Conversations as a Holistic Assessment Tool: An Action Research Report

by
Branden KIRCHMEYER*

Abstract

As Japanese educational policy places greater emphasis on spoken communication skills (MEXT, 26 September 2014), a greater need arises for appropriate assessments that can provide meaningful information to all stakeholders. This paper presents data collected from student performances on an experimental spoken English assessment task conducted at a Japanese university. Designed as the foundation for a reconstructed English communication curriculum for first-year university students enrolled in compulsory English classes (Tempest, 2018), the assessment task elicited free-form conversation, which was then transcribed and analyzed for specific metrics of spoken L2 output. Data deriving from student-generated transcriptions of recorded conversations included total words spoken, total turns taken, average turn length, and longest turn length. Preliminary analysis of the data indicates positive student gains over the course of the semester. As an investigative report of the assessment task trial period, few claims can be made at this point. However, future studies are forthcoming with support from JSPS KAKENHI grant number 19K13309.

Key Words: L2 speaking assessment, formative assessment, CALL

1. Introduction

Despite a dedicated effort to improve the overall English proficiency of Japanese students, government initiatives have continued to fall short of meeting stated aims (Tahira, 2012). This is especially true as it applies to the implementation of communicative language teaching (CLT) which prioritizes communicative tasks aimed at developing learners' pragmatic competence. Researchers investigating these shortcomings have cited ambiguity (Sakui, 2004) and lack of confidence (Nishino & Watanabe, 2011) as existing challenges. Importantly, much of this research has been conducted at the primary and secondary levels, where teachers are often pressured to prepare students for traditional high-stakes examinations. In comparison, tertiary level English education contexts can provide teachers with a greater deal of freedom, resulting in classroom practices that employ CLT approaches more readily. This is especially true for non-major university students enrolled in compulsory English education, where the pressure to perform on standardized tests can be replaced with an invitation to communicate face-to-face in relaxed environments (Rowberry, 2010). Moreover, university language teachers in Japan are often encouraged to explore progressive methodologies and conduct research through government funding that can drive the field into new territories (JSPS,

＊Senior Assistant Professor, Sojo University

2019). These conditions provide an ideal testing ground for language teaching and assessment strategies that may apply to a greater range of contexts and help bridge the gap between policy and practice throughout English education in Japan.

This paper presents preliminary findings of one such attempt. An action research approach was adopted to assess the viability of a sequential speaking-transcribing task both as a benchmark assessment tool and as a means of collecting data for research into conversations-as-learning tasks. The aim of this report is to present preliminary evidence of this alternative assessment task as a practical tool for English language education and research in Japan, and to explore the possible benefits of its implementation in various EFL settings.

2. Assessing Spoken English in Japan

In the mid-1980s, language assessment experts worldwide launched a "quest for authenticity" in communicative language assessments that stressed pragmatic competence, i.e. the ability to use language in authentic tasks (H. D. Brown & Abeywickrama, 2010). Consequently, many trusted language examinations now include speaking tasks that are scored using holistic scales such as the Finnish National Certificate Scale, the American Council for the Teaching of Foreign Languages Speaking Scale, the Test of Spoken English scale, and the Common European Framework speaking scales (Luoma, 2004). These rubric-based scales involve varying levels of subjective scoring and provide overall impressions of a speaker's ability. One downside of these assessment tools is that they are often impractical for reliable implementation on a local scale, such as a high school English class with one teacher responsible for administrating and scoring hundreds of examinees. Thus, while many educational testing and assessment organizations have integrated speaking tasks into their examinations, individual schools and teachers may not have the resources or incentive to do so.

This is certainly the case in Japan, where traditional standardized examinations reign and educational institutions have been slow to adopt CLT approaches. In fact, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) only used the term "communication" for the first time in their updated 1989 Course of Study directive (Yoshida, 2003). Implementation of alternative assessment strategies that reflect a prioritization of communicative skills has been equally slow — the first iteration of assessments with speaking tasks for junior high school students is set to commence in 2020 for students entering Tokyo Metropolitan high schools (Ichikawa, 2019).

