

Assessing students' English presentation skills using a textbook-based task and rubric at a Japanese senior high school

Rie Koizumi¹ and Ken Yano²

rkoizumi@juntendo.ac.jp

1. Juntendo University, Chiba

2. Taga Senior High School, Ibaraki

Abstract

Assessing as well as teaching speaking English as a second language (L2) is encouraged in the classroom because there are potential opportunities outside the classroom for native and nonnative speakers of English to interact in English. However, speaking assessment is not conducted regularly in Japanese senior high schools (SHSs). One measure to promote speaking assessment is to introduce an instance of speaking assessment in the classroom to show detailed procedures and outcomes based on the analysis of the test data. This study reports on such an attempt to conduct speaking assessment of oral presentations based on a textbook task and a rubric at an SHS. Presentations of 64 students were evaluated by two raters using two rating criteria. Analysis of scores using many-facet Rasch measurement showed that the test functioned well in general, and the results of a posttest questionnaire suggested that students generally perceived the test positively.

Keywords: speaking assessment, analytic rubrics, raters, many-facet Rasch measurement, students' perceptions

Despite the wide recognition that speaking assessment is essential, it is not conducted regularly in Japanese senior high schools (SHSs). Since 2013, the Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2018a) has conducted an annual survey of English language teaching, asking public schools whether they conducted performance tests (including speaking and writing tests) and if so, how many times they conducted speaking tests (including speeches, interviews, presentations, discussions, and debates) and writing tests. The results are useful for understanding the current state of speaking test administration, although they are based primarily on teachers' self-reports (see also MEXT, 2018b, for the average number of times speaking tests are conducted in each prefecture and plans to improve the current situation). Figure 1 summarizes the general upward trends of conducting performance tests. In the case of junior high schools (JHSs), it was reported that the percentage of performance tests conducted increased from approximately 92.33% to 96.76% (almost all schools) in third-year classes at JHSs (MEXT, 2014, 2018a). When third-year JHS teachers conducted speaking tests in 2017, they did so 3.20 times, on average (i.e., 29,040/9,070); popular test formats used were speech (36.45%), interview (34.40%), and presentation (20.08%; MEXT, 2018a).

In contrast, at senior high schools (SHSs), in the case of Communication English I classes in general courses (*futsuka*), the percentage of performance tests conducted increased from 54.01% to 69.57% (MEXT, 2014, 2018a), which is a relatively constant increase but far behind the JHS situation. When such class teachers conducted speaking tests in 2017, they did it 2.46 times, on average (3,908/1,591). Popular test formats were speech (33.44%), interview (30.53%), and presentation (28.76%), in the same order as JHSs (MEXT, 2018a). In sum, while the percentage of performance tests conducted gradually increased at JHSs and SHSs, almost all JHSs conducted performance tests, whereas only two-thirds of SHSs did so; the frequency at which tests were conducted was not very high. Additionally, MEXT's survey did not examine the quality of the performance assessment, and it remains to be investigated to what extent performance assessment was properly conducted.

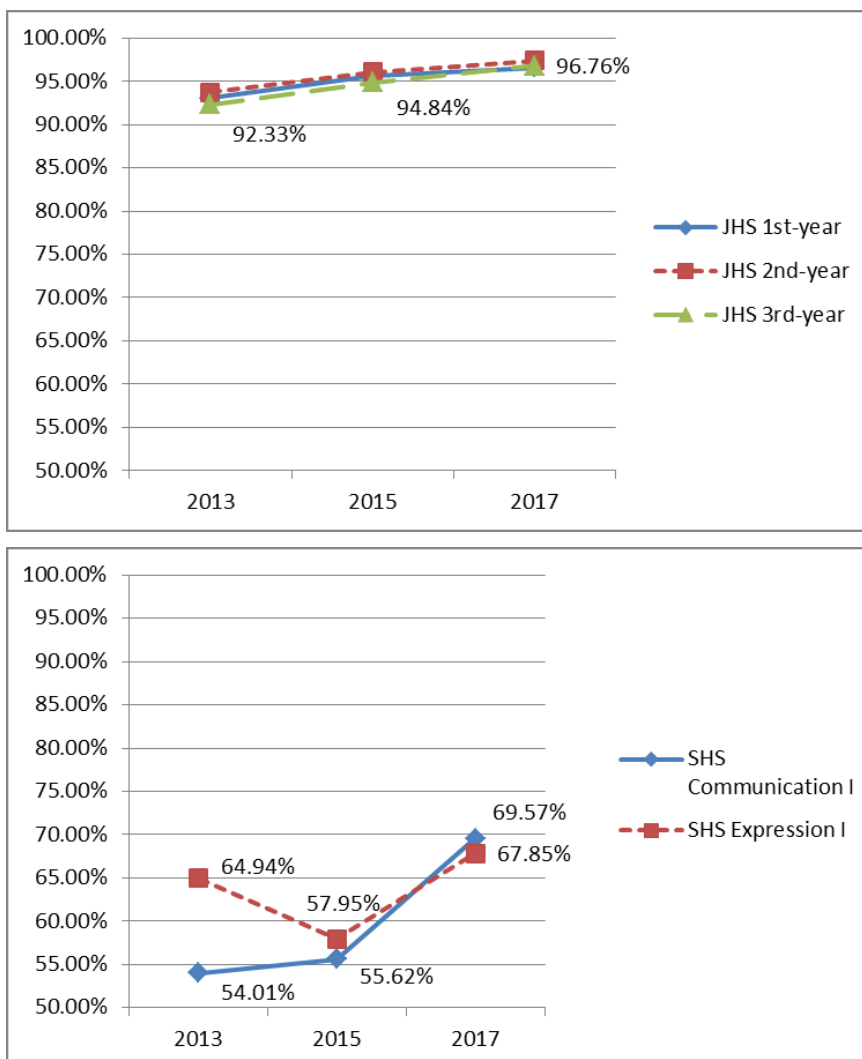


Figure 1. Percentages of performance tests conducted at JHSs (above; with the numbers displayed for the third-year) and SHSs (in general course; below). Percentages were calculated by (a) “the number of schools that conducted (including those that said they would conduct) performance tests” / (b) “the number of all schools that responded” * 100 for JHSs; (a) / (c) “the number of schools that have a certain class (e.g., Communication English I)” * 100 for SHSs. Data for this figure were derived from MEXT (2014, 2016, 2018a).

Attempts to Increase and Improve Speaking Assessment Practices

To increase and improve speaking assessment practices at Japanese schools, various measures have been planned and implemented. At the national level, knowledge and skills of English assessment will be incorporated as essential components in the Core Curriculum in pre-service and in-service teacher training programs for JHS and SHS teachers (Tokyo Gakugei University, 2017). In this context, teacher training sessions with a special focus on speaking assessment have recently been held. Books on the theory and practice of speaking assessment are available for a worldwide audience (e.g., Fulcher, 2003; Luoma, 2004;

Taylor, 2011), including English instructors in Japan (e.g., Koizumi, In'nami, & Fukazawa, 2017; Talandis, 2017).

Furthermore, previous studies provide useful hints that can help SHS teachers learn about speaking assessment. For example, Akiyama (2003) conducted a speaking test that consisted of speech, role-play, picture description, and interview tasks at Japanese JHSs and analyzed student performance using many-facet Rasch measurement (MFRM). Nakatsuhara (2013) conducted a group oral test at five Japanese SHSs, in which a group of students interacted using information gap, ranking, and free discussion tasks. She analyzed the effects of students' personalities and the number of members in a group on speaking performance. Ockey, Koyama, Setoguchi, and Sun (2015) used various test formats (e.g., oral presentation, group oral discussion, picture and graph description, and the Test of English as a Foreign Language Internet-based Test [TOEFL iBT] tasks) at a Japanese university. They reported that there are strong or moderate relationships ($r = .76$ at most) between them and that different test formats assess shared speaking ability, but that each also assesses different aspects of ability. These instances of speaking assessment provide teachers with practical information on how to develop and administer a speaking test.

However, these resources are not clearly linked to the textbooks authorized by MEXT and used in daily lessons at SHSs. An explicit association between instruction and assessment is needed for formative and summative speaking assessment in the classroom. Furthermore, rubrics with detailed descriptions and various (oral or transcribed) speech samples are not sufficiently provided. Therefore, the study of the development and examination of a speaking assessment task and a detailed rubric based on an authorized textbook would be helpful to fill this void. The current study attempts to address this issue.

Current Study

We pose the following five research questions (RQs) to examine a speaking test in detail. Specifically, we examine aspects of the validity of interpretations and uses based on scores of an oral presentation assessment, using many-facet Rasch measurement (MFRM).

RQ1: To what degree are student responses consistent with the responses predicted by MFRM?

RQ2: To what degree do raters score similarly and consistently?

RQ3: To what degree do rating criteria function as intended?

RQ4: To what degree are there biased interactions between students, raters, and rating criteria?

RQ5: To what extent do students respond positively to potential effects of the speaking test on students' learning?

A term used in RQ1 may require further explanation. As for the "responses predicted by MFRM," MFRM predicts that students will provide response patterns in the following manner: Students with higher speaking abilities are likely to get higher scores on evaluation criteria from raters, and students with lower speaking abilities are likely to get lower scores on evaluation criteria from raters. RQ1 examines if actual student performances conform to this predicted pattern.

The results in relation to RQ1 to RQ5 are expected to provide evidence for inferences when we make a validity argument for this test using an argument-based approach to validity (Chapelle, Enright, & Jamieson, 2008). To be specific, the RQs correspond to inferences as follows: RQ2 to evaluation and generalizability inferences, RQ3 to evaluation and decision inferences, RQ4 to an evaluation inference, RQ5 to a consequence inference, and RQ1 is not directly related to any inferences (see Knoch & Chapelle, 2018; Koizumi, 2018, for details). A detailed analysis of the assessment features should contribute to an understanding of how to construct effective speaking classroom assessment involving tasks, rubrics, and raters.

Method

Participants

Second-year students in two classes at a public SHS took a speaking test as part of their Communication English II subject ($N = 64$) in January and February 2017. This was a compulsory English class required for graduation, and all students in two classes (except for those who were absent) took the test. One class was for students in a humanities course ($n = 36$), whereas the other was for those in a science course ($n = 28$). Some students were highly motivated, whereas others were only slightly or not very much motivated to speak English. Overall, their English proficiency level was fairly high for second-year SHS students. While each class had 40 students, 16 students were absent because of a flu epidemic.

