
Tasks difficulty was appropriate, "easy"	6 (10)	5 (24)
Tasks were realistic	3 (5)	3 (14)
Topics related to real life	6 (10)	5 (24)
Negative		
Preparation or response time not enough	15 (24)	9 (43)
Task difficulty too hard or easy	2 (3)	2 (10)
Business topics not appropriate	3 (5)	2 (10)
Technology, innovation		
Positive		
Tasks/format innovative	3 (5)	3 (14)
Negative		
Technology problems	5 (8)	3 (14)
Tasks were similar to other tests	4 (6)	4 (19)
Interactivity, discussing with others		
Positive		
Interactive, liked communicating with others	6 (10)	5 (24)
Negative		
Problems sharing floor, anxious	3 (5)	2 (10)
Other		
Positive		
Experience was fun, interesting	8 (13)	8 (38)
Prepare presentation in advance	2 (3)	2 (10)
Good practice for speaking	4 (6)	4 (19)
Felt relaxed	6 (10)	4 (19)
Negative		
Feasibility of using a prepared talk	2 (3)	1 (5)
Other	3 (5)	3 (14)

[a] $n = 62$. [b] $n = 21$.

otherwise preferred the online format. Overall, opinions expressed were in keeping with their comments in the written survey.

Conclusion, Outlook, and Future Research

The construct of oral ability is multifaceted, but computer-delivered oral assessments typically are designed to assess only certain aspects of this ability. This study aimed to determine if video-mediated technology in conjunction with an ETS-developed platform could be used to deliver an assessment that would provide test takers with opportunities to more fully demonstrate their oral proficiencies, including interactional competence. The results provide preliminary evidence that the reliability of computer video-mediated technology varies considerably, with technical disruptions being relatively few in the U.S. trial but very frequent in the China context.

The fact that approximately two-thirds of the sessions for the U.S. context functioned without technological problems suggests that for the U.S. context, computer video-mediated technology seems to have almost reached a stage of being stable enough for high-stakes testing situations. Nonetheless, it is likely that occasionally, test takers would need to be retested due to technological failure. On the other hand, every session in the China trial encountered technological problems of some kind, typically with repeated instances of several different kinds of problems, such as dropped calls, failure to deliver and/or display content, and time delays in delivery of content. The full impact of such disruptions on the performance of participants is difficult to gauge, but at the very least, major problems, such as dropped calls or missing task content, would limit the opportunity for speakers to demonstrate their abilities. More subtle disruptions, such as delays in the audio signal or dropped video, have the potential to disrupt turn taking and create misunderstandings between speakers, complicating scoring and again impacting inferences regarding test takers' abilities. The China trial was intentionally designed to be a challenging test of the technology, and these results indicate that central administration of such a test

across international locations is likely to be problematic given current technological limitations. At the very least, considerable care will need to be taken in designing and implementing the technology that supports a test to ensure adequate reliability in test delivery.

Given that the video was transmitted from a server in Princeton, New Jersey, and time delays varied across locations and appeared related to the local Internet connection speed, it remains to be investigated if the challenges encountered in the China administration might be solved by employing a different design, such as delivering materials and running sessions from computers located more locally. Moreover, the study showed that human error (e.g., accidentally disconnecting cables), as indicated by human-related issues (25% of the sessions in the United States), can lead to major challenges when technology and humans are involved, suggesting the need for careful training in how to use and troubleshoot problems when they arise.

With regard to the tasks, test takers in both trials viewed the activities of giving a short response to a question asked by an interlocutor, summarizing a proposal, defending a position in a group/paired discussion, and giving a prepared presentation and responding to questions from other test takers as generally representing the kinds of tasks that they encounter in the oral language use domain. Test takers generally felt that the tasks afforded them the opportunity to demonstrate their oral abilities and that these tasks can be effectively completed in a computer video-mediated environment. Additionally, participants regarded the experience of completing the tasks as enjoyable, relaxing, and good practice for speaking.