In contrast, CLT approaches and assessments are well-established in language departments at many Japanese universities. One such example is the Kanda English Proficiency Test, an in-house assessment at Kanda University which has been used since 1989 (Lockley & Farell, 2011). In this exam, students speak for eight minutes after viewing a short topic prompt and are subsequently rated using a holistic scale (p. 189-190). Research conducted on this particular assessment tool by Van Moere (2006) concluded that this type of assessment could be used to reliably place students, evaluate student progress, and make informed decisions about curricula.

In a guidebook on testing speaking skills in Japan, Talandis (2017) identifies negative washback as a key contributor to the relatively poor level of Japanese students' English speaking skills. He stresses, however, that carefully constructed assessments reflecting ideal teaching and learning conditions (i.e. CLT approaches) can create positive washback and induce a "virtuous cycle of learning" (p. 15). It is this concept that has informed the current action research project and the assessment task explained in the following section.

2.1. Conversations as Holistic Assessment Tool (CHAT)

Put simply, this task asks students to make audio recordings of a conversation they conduct in English

before transcribing their individual contributions to the conversation. Variable conditions of the task can include duration and topic of conversation, as well as the number of students engaging in a single conversation. Based on the task's role as a complementary tool set within a holistic approach to evaluating students, this task was named Conversation as a Holistic Assessment Tool (CHAT). The curriculum was structured around the CHAT, which was configured in two ways as summarized in Table 1.

Table 1. *Speaking task design*

Condition	**P-CHAT**	**CHAT**
assessment type	formative	summative
duration	5 min.	10 min.
topic	restricted	unrestricted
group size	2~3	2~3
transcriptions	paper	electronic
scoring	in/complete	0~100
occurrences	7	3

As a norm-referenced summative assessment task used for course evaluation, three CHATs served as benchmarks of student progress. Shorter practice versions called P-CHATs were utilized as formative assessment tasks and provided students with valuable practice and timely opportunities to reflect on their performance with an eye toward improvement.

In accordance with the main objectives of the CHATs and P-CHATs, which were to provide students and teachers with tangible metrics of their speech, several quantifiable aspects were targeted: total words spoken, total turns taken, average turn lengths, and longest turn length. Curriculum designers reasoned that simple numbers correlating to specific features of a student's contribution would provide an accessible means for goal setting and focused practice. For P-CHATs, these numbers were calculated and graphed on a tracking chart by students (see Figure 1), while the same data was collected, calculated, and reported to students by teachers on the CHATs (see Appendix A).

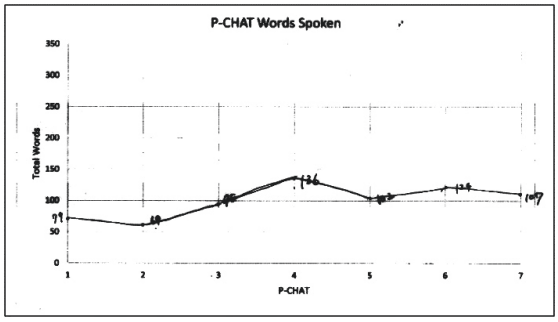

Figure 1. P-CHAT tracking chart for students

3. Research Context

This action research project took place at a small science and technology university in Southern Kyushu, Japan. With the exception of foreign students (who take Japanese language courses), all first- and second-year students enrolled in the university take compulsory courses focused on developing communicative skills in English. Due to scheduling, first-year students are assigned to classes based on their academic departments and are further divided into tiers based on their performance on an in-house English proficiency exam taken during the first week of the semester. The placement test is not yet mapped to any established English proficiency scale, though English proficiency among the participants ranged from A1 to B2 according to CEFR guidelines.