Materials

Speaking Assessment Task and Rubric

Students used the textbook *Genius English Communication II* (Muranoi, Mano, Hayashi, Hatano, Yamaga, Uekusa, & Taishukan, 2016) in the Communication English II class where they learned from one teacher. One chapter deals with the topic of animal emotions (Chapter 8: Emotions Gone Wild) and has a project task at the end. The speaking assessment was planned based on this project.

The textbook was selected by school teachers who found the overall structure, topics, and tasks included interesting and relevant to students' English proficiency levels and interests. It has ten chapters that cover a variety of topics. Each chapter starts with warm-up activities involving listening and key vocabulary learning. In addition to a reading passage, it includes text-based communicative tasks (i.e., summarizing, answering questions, discussion, and a project), a summary of grammatical points and activities using the target grammar, and further reading material. We considered the textbook well balanced, and that it would enable teachers to conduct both form-focused and meaning-focused instruction. We focused on a project task, one of the tasks in a chapter, because this task provides students with an opportunity to relate the content of a reading passage to their interests. By using it as an assessment task, we expected that we could enhance learning as much as possible by having students seriously prepare and present their learning in the form of an oral presentation, which would help them to acquire useful presentation skills and vocabulary.

In planning the speaking assessment, we first created test specifications and an oral presentation task (see Tables 1 and 2). Oral presentation is one of the many useful formats that can assess speaking ability, which is sometimes used at the SHS level (MEXT, 2018a). This format can elicit a prepared monologue based on a student's interests, but one weakness is that students tend to simply read aloud their scripts, which also may not have been written by themselves. While keeping such factors in mind, we conducted the oral presentation assessment based on the course objectives and textbook. In the presentation task, students were asked to create a poster and make a script to be presented in a group. Additionally, a rubric was created to assess a one-minute presentation (see the English version in Table 3; see Appendix A for the original Japanese version and Appendix B for samples of presentations with rationales of getting certain scores).

We selected an analytic rubric type that had two criteria based on the class objectives and test construct: Task achievement and Fluency. Although it was possible to have a holistic rubric including multiple aspects in one criterion, we used an analytic scale because the test purposes were not only to check the achievement but also to give students feedback.

Table 1

Test Specifications for the Oral Presentation

Test purposes:	To check achievement and give students feedback
Lesson objective:	Can fluently make a presentation about animal emotion and one's opinions based on the textbook and research
Test construct (ability to be measured):	The ability to speak appropriately and fluently about information and opinion
Task:	An individual oral presentation
Rubric:	Analytic type; criteria: task achievement and fluency; three levels

Table 2

Presentation Task

By using websites and other resources, find scientific studies on animal emotions and summarize their findings. Make a speech for one minute. Include the following points:

- (1) What animal is it?
 - (2) What emotion did it show?
 - (3) In what situation? Explain concretely.
 - (4) Explain the emotion scientifically (Primary? Secondary?)
 - (5) Your comment
-

Note. Based on the Project task in Muranoi et al. (2016, p. 117). In terms of (4), two types of emotions were described in the textbook: Primary emotions are “basic, inborn emotions,” which “require no conscious thought” such as “anger, happiness, sadness, disgust, fear, and surprise”; secondary ones are “more complex” and involve “conscious thought” such as “regret, longing, or jealousy” (pp. 110–113).

Table 3
Rubric for the Presentation Test (in English)

	Task achievement	Fluency
A (Satisfies to a large degree)	The presentation (a) describes (1) a situation in which an animal has a certain emotion, (2) a scientific explanation, and (3) an opinion; and (b) is fully comprehensible and detailed.	There are no long pauses (five seconds or more). Repetition and correction do not hamper comprehension. The presentation is conveyed smoothly. The student does not look at the script most of the time.
B (Mainly satisfies)	The presentation satisfies only (a). ^a	There is one long pause. Relatively many repetitions and corrections sometimes hamper comprehension. The presentation is conveyed relatively slowly. The student sometimes reads the script aloud. The presentation has characteristics of the descriptions of Level B.
C (Requires more effort)	The presentation does not satisfy (a).	There are two or more long pauses. Comprehension is difficult owing to many repetitions, corrections, and/or slow speed. (x) The student reads the script aloud most of the time. The presentation has characteristics of the descriptions of Level C. If (x) is observed, the rating is always C.

Note. ^aThe following will be added to future scoring: Most of (a) is satisfied AND (b) is fully satisfied.

The first criterion focused on whether students could convey their presentation contents clearly. The task achievement criterion looked at two components, (a) whether the presentation described specified content (consisting of three elements: (1) a situation, (2) a scientific explanation, and (3) an opinion) and (b) whether the presentation content was comprehensible and detailed. As long as the presentation was comprehensible, minor errors were ignored. The second criterion focused on whether their delivery was fluent and also included the element of looking at their script. If students read it aloud, they got a lower score. This is because students who tend to read out the prepared script of a presentation are likely to give an apparently fluent presentation.¹ The criteria had three levels: A, B, and C. C was the level of not fulfilling the required performance, B of minimally fulfilling it, and A of going well beyond the required level. This three-level system is based on assessment guidelines for SHS teachers from the National Institute for Educational Policy Research (2012). While it may be possible to distinguish levels further, within a limited time for assessment conducted in class, three levels were judged to be sufficient (see Muranoi et al., 2017, 2018, for other rubric examples of various discussion and project tasks).

Questionnaire

A posttest questionnaire aimed at learning about students' perceptions of the test, especially how they felt the test affected their speaking ability and learning (Q9 to Q11; see Table 4). It included 11 Likert-type questions, each of which was judged using a scale of five, with 1 being "No, not at all," 2 "No, not much," 3 "Neither no or yes," 4 "Yes, a little," and 5 "Yes, very much." The questionnaire also involved open-ended questions that elicited their impressions to improve the test (Q12 to Q15).

Table 4

Questions in the Posttest Questionnaire

1. Do you think you performed well on the test?
2. Do you think this test assesses English speaking ability, in general?
3. Did you prepare enough for this test?
4. Was there enough time to say what you prepared?
5. Were you able to convey your message in the question and answer session after the presentation?
6. Were you anxious or nervous during the test?
7. Were you disturbed by people or noises around you during the test?
8. Were you motivated to speak much during the test?
9. Do you think this test helps you develop English ability?
10. Did this test make you feel that you should study English more?
11. Do you think repeated activities like this test will enhance your English ability?
12. Did you have any problems while taking this test? If so, write about them in detail (e.g., I was not able to hear other students' presentations).
13. What was difficult when you took this test (e.g., I did not know the pronunciation of some words)?
14. Have you ever taken a speaking test before? If yes, how do you compare this test with tests you took previously?
15. Do you have other opinions or impressions?

Note. Q1 to Q11: Likert-type questions on a scale of 1 to 5. Q12 to Q15: open-ended questions.

Procedure*Before the test*

While reading the textbook chapter, students were instructed, as a brainstorming activity, to write examples of animals that are not described in the textbook but that have similar characteristics. After reading through the chapter, students started to work on a presentation project. They spent one lesson (i.e., one class period) preparing for and practicing the presentation. First, they were given a worksheet that included a table with information on young bears that was covered in the textbook (see Appendix C). While filling out the worksheet individually, they were asked to summarize their findings and opinions based on their search for relevant cases, while comparing the textbook information with their data. They then created a poster that included a picture of an animal they would talk about, as well as keywords for their presentation. During the preparation stage, all activities consisted of independent work. Then, the students formed pairs and practiced giving their presentations with each other until the lesson was over.

On the test day, the teacher explained how the presentation would be evaluated by showing the rubric to be used. Students were also informed that the teacher and a guest teacher (i.e., an external researcher) would serve as raters. In the first lesson, a test administrator was also in the classroom for group formation and time management (see the *During the test* section below).

During the test

The assessment was conducted using two 60-minute lessons (i.e., two class periods). There were five students in a group, each of whom took turns and made a presentation and answered questions from other students in the group. During and after the presentations, students were asked to take notes and select the

best presentation of the day, while filling out the worksheet (see Appendix C). When all five members finished the presentation, they dissolved the group and made another group. A group formation plan was presented in advance by the teacher (see Figure 2).

Rater 1	Rater 2	Teacher's desk								
Group 1	Group 1	Group 5	Group 5	Group 5	Group 6	Group 6				
Group 1	Group 1	Group 4	Group 5	Group 5	Group 6	Group 6				
Group 1	Group 2	Group 4	Group 4	Group 7	Group 7	Group 6				
Group 2	Group 2	Group 4	Group 4	Group 7	Group 7	Group 7				
Group 2	Group 2	Group 3	Group 3	Group 8	Group 8	Group 8				
	Group 3	Group 3	Group 3	Group 8	Group 8					

No.	Name	Teachers' evaluation	1st trial	2nd trial	3rd trial	4th trial	5th trial	6th trial	7th trial	8th trial
1	Student 1	1st trial	Group 1	Group 2	Group 3	Group 8	Group 3	Group 3	Group 3	Group 3
2	Student 2	1st trial	Group 1	Group 7	Group 2	Group 3	Group 2	Group 5	Group 6	Group 7
3	Student 3	1st trial	Group 1	Group 4	Group 8	Group 4	Group 7	Group 2	Group 2	Group 5
4	Student 4	1st trial	Group 1	Group 3	Group 5	Group 7	Group 5	Group 7	Group 5	Group 2
5	Student 5	1st trial	Group 1	Group 6	Group 4	Group 5	Group 4	Group 4	Group 4	Group 4
6	Student 6	4th trial	Group 2	Group 6	Group 8	Group 1	Group 5	Group 5	Group 8	Group 4
7	Student 7	2nd trial	Group 2	Group 1	Group 5	Group 3	Group 2	Group 7	Group 5	Group 5
8	Student 8	7th trial	Group 2	Group 3	Group 3	Group 5	Group 4	Group 2	Group 1	Group 7
9	Student 9	8th trial	Group 2	Group 4	Group 2	Group 7	Group 8	Group 3	Group 4	Group 1
10	Student 10	5th trial	Group 2	Group 2	Group 6	Group 2	Group 1	Group 6	Group 6	Group 3
11	Student 11	3rd trial	Group 3	Group 2	Group 1	Group 5	Group 5	Group 7	Group 4	Group 4
12	Student 12	5th trial	Group 3	Group 6	Group 3	Group 8	Group 1	Group 8	Group 8	Group 2
13	Student 13	6th trial	Group 3	Group 5	Group 4	Group 6	Group 6	Group 1	Group 6	Group 8
14	Student 14	4th trial	Group 3	Group 4	Group 7	Group 1	Group 2	Group 4	Group 2	Group 5
15	Student 15	7th trial	Group 3	Group 7	Group 8	Group 7	Group 4	Group 2	Group 1	Group 7
...

