Using Technology to Assess Interactive Skills in a Speaking Test

Jacob B. Petersen
Iwate University
Daniel Newbury
Fuji University
Natsumi Onaka
Iwate University

Reference Data:

Petersen, J. B., Newbury, D., & Onaka, N. (2022). Using technology to assess interactive skills in a speaking test. In P. Ferguson, & R. Derrah (Eds.), *Reflections and New Perspectives*. JALT. https://doi.org/10.37546/JALTPCP2021-25

A pilot test was created to measure Japanese student English conversation abilities through a grant-funded study. One of the goals of this project was supporting the administration and scoring of the test through the innovative use of technology. We describe how the test was developed using backward design, the use of multimedia in the test questions, and how assessment was conducted by watching recordings of student tests. The future direction of the project goals are also explained.

日本人の英会話能力を測定するための試行テストが、助成金を受けた研究によって作成された。このプロジェクトの目標の一つは、テクノロジーの革新的な利用によって、テストの実施と採点を支援することである。本論文では、バックワードデザインを用いたテストの開発方法、テスト問題へのマルチメディアの利用、録画された学生のテストの視聴による評価の実施方法について説明する。また、プロジェクト目標の今後の方向性について説明する。

As Japanese foreign language education evolves towards a more communicative approach (MEXT, 2018) and applicable technology becomes more accessible, teachers have the opportunity to modernize some of the processes that were once only possible in face-to-face settings. One integral component of language teaching that may benefit from more recent technology innovations is that of assessment. As discussed here, it is possible to keep assessment aligned with the ever-growing focus on communication skills using currently available technology. While several of the widely used standardized speaking tests have been digitized, they may not fully measure test takers' communicative abilities. For example, some of these tests feature a prompt to which examinees respond orally. A recording of that response is then distributed remotely to raters for assessment. The output produced from this kind of task is conducive to scoring based on criteria traditionally used in written proficiency tests like grammatical accuracy (Patharakorn, 2018). Such metrics are important in measuring oral language proficiency, but they do not provide any insight on test takers' interpersonal communicative skills.

Tests may be administered online in an interview format and, as such, provide some evidence of test takers' interactional ability. However, past studies have shown that interview-led dialog limits the kinds of output test takers produce in comparison to tests that incorporate a peer-to-peer element (Ducasse & Brown, 2009). Many speaking tests are still paper based, and scores are awarded in real time. This is not inherently bad, but the process could be made more flexible and robust through the incorporation of relatively simple-to-use technology. One example is video recordings uploaded to an online server for remote viewing by offsite raters. This kind of system negates time and location constraints and supports the assessment of test takers' non-verbal communication strategies, which are known to be an important factor in assessing interactional skills (Ducasse & Brown, 2009).

In light of these observations, a novel approach to assessing English communicative ability was designed and piloted at a national university in the Tohoku region of

Japan. The initial implementations of the test incorporated both face-to-face and online elements, with test administration conducted in person, and raters evaluating performances remotely. We are hopeful that our innovative approach to using videos and online tools will evolve into a decentralized test that can be conducted completely online, supporting the needs of all types of schools and communities. Since our test focused on measuring communicative English ability per MEXT directives, it was clear that technology is crucial in supporting the test and its usability in future administrations. By applying backward design (Richards, 2013), we came up with four main themes that we recognized as important and that would benefit from technological support, as follows:

- We wanted to test both verbal and nonverbal communication. Because of that, it was useful to record students taking the test to watch and assess later.
- The English level of students entering the university was widely varied. To be as accessible as possible for all students ranging from the low level to high level, we decided to use images and video in the question prompts.
- When the project was initiated, we wanted to use a tablet computer for image and video questions, but when COVID-19 became an issue, safety precautions were easily manageable.
- We wanted an all-in-one system of test recording distribution and evaluation. Moodle was a system that supported these efforts.

Since technology was deemed important in student recordings, question prompts, tablet use, and video distribution, the literature review discusses the theoretical background in connection to these themes.

Literature Review

When designing a test measuring interactional ability, we had to envision how the test would be administered and assessed. For spoken language proficiency tests, it is not uncommon to have an interactive assessment with pairs of test takers (Galaczi, 2010; Kao, 2020; Köroglu, 2019). An interactive assessment can be defined as one that includes deliberate and planned engagement of test takers and assessment of their performance (Haywood & Tzuriel, 2002). From the planning stage, the test was based on the model of having pairs of students respond to a question together, i.e., an interactive pair assessment. With the long-term goal of migrating the assessment online, we knew the test had to be flexible enough to support peer-to-peer assessment while being administrable in an online format, unlike traditional speaking tests which often require examiners and examinees to meet in person (Isaacs, 2016). Therefore, we had to consider new ways of testing interaction through the implementation of backward design for test creation, development of multimedia test prompts, and the use of video recordings of test-taker performances for later assessment.