Other advantages and challenges relating to the use of computer-mediated technology to deliver speaking assessments emerge from this study. For instance, a major potential benefit of this technology is that this approach may achieve some of the convenience and consistency of computer-based oral tests by removing the need for examiner and test takers to be co-located. Another possible advantage is that the online format may be perceived as less threatening than facing others in person, resulting in reduced test anxiety. In the China focus groups, examples of such perceptions included comments like "I feel more relaxed this way since I don't need to face the interviewer and other test takers face-to-face" and "I like this video format—it makes me relaxed, since you can only see part of my body" (paraphrased from the original Chinese). Finally, it is conceivable that when speakers are no longer in each other's physical presence, the impact of a variety of interlocutor characteristics, such as power, gender, or personality, may change, with either positive or negative effects.

The potential for differences in language behavior across face-to-face and video-mediated conditions could also suggest a new way of thinking about speaking ability. That is, should video-mediated communication be considered a separate domain of language use? The answer to this question will likely depend on the types of uses test designers want to support and the way the speaking construct is conceptualized to support such uses. In any case, as video-mediated communication becomes more ubiquitous in academic, business, and social contexts, it seems likely that video-specific features of communication will emerge and that assessment users may specifically want to know about the ability of test takers to navigate video-based contexts.

Further research is needed before video-mediated test tasks such as these can be employed in operational assessments that aid in making high-stakes decisions. Maybe most importantly, micro-level analyses should be conducted to investigate whether these tasks actually elicit discourse that can be used to assess abilities not typically measured by computer-delivered assessments, such as interactional competence. Another aspect that needs to be explored further is how certain phenomena that were noticed in the trials, such as partner effects (e.g., participant pairings of similar vs. differing oral proficiency), impact oral performance on the interactive oral tasks in a computer-mediated video environment. It also needs to be examined how we can account for these phenomena when evaluating and scoring test-taker performances. It is crucial that we do not simply assume that known influences in face-to-face oral assessments, such as interlocutor familiarity, will have the same effects on video-mediated oral assessments.

Our research gives us bridled optimism for the potential use of video-mediated technology for assessing a broader construct of oral ability than is currently assessed by most computer-delivered assessments. However, we recommend that the field move in this direction cautiously as technology continues to improve.

Notes

1 Note that one participant did not answer the biographical questionnaire.

2 Note that, in this report, all quotations taken from the survey data are direct representations of the original responses, which are presented verbatim, and may thus contain linguistic inaccuracies.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the L2 group oral discussion task. *Language Testing, 20*, 89 – 110. https://doi.org/10.1191/0265532203lt245oa

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*, 341 – 366. https://doi.org/10.1177/0265532209104666

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT TM and real-life academic speaking activities. *Language Assessment Quarterly, 11*, 353 – 373. https://doi.org/10.1080/15434303.2014.947532

Clark, J. L., & Hooshmand, D. (1992). "Screen-to-screen" testing: An exploratory study of oral proficiency interviewing using video teleconferencing. *System, 20*, 293 – 304. https://doi.org/10.1016/0346-251X(92)90041-Z

Cohen, K. M. (1982). Speaker interaction: Video teleconferences versus face-to-face meetings. In L. A. Parker & C. H. Olgren (Eds.), *Proceedings of Teleconferencing and Electronic Communications* (Vol. *2*, pp. 189 – 199). Madison, WI: University of Wisconsin Press.

Davis, L., Timpe-Laughlin, V., Gu, L., & Ockey, G. (2018). Face-to-face speaking assessment in the digital age: Interactive speaking tasks online. In J. M. Davis, J. Norris, M. Malone, T. McKay, & Y. A. Son (Eds.), *Useful assessment and evaluation in language education* (pp. 115 – 130). Washington, DC: Georgetown University Press. https://doi.org/10.2307/j.ctvvngrq.10

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*, 423 – 443. https://doi.org/10.1177/0265532209104669

Ferris, D., & Tagg, T. (1996). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL Quarterly, 30*, 297 – 320. https://doi.org/10.2307/3588145

Frost, K., Elder, C., & Wigglesworth, G. (2011). Investigating the validity of an integrated listening – speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*, 345 – 369. https://doi.org/10.1177/0265532211424479

Fulcher, G. (2003). *Testing second language speaking*. Harlow, England: Longman.

Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Nakatsuhara, F. (2012). The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking test. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers: Vol. 2. Research in reading and listening assessment* (pp. 519 – 573). Cambridge, England: Cambridge University Press.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2015). *Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery* (IELTS Partnership Research Papers 1). Retrieved from https://www.ielts.org/teaching-and-research/research-reports/ielts-partnership-research-paper-1

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*, 161 – 186. https://doi.org/10.1177/0265532208101005

Ockey, G. J. (2014). The potential of the L2 group oral to elicit discourse with a mutual contingency pattern and afford equal speaking rights in an ESP context. *English for Specific Purposes*, *35*, 17 – 29. https://doi.org/10.1016/j.esp.2013.11.003

Ockey, G. J. (2018). Oral language proficiency tests. In *The TESOL Encyclopedia of English language teaching*. John Wiley & Sons. Retrieved from https://doi.org/10.1002/9781118784235.eelt0234

Ockey, G. J., Gu, L., & Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Language Assessment Quarterly, 14*, 346 – 359. https://doi.org/10.1080/15434303.2017.1400036

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*, 39 – 62. https://doi.org/10.1177/0265532214538014

Ockey, G. J., & Li, Z. (2015). New methods and not so new methods for assessing oral communication. *Language Value,* 7, 1 – 21. https://doi.org/10.6035/LanguageV.2015.7.2

O'Conaill, B., Whittaker, S., & Wilbur, S. (1993). Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human Computer Interaction, 8*, 389 – 428. https://doi.org/10.1207/s15327051hci0804_4

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*, 277 – 295. https://doi.org/10.1191/0265532202lt205oa

Rouhshad, A., Wigglesworth, G., & Storch, N. (2015). The nature of negotiations in face-to-face versus computer-mediated communication in pair interactions. *Language Teaching Research, 20*(4). https://doi.org/10.1177/1362168815584455, 514, 534

Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human Computer Interaction, 10*, 401 – 444. https://doi.org/10.1207/s15327051hci1004_2

Theodoropulos, C. (2012, April). *Investigating the "interactive" in the Interactive Performance Test: A conversation analysis approach*. Paper presented at the Language Testing Research Colloquium, Princeton, NJ.

Van Moere, A. (2006). Validity evidence in a group oral test. *Language Testing, 23*, 411 – 440. https://doi.org/10.1191/0265532206lt336oa
Zappa-Hollman, S. (2007). Academic presentations across post-secondary contexts: The discourse socialization of non-native speakers. *The Canadian Modern Language Review, 63*, 455 – 485. https://doi.org/10.3138/cmlr.63.4.455

Appendix A

Example Script of Video Proposals Used in Task 2

Topic: Recruiting Talent

Proposal 1: Benefits

Hi, I think the best way we can recruit good people to work at our company is by offering better benefits and training. These days, a job offer is not just about the salary. Job candidates are thinking about the whole package! Benefits like providing health care are important so employees do not have to spend a lot of their own money on medical expenses. And paid vacation time is really important for people who want to maintain a good work – life balance. My friend told me that his company gives employees 4 weeks of vacation a year, and in addition, they get national holidays off too. Now he has enough time to visit his family abroad every year. Finally, we could also offer more job training for workers who want to learn new skills or advance in their careers. Opportunities such as leadership programs or advanced training would be important to attract younger workers who are ambitious and want to improve their abilities. For some people, good benefits are more important than a high salary. We need to think about everything a job can offer if we want to recruit talented employees!

Appendix B

Guidelines for Making Presentations

Guidelines for Test Takers

Preparing Your Presentation

You are asked to select a topic and prepare a one-page presentation slide *prior* to taking the test. Here are some guidelines for the presentation. (See Appendix 1 for template.)

1. Topics

   You can select any business-related topic for your 2-minute presentation. The following list provides example topics and slides. (See Appendix 2 for example.)

- Social media impacting business
- Going green! — Enhancing a company's public image
- Creative team-building methods

Please note that these slides are demonstration material. They must not be used in your presentation, as you are expected to provide your own individually created presentation.