As reported by Tempest (2018), a new curriculum focused on speaking fluency was developed featuring a cyclical sequence of tasks designed to improve student performance on a conversation-based speaking task. In summary, the semester-long course was organized into seven thematic units of study consisting of three lesson types: preparation, practice, and performance. In the preparation lessons, target language including vocabulary and grammatical frames were introduced, and students were encouraged to develop unique texts within the unit's theme. In the practice lesson, students expanded on their texts and practiced engaging with classmates through structured and authentic

dialogues. In the final performance lesson, students reviewed the unit content and self-generated texts before engaging in the shorter, formative P-CHAT.

3. Methods

During the period of data collection, each class met for ninety minutes twice weekly for a total of thirty lessons. Identical CHATs served as pretest, midterm, and final exam for the course. For these three assessments students recorded their conversations directly to a learning management system (LMS) on a school computer and their personal devices simultaneously. In cases where the LMS or the school computer failed to capture a usable audio recording, the students' personal devices were used for playback. After uploading the audio file, students created transcriptions of their own contributions to the conversation by typing directly into a text input field on the LMS. Students were provided with, and regularly reminded of, several key policies and procedures regarding data entry in order to create as accurate a dataset as possible. These transcriptions were then checked by classroom teachers, who assembled the data into preformatted Microsoft Excel spreadsheets before submitting class data to the researcher. Data were compiled by the researcher into a custom-built Microsoft Access database for analysis and reporting.

An entire first-year cohort of students was included in the sampling pool, as each student was enrolled in the same course and subject to the same CHATs. Only students who completed all three CHATs (n=661) were included in the data set.

4. Findings and Discussion

As listed in Table 2, data collected from the three CHATs showed, on average, a gain in total words spoken, turns taken, and average turn length over the semester.

Table 2. *Oral production metric means*

	Pretest	Midterm	Final
words spoken	77	138	148
turns taken	17	26	28
turn length	4.77	5.66	5.70

Individual students' data were reported back to students as early as possible, which was in most cases a full week after students had submitted their transcriptions. The purpose for reporting these numbers was to help students set tangible goals for improvement on subsequent tasks. Teacher feedback and reflective prompts often accompanied the results, encouraging students to review their transcriptions and highlight areas that they could focus on, such as diversifying their vocabulary or increasing the length of their turns by uttering fuller sentences and adding detail to support their basic declarations.

Figures 2 and 3, which show the distributions of words spoken and average turn lengths, illustrate positive trends over the three assessment tasks. The symmetry of these bell curves indicates mostly normal distribution, meaning that similar numbers of students produced scores both above and below the averages. As can be seen in both figures, the right tails of each curve stretch out much farther than those on the left. This is due to outliers who were able to utter more words and longer turns than average, and the inability of any student to say fewer than zero words during a conversation.

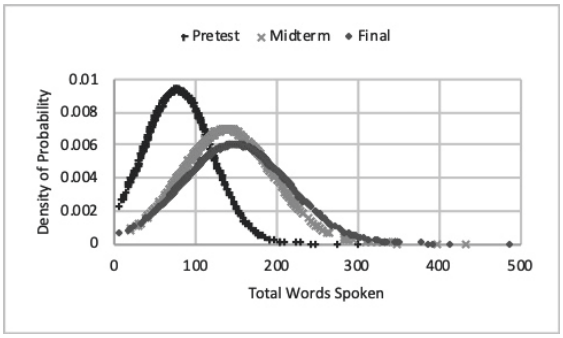

Figure 2. Total words spoken

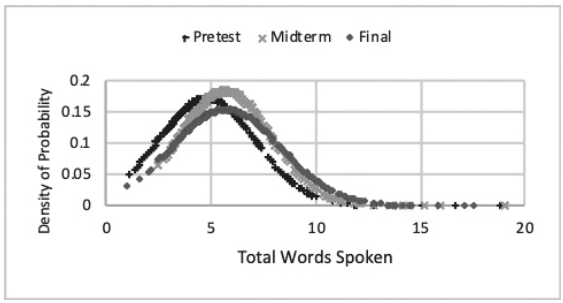

Figure 3. Average turn lengths

For teachers, these graphs helped visualize an important point—the symmetry of the curves implies that the assessment task might be useful for assigning standardized scores for course evaluation. Taken in conjunction with holistic rubric-based evaluations, these numbers might provide a fuller description of student progress and ability.