Figure 2. Group formation plan. The above shows how raters and students were seated in the activity. For example, five students in Group 1 were seated in the left, front area of the classroom. Five students in Group 8 were seated in the right, back area of the classroom. The table below shows how each student was assigned to a group. Yellow cells indicate when students were assessed by the raters (when they were in Group 1). For example, Student 1 belonged to Group 1 and was assessed in the first trial, and s/he belonged to Group 2 in the second trial, Group 3 in the third trial, Group 8 in the fourth trial, and so forth. Student 6 was evaluated in the fourth trial, whereas Student 7 was evaluated in the second trial.

During the assessment, time was managed as follows. One student was required to speak for 1 minute; then 30 seconds were given for a question and answer session; thus, 1.5 minutes were needed for each presenter. Since a group had five members, 7.5 minutes were needed in total (1.5 min x 5). There were eight groups, so 60 minutes were required in all. In addition, between the group presentations, 3 minutes were given to make a new group; thus, 21 minutes were needed in total (3 min x 7 intervals). Further, instructions were given before the test for 5 minutes in each lesson (because two lessons were used and there were some students absent from the first lesson). Therefore, 91 minutes (60 min + 21 min + 5 min x 2 times) were used overall.

When students belonged to Group 1 in a session, they were scored by two raters (marked by yellow highlights in Figure 2). To record their performance, there was a voice recorder near the student who was being scored, and a video camera, for cases where the raters needed to check students' performance later.

After the test

Students submitted their worksheet and poster and answered a questionnaire. In two weeks, they received feedback on a score report (see Appendix D).

Scoring

Before the assessment day, an external researcher (Rater 1) and a teacher rater (Rater 2) decided the rubric first. They then watched videos of students (who were different from the current study but with similar abilities) engaging in similar activities, and independently evaluated their performance using the rubric. They discussed their scores (i.e., ratings) until they reached agreement.

During the presentations, the two raters initially rated the first three students in a class and then discussed their ratings and adjusted their criteria. It may be ideal to adjust the criteria before the assessment starts, but without actual presentation samples of the same task, it was considered to be safe to discuss the rating in the beginning.

Then, based on the agreed-upon criteria, they marked the presentations independently. The two raters were not asked to evaluate the presentations in the same manner, but to bring their perspectives into alignment and use their expertise in judging the presentation quality while maintaining consistency based on the criteria, as assumed in MFRM (Eckes, 2015; McNamara, Knoch, & Fan, 2019). The independently rated scores were used for MFRM. After the scoring, the researcher checked where and how the two raters diverged, and the two raters discussed diverging points for improvement.

For the analysis, three levels (i.e., A, B, and C) were converted into 3, 2, and 1. The final scores were calculated by averaging the two raters' scores. These scores were used for giving feedback as well as grading. The presentation test scores were combined with scores of other performance tests such as a writing test and a recall test (i.e., a one-on-one test with a teacher in which a student is required to summarize the textbook content using keywords and answer questions from the teacher). These performance test scores accounted for approximately 30% of the total grades. The use of speaking test scores for student grades was explained in advance. The assessment in the present study was used for both formative and summative purposes, but the stakes were not high since it was only one portion (30%) of their final grades.

Analysis

For RQ1 to RQ4 using presentation scores, the partial credit model of many-facet Rasch measurement (MFRM) was used, employing FACETS (Linacre, 2019), with three facets included: test-taker ability, rater severity, and rating criterion difficulty. Agreement ratios were also calculated for RQ2 using langtest

(<http://langtest.jp/shiny/kappa/>). To analyze responses to the Likert-type questionnaires (RQ5), the percentages of student responses were computed.

MFRM is a statistical model of test analysis that enables researchers to translate ordinal scores into interval scores located on a logit scale, where multiple facets can be compared (Bond & Fox, 2015; Eckes, 2015; Engelhard, 2013; Engelhard & Wind, 2018; McNamara et al., 2019; Sick, 2009). MFRM produces detailed output that allows researchers to examine test scores from various perspectives. Although the number of the current participants is not very large for MFRM ($N = 64$), we considered it sufficient for our low-stakes purpose, based on Linacre (1994), who claims that 50 participants are minimally necessary for polytomous data, when variations within 1 logit (99% confidence interval) are considered acceptable. We considered standard errors to assess the effects of a small dataset later.

The fit of the data to the Rasch model was examined using infit mean squares. According to Linacre (2018), values between 0.50 and 1.50 are considered acceptable and fitting to the Rasch model, whereas values of less than 0.50 are considered overfitting (meaning that the data behaved more predictably than expected) and values of more than 1.50 are considered underfitting (meaning that the data behaved more unpredictably than expected). While cases of less than 0.50 and more than 1.50 indicate possible problems, values of more than 2.00 are considered to seriously impact the results (Linacre, 2018).

Results

Figure 3 shows a variable map (Wright map) illustrating locational relationships between students, raters, and rating criteria on a logit scale. It indicates that students' ability ranged widely, that the two raters scored in a similar manner in terms of severity, and that Fluency was much more difficult than Task achievement. Table 5 shows means and other descriptive statistics. Despite students' wide-ranging ability, the students' separation and strata were small, 1.46 and 2.29, respectively. These values suggest that students were differentiated into one or two levels at most. The small separation and strata were derived because Model standard errors were large (Mean = 1.18 in Table 6). These values are calculated using "Separation = True SD / Average measurement error" and "Strata = $(4 * \text{Separation} + 1) / 3$." The reliability of students was a little low (.68) because of the small separation. Although there were three levels in the rating criteria, the test was intended to be criterion-referenced and to assess achievement, and the small student separation was considered acceptable in this context.

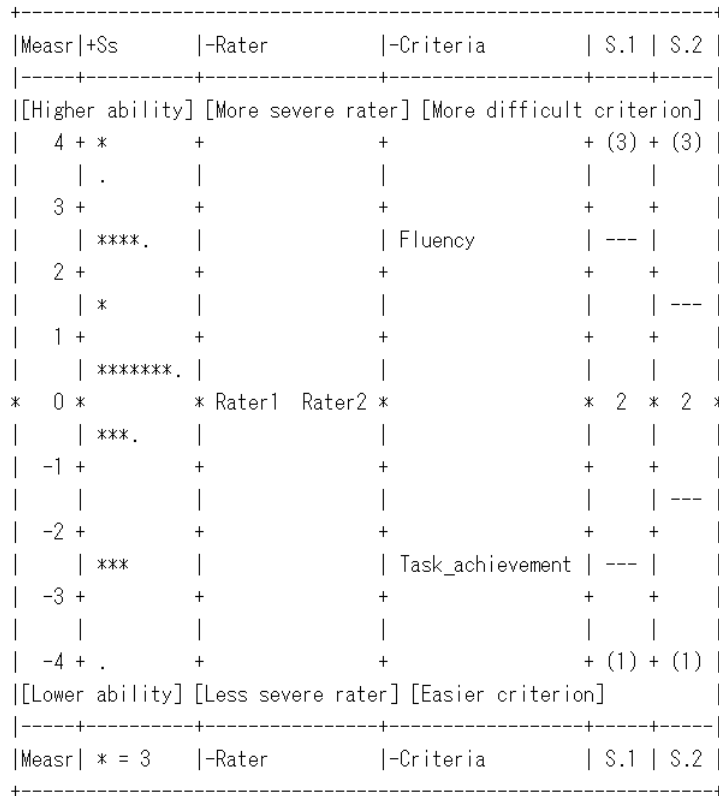


Figure 3. Variable map for students, raters, and rating criteria. S.1 = Task achievement. S.2 = Fluency.

Table 5
Descriptive Statistics of Three Facets

	Mean (SD)	Min to Max	Range	Separation	Strata	Reliability
Students	0.48 (2.16)	-6.40 to 5.20	11.60	1.46	2.29	.68
Raters	0.00 (0.14)	-0.14 to 0.14	0.28	0.00	0.33	.00
Criteria	0.00 (2.36)	-2.36 to 2.36	4.72	11.70	15.94	.99

Table 6

Extract of Student Measurement Report

Obsvd	Fair(M)		Model	Infit		Outfit		Estim.	Corr.
Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	Discrm	PtBis
3.00	2.97	5.20	1.92	Maximum					.00
2.75	2.87	3.71	1.20	2.05	1.20	9.00	2.40	-1.81	-.44
2.50	2.68	2.52	1.02	1.90	1.10	3.48	1.60	-0.63	-.01
2.50	2.68	2.52	1.02	1.66	0.90	3.35	1.50	-0.43	.03
2.50	2.68	2.52	1.02	0.03	-2.40	0.05	-1.00	1.87	.70
2.25	2.44	1.54	0.98	3.02	2.10	3.62	2.30	-2.20	-.39
2.25	2.44	1.54	0.98	0.87	0.00	1.26	0.50	0.61	.42
2.25	2.44	1.54	0.98	0.69	-0.20	0.53	-0.40	1.53	.62
2.00	2.17	0.57	1.00	9.00	6.40	9.00	6.60	-10.00	-.72
2.00	2.17	0.57	1.00	1.71	1.10	1.74	1.20	-0.81	.00
2.00	2.17	0.57	1.00	1.24	0.50	1.29	0.60	0.36	.46
2.00	2.17	0.57	1.00	1.10	0.30	1.11	0.30	0.64	.49
2.00	2.17	0.57	1.00	0.50	-0.80	0.50	-0.90	2.09	.70
1.75	1.84	-0.56	1.16	1.43	0.70	2.10	1.20	0.24	.38
1.75	1.84	-0.56	1.16	1.34	0.60	1.62	0.80	0.43	.42
1.75	1.84	-0.56	1.16	0.87	0.00	0.69	0.00	1.24	.62
1.75	1.84	-0.56	1.16	0.68	-0.20	0.54	-0.20	1.43	.64
1.50	1.33	-2.48	1.62	0.01	-1.10	0.01	0.00	1.45	.70
1.25	1.05	-4.69	1.38	0.83	-0.30	0.42	2.30	1.52	.42
1.00	1.01	-6.40	1.97	Minimum					.00
2.04	2.13	0.48	1.18	0.81	-0.50	1.04	0.00	0.49	Mean
0.42	0.50	2.16	0.29	1.25	1.40	1.75	1.30	0.34	S.D. ^a
0.42	0.50	2.18	0.29	1.26	1.40	1.77	1.30	0.34	S.D. ^b

