Approaches and Challenges to Assessing Oral Communication on Japanese Entrance Exams

Gary J. OCKEY  
Iowa State University

Abstract
The Japanese university entrance exam does not include an oral communication section, and as a result, teachers are often conflicted as to whether they should spend class time on oral communication activities as is mandated by the national curriculum, or time focusing on preparing their students for the entrance exams. In addition, it is unlikely that universities are able to accurately select students with the strongest English abilities when they are not given information about students’ abilities to orally communicate. It would seem obvious that to solve this problem, an oral communication section could simply be included on the exam. However, valid oral communication assessments are immensely challenging to implement in large-scale high stakes contexts.

This paper discusses some of the possible approaches that have been used to assess oral communication in large-scale high stakes contexts and their appropriateness for the Japanese entrance exam. It also explores some of the current areas of research that might help to alleviate some of the challenges of including an oral communication section on the Japanese entrance exam. It focuses on three approaches currently used: computer-delivered and computer-scored, computer-delivered and human-scored, and human-delivered and human-scored.

Keywords: entrance exams, oral assessment

For decades, there has been tension between Japanese classrooms, where oral communication is seen as important, and the Japanese national university entrance exam, which has not directly assessed oral communication. One result has been that teachers have been forced to make the difficult decision of choosing whether they should spend class time preparing students for the university entrance exam, or following the mandated curriculum which emphasizes helping students develop oral communication skills.

Given the aims of English education in Japan, the importance of including an oral communication section on the university entrance exam appears obvious. However, numerous challenges have made its inclusion difficult. This paper discusses some of the approaches that could be taken to achieve this purpose. It begins with a discussion of the aims of a university exam and describes a framework that could be used to guide the development of an
appropriate exam. Next, it describes and evaluates three approaches currently used, along with possible variants of these approaches, in high stakes large-scale assessments for assessing oral communication: computer-delivered and computer-scored, computer-delivered and human-scored, and human-delivered and human-scored. It concludes with a plea for decision makers to weigh the options and adopt one or a combination of the options, in spite of the challenges that will need to be overcome.

**Aims of a University English Entrance Exam**

Two important aims of a university English entrance exam are to promote effective English education and to identify the most capable English users for university studies. To achieve these two major aims, assessments can be evaluated based on their instructional validity, construct validity, reliability, and practicality. In this next section, these concepts are defined in the context of speaking assessments.

Construct validity refers to, “...the extent to which we can interpret a given test score as an indicator of the ability(ies), or constructs, we want to measure” (Bachman & Palmer, 1996, p. 21). The construct of L2 speaking has been defined as “the verbal use of language to communicate with others” (Fulcher, 2003, p. 23). Ockey and Li (2015) have divided this more general construct of speaking into (a) interactional competence, (b) appropriate use of phonology, (c) appropriate and accurate use of vocabulary and grammar, and (d) appropriate fluency. They defined interactional competence, or the ability to respond appropriately, as, “An individual’s underlying ability to actively structure appropriate speech in response to incoming stimuli, such as information from another speaker, in real time” (p. 5). They consider effective use of both segmental and super-segmental aspects of language as part of phonology, both depth and breadth of vocabulary and grammar, and naturalness of speech, pausing, and repetition as aspects of fluency.

The concept of instructional validity relates to the extent to which the test tasks and content of the test align with the curriculum and instructional practices for a given context (Pellegrino, DiBello, & Goldman 2016). For the Japanese Entrance Exam, this instructional context would be Japanese primary and secondary schools. Thus, for the Japanese Entrance Exam to have high instructional validity, the test tasks would need to align with the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) mandated curriculum for Japanese primary and secondary schools.

A major aim of Japan’s English curriculum is for students to be able to talk about their own thoughts. More specifically, they should be able to (a) use English speech sounds, including stress, intonation, and pauses correctly, (b) speak accurately to a listener about thoughts and feelings or facts, (c) discuss an opinion about written or oral input, (d) speak continuously, and (e) be able to give a simple speech on a provided theme (Second School,
Section 9: Foreign Languages\(^1\). This description indicates the importance of accurately using English to discuss one’s opinion about a stimulus as well as to give a short speech. Thus, test tasks with high instructional validity would require test takers to engage in oral discussions and give short talks.

Reliability has been defined as consistency of measurement (Bachman & Palmer, 1996). It refers to the extent to which an assessment provides the same score to a given test taker over repeated administrations. Reliability is an important quality of an assessment, but achieving this quality in speaking assessments, which often rely on human judgements, is an age-old challenge. Research shows a great deal of variation in human judgements of the same test taker’s ability (Lunz, Wright, & Linacre, 1990).