Though the procedures of the task produced quantifiable metrics that were useful in directing student effort, there are many issues that need to be addressed if the task is to be routinely implemented, of which reliability and validity top the list. Alternative assessments have raised concerns regarding reliability and validity since their early days (J. D. Brown & Hudson, 1998), and this task inherits a variety of complicating variables including conversation participants, topic, and repeatability. A second issue concerns the data collection methods, which are far from infallible. Despite attempts to protect data integrity, there is a non-zero possibility that a student may incorrectly transcribe their recording, and that the error is not picked up by the classroom teacher. It was beyond the scope of this report to audit the data to determine the existence and extent of this concern, though it will be addressed in subsequent research. Certainly, much work remains both procedurally and conceptually if CHATs and P-CHATs are to be routinely implemented.

5. Conclusion

The aim of this action research project was to examine the feasibility of using recorded conversations and student transcriptions as one part of a holistic approach to language assessment. The potential for such an assessment task to produce authentic speaking practice, quantifiable data for learning, and positive washback is encouraging, though its current limitations should not be taken lightly. Further research into this task has been approved and funded by a government research grant, and will integrate machine translation and automated data analysis through the development of a new interactive online assessment module. At the time of publication of this report, the module will have been developed and is being trialed.

Acknowledgements

I would like to thank my industrious colleagues who helped implement CHATs and P-CHATs in their classes, laboriously assembled data, and contributed their thoughtful ideas throughout the process; Rachel Barington, Yoko Kinoshita, Elton LaClare, Tina Shu-wen Lin, Benjamin Snyder, Christopher Tempest, Simon Wally, and Larry Xethakis. Also, I would like to thank the students who contributed to the database for their commendable efforts and patience during these tasks.

References

Brown, H. D., & Abeywickrama, P. (2010). *Language Assessment, Principles and Classroom Practices* (2nd ed.). White Plains, NY: Pearson Education.

Brown, J. D., & Hudson. (1998). The alternatives in language assessment. *TESOL Quarterly*, *34*(4), 653-675. Retrieved from http://www.personal.psu.edu/kej1/APLNG_493/old_site/brown_hudson.pdf

Ichikawa, A. (2019, 14 February 2019). Tokyo to introduce English-speaking test in high school entrance exams. *毎日新聞*. Retrieved from https://mainichi.jp/english/articles/20190214/p2a/00m/0na/029000c

JSPS. (2019). Handbook on the Grants-in-Aid for Scientific Research (KAKENHI) Program. In.

Lockley, T., & Farell, S. (2011). Is grammar anxiety hindering English speaking in Japanese students? *JALT Journal*, *33*(1), 175-189.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

MEXT. (26 September 2014). Report on the Future Improvement and Enhancement of English Education (Outline): Five Recommendations on the English Education Reform Plan Responding to the Rapid Globalization. Retrieved from http://www.mext.go.jp/en/news/topics/detail/1372625.htm

Nishino, T., & Watanabe, M. (2011). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly*, *42*(1), 133-138.

Rowberry, J. (2010). A new member of the family: The Sojo International Learning Center. *Studies in Self-Access Learning Journal*, *1*(1), 59-64.

Sakui, K. (2004). Wearing two pairs of shoes: language teaching in Japan. *ELT Journal*, *58*(2), 155-163. Retrieved from http://sakaienglishteachers.pbworks.com/f/Wearing%2Btwo%2Bpairs%2Bof%2Bshoes%2Blg%2Bteaching%2Bin%2BJapan.pdf

Tahira, M. (2012). Behind MEXT's new Course of Study Guidelines. *The Language Teacher*, *36*(3), 3-8.