Backward Design

Backward design, defined by Richards (2013) as starting "with a careful statement of the desired results or outcomes: appropriate teaching activities and content are derived from the results of learning" (p. 20), was popularized by Wiggins and McTighe (1998) in the United States by connecting it to curriculum design. The idea is that the teacher should have the final assessment in mind before planning course and unit activities. According to Wiggins and McTighe, there are three stages, which are (1) identify desired results, (2) determine acceptable evidence of learning, and (3) design learning experience and instruction. The concept was used by Jensen et al., (2017) as a heuristic for research projects by following and adapting the three stages. Also, the CEFR is an example of using this design method in language learning (Richards, 2013). In our project, backward design was used to clarify the steps we needed to take, as follows: identify the range of student abilities using the CEFR; determine what language use demonstrates student abilities within a range; and, design test questions that elicit the targeted language.

Backward design was useful in this pilot project because the end goals were identified prior to the design stage and acted as end points for determining success at each stage. For example, we wanted to observe how pairs utilized nonverbal communication during interaction. Because we knew what was going to be measured, we understood that we had to video record the test, as opposed to just an audio recording. Therefore, backward design was used in the development of all components of the test, technological and otherwise.

Multimedia Prompts in Assessment

The use of multimedia in instructional contexts is well supported empirically and theoretically (Mayer, 2009). However, relatively little supporting research describing best practices in its use in assessment contexts exists (Kirschner et al., 2017). One possible advantage for using multimedia, such as images and videos in the assessment process, is to strengthen the fidelity of task prompts to the real-world context they are meant to represent. In their paper describing a framework for assessing the use of multimedia in assessment contexts, Kirschner et al. (2017) describe ecological validity as the similarities between the task demand of an assessment and those of the target. In an

English assessment context, such considerations regarding attempts to simulate real-life situations in task design resemble Bachman and Palmer's (2010) description of using target language use (TLU) domains to design tasks that promote the use of language input for a task so that it accurately represents the target environment. Kirschner et al. (2017) also points out that the use of video may be one way to align tasks to their respective domain(s). With that in mind, one of the goals of this project was to explore the use of technology in speaking tests.

Video as a Medium for Assessing Test Taker Performances

Focal to the approach in the current project is the use of video recordings of test-taker performances, which supports offsite and multiple viewings by graders. In speaking tests, test-taker output is generally viewed live, or captured in either audio-only or audiovisual format. Becce and Hennessy (2015) conducted a face-to-face speaking test with Japanese university students using the CEFR-J descriptors for differentiating test-takers' levels. The researchers reported that due to the number of test takers and limited time, the assessment needed to be conducted as quickly as possible. This prompted us to video record the performances to relieve the time sensitivity that results from assessing test taker performance in real time. Another benefit of recording is that the raters can re-watch or pause the student videos at any time to make notes or reference a rubric.

One potential issue with the introduction of video-recorded performances may be score variations due to reliance on this medium. Therefore, understanding if there are differences in scores awarded to performances caused by differences in the format used in the assessment process is vital to the integrity of the assessment. Nakatsuhara et al. (2021) investigated whether scores differed on IELTS speaking tests depending on whether the performances were graded live or using audio-only or video recordings. Their research showed that a video format supported the assessment process by providing a more holistic account of performance when compared to assessments based on audio-only performance data and that scores awarded to performances assessed through video were comparable to those given to the live version of performances. These findings suggest that video can act as a suitable alternative to live performances in spoken language testing.

In short, the test was structured using backward design, specifically with attention to the following aspects: the range of abilities; students' ability to demonstrate their level within the range; and test questions that elicit student production of such language ability. We had to theoretically understand the role of multimedia in testing and practically apply the theory in making our own image and video question types. Finally, we realized that raters would benefit from being able to watch test takers without the time and location constraints of grading the test live, so we decided to record each test.

Test Setup

Learning Management System

In 2010, an information and communications technology (ICT) program was created at the university for the purpose of supporting and expanding the capabilities of foreign language instruction. As a part of this, Moodle, which is a learning management system (LMS), was installed for use by departmental instructors (Petersen et al., 2020). To support remote grading in the current project, videos of test taker performances were stored in the LMS and accessed by raters through their own devices.