2. Structure

   The one-page slide is a visual support for your oral presentation. It should be clear, concise, and easy to follow. To provide for a clear setup and structure, your presentation needs to conform to the following guidelines:

a. **A smooth background and clear setup**

   Backgrounds should be unobtrusive.

b. **A maximum of *three* bullet points with a maximum of *five* words per bullet point**

   A bullet point is a short summation of the key point that you want to make. It is not a complete sentence. For each idea you want to convey, consider what the key point is and put that as a bullet point.

c. Visual

   *The graphic displayed in your presentation needs to present **data** to support the argument or point you are making. The image can either be downloaded from the Internet, provided a reference to the original source is given on the slide, or it can be a graph, chart, or figure from data that you have available and are allowed to share.*

3. Technical requirements
   The one-page presentation slide needs to be saved as a pdf or jpg file.

During Your Presentation

Oral presentation skills include verbal and nonverbal communication. Your score will be based on your words, the delivery of this information, and timing.

1. **Verbal communication** will be considered when scores are assigned. It includes words as well as voice quality, so when speaking, remember the following:

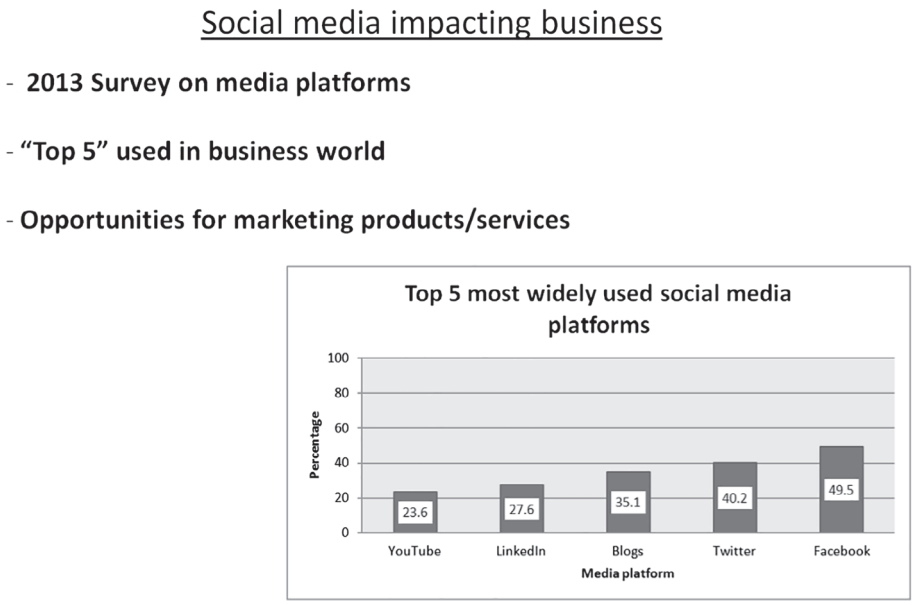
- **Words**
  - **Attention should be paid to structure, organization, and vocabulary.** A well-structured and organized presentation helps the audience to follow along and maintain interest.
  - **Words are pronounced correctly.**
- **Voice**
  - **Voice is loud enough for everyone to hear.**
  - **Speaking is appropriately paced and clear.** Speech should be fluent but include pauses appropriate for emphasizing important points and not so fast that it becomes unclear.
  - **Tone of voice is appropriately varied.** Pitch and volume should be used to emphasize points and to keep the audience engaged and interested.

2. **Nonverbal communication** includes *eye contact, facial expressions, posture, movements,* and *gestures*. The following will be considered in the assessment.

- **Eye contact.** Eyes are used to effectively engage with audience.
- **Use body movements and hand gestures effectively.** Body posture, facial expressions, movements, and gestures should be used to emphasize important points.

3. **Timing** is a key factor in giving a presentation and will be considered in scoring. Make sure to use your 2 minutes effectively.

Example visual (full guidelines included two other examples of visuals).



Appendix C

Biographical Data Questionnaire

Please answer the following questions by indicating the response that best describes you. Your responses will not be used to identify you.