Practicality refers to the extent to which the test administration can be feasibly completed given its context. Practicality includes the cost of test development and maintenance, ease of marking, time required for administration, and availability of resources needed (Davies et al., 1999). For oral communication assessments, practicality would relate to the availability and cost of human administrators, raters, or the technology needed for machine scoring, time, and the availability of space and computers to administer the assessments.

**Assessment Delivery Options for Assessing Oral Communication**

**Human-Delivered and Human Scored**

For decades, the gold standard for assessing oral communication has been face-to-face one-on-one oral interviews. A number of large-scale high stakes assessments use this approach, including Cambridge’s International English Language Testing System (IELTS) Academic. The speaking section of IELTS academic has three parts. In the first part, an interviewer asks the test taker a list of questions, and the test taker is expected to provide an appropriate response. In the second section, the test taker is asked to speak about a specific topic for two minutes, and is then asked two questions about the content of the response. Finally, the interviewer asks the test taker further questions about the topic from section two along with follow-up questions to the test taker’s responses (https://www.ielts.org).

The construct validity of a one-on-one oral interview such as IELTS Academic is quite high. The assessment is likely to assess all or almost all of the abilities indicated in Ockey and Li’s (2015) oral communication construct. The possible exception to this might be some aspects of interactional competence, since test takers may not have opportunities to demonstrate all of the aspects of this ability, including initiating and closing topics.

The instructional validity of oral interviews might be considered moderate. On one hand, test takers are interacting with a person as is expected by the Japanese mandated curriculum. However, in the communicative classes mandated by MEXT, students are

---

expected to discuss a topic, not just respond to an interviewer’s questions.

The reliability of one-on-one oral interviews has been shown to rely heavily on the training of the raters and can be quite high. The practicality on the other hand, is very low, given the need for detailed scheduling, a large number of trained interviewers/examiners, and space and time to conduct the tests.

In an effort to address some of the limitations of one-on-one oral interviews, paired and groups assessments have gained popularity. In this format, two or more test takers discuss an assigned topic, while a trained examiner provides a rating, or, in certain cases, might join the discussion. An example of a paired test is the third section of Iowa State’s English Placement Test (EPT). After retelling information about two speakers’ points of view on a topic, the test takers are asked to defend one of the points of view during a four-minute discussion with each other. Test administrators do not join the conversation between the two test takers. The raters listen to the discussion and rate the performance (https://apling.engl.iastate.edu/english-placement-test).

Recently, researchers have experimented with the use of computers to help deliver speaking assessments. For example, computer-mediated video (e.g., Skype, google hangouts) and virtual environments have been used to deliver real-time interactive human-to-human speaking assessments. Davis, Timpe-Laughlin, Gu, and Ockey (2017) developed a video-mediated speaking assessment. They used Skype to provide visual and verbal communication among three test takers and a test administrator and a custom interface to help deliver four test tasks. The tasks included the examiner asking each test taker three questions, test takers summarizing video input, the three test takers discussing a topic, and each test taker giving a prepared presentation followed by questions about the content of the presentation from the other test takers. The researchers piloted the technology across four locations in the United States, and across three locations in China with the test administrator in the United States. Of the 28 sessions conducted in the United States, 18 had no technological problems, and of the other 10 sessions, only about half had major problems. However, all 24 sessions in China had some technological problem, and most of these issues were major (e.g., dropped calls, presenters slides not visible to other test takers). The researchers concluded that further technological developments are needed before high stakes assessments can adopt a video-mediated approach for aiding in delivering speaking tests.

Better technological success has been reported with the use of computer-mediated virtual environments than with computer-mediated video. Ockey, Gu, and Keehner (2016) used a custom-made virtual environment to help deliver a group oral discussion task to three test takers and a test administrator, who were located in three places in the eastern United States. The web-based desktop VE technology presented a virtual space that represented a university library. The test takers could see representations of the other test takers in the form of avatars immersed in the 3D environment. They could speak to each other in real time, and the voice activated avatars moved to indicate that a test taker was speaking. In six four-minute sessions, there were only two instances (of between three and four seconds each)
that the technology failed. Sound was temporarily not audible, but within seconds the sound returned, and the test continued with little effect. The researchers concluded that VE technology could be reliably used to help deliver oral communication assessments. Test takers had mixed views about the technology. Some wanted to be able to see the actual people that they were conversing with, while others reported that they felt less anxiety while taking the test because they could not be seen.