Since the school already had an LMS in place it was initially decided to use it as part of the pilot test. Using backward design, we considered the way the tests would be evaluated by the raters and then tried to realize that in Moodle. Our initial design was to have each rater with their own Moodle course page specifically created for them with each student test video uploaded as an assignment. We tested this method with three raters, who logged into their Moodle course, watched the video, and graded the performance of either Student A or Student B (Student A was always the student sitting on the left side). As the raters watched the videos, they recorded scores in the Moodle assignment tool to grade the performances. Raters had the option to pause the video at any time, and did not need to refer to other documents, such as a printed handout or a different web browser tab, to complete the grading process.

Testing Room

As stated previously, the long-term goal of this project is to make the test fully online, but to avoid issues that invariably arise in the initial stages of any project, it was decided to do the pilot in a hybrid fashion in order to make the best use of the technology we had selected, and consideration was given to test room setup. Initially, a large space was used to reduce risks associated with COVID-19, but upon a review of video data it was decided that a well-ventilated smaller space with a high-quality microphone (Felyby BM-800) and amplifier (48V Phantom Power) would improve the sound quality and facilitate more accurate assessments. This change was needed because of the use of plastic shield between the participants and their use of masks.

Test-taker performances were recorded using the camera on a Microsoft Surface Pro 7 to reduce the potential anxiety caused by a large camera on a tripod. A portable monitor (ASUS ZenScreen M16A) was used to present a Microsoft PowerPoint slideshow containing the testing prompts to the test takers, who were seated about 2 meters in front of the examiner.

As mentioned, this test was done in a hybrid fashion, so the five raters online participated across multiple locations and prefectures acting as graders. Since there were no issues for the raters, the potential of having the grading process fully online appears possible, negating the need for locally based raters.

Test Prompts

Prompts were presented sequentially on different slides to the test takers. The image and video prompts were placed at the end of the test because it was expected for students to be somewhat relaxed after doing an initial warm up using interview type questions. The PowerPoint presenter view provided the benefit of allowing test takers to see the test prompt. The examiner could see the same test prompt accompanied by the test-delivery script located in the notes section of the slide. We found the hyperlink tool in PowerPoint to be helpful at first because we were able to make inconspicuous links to different test prompts in the slide deck. The idea was to scale questions in terms of difficulty based on the initial assessment of the English proficiency of the test taker by the examiner, but we noticed that this increased test time leading to unacceptably long silences. This was problematic because we wanted to keep test timing consistent. Therefore, future implementations of the test will include role-plays and open questions that support grading different levels of test takers without the need for questions of varying difficulty.

Test Implementation

Test Participants

The onsite testing process included one examiner and two test takers in person, with the raters doing their work online watching and scoring recorded videos. The test was conducted using a Microsoft Surface to show test prompts on PowerPoint slides, while the camera on the Surface was used to record the students. The original plan was to not have feedback from the participants because we had envisioned doing many tests within a set amount of time, but the students offered unsolicited feedback, which caused us to further reflect on and improve the test. Other than comments about specific questions, a common theme was that students stated that they forgot they were being recorded because they did not see a camera.

Timing

At the beginning of the pilot, each assessment was scheduled to last 15 minutes with time between assessments if needed. Upon analyzing the first round of assessments, the researchers noticed that the low-level speakers, when faced with communication failure or the lack of ability to express themselves, stopped speaking completely and looked at the examiner. High-level speakers slowed down after 1-2 minutes of continual dialog and then ask the examiner if they should keep going or if they were allowed to finish. Since disfluency among lower-level test takers increased the assessment time and the high-level speakers consistently ended their interaction within 1 or 2 minutes, it was decided that the examiner would end the conversation after 2 minutes in future tests.

Question Prompt Interaction

When students were provided with image and video prompts, they had to discuss what option they thought was best with their partner. This produced interaction between the two test takers as they discussed the prompts. The ability of the students to engage with each other was encouraging and we hope to develop the question types further. The video questions appear to be effective based on the amount of time of engagement elicited. Many paper-based proficiency tests rely on images and texts for discussion, but in this pilot test, which used both image prompts and video prompts, the test takers produced similar amounts of conversation. The average time taken for image-based prompts took 2:40 (minutes and seconds), while the video prompts elicited an average time of 2:22.