With extremes, Model, Populn: RMSE 1.22 Adj (True) S.D. 1.78 Separation 1.46 Strata 2.29
Reliability .68

With extremes, Model, Sample: RMSE 1.22 Adj (True) S.D. 1.80 Separation 1.48 Strata 2.31
Reliability .69

Without extremes, Model, Populn: RMSE 1.15 Adj (True) S.D. 1.31 Separation 1.14 Strata 1.85
Reliability .56

Without extremes, Model, Sample: RMSE 1.15 Adj (True) S.D. 1.33 Separation 1.16 Strata 1.87
Reliability .57

With extremes, Model, Fixed (all same) chi-square: 141.3 d.f.: 63 significance (probability): .00

With extremes, Model, Random (normal) chi-square: 60.7 d.f.: 62 significance (probability): .52

Note. Twenty students' results are presented as examples. ^aPopulation. ^bSample.

RQ1: To what degree are student responses consistent with the responses predicted by MFRM?

As explained above, MFRM predicts that students will respond such that higher-ability students tend to perform better and achieve higher scores on the evaluation criteria from raters, particularly in the Task achievement criterion, rather than Fluency (because the former is easier, as shown in the variable map in Figure 3); the opposite is also predicted by MFRM. When actual ratings (and response patterns indicated by ratings) differ from the patterns predicted by MFRM, this is indicated by high or low fit statistics.

The results regarding student model fit suggest that student responses were consistent with the responses predicted by MFRM, to a limited degree (see Table 6, which shows a sample of student results). Half (50.00%) of students had infit mean squares between 0.50 and 1.50 ($n = 32$ out of 64). On the other hand, 15.63% of the students ($n = 10$) had infit mean squares of more than 1.50, showing underfit to the Rasch model, whereas 28.13% ($n = 18$) had lower than 0.50, showing overfit. There were three students (4.69%) with more than 2.00, which indicates that their responses were highly unpredictable.² The analysis of these three students' responses (see Table 7) showed that they had lower Task achievement than Fluency, which contrasted with the overall result that Task achievement was much easier than Fluency (see Figure 3). The reason for lower Task achievement was that these three students did not include or were not able to convey all the required elements for the presentation: The first student's talk was not very comprehensible (see Appendix B, Sample 4 for the transcription), while the second and third students did not include scientific explanations for animal emotion but provided fair details. The content was not complete because two of the students forgot to bring the poster and made a spontaneous presentation. Although the MFRM results showed that their responses were unexpected, it does not seem to be a problem related to this test. In the case of overfitting students (those with infit mean squares of less than 0.50), all the students had one point higher in Task achievement than in Fluency (e.g., 3 vs. 2, respectively), and their responses conformed to the expectations from the Rasch model to a very high degree. These high percentages of underfitting and overfitting students were probably caused by the small dataset (data point = 256), as Linacre (1994) suggests a small sample affects estimates and fit statistics. Future analyses should increase the quantity of data, as suggested by one of the reviewers, which will also be touched upon in the Discussion and Conclusions section.

Table 7
Three Students' MFRM Results and Ratings

No.	Students' ability measure	Standard error	Infit mean squares	Task achievement		Fluency	
				Rater 1	Rater 2	Rater 1	Rater 2
1	0.57	1.00	9.00	1	1	3	3
2	1.54	0.98	3.02	2	2	3	2
3	3.71	1.20	2.05	2	3	3	3

Note. Student 1's presentation can be seen in Appendix B, Sample 4.

RQ2: To what degree do raters score similarly and consistently?

The MFRM results showed that raters score similarly and consistently. Table 8 shows that the exact agreement of the two raters was high and higher (78.90%) than the agreement that MFRM predicted (65.60%). Rater severity was quite similar (0.14 and -0.14). The raters scored consistently, with infit mean squares of 0.91 and 0.94. In other words, both the teacher rater and the external researcher rated similarly and consistently based on the same criteria. Although some may argue that similar and consistent

ratings across raters are not uncommon after rater training, previous studies on rater-mediated assessment suggest that it is sometimes difficult for raters to agree even after formal rater training and individualized feedback (e.g., Eckes, 2015; Knoch, 2011). The result that a teacher rater who practiced rating could score the presentation effectively should be encouraging to teachers who are interested in this type of assessment. To examine details of rater disagreement in some scores, we also computed rater agreement for each criterion. Table 9 indicates that the raters agreed, to a fairly strong degree.

Table 8
Rater Measurement Report

Obsvd Average	Fair(M) Average	Model Measure	Infit S.E.	Outfit MnSq	Estim. ZStd	Corr. PtBis			
2.01	2.07	0.14	0.20	0.91	-0.60	1.10	0.40	0.97	.34
2.06	2.15	-0.14	0.20	0.94	-0.40	1.18	0.70	1.04	.34
2.04	2.11	0.00	0.20	0.93	-0.50	1.14	0.60	0.34	Mean
0.03	0.04	0.14	0.00	0.02	0.10	0.04	0.10	0.00	S.D. ^a
0.04	0.06	0.20	0.00	0.02	0.20	0.06	0.20	0.00	S.D. ^b

Model, Populn: RMSE .20 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability (not inter-rater) .00
 Model, Sample: RMSE .20 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability (not inter-rater) .00
 Model, Fixed (all same) chi-square: 1.0 d.f.: 1 significance (probability): .33
 Inter-Rater agreement opportunities: 128 Exact agreements: 101 = 78.9% Expected: 83.9 = 65.6%

Note. ^aPopulation. ^bSample.

Table 9
Agreement Ratio Between Two Raters

	Task achievement	Fluency	Total ^a
Agreement ratio	71.88%	84.38%	68.75%
Cohen's weighted kappa	.47	.84	.68
(95% confidence interval)	(.23, .72)	(.74, .93)	(.52, .84)
	[Moderate]	[Almost perfect]	[Substantial]

Note. [] = Interpretation based on Landis and Koch (1977), which have the criteria of Cohen's weighted kappa as follows: "< 0.00 Poor; 0.00–0.20 Slight; 0.21–0.40 Fair; 0.41–0.60 Moderate; 0.61–0.80 Substantial; 0.81–1.00 Almost perfect" (p. 165). ^aRatings of Task achievement and Fluency were added and analyzed.

We examined where our ratings diverged to identify possible explanations. In Task achievement, three prominent reasons were observed. First, the raters differed in judging whether the content had sufficient details and whether the speech was comprehensible, and borderline performances received higher or lower ratings (three occasions observed). Second, presentations did not sometimes include the three required elements ((1) to (3), in Table 3), but they contained details. The raters sometimes diverged in the way details supplemented insufficient information (three occasions). A third reason for rater divergence in Task achievement was that some students had effective content but ineffective voice volumes (five occasions). This tended to occur when there was a lot of noise around students (because other groups were also making their own presentations), when the teacher rater knew the content in advance owing to prior instruction, or when students' fluency was excellent. For example, when the researcher rater listened to these students for the first time, she found it difficult to comprehend the presentation and rated lower. On the other hand, when the teacher rater was familiar with their presentation content and/or speaking style, including voice volume, he thought they were comprehensible enough to give higher ratings.

In terms of Fluency, first, raters perceived students' degree of fluency and dependence on scripts differently (seven instances). For example, one rater argued that although one presenter's eyes were sometimes on the script, she did not read it aloud, so her fluency rating was A. The script reading judgment was also affected by delivery, where very articulate, smooth speakers with moderate glances at scripts received higher scores. Second, raters were sometimes affected by factors not described in the rubric such as voice volume and students' attitudes. These can be interpreted as halo effects, defined as "the distorting influence of early impressions on subsequent judgements of a subject's attributes or performances" (Davies et al., 1999, p. 72). This sometimes happened unconsciously when raters were judging borderline cases (two instances). These points for potential rater divergence have implications for future rater training, where these issues should be included and discussed, while raters listen to past-year students' videos or recordings and examine the corresponding levels assigned.

RQ3: To what degree do rating criteria function as intended?

As shown in Table 10, Fluency (2.36) was found to be more difficult than Task achievement (-2.36). It was also concluded that the Task achievement and Fluency criteria functioned as intended to a moderate degree. Bond and Fox (2015) summarize five conditions for rating criteria to function effectively: (a) Average measures and Rasch-Andrich thresholds measures increase as levels increase. (b) Each level has more than 10 data. (c) The probability curve has a clear top. (d) The fit statistics should be less than 2.00. (e) Distances (i.e., differences between thresholds) should be between 1.40 and 5.00 logits. In the case of Task achievement (see Table 11 and Figure 4), (a), (c), and (e) were satisfied. For example, the distance between Scores 2 and 3 (Levels B and A) was 4.60 (i.e., 2.30 - (-2.30)). For (b), Score 1 (Level C) had only three observed counts. This test is criterion-referenced and aims to assess achievement, and the infrequent use of Score 1 does not seem to be problematic if students fulfil the minimum required level. However, it can also be argued that Score 2 (Level B) might have been too easy to assess whether the learning objective was accomplished. If this is the case, further exploration and revision may be needed. For (d), Score 2 had an Outfit mean square of 2.00. This may have been observed owing to extreme underfitting responses from three students, and the reasons should be explored further. In the case of Fluency (see Table 12 and Figure 5), all the conditions except for (b) were satisfied. Only nine students were assigned Score 3 (Level A). Three potential problems of Task achievement and Fluency were found, and this may require future revision.