1. What is your gender?
   - Male
   - Female
2. Please indicate your age range
   - Age 46 or above
   - Between 31 and 45
   - Between 22 and 30
   - Less than 21
3. What country are you from? _____
4. What is your first language? _____
5. How long have you been living in the United States?
   - 3 years or more
   - 1 – 2 years
   - 6 months to 1 year
   - Less than 6 months
6. What is your student status at your university/college?
   - Graduate
   - Undergraduate
   - Pre-university
   - Other_____
7. What is your (desired) area of interest or major at your university/college?
   - Humanities (history, language, culture)
   - Business (economics, computer science, finance)
   - Social Sciences (psychology, education, linguistics)
   - Natural Sciences (math, physics, biology)
   - Other _____
8. How long did you study English in your own country before coming to the United States?
   - 6 years or more
   - 3 – 5 years
   - 1 – 2 years
   - Less than 1 year
9. How often did you have English class in your own country?
   - More than 5 hours per week
   - 1 – 5 hours per week
   - 1 hour per week
   - Never
10. How long have you been studying English in the United States?
    - 6 years or more
    - 3 – 5 years
    - 1 – 2 years
    - Less than 1 year

11. How many years have you used English in an English-speaking environment (both in your own country and in the United States)?
    - 3 years or more
    - 1 – 2 years
    - 6 months to 1 year
    - Less than 6 months
12. How would you describe your English speaking ability?
    - I can easily communicate even complex ideas in English.
    - I can communicate complex ideas in English, but I have to work hard to do it.
    - I can communicate simple ideas pretty easily, but I cannot express complex ideas.
    - I have to work very hard to communicate even basic ideas in English.
13. Approximately how many hours do you use a computer in a week?
    - More than 4 hours
    - 1 – 3 hours
    - Less than 1 hour
    - Never
14. Have you ever used a laptop computer?
    - Yes, many times.
    - Yes, but not often.
    - No
15. How often do you use the computer to video chat (using Skype, Google Hangouts, FaceTime, etc.)?
    - Every day
    - More than once a week
    - Once a week
    - Once a month
    - Never
16. If you use video chats, how many people do you usually video chat with (check all that apply)?
    - More than 3 people
    - 3 people
    - 2 people
    - 1 person
17. If you use video chats, what is the purpose of using it (check all that apply)?
    - For work
    - For school work
    - To video chat with friends
    - To video chat with family
    - Other: _____
18. Please describe your comfort level with learning new computer programs and technologies.
    - Very comfortable
    - Comfortable
    - Somewhat comfortable
    - Not at all comfortable
19. Which of the following best describes your interest in learning new computer programs and technologies for communicating with others?

- Highly interested
- Somewhat interested
- Not at all interested

20. How often do you talk to an automated computer on the phone (e.g., when you call the electricity company or cell phone company)?

    - More than once a week
    - Once a week
    - Once a month
    - Once a year
    - Never

21. For each of the three rows, please select the option that best describes you. When I speak English with a group of people that I don't know, I:

feel very shy	feel fairly shy	feel a little shy	don't feel shy at all
am never a leader hate to talk	am not usually a leader don't like to talk	am often a leader like to talk	am almost always a leader love to talk

Appendix D

Questions About Tasks

Questions after all task types

1. Activity X was engaging/interesting.
2. Activity X gave me a good chance to show my ability to answer questions in English without preparation time.
3. Activity X gave me the chance to get accustomed to the other test takers' language varieties and accents.
4. Activity X gave me a good chance to show how well I can speak English.
5. Because of the technology, there were inappropriate interruptions when responding to the questions.
6. Because of the technology, there were unnatural and uncomfortable pauses when responding to the questions.

Questions specific to Task 1

1. I felt comfortable speaking in front of the other test takers.
2. In real life, I use English to answer brief questions.

Questions specific to Task 2

1. I was able to understand the proposals.
2. Taking notes while listening would have helped me summarize better.
3. Seeing the video helped me understand the information better than just listening.
4. There was enough time to summarize the proposal.
5. In real life, I use English to summarize ideas to others.