Talandis, J. (2017). *How to Test Speaking Skills in Japan: A Quick-Start Guide*. Kyoto: Alma Publishing.

Tempest, C. (2018). *Using recordings and speaking fluency tasks to enhance spoken interactions*. Paper presented at the KOTESOL International 2018, Seoul, Korea.

Van Moere, A. (2006). Validity evidence in a group oral test. *Language Testing*, *23*(4), 411-440.

Yoshida, K. (2003). Language education policy in Japan ― the problem of espoused objectives versus practice. *Modern Language Journal*, *87*(2), 291-293. Retrieved from http://pweb.cc.sophia.ac.jp/1974ky/Language%20Education%20Policy+in%20Japan.pdf

Appendix A

CHAT Comparison Report
スピーキングテストの比較レポート

Student ID [redacted]
Family Name [redacted]
Given Name [redacted]
Class [redacted]

	Pretest 予備	Midterm 中間	Final 期末
❶ Total Words	79	162	294
❷ Total Turns	18	10	25
❸ Average Turn Length	4.4	16.2	11.8
❹ Longest Turn Length	9	63	38

説明
❶ は、あなたが話した単語の数のことです。
❷ は、あなたが話した回数のことです。
❸ は、あなたの番にあなたが発した言葉の平均数です。
❹ は、一番長かった発言した時の単語数です。

Lesson #1 CHAT

What food do you like? I like omrice. When did you eat sushi. Yesterday. Very delicious. School lunch. Me too. I don't like... No! I don't have dislike food. Child. Next question. What is your favorite movie in YouTube. Channel? I like Arashi. I know. I know. I love Ninomiya Kazunari. Yes. I love. He is cool. Very cool. My hometown is Kumamoto. I belonged to gymnastics club. Kohei Uchimura. How about you? What did you do? Me too.

Midterm CHAT

My name is [redacted]. How many people in your family? My family has four members. My father, my mother and older brother. My father works for a company. He is fut. My mother is a homemaker. My mother is very kind. Her dish is very delicious. My older brother is twenty four years old. He likes games. And he played tennis in high school but he took up an instrument in university student. What is your hobby? What is your favorite comic? My hobby is playing sports. I practiced gymnastics from two years old to eighteen years old. Gymnastics's famous player is Kohei Uchimura. I have quit it but I love gymnastics very much. My favorite genre of music is J-pops. My favorite J-pop artist is Arashi. They are very cool. How about you? I like Kazunari Ninomiya. He is very cool. I love him. His voice is very beautiful. My favorite actor is Okada Masaki. I like his face very much.

Final CHAT

Hello. What's your name? My name is [redacted] Nice to meet you. Let's talk. What is your hobby? What is your favorite animal? Otter? What is otter? My favorite animal is rabbit. Rabbit is very cute and my cousin's family had rabbit. It was very cute. I want to have rabbit someday. I don't have pets. But about five years ago, I had a goldfish. I got it omatsuri, Kingyosukui in festival. Do you have pets? What is dog's name? Choko? Cute. I want to meet your dog. Understand. I think Miniature Schnauzer is same Miniature Dachshund. Different. Ok ok. My hobby is listening to music. I often listen to music. I listen to music when I go to school and I studying. Before sleeping? I love Arashi. I love them since I was elementary school student. I will go to their concert in December. I'm really looking forward to meeting them. Do you know Arashi? Monster. It is very cool. My favorite song is One love. Do you know? I want you to listen it. I'm from Kumamoto. I live in Kumamoto now. How about you? Where is Tokushima's good place or Nara's good place? I have been to there in my junior high school school trip. It was very fun. My family has four people, my mother and my father and my older brother and me. My brother is university student. He likes to play game and to play tennis. How many member do you