Table 10

Criteria Measurement Report

Criteria	Obs Ave	Fair Ave	Logit Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discr	Corr. PtBis
Fluency	1.52	1.30	2.36	0.19	0.75	-1.80	0.69	-0.80	1.28	.32
TA	2.55	2.60	-2.36	0.21	1.15	1.20	1.59	2.60	0.73	.16
<i>M</i>	2.04	1.95	0.00	0.20	0.95	-0.30	1.14	0.90	0.24	
S.D. ^a	0.52	0.65	2.36	0.01	0.20	1.60	0.45	1.80	0.08	
S.D. ^b	0.73	0.92	3.33	0.02	0.28	2.30	0.64	2.50	0.11	

Model, Populn: RMSE .20 Adj (True) S.D. 2.35 Separation 11.70 Strata 15.94 Reliability .99

Model, Sample: RMSE .20 Adj (True) S.D. 3.33 Separation 16.58 Strata 22.44 Reliability 1.00

Model, Fixed (all same) chi-square: 275.9 d.f.: 1 significance (probability): .00

Note. TA = Task achievement. ^aPopulation. ^bSample.

Table 11

Category Statistics of Task Achievement

Score	Counts Used	%	Ave Meas	Exp Meas	Outfit MnSq	Rasch-Andrich Threshold	S.E.	Exp Measure Category	At -0.5	Most Probable from	Rasch-Thurstone Threshold
1	3	3%	1.13	-0.65	1.80			(-3.37)		low	low
2	47	39%	1.65	1.65	2.00	-2.30	0.65	0.00	-2.30	-2.30	-2.30
3	70	58%	3.51	3.59	0.90	2.30	0.23	(-3.38)	2.31	2.30	2.30

Table 12

Category Statistics of Fluency

Score	Counts Used	%	Ave Meas	Exp Meas	Outfit MnSq	Rasch-Andrich Threshold	S.E.	Exp Measure Category	At -0.5	Most Probable from	Rasch-Thurstone Threshold
1	75	63%	-2.91	-2.75	0.60			(-2.17)		low	low
2	36	30%	-0.56	-0.94	0.50	-1.01	0.23	0.00	-1.26	-1.01	-1.12
3	9	8%	-0.07	0.07	1.30	1.01	0.39	(-2.17)	1.27	1.01	1.11

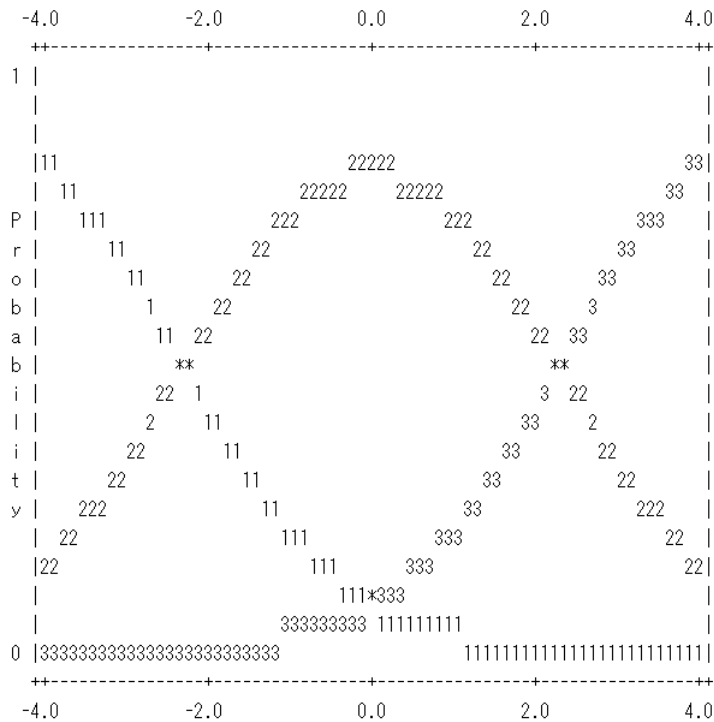


Figure 4. Probability Curves of Task achievement.

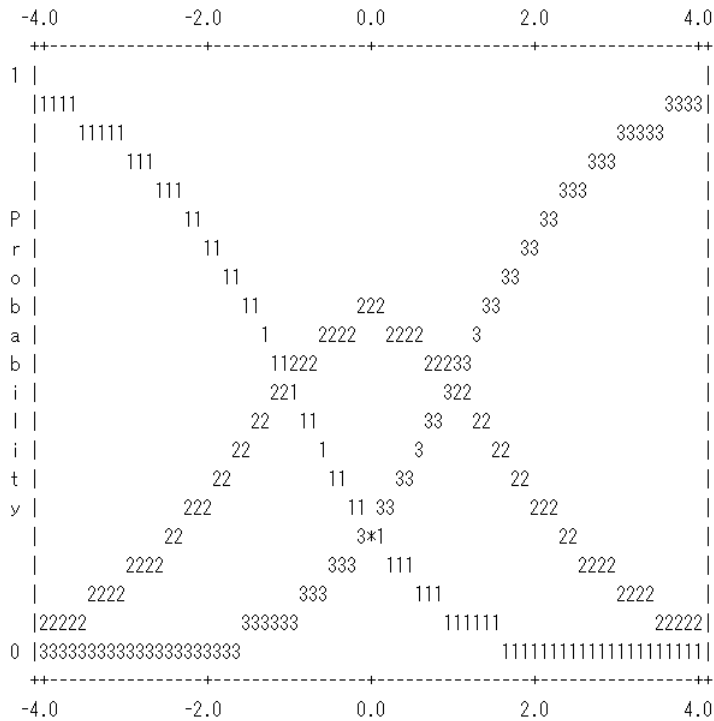


Figure 5. Probability Curves of Fluency.

RQ4: To what degree are there biased interactions between students, raters, and rating criteria?

We concluded that there were few biased interactions among the three facets. This research question examines if there are systematic and unexpected patterns between the three facets. We used *t* values of equal to or more than $|\pm 2.00|$ as a benchmark for detecting bias (Linacre, 2018). There were no biased patterns between students and raters and between raters and criteria. There were four biased interactions (3.33%, 4/120) between students and criteria. This suggests that some students were marked more strictly or more leniently in the rating criterion than expected. However, the four cases were from two of the three extreme underfitting students (see Table 7), and they were found not to indicate problems with the assessment itself.

RQ5: To what extent do students respond positively to potential effects of the speaking test on students' learning?

The questionnaire responses from the students about the speaking test show they generally responded positively to the potential washback effects. On the Likert scale, 4 means “Yes, a little” and 5 means “Yes, very much.” Thus, a combined percentage of students selecting 4 or 5 indicated a positive response (see the “4 + 5” column in Table 13). In general, in the following questions, more than half of the students responded positively: apparent test construct (Q2 = 65.63%; i.e., what the test seems to assess to lay people), presence of enough preparation time (Q4 = 73.44%), anxiety during the test (Q6 = 56.25%), possible effects of the test on students' English ability (Q9 = 60.94%) and on students' motivation to study English (Q10 = 51.56%). Thus, in terms of washback effects of speaking on the students' learning, the students responded positively in general. However, it should be noted that the question regarding the possible effects of the repetition of presentation activities on students' English ability (Q11 = 42.19%)

received limited positive responses, and further exploration is necessary to enhance positive washback on students' learning.

Table 13

Responses to the Speaking Test: Percentages of Students Who Selected Each Option (N = 64)

Question	1	2	3	4	5	No answer	4 + 5
1. Do you think you performed well on the test?	10.94	17.19	40.63	25.00	1.56	4.69	26.56
2. Do you think this test assesses English speaking ability, in general?	0.00	7.81	21.88	46.88	18.75	4.69	65.63
3. Did you prepare enough for this test?	3.13	20.31	26.56	42.19	3.13	4.69	45.31
4. Was there enough time to say what you prepared?	0.00	4.69	17.19	20.31	53.13	4.69	73.44
5. Were you able to convey your message in the question and answer session after the presentation?	20.31	15.63	42.19	10.94	3.13	7.81	14.06
6. Were you anxious or nervous during the test?	9.38	15.63	14.06	23.44	32.81	4.69	56.25
7. Were you disturbed by people or noises around you during the test?	26.56	25.00	17.19	14.06	12.50	4.69	26.56
8. Were you motivated to speak much during the test?	3.13	17.19	39.06	26.56	9.38	4.69	35.94
9. Do you think this test helps you develop English ability?	3.13	7.81	23.44	53.13	7.81	4.69	60.94
10. Did this test make you feel that you should study English more?	6.25	7.81	29.69	39.06	12.50	4.69	51.56
11. Do you think repeated activities like this test will enhance your English ability?	3.13	7.81	42.19	28.13	14.06	4.69	42.19

Note. Option 1 = "No, not at all"; 2 = "No, not much"; 3 = "Neither no or yes"; 4 = "Yes, a little"; and 5 = "Yes, very much." Q1 to Q11 concern students' perceptions of the test, and Q9 to Q11 especially concern how students felt the test affected their speaking ability and learning.

Open-ended comments (Q12 to Q15; see Table 4) were also analyzed to detect assessment issues (number of respondents: $n = 38$ to 59 for Q12 to Q15). One frequent issue mentioned by students was the presence of a video and voice recorder that made them nervous, which was consistent with the result of Q6 (56.25%, with 4 and 5 combined). Nevertheless, recording seems unavoidable to enable raters to check scores after the test. We may be able to consider where and how they should be placed or the possibility of using such recordings more frequently in regular lessons to help students get used to recording, since this type of student anxiety may be reduced after multiple practices and assessments. Another issue was how students perceived different test formats. We asked what students thought of the current test in comparison with tests they had previously taken (Q14). Here are six types of responses:

1. *I was tenser during this test than previous tests (n = 6).*

I have many experiences of one-one-one presentations with a teacher. This time, there were many classmates listening to me and I became tense ($n = 3$).

My teacher evaluated my speech before, but it was not full-fledged assessment.

2. *I was less tense during this test than previous tests (n = 1).*

I was able to enjoy this test because the atmosphere was less tense.

3. *This test was more difficult than previous tests (n = 2).*

Previous tests were easier, and the procedures were simpler.

4. *This test was easier than previous tests (n = 1).*

It was relatively easy for me to talk to a group, not to a whole class.

5. *This test requires more practice and active participation (n = 2).*

I had to talk much more than previous tests, so I practiced a lot.

Previous tests were one-on-one interactive types with a teacher, but in this test, students were required to ask questions more actively and elicit questions from listeners, to arrive at deeper thoughts.