Reflections

As noted above, the students often commented that they forgot about, or did not notice, the camera. This could potentially be a good sign for online-only tests, in that the recording technology was not intrusive, suggesting that students may be more likely to engage in natural interaction, and that if an online test is deployed, screen recording technology is less obtrusive than what we are currently using. Additionally, the test was done in person following COVID-19 safety precautions, which required us to consider the use of professional microphones in order to evaluate the student performances. However, in the future, if this test is online, the use of online conferencing tools could remove the need for extra equipment.

Feedback

As stated earlier, the students voluntarily provided feedback on the test which showed that they were happy to offer comments, so we plan to set up a post-test questionnaire for the next iteration of the test. The questionnaire will ask students what they thought about the question types, camera usage, anxiety, and technology.

Video Prompts

Upon review of the pilot, it was decided to make new videos in the hope of expanding on what we had. When we designed the test using backward design, we noted expected content and vocabulary appropriate for students at levels comparable to English commonly taught in Japanese public high schools. Therefore, we expected students at a minimum to be able to converse at that level and designed the test questions to target that language using the CEFR-J as a guide. A QR code link to an example video that promoted student interaction is provided in Figure 1.

Figure 1
Video Question

The short silent video, called "Bump", shows a student looking at their phone and walking into another student carrying papers. The student holding the papers drops them and they both pick up the papers together. When the video was made, it was expected that the students would discuss the cause of the situation, what happened next, and perhaps point out that the characters should be more careful. While this was usually the case, interactions of some pairs developed in completely different ways than anticipated. For example, two different pairs discussed the situation in the video as the beginning of a love story from a TV program. Although this in itself is not a bad thing, it became obvious that topics like these, which were beyond the range of test-takers' competence, hurt the scores of lower-ability students. These students suffered communication breakdowns as they tried to discuss situations for which they seemed to have no English frame of reference.

Moodle

It was decided to alter the grading system because of the way that the Moodle assignment tool displayed the rubric made grading difficult for the raters. In the previous version, we needed to go into the course page of each rater and download the results of each specific test, which took a lot of extra time to click through every page and to compile the results. The new method now requires the rater to log into Moodle as a student and take a "test," with scores being input as "answers," by typing the score into a fill-in-the-blank-style test question representing the different scoring criteria in the rubric. This system has proven to be useful because each scoring session can be done on one screen, all the tests are in one location, and it is easy to compile all scores into one worksheet for review.

Overall, there have been no issues using Moodle to make videos accessible to raters, but we still want to improve this delivery method. One option is a video plug-in tool in Moodle, which allows better video distribution and score tabulation. The current method of having a "test" that the raters answer is an improvement, but there are still issues with how the test is displayed to raters, the video embedding process, and the time it takes to set up the test. Ideally, implementing the Moodle video assessment tool will help streamline the process in the future.

Conclusion

The test, in its current state, is promising. Using backward design, we were able to make a test of interactional English proficiency. Not only is the test practical, but the utilization of technology was integral in its production. The use of a LMS to allow raters the ability to evaluate the participants' videos helped to address the issue of evaluation time reported by Becce and Hennessy (2015). As described above, video test delivery medium resulted in similar lengths of speaking time when compared to the more traditional method of using images. Video, as a question type in tests, has been useful in prompting students to discuss what they watched from a point of view that is more relatable to them which connects to better ESL assessments (Kirschner et al., 2017).

At the same time, the test needs further development to meet our goals. The current iteration of the test is a hybrid of online and face-to-face, with improvements needed to

migrate it to a completely online format. In order to do this, we need to collect student feedback, revise video question prompts, and reconsider how Moodle is used. With that in mind, what we have done so far shows that a completely decentralized test in which all participants may be in different locales is possible. This would be beneficial for rural schools and students who want to take a proficiency test but face difficulties due to travel and costs.

As emphasized by Suzuki (2019) rural areas in Japan are disadvantaged in many ways, including access to educational opportunities. Having the test completely online would make it more inclusive, and in the future students in rural communities could be aided by having more educational resources online (Mehta & Kalra, 2006). For these reasons, it is our intention to increase the scale of this project to allow for test taking to be available fully online in the future to help provide access to test takers in more rural areas.

With a fully online version of this test appearing to be feasible, the test shows promise at the time of this writing, but more work is needed. It has been demonstrated that communicative testing using image and video prompts are practical and engaging for students. Considering current limitations in speaking assessments, this system offers some promise to advance the state of spoken language assessment in Japan and beyond.