The construct validity of both the face-to-face and the computer-delivered paired or group oral assessments are both potentially quite high. All aspects of Ockey and Li’s (2015) construct could be assessed by this task type. These task types are designed to elicit aspects of interactional competence via the interaction among the test takers. Paired and group oral tasks typically include a rating scale for assessing interactional competence. For example, the Iowa State University EPT includes such a band with the focus being on the ability to respond appropriately (https://apling.engl.iastate.edu/english-placement-test). Research has shown that paired tasks provide test takers opportunities to demonstrate their ability to engage in complex interaction (Brooks, 2009), and the group task has been shown to provide opportunities for test takers to initiate and close topics (Ockey, 2014).

However, the paired/group oral format may have potential to bias test results. Research has shown that personality (Berry, 2004; Ockey, 2009b), familiarity with other test takers in one’s group (O’Sullivan, 2002), and number of peers in one’s group (Nakatsuhara, 2011) among other factors can affect a test taker’s score when the paired/group oral format is used.

The instructional validity of the paired/group oral approach would be quite high. Completing the test task would require test takers to use English speech sounds, speak to a listener about thoughts and feelings or facts, discuss an opinion, and speak continuously, which constitute four of the five major aims of the Japanese speaking curriculum. The fifth major aim, which is to be able to give a simple speech on a provided theme, may also be part of a paired or group test if each speaker is asked to begin with a short speech about their opinion on the topic.

Like the one-on-one oral interview, the reliability of paired/group oral tasks delivered by humans depends not only on the task type, but on the effectiveness of rater training. Research has shown that group test tasks can be moderately reliable (Bonk & Ockey, 2003; Ockey, Koyama, Setoguchi, & Sun, 2015), but, in part, because the discourse is controlled by the test takers, reliability is unlikely to be high.

Human-delivered paired or group tests are not practical compared to selected response item types that can be delivered and scored by computers. However, they are more practical than human-delivered and human-scored one-on-one oral interviews. Unlike one-on-one oral interviews, administrators only need to judge performances—not judge performances and ask questions or engage with the test taker, and more test takers may be assessed in less time.

**Computer-Delivered and Computer Scored**

A growing number of computer-delivered and computer-scored systems are being used
to assess oral communication. Tasks can be delivered via computer, tablet, mobile phone, or other electronic devices. Duolingo is an example of a test which is computer-delivered and computer-scored. It uses a read aloud task for assessing speaking. Test takers are shown a written sentence, such as, “She is eating an apple,” and are then asked to “Speak this sentence” (Wagner & Kunnan, 2015; Ye, 2014). Wagner and Kunnan (2015) assume that the test is scored based on speech rate, length of pauses, phoneme accuracy, intonation, and other related speech features that can be assessed by automated speech scoring systems (Bernstein 2013). Versant, another large-scale high-stakes assessment that uses computers to deliver and score their oral communication assessments also uses read aloud tasks. Additionally, they use repeat tasks, where test takers are asked to repeat a sentence that they hear; questions, where test takers give a simple answer to a question, such as; “Would you get water from a bottle or a newspaper?”; sentence builds, where test takers are asked to rearrange some words they hear to create a sentence; and story retelling, where test takers listen to a short story and then attempt to retell it (https://cdn2.hubspot.net/hubfs/559254/versanttest-aug2015/pdf/Sample-TEST-PAPER-Versant-English-Test-watermark.pdf?t=1498671761325). The common thread that connects these tasks is that they lead to predictable speech, which makes it possible for the computer to score it by comparing it to other speech elicited by the tasks that the test developer judges to be effective.

The construct validity of computer-delivered and computer-scored oral communication assessments is quite low compared to other direct assessments of oral communication. The task types that are typically used with this format have been criticized for assessing a narrow construct of oral communication (Chun, 2006; O’Sullivan, 2013; Ockey & Li, 2015), which does not include interactional competence. Computer-scored speaking assessments typically rely on measuring features such as speech rate, latency of response, length and position of pauses, word stress, segmental forms of words, and a comparison of the vocabulary with words used by highly proficient speakers (Ockey, 2009a), and, as a result, fail to assess interactional competence to a substantial extent. In addition, tasks, such as sentence repeat have been criticized on the grounds that test takers may be able to imitate spoken sentences without being able to segment the speech into meaningful units, calling into question their efficacy in assessing pronunciation and fluency, two important sub-constructs of oral communication ability (Ockey & Li, 2015).