Questions specific to Task 3

1. The talking time was shared equally.
2. Defending my position made me feel uncomfortable.
3. In real life, I use English to defend positions when I speak with others.
4. Activity 3 gave me a good chance to show my ability to interact with others.

Questions specific to Task 4 (presentation part)

1. Using Skype was appropriate for activity 4
2. Standing during the presentation was appropriate
3. In real life, I give presentations in English

4. Activity 4 gave me a good chance to show my ability to give a presentation in English.

Questions specific to Task 4 (question-and-answer part)

1. The Q and A part of the presentation was similar to my experience with presentations I was given before.
2. Asking questions to others after each presentation was fair.
3. I felt comfortable asking questions to other group members after each presentation.
4. Asking questions after each presentation gave me a good chance to show how well I can speak English.
5. Having others ask questions after each presentation was fair.
6. I felt comfortable answering questions after my presentation.
7. Responding to questions after my presentation gave me a good chance to show how well I can speak English.

Note. All selected response items used a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). A section labeled "Comments" followed each section.

Appendix E

Questions About Technology

1. I found Skype to be user friendly.
2. Seeing other people through Skype made the activity engaging/interesting.
3. I could clearly see the other participants.
4. I could easily identify the person who was talking.
5. Being able to see others through Skype made it easy to understand their reactions/feedback.
6. I could clearly hear the other participants.
7. Skype made it easy to express myself.
8. Skype made it easy to respond to others.
9. Overall, there were inappropriate interruptions because of the technology.
10. Overall, there were unnatural and uncomfortable pauses because of the technology.
11. It was clear and useful to have the instructions on the screen.
12. It was useful to have the countdown timer on the screen.
13. I could see the presentation slides clearly.
14. It was easy to upload my presentation slide.
15. It was clear to use the function of the buttons on the screen.
16. Skype allowed us to interact well.
17. Skype allowed us to complete the activities as instructed.
18. Skype made me feel like I was interacting with people like I often do in real everyday life.
19. Using Skype was similar to my communication through technology (e.g., Skype, FaceTime, Google Hangouts, etc.). If you do use these technologies, please check here.
20. If there were no technology problems throughout the four activities (e.g., frozen screen, time delay, etc.), I think there is a difference between communicating through technology and face-to-face.

Note. All selected response items used a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). A section labeled "Comments" followed each section.

Appendix F

Focus Group—Protocol

After completing all speaking tasks, the moderator will say in **English**:

Now, we would like to ask you some questions about what you just finished. The three of you will discuss these questions for 10 minutes. After that, you will each individually answer some written questions. This discussion will be done in Chinese, so I will now ask Professor L. to take over.

Professor L. administering the speaking tasks will then begin saying the following in **Chinese**:

In this discussion I will ask you a few questions about your reactions to the session you just finished. We are doing this in Chinese so that you can easily and fully express your ideas, so please feel free to say whatever you think. We are interested in what you have to say, so that we can improve the tasks and technology you saw today.

Professor L. should then ask Question 1.

- When discussion of question 1 seems to be finished, then ask question 2.
- Continue like this until the time is finished.
- **Note:** If there is less than 1 minute left, go directly to question 7 and skip any other remaining questions.

The moderator (in Princeton) will watch the time, and will let everyone know when the time is finished.

Questions

Question 1: What do you think of the activities overall?

Question 2: What things did you especially like or dislike? Why do you say that?

Question 3: How do these tasks compare to other English speaking tests you have taken?

Question 4: How well did these tasks allow you to show your ability to speak in English?

Question 5: These tasks were done online. How does this compare to speaking in person? Which do you like better?

Question 6: If you were going to take a test like this, would your preparation be any different compared to another speaking test like TOEFL or IELTS?

Question 7: Any other comments about the tasks?

Note: Questions should be translated into Chinese.

Suggested citation:

Ockey, G. J., Timpe-Laughlin, V., Davis, L., & Gu, L. (2019). *Exploring the potential of a video-mediated interactive speaking assessment* (Research Report No. RR-19-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12240

Action Editor: Keelan Evanini

Reviewers: Ching-Ni Hsieh and Jackson Tanner