Jacob B. Petersen is an associate professor in the International Education Center at Iwate University. His research interest is the use of educational technology in the classroom. <jacobp@iwate-u.ac.jp>

Daniel Newbury is an associate professor at Fuji University. His research interests are in the use of technology in language learning and the processes integral to oral communication. <daniel@fuji-u.ac.jp>

Natsumi Onaka is a professor and assistant director in the International Education Center at Iwate University. Her current interests are assessment of English-speaking skills for meaningful interaction and intercultural co-learning. <onaka@iwate-u.ac.jp>

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20K00855.

References

Bachman, L. F., & Palmer, A.S. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford University Press.

Becce, N., & Hennessy, C. (2015). Perceptions of Japanese EFL student oral language ability: Learner self-assessment versus interviewer assessment using CEFR descriptors. *Kokusai Kyoiku Koryu Kenkyu, 2,* 1-12. Retrieved from https://www.u-fukui.ac.jp/wp/wp-content/uploads/Nicolangero-Becce-Christpher-Hennessy.pdf.

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

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities, In L. Araújo (Ed.), *Computer-based assessment of foreign language speaking skills, 29*, 51. https://doi.org/10.2307/j.ctvvngrq.10

Haywood, H. C., & Tzuriel, D. (2002). Applications and challenges in dynamic assessment. *Peabody Journal of Education, 77*(2), 40-63 https://doi.org/10.1207/S15327930PJE7702_5

Isaacs, T. (2016). Assessing speaking. *Handbook of second language assessment*, *12*, 131-146. https://doi.org/10.1515/9781614513827-011

Jensen, J. L., Bailey, E. G., Kummer, T. A., & Weber, K. S. (2017). Using backward design in education research: A research methods essay. *Journal of Microbiology & Biology Education, 18*(3), 18-3. https://doi.org/10.1128/jmbe.v18i3.1367

Kao, Y. T. (2020). A comparison study of dynamic assessment and nondynamic assessment on EFL Chinese learners' speaking performance: Transfer of learning. *English Teaching & Learning*, *44*(3), 255-275. https://doi.org/10.1007/s42321-019-00042-1

Kirschner, P. A., Park, B., Malone, S., & Jarodzka, H. (2017). Toward a cognitive theory of multimedia assessment (CTMMA). In M. Spector, B. Lockee, & M. Childress (Ed.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy,* 1-23. https://doi.org/10.1007/978-3-319-17727-4_53-1

Köroglu, Z. Ç. (2019). Interventionist dynamic assessment's effects on speaking skills testing: Case of Elt teacher candidates. *Advances in Language and Literary Studies, 10*(3), 23-31. https://doi.org/10.7575/aiac.alls.v.10n.3p.23

Mayer, R. E. (2009). *Multimedia learning.* Cambridge University Press. https://doi.org/10.1017/cbo9780511811678

Mehta, S., & Kalra, M. (2006). Information and communication technologies: A bridge for social equity and sustainable development in India. *The International Information & Library Review, 38*(3), 147-160. https://doi.org/10.1080/10572317.2006.10762716

MEXT. (2018). 高等学校学習指導要領（平成30年告示）解説 外国語編 英語編 [High school curriculum guidelines (2018 notice) commentary: Foreign language edition English edition.]. Retrieved October 17, 2021, from https://www.mext.go.jp/content/1407073_09_1_2.pdf

Nakatsuhara, F., Inoue, C., & Taylor, L. (2021). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly, 18*(2), 83-106. https://doi.org/10.1080/15434303.2020.1799222

Patharakorn, P. (2018). Assessing interactional competence in a multiparty roleplay task: A mixed-methods study. Doctoral dissertation, University of Hawai'i at Manoa.

Petersen, J. B., Townsend, S. D., & Onaka, N. (2020). Utilizing flipgrid application on student smartphones in a small-scale ESL study. *English Language Teaching, 13*(5), 164-176. https://doi.org/10.5539/elt.v13n5p164

Richards, J. C. (2013). Curriculum approaches in language teaching: Forward, central, and backward design. *RELC Journal, 44*(1), 5-33. https://doi.org/10.1177/0033688212473293

Suzuki, K. (2019). Social equity in Japan. In M. Johansen (Ed.), *Social equity in the Asia-Pacific region* (pp. 159-175). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-15919-1

Wiggins, G., & McTighe, J. (1998). What is backward design. *Understanding by Design, 1*, 7-19.