The instructional validity of computer-delivered and computer-scored oral communication assessments is also poor compared to other direct approaches. The MEXT mandated speaking curriculum does not require students to read aloud or to repeat sentences. Although completing tasks typically used for assessing oral communication with this format would require students to use English speech sounds, and maybe speak continuously, which are aims of the Japanese secondary school curriculum, they would not require students to speak to a listener about thoughts and feelings or facts, discuss an opinion, or give a simple speech. The likely effect of using a test with such poor instructional validity would be that teachers would remain torn between teaching to the test, or using the MEXT mandated
curriculum. Thus, one desired impact, encouraging teachers to teach the mandated English curriculum, may not be realized if a computer-delivered and computer-scored test were used to assess oral communication on the Japanese entrance exam.

The reliability and practicality of computer-delivered and computer-scored oral communication assessments are likely quite high, once the technology is developed fully. However, the technological requirements for test delivery and scoring can be immense for a large-scale test. With current technology, the number of test takers who can take a test at one time may be fairly limited. However, aside from the challenges associated with delivering a large number of tests in a short time period, compared to other direct tests of oral communication, computer-delivered and computer-scored tests are very practical.

**Computer-Delivered and Human Scored**

Currently, a number of large-scale assessments are delivered by computers and scored by human raters. The large majority of these task types are non-interactive, but recently, a few tests that use this format are beginning to include interactive tasks. Interactive tasks require communication between two or more language users and must include negotiation of meaning between the sender and the receiver, while non-interactive tasks are individual language use tasks and do not require communication between two or more users (Davies et al., 1999).

For non-interactive tasks, it is common for the computer to deliver a recorded stimulus, which can be audio only, or it can include both audio and video or still photo images. Test takers are expected to respond to the recorded stimulus and the responses are recorded by the computer. These recorded responses are then sent to raters, who provide ratings of the test takers’ abilities based on the discourse samples. An example of a large-scale test that uses this model is the Educational Testing Services’ (ETS’), Test of English as a Foreign Language internet-based test (TOEFL iBT). The test includes six tasks, all of which are based on recorded stimuli that are presented via the computer to the test taker. After each stimuli, the test taker is given planning time, and then is expected to produce a short response (roughly a minute) for each of the six tasks. The first two tasks require test takers to use their own background knowledge and experience for their responses. An example prompt is:

“Talk about a pleasant and memorable event that happened while you were in school. Explain why this event brings back fond memories” (https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf).

The third and fourth tasks require the test taker to read a paragraph about a topic, listen to a recorded speech or dialog about the same topic, and then summarize or provide an opinion about the topic. An example prompt is:
“The man expresses his opinion of the Student Association’s recent purchase. State his opinion and explain the reasons he gives for holding that opinion” (https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf).

For the last two tasks, test takers listen to a recorded stimuli, are given planning time, and are asked to talk about something from the stimuli. An example prompt is:

“Using points and examples from the talk, describe the two different definitions of tools given by the professor” (https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf).

The construct validity of these non-interactive task types could be considered moderate. They have been shown to measure many aspects of oral communication, but have failed to adequately measure interactional competence. For instance, Ockey et al. (2015) found that these task types sufficiently measured fluency, pronunciation, and grammar/vocabulary, but did not adequately measure interactional competence. Similarly, Wagner (2016) found that scores on these task types correlated moderately with measures of language use ($r = .40$), but poorly with measures of interactional competence ($r = .15$). Brooks and Swain (2014) came to a similar conclusion finding that grammatical and lexical features were measured by these task types, but characteristics of interaction, such as use of questions were not elicited by these tasks.

A strength of computer-delivered human-scored systems is that ratings can be completed by a central pool of raters. Recorded performances can be sent to a central bank of raters, who are unlikely to be biased toward any particular speech variety. This is in contrast to human-delivered tests scored by humans in real time, which have been found to have potential for bias depending on the experiences of the local raters. For instance, Carey, Mannel, and Dunn (2011) found that speaking assessments rated by local raters who were familiar with the speakers’ speech variety, assigned much more lenient ratings than raters who were not familiar with the local speech variety. Limiting possible local bias with a central pool of raters model is a major advantage of a computer-delivered human-scored approach over a human-delivered and human-scored in real time approach.

The instructional validity of these task types could be considered moderate. Students might be expected to read and listen to stimuli and then give their opinions about this content in Japanese high schools. However, the Japanese government mandated curriculum requires high school English language programs to discuss and exchange opinions about information and ideas, and listen to, understand, and respond to others. These non-interactive tasks do not require test takers to engage in such activities.