6. *This test was interesting (n = 1).*

I was glad to be able to listen to others' talk this time, which was not available for the one-on-one test. It was a good learning experience.

These responses suggest that some students perceived the test as not only the presentation and scoring of the raters but also all the activities in the lessons, including talking to and answering questions from classmates, even when raters were not present in the group. Moreover, the results indicate that the test gave different impressions to students (e.g., more tense vs. less tense), but their perceptions were not negative. Each speaking test format has its own characteristics, and students' reactions to each format vary. Thus, it seems desirable to use various types of speaking tests to cater to diverse students' needs and elicit different types of speaking performance.

Discussion and Conclusions

The current study examined an oral presentation test for senior high school students. All the research questions were generally answered positively, except for RQ1 and RQ3. First, RQ1 asked to what degree student responses were consistent with the responses predicted by MFRM. Only half of the students had responses consistent with our predictions. Sixteen students' responses were unexpected but three extremely underfitting students' responses could be explained due to their irregular performances (i.e., obtaining lower scores for Task achievement than those for Fluency, which is expected to be the opposite due to the difficulty of the two criteria). Second, in RQ3 (To what degree do rating criteria function as intended?), the Task achievement and Fluency criteria had three levels, but the lowest one (Level C) in Task achievement and the highest one (Level A) were not used often, and the second level (Level B) in Task achievement showed underfit.

The research questions that were related to our validity argument (see the Current Study section above) and answered positively were RQ2, RQ4, and RQ5. The results of RQ2 (To what degree do raters score similarly and consistently?) suggest that rater severity was similar across raters, they rated presentations consistently, and their agreement was considered fairly high. These results can be used as evidence that

the students' performances were adequately evaluated and that the results can be generalized across raters. Thus, they can be employed for the evaluation and generalization inferences in the validity argument. In terms of RQ4 (To what degree are there biased interactions between students, raters, and rating criteria?), there were few biased interactions between the students and raters, raters and criteria, and students and rating criteria, which suggests that the ratings were properly conducted, which helps test users justify their claim that the student presentations are appropriately evaluated (in the evaluation inference in the validity argument). The results of RQ5 (To what extent do students respond positively to potential effects of the speaking test on students' learning?) indicate that most students responded favorably in terms of the potential effects of the speaking test on their learning, suggesting that we can expect positive results from using the speaking test on student learning. Therefore, these results help justify making a consequence inference in the validity argument.

Based on the assessment reported in the current study, we will plan to develop and implement a better speaking test in future research. For this purpose, we will summarize six areas for improvement. First, explanations of rubrics and example performances for each level of the rubrics should be presented before or during the test preparation stage. In the present study, the teacher did not explain the rubric until the test day itself. Thus, there was no time for students to prepare based on that. Receiving the explanation before and during the preparation stage may have led students to practice effective presentation delivery more and improve their performances in terms of Task achievement and Fluency.

Second, practice time (e.g., 5 minutes) should be given before the start of teachers' evaluations to alleviate an order effect. In the present study, the raters evaluated a group of five students first and another group next; (a) students in the first group were evaluated when they made their first presentation, whereas (b) other students were evaluated after giving their presentations several times. No students complained about this order and if they had done so, the raters would have evaluated their performance again. However, providing prior practice time would make students in the first group unlikely to feel they could have done better. Furthermore, when similar presentation assessment is conducted, we can use different orders in a way in which students assessed earlier will be evaluated later.

Third, the sample poster on the worksheet given to the students for presentation preparation should be revised (see Appendix C). The poster had a sample picture, sample key points, and sample comments, and these key points and comments were written mostly in sentences. This example could have encouraged the students to write the key words in sentences and simply read them aloud in the presentation. Further, some students wrote the script on the back of the poster. To avoid students reading aloud scripts and sentences, instructors should specify whether students are allowed to do this, before they start to prepare for the presentation.

Fourth, we should improve the question and answer sessions following student presentations. In some groups, it seemed that no one asked questions. To remedy this situation, teachers can allocate the role of asking a question to one of the listeners in the group for each presentation.

Fifth, we can provide more detailed feedback for students. In the present study, score reports with averaged scores were given. While it may not be possible to give individualized feedback, we can use some time in class to explain what Level A means and how it differs from Levels B and C using actual examples. Teachers can also iteratively touch on points for improvement to assist students with future speaking activities.

Sixth, we can make an annual or long-term plan on how speaking assessment will be implemented over a length of time. The plan should include when and how speaking tests are administered, based on teaching objectives and textbooks. While there may be cases where one format of speaking assessment is repeatedly conducted, it is better to use a variety of formats, such as presentations, interviews, and paired or group

oral tests. We can develop a textbook-based task and rubric in each test administration, as we did in the current study. We can also consider how a teacher or school can put into practice long-term, sustainable, repeated administration of low-stakes speaking tests (as suggested by one of the reviewers) and how such speaking assessments and feedback based on them can be used to implement assessment *for* learning and assessment *as* learning, and how they can be fully integrated into a curriculum and learning cycle (see Chong, 2018).

Besides ways to improve the current assessment method, there are points to consider in future research. First, to conduct rigorous analysis and obtain stable results, more data should be collected by increasing the number of students, raters, criteria, and the number of levels in each criterion. In terms of standard errors, Model standard errors were small for the two raters (0.20 for each in Table 8) and two rating criteria (0.19 to 0.21 in Table 10), and these results seem to be relatively stable. However, Model standard errors were relatively large for students (see Table 6; Mean = 1.18, $SD = 0.29$, Min = 0.98, Max = 1.97); person reliability was also not high, at .68. Although these results were not considered very problematic in the current relatively low-stakes assessment, it may be worthwhile to point out that relatively low reliability and large errors may be likely to be observed in one-shot classroom assessment, as mentioned by one of the reviewers. We can speculate that these results of relatively low reliability were likely derived because the numbers of raters, criteria, and tasks were limited (Engelhard, 2013). Therefore, we should be cautious when making strong claims about student presentation skills based only on the presentation ratings obtained in the present study. For future attempts, one way to increase the number of students and raters is to ask teacher and researcher colleagues or students to join an assessment project, or to ask the same raters to rate multiple times after long intervals (i.e., after raters forget what they have rated) or to have students complete self- and peer assessment.³ Having a wider range of speaking proficiency levels may also help, but the present study had students with maximum and minimum scores (see Table 6), so this would not be helpful. It is also possible to increase the number of criteria (e.g., adding Delivery, and subdividing Fluency into Speed, Pauses, and Repair, as seen in the previous fluency literature; see e.g., Tavakoli & Skehan, 2005), and further subdividing each criterion into more than three levels. However, this requires a careful approach because the use of elaborate criteria with detailed levels in one test administration can exhaust teachers and may prevent regular implementation of speaking assessment in practice. While taking possible human and time resources into account, a plan for collecting more data should be made.

A second point to consider is to include questionnaire questions related to the feedback given on the score report, which the current questionnaire did not include. Extra questions such as “Do you think that the feedback you received was easy to understand?” and “Do you think that the feedback you received helped you study?” would allow us to check on a consequence inference in more detail, which would be useful in constructing a future validity argument. Moreover, methods other than questionnaires to examine washback effects should be explored for refined investigations (see Watanabe, 2004).

Despite some limitations, in the present study, a presentation test and rubric were created based on the textbook, student performances were evaluated over two lessons, and assessment results were positive, overall. The procedures of test development, assessment, and examination of test scores would be useful for similar classroom assessment contexts. This study involved an external researcher as a rater, but an assistant language teacher or Japanese teacher of English, or even the same teacher scoring after a certain time interval, could serve as a second rater. Moreover, MFRM was used to examine test details in the current study, but in practice, raw scores could be used for grading and giving feedback. Positive overall results to the speaking test and test scores in the current study, in combination with more studies using different types of speaking test formats and rubrics, would help English teachers feel that it is feasible to conduct tests and assess speaking effectively.

Notes

¹ This Fluency criterion is adequate, and the current study showed that it worked effectively, as intended. However, some may worry that this criterion does not differentiate between those who always looked at scripts but gave a fluent presentation and those who always looked at scripts but had poor fluency, and achieving the highest level A became difficult. Having separate criteria of fluency and script reading may be one idea to address these concerns.

² We removed the three underfitting students' responses and reanalyzed the data ($n = 61$), finding that our measurement worsened, with more students showing underfit and overfit: 16.39% (10/61) with infit mean squares of more than 1.50, 11.48% (7/61) with more than 2.00, and 59.02% (36/61) with lower than 0.50. Thus, we decided not to remove the three underfitting students.

³ One of the reviewers suggested that "time" could be included as a facet of MFRM, to examine how student presenters, student raters, as well as teacher raters behave, as presentations continue and as raters get used to using the rubric. This would be a complex but worthwhile research topic.

Acknowledgment

This work was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (C), Grant Number 26370737. We are deeply indebted to two anonymous reviewers for insightful comments, and Yumi Koyamada and Yo In'nami for their strong support for this project.

References

- Akiyama, T. (2003). Assessing speaking: Issues in school-based assessment and the introduction of speaking tests into the Japanese senior high school entrance examination. *JALT Journal*, 25, 117–141. Retrieved from <http://jalt-publications.org/jj/articles/2627-assessing-speaking-issues-school-based-assessment-and-introduction-speaking-tests-j>
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.
- Chong, S. W. (2018). Three paradigms of classroom assessment: Implications for written feedback research. *Language Assessment Quarterly*, 15, 330–347. doi:10.1080/15434303.2017.1405423
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press.
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd revised and updated ed.). Frankfurt am Main, Germany: Peter Lang.
- Engelhard, Jr. G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.
- Engelhard, Jr. G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales*. New York, NY: Routledge.
- Fulcher, G. (2003). *Testing second language speaking*. Essex, U.K.: Pearson Education Limited.

- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—A longitudinal study. *Language Testing*, 28, 179–200. doi:10.1177/0265532210384252
- Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, 35, 477–499. doi:10.1177/0265532217710049
- Koizumi, R. (2018). *Eigo yongino tesuto no erabikata to tsukaikata: Datousei no kanten kara* [How to select and use English four-skill tests: From a perspective of validity]. Tokyo: ALC.
- Koizumi, R., In'nami, Y., & Fukazawa, M. (Eds.). (2017). *Jitsurei de wakaru eigo tesuto sakusei gaido* [A practical guide for developing English tests]. Tokyo: Taishukan.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. doi:10.2307/2529310
- Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions*, 7(4), 328. Retrieved from <https://www.rasch.org/rmt/rmt74m.htm>
- Linacre, J. M. (2018). *A user's guide to FACETS: Rasch-model computer programs (Program manual 3.81.0)*. Retrieved from <https://www.winsteps.com/manuals.htm>
- Linacre, J. M. (2019). FACETS: Many-facet Rasch-measurement (Version 3.81.2) [Computer software]. Chicago: MESA Press.
- Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.
- McNamara, T., Knoch, T., & Fan, J. (2019). *Fairness, justice, and language assessment*. Oxford, UK: Oxford University Press.
- Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2014). *Heisei 25 nendo eigo kyoiku jisshi jokyo chosa no kekka nitsuite* [Results of the survey of the state of implementation of English language education in the academic year 2013]. Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1351631.htm
- MEXT. (2016). *Heisei 27 nendo eigo kyoiku jisshi jokyo chosa no kekka nitsuite* [Results of the survey of the state of implementation of English language education in the academic year 2015]. Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1369258.htm
- MEXT. (2018a). *Heisei 29 nendo eigo kyoiku jisshi jokyo chosa no kekka nitsuite* [Results of the survey of the state of implementation of English language education in the academic year 2017]. Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1403468.htm
- MEXT. (2018b). *Heisei 30 nendo eigo kyoiku kaizen puran* [Plans for improving English language education in the academic year 2018]. Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1407569.htm
- Muranoi, H., Mano, Y., Hayashi, B., Hatano, S., Yamaga, Y., Uekusa, T., & Taishukan (2016). *Genius English Communication II*. Tokyo: Taishukan.
- Muranoi, H., Tsutsui, M., Narita, K., Koizumi, R., Matsuo, M., Yano, K., . . . Mito, N. (2017). *Genius English Communication I Revised kyojuyo shido shiryō* [teacher's manual]. Tokyo: Taishukan.
- Muranoi, H., Tsutsui, M., Narita, K., Koizumi, R., Yano, K., Yamaga, Y., . . . Mito, N. (2018). *Genius English Communication II Revised kyojuyo shido shiryō* [teacher's manual]. Tokyo: Taishukan.
- Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Frankfurt am Main, Germany: Peter Lang.

- National Institute for Educational Policy Research. (2012). *Hyoka kijun no sakusei, hyoka hoho to no kuhu kaizen no tameno sanko shiryō (koko gaikokugo)* [Reference documents for Japanese senior high school foreign language studies for the development of assessment criteria and improvement of assessment methods and others]. https://www.nier.go.jp/kaihatsu/hyouka/kou/11_kou_gaikokugo.pdf
- 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. doi:10.1177/0265532214538014
- Sick, J. (2009). Rasch measurement in language education Part 3: The family of Rasch models. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 13, 4–10. Retrieved from <http://jalt.org/test/PDF/Sick3.pdf>
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 238–273). Amsterdam, the Netherlands: John Benjamins.
- Tokyo Gakugei University. (2017). *Monbu kagakusho itaku jigyo eigo kyoin no eigoryoku shidoryoku kyoka notameno chosa kenkyu jigyo, Heisei 28 nendo hokokusho* [Report for the MEXT's Contract Research Project in the 2016 academic year: Survey study for strengthening English proficiency and instruction of English teachers]. Retrieved from <http://www.u-gakugei.ac.jp/~estudy/report/>
- Taylor, L. (Ed.). (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge University Press.
- Talandis, J. Jr. (2017). *How to test speaking skills in Japan: A quick-start guide*. Kyoto: Alma Publishing.
- Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng & Y. Watanabe, with A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 19–36). Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix A

Rubric for the Presentation Test (in Japanese)

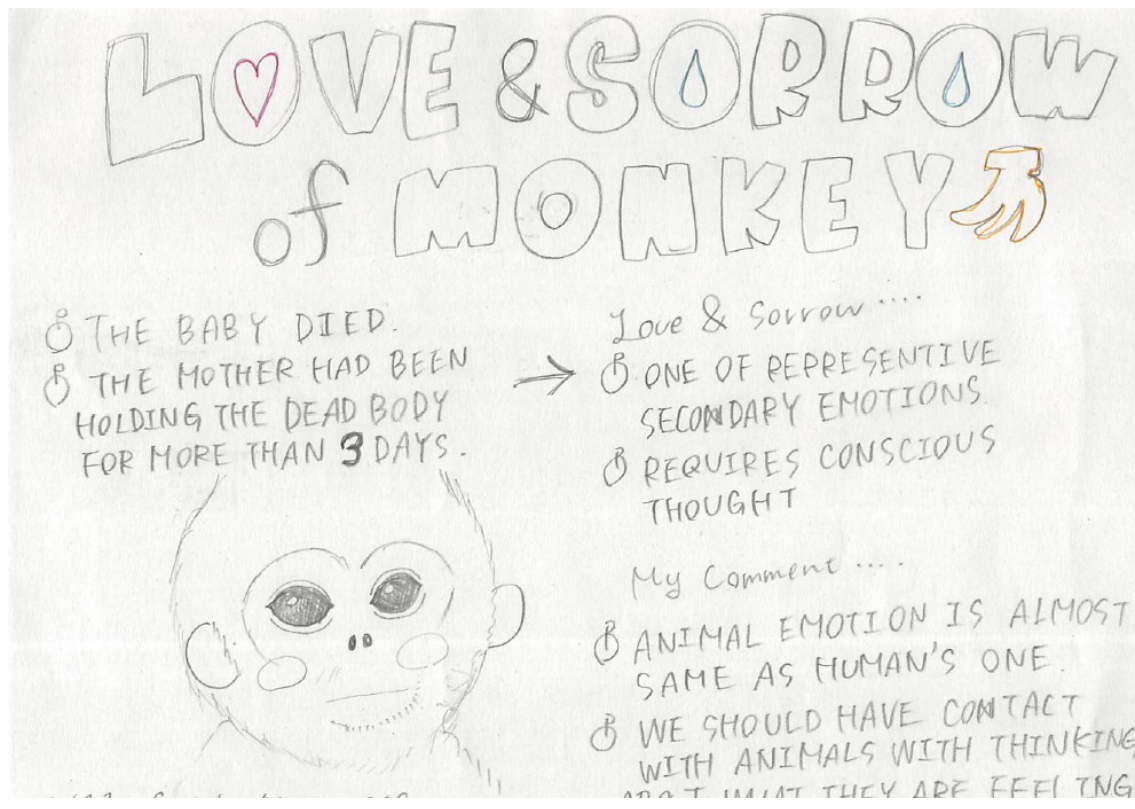
タスク達成度	流暢さ
A (十分満足できる) (ア) ①ある動物が、どのような状況で、どんな感情を持ったのか、②その感情についての科学的な説明、③それに関する自分の意見について述べている。かつ(イ)内容が十分伝わり、詳細である。	5秒以上の長い沈黙がない。言い直しがあっても気にならない程度である。スムーズに話している。かつ原稿をほとんど見ていない。
B (おおむね満足できる) (ア)のみを満たしている。 ^a	長い沈黙が1回ある。言い直しが多めで少し気になる程度である。話すスピードが遅めである。または原稿をたまに読み上げている。ほとんどがBの記述に当てはまる。
C (努力を要する) (ア)を満たしていない。	長い沈黙が2回以上ある。言い直しが多し。話すスピードが遅く、理解に影響がある。原稿をほとんど見ている。ほとんどがCの記述に当てはまる。「原稿をほとんど読み上げている」に当てはまる場合は常にC

Note. ^a本研究時には入れていなかったが、合意が取れていた点について、以下のように今後入れる予定：(ア)をほとんど満たし、かつ(イ)内容が十分伝わり、詳細である。

Appendix B

Sample Presentation and Poster

Sample 1



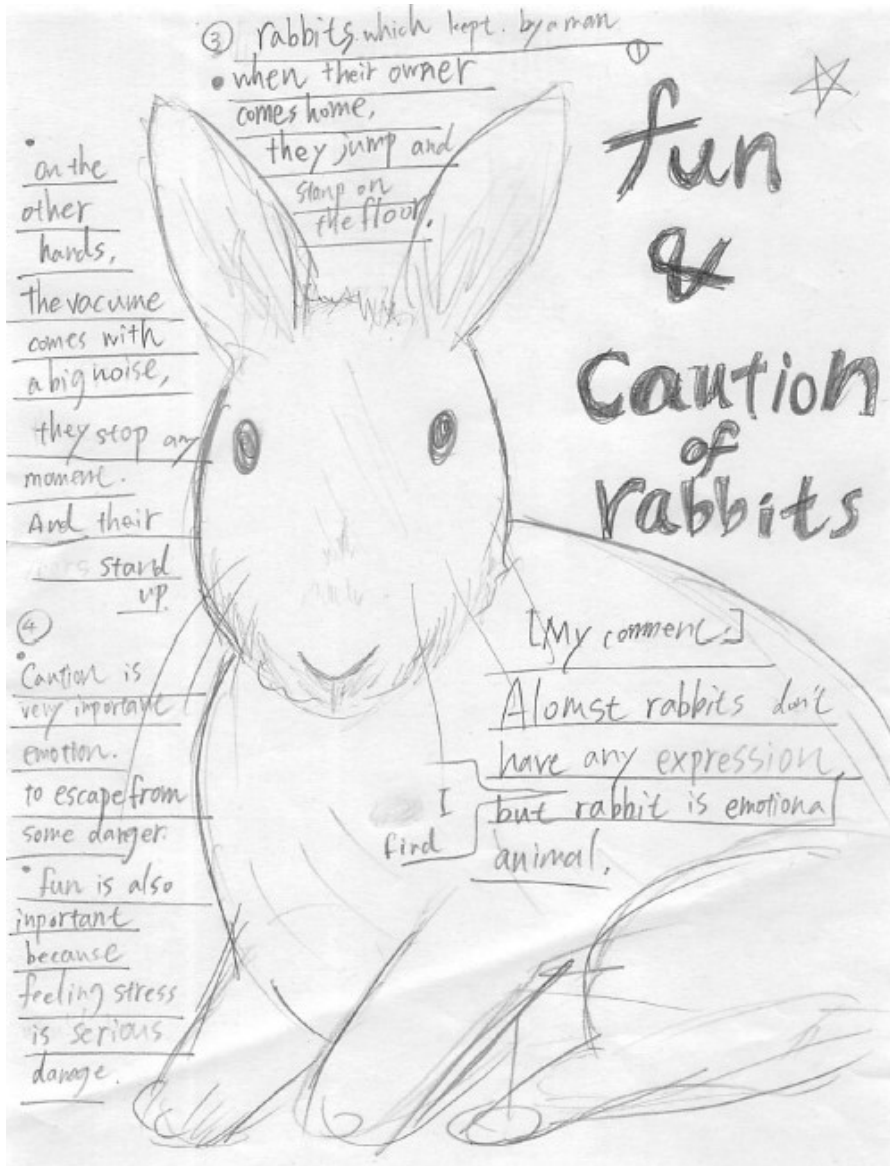
I'll talk about animal emotions. Ah I watched a TV program about a mother and a baby monkey in China. And the mother monkey showed me sorrow and love. The situation was like this: One day the baby monkey died because ah that that is very difficult to grow up in severe nature. Then, can you guess what the mother monkey did? In fact, she had been holding the dead body of her baby for more than 3 days. Ah This action is a sign of love for baby and sorrow for the death, I guess. Sorrow and love is one of representative secondary emotions requires conscious thought. When I knew this fact, I thought animal emotions is almost same as human's one. I thought we should contact with animals with thinking about what they are feeling. Thank you.