The reliability of computer-delivered and human scored assessments, which are non-interactive depends to a large extent on the effectiveness of the rater training. Moderate to high reliability has been achieved with these task types. For instance, Ockey et al. (2015)
report a Cronbach’s alpha reliability of .82 for 222 Japanese university students on the six TOEFL iBT tasks. ETS reports even higher measures of reliability for their more diverse test taker population.

The practicality of computer-delivered and human-scored assessments could be considered low, even after the systems have been developed. For large-scale assessments, they require a large number of computers as well as a great deal of delivery capacity, as was discussed in the section about computer-delivered and computer-scored tasks. Human raters are expensive, and a large number are needed when many tests need to be rated in a short period of time.

Concerns that the non-interactive computer-delivered assessments fail to assess test takers’ interactional competence have led to attempts to create computer systems that can “interact” with test takers. These systems, referred to as chatbots, artificial intelligences, artificial conversation entities, and spoken dialog systems (SDSs), provide general responses to what test takers say. They do not actually interact with test takers. Instead their aim is to provide an opportunity for test takers to be able to demonstrate their interactional competence by keeping a discussion going. They may rely on such conversational features as key word recognition, prosody, hesitations, and even gestures and body movements.

While a growing body of literature exists on the use of SDSs for language learning, little has been written about their use for language assessment. A notable exception is the research of Litman and colleagues (Litman et al., 2016). An aim of their study was to investigate the extent to which the elicited discourse from an SDS could be scored. Three interactive tasks, searching for a laptop, finding a restaurant, and giving directions, were used in the study. Sixty-seven non-native speakers of English completed the tasks in English and were assigned holistic scores on the CEFR proficiency-level descriptors by human raters. A major finding of the study was that the SDS did not elicit speech that was representative of natural discourse. In some cases, the test takers had little opportunity to contribute to the conversation, while in others, the test takers’ discourse was found to be quite unnatural, likely, because they were trying to accommodate to the unnatural computer discourse.

Ramanarayanan et al. (2016) explored the potential of HALEF (Help Assistant-Language-Enabled and Free), an SDS that uses rule-based dialog flows. Test takers respond to questions, and based on the semantic features of their responses, the computer is directed to a branch of a dialog tree. The computer’s response depends on the branch to which it is directed. The researchers used the system to deliver four tasks, negotiating the schedule of a meeting, responding to an offer for food, interviewing for a job, and taking an order in a restaurant, to 676 English speakers (14 self-identified as L2 users of English). The system worked reasonably well when the speakers had fewer and shorter turns, but less well for the longer more frequent turns. The test takers reported a wide range of satisfaction with the system’s performance. About half rated the system four or five with five indicating most satisfactory and one indicating least satisfactory, while about a quarter rated the system’s performance one or two. Test takers expressed similar opinions of the system’s ability to
understand them.

SDSs have potential for high construct and instructional validity. They aim to assess a broad construct of oral communication, including interactional competence, and they could conceivably be designed to simulate the tasks used in communicative language classrooms. Their reliability and feasibility would presumably be similar to non-interactive human-delivered and human scored tasks. Currently, however, the research indicates that SDSs need to be further developed before they can be trusted for use in large-scale high stakes contexts.

Conclusion

Revising the Japanese entrance exam to directly assess oral communication would likely be a great step forward in helping to improve the quality of English education in Japan. Both construct and instructional validity of the entrance exam would be greatly increased, and the likely result would be more accurate selection and placement of students into appropriate universities and positive washback on primary and secondary instructional practices. Moreover, secondary school teachers would not be conflicted about whether to prepare students for the entrance exam or to follow the mandated government curriculum. Instead, the test would encourage them to follow the mandated curriculum, which aims to help prepare students to function in a global English environment.

Unfortunately, administering and scoring a direct assessment of oral communication comes with practical challenges, and there is no perfect solution to overcome these challenges. This paper provided some of the options that are available and discussed the advantages and disadvantages of each. It discussed some of the high-stakes large-scale tests that have been used as English university entrance and placement exams. It also provided insights into some current and future directions that may help to minimize some of the challenges associated with oral communication assessments. Needless to say, for Japan to adopt or develop its own oral communication university entrance exam will require a great deal of effort and cooperation on the part of many. However, this step needs to be taken to move English education to a higher level in Japan.

References


Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better


Ockey, G. J., Gu, L., & Keehner, M. (2016, June). *The use of live voice communication*
virtual environments to assess L2 oral communication. Paper session presented at the Language Testing Research Colloquium, Palermo, Italy.