Note. (135 words).

Task achievement = A: This includes all (a) (1) to (3) and (b). There were some errors (e.g., *Sorrow and love is one of representative secondary emotions requires conscious thought --> Sorrow and love are two of the representative secondary emotions that require conscious thought*). There were sometimes unclear words that were not pronounced clearly, but they did not impede comprehension.

Fluency = A: Although there were repetitions (e.g., *the mo*) and dysfluency markers (e.g., *ah*), they did not prevent comprehension. The presentation was generally conveyed smoothly, and the presenter did not look at the script most of the time.

Sample 2



I will introduce about ... rabbits, ... fun and caution of rabbits. Rabbits ... have secondary emotion. Ah When its owner comes home, it jumps vertically and ... jump and jump. Eh the When the vacuum comes with a big noise, its ears mm ... its ears st stand up. ... Fun is an important secondary emotion. ... When ... when his workers owners comes home and play with them ... play with he or ... he or she the rabbit looks very fun. And the caution is an auto ... automatic one. It's primary emotions. Eh ... They need ... caution ... when they ... face ... they faces when they face ... ah some danger. ... I think ...

Note. (80 words). ... (pause of less than 5 seconds).

Task achievement = B: The presentation included (a) (1) and (2), but (3) the opinion was not clear. The details were insufficient. The content was not always comprehensible because of many choppy expressions.

Fluency = C: The student almost always looked at the script. Other features were acceptable, with short pauses, repetitions that were not extensive, and smooth speed.

Sample 3

The image shows a handwritten presentation script on a piece of paper. At the top, the title 'Crows can enjoy the thrill of being alive!' is written in large, bold letters. Below the title, the text 'In Scotland' is followed by a small drawing of the Scottish flag. A box contains the text 'Crows slid down snowy hillsides and then returning to do it again', with an arrow pointing to another box that says 'thrill' and 'They don't demand evolutionary benefit, but they just enjoy themselves for the thrill of being alive.' Below this, there is a drawing of two crows and a box labeled 'My Comment' containing the text 'I was amazed that animals have emotions humanly. I want to share the emotions with animals.' The entire script is written in a casual, handwritten style.

I will talk about crows' emotion. ... eh Crows can feel the thrill of being alive. Eh In Scotland, crows sild down snowy eh ... hillsides and the and then returning to do it again. They don't ... eh demand evolutionary benefit, but they just enjoy themselves of the thrill of eh being alive. ... eh I was amazed that animals have emotions humanly. Eh I want to share the emotions with animals. That's all.

Note. (62 words).

Task achievement = B: This student did not include details or a scientific explanation of animal emotion. His pronunciation of *crows* sounded like *clothes*, pronunciation of some words (e.g., *slid down snowy*) was not clear, and he often inserted vowels after consonants. These features made his presentation difficult to hear, but the presentation was acceptable, overall.

Fluency = B: The student initially tried to speak without the script, but later, he almost always looked at the script. Other features were acceptable, with few short pauses, repetitions that were not extensive, and smooth speed.

Sample 4

Mice can show sympathy!

In Chicago University ...

- ① Mouse practice releasing a trap
- ② Another mouse was trapped
- ③ The first mouse helped



[Sympathy]

- One of the important secondary emotions, not automatic
- Sharing the fellow feeling

* the mouse helped even if he don't have all to himself monopolize chocolate.

chd



[My Comment]

- Animals treasure friends like human

I am going to about mice's sympathy. In instrument is American Chicago university team. eh First mouse practice escaping from trap. Second, mouse be was trapped and mouse a found. Then mouse helped mouse. ... Another experiment mouse also had evening see to get a break. This is so. Sympathy is one of the important secondary emotions, not automatic. I think mouse must feel feeling like a human. Thank you for listening.

Note. (71 words).

Task achievement = C: It is difficult to understand what the speaker said, especially concerning the content of the experiment. *Mouse* sounded like *must*, and other words were not very clearly pronounced. According to the teacher, she forgot to bring her script and spoke on the spot.

Fluency = A: The delivery was fluent, with few pauses and repetitions. She had natural eye contact and did not read the script.

Appendix C

Worksheet for Preparing the Presentation (1. and 2.), Taking Notes (3.), and Evaluating Classmates' Presentations (4.), With Sample Student Responses

1. Web research

By using websites and other resources, find scientific studies on animal emotions and summarize their findings.

Make a speech for one minute. Include the following 5 points:

Points	Example	Your Case
(1) What animal is it?	Grizzly bear cubs in Alaska	Household dogs
(2) What emotion did it show?	Empathy and compassion	jealousy
(3) In what situation? Explain concretely.	<ul style="list-style-type: none"> ➤ The mother bear was shot and killed ➤ Two bear cubs are orphaned ➤ One of them are wounded and limped ➤ The other female cub helped the wounded bear 	- Their owner displayed affectionate behavior to what appeared to be another dog (as compared to nonsocial objects.) • dogs showed behaviors such as
(4) Explain the emotion scientifically (Primary? or Secondary?)	<ul style="list-style-type: none"> ➤ Compassion is an important secondary emotion. ➤ Requires conscious thought ➤ not automatic ➤ require a good deal of conscious thought 	snapping, getting between the owner and object, pushing/ touching the object, owner. • Jealousy is an secondary emotion, works better on social objects than nonsocial ones,
(5) Your comment	I learned that animals can express more complex emotions. I want to treat them more like humans.	• hate to prefer something to itself. people try to hide the emotion, because they are ashamed of it. I was surprised that dogs have emotions as if they were humans.
Memo	I think we may share emotion between We need to think about animals' action more deeply.	

2. Prepare for your presentation

Using the information in the box above, make a presentation poster and the script.

(1) Write your speech script in the box below.

I researched rats show regret. Regret is a cognitive behavior once thought to be uniquely human. It's the recognition that you made a mistakes, that if you had done something else, you would have been better of. The researchers do "Restaurant Row". It's rat is presented with a series of food options but has limited time at each restaurant. It shows rats' indicators in the orbitofrontal cortex, represent what the rat should have done, not the missed reward. I learned the animals' model of regret may teach us how regret affects the decisions we make. Maybe, other animals such us dogs, monkeys, sometimes regret.

(2) Make a small poster with a small piece of paper given from your teacher. You can draw a picture or put a printed photo on it.

(1) (2) 動物の名前を含めたタイトルを考える

(Sample) **Bears can show compassions!**

In Alaska...

(3) 事例の具体的な説明


- Bear cubs became orphans
- One cub was shot and wounded
- The other cub remained and helped her

(4) Emotions についての科学的説明

【Compassion】

- One of the important secondary emotions
- Not automatic
- Requires a good deal of conscious thought

(5) 自分の感想など



【My Comment】

I learned that animals can express more complex emotions. I want to treat them more like humans.

2141 Ken YANO

3. Let's make a group presentation!

Make a group of 5 and talk about the results of your web research. While you are listening to speeches of other students, take notes.

Session	No.	Speaker	Animals	Emotions	Other useful information
1	1				
	2				
	3				
	4				
2	5				
	6				
	7				
	8				

4. Choose the best presentation of the day

Choose the three best presentations. Write reasons you chose them.

Speaker's name	Topic of the presentation	What are the reasons you chose?
Mr. [redacted]	Snake's consideration	He spoke frequently.
Ms. [redacted]	Cat's compassion	Her voice was easy to listen and her speech was easy to understand.
Ms. [redacted]	Elephant's compassion	Her speech's script was excellent.

Appendix D

Score Report of Students (above in English; below in Japanese)

Speaking test results (Date:): 2nd-year () class, No. (), Name ()

A (Satisfies to a large degree)	The presentation (a) describes (1) a situation where an animal has a certain emotion, (2) a scientific explanation, and (3) an opinion; and (b) is fully comprehensible and detailed.	There is no long pause (5 seconds or more). Repetitions and corrections do not hamper comprehension. The presentation is conveyed smoothly. The student does not look at the script in most of the time.
B (Mainly satisfies)	The presentation satisfies only (a).	There is one long pause. Relatively many repetitions and corrections sometimes hamper comprehension. The presentation is conveyed relatively slowly. The student sometimes reads aloud the script. The presentation has characteristics in the descriptions at Level B.
C (Requires more efforts)	The presentation does not satisfy (a).	There are two or more long pauses. Comprehension is difficult due to many repetitions, corrections, and/or slow speed. (x) The student reads aloud the script in most of the time. The presentation has characteristics in the descriptions at Level C. If (x) is observed, the rating is always C.

☆Use this test result and become prepared to learn and use English in the future.

☆Based on your test and questionnaire responses, we will improve the test and assessment methods.

Note. Their scores were marked with a circle for each rating criterion. When scores had numbers with a decimal point, a circle was placed between the two levels (e.g., 2.5 = between Levels 2 and 3).