【学術論文】（査読あり）
京都工芸繊維大学学術報告書　第 14 巻（令和 4 年 2 月）
BULLETIN OF KYOTO INSTITUTE OF TECHNOLOGY　14 (Feb. 2022)

Development and Administration of a Skype-based English Speaking Test in a Japanese High School

Katsunori Kanzawa[1], Haruhiko Mitsunaga[2], Glen Edmonds[3], Yumi Hato[1], Yasushi Tsubota[1], Masayuki Mori[1], Yuko Shimizu[4]

[1]Kyoto Institute of Technology, [2]Nagoya University, [3]Kyoto Sangyo University, [4]Ritsumeikan University

kanzawa@kit.ac.jp

（2020 年 10 月 6 日原稿受理　2020 年 11 月 20 日採用決定）

Summary

How to incorporate English speaking tests into entrance examinations is currently much discussed, and Kyoto Institute of Technology's (KIT) experience doing so at tertiary level has highlighted the importance of preparing the ground with daily and term-end tests. Existing standardized tests are not well suited for this, though, and at secondary level, the limited number of assistant language teachers (ALTs) makes administrating face-to-face tests difficult. Therefore, KIT in partnership with Kyoto Kogakuin High School (KGH) and QQEnglish (QQE), a business based in Cebu, the Philippines, that provides English lessons via Skype, developed a Skype-based speaking test. This was designed to relate to KGH's English syllabi, test English as a lingua franca (ELF), and have positive washback. Regarding washback, surveys showed students who felt hesitant about speaking English improved when they overcame this hesitancy.

Keywords: video-conferencing, speaking test, lingua franca, washback effect, survey

1. Introduction

English education in Japan has long been centered on reading and grammar. Since the 2000s, however, the Ministry of Education, Culture, Sport, Science and Technology (MEXT) has sought to shift the focus to improving students' communication skills (MEXT 2003). Of the 'four skills' (speaking, listening, reading and writing) speaking is the most difficult to measure in practice. Educational institutions have found it difficult to administer face-to-face tests on a sufficient scale, and a method of assessment speaking that matches Japanese students has yet to be sufficiently established.

KIT's experience with developing English speaking tests goes back to the institute's introduction of its own computer-based test (the KIT Speaking Test) in the 2014 academic year. This has been administered to all first-year students, approximately 600 in each cohort, on a yearly

basis ever since. From the early stages of development, KIT planned also to incorporate the test into the graduate school entrance examinations that the majority of undergraduate students take. However, questionnaires conducted after test administrations showed that although it had face validity among students, many felt insufficiently prepared to take such a test as part of a high-stakes entrance examination. Consequently, KIT decided to postpone the integration of the test into the graduate school entrance examination. Instead, attention was turned to end-of semester tests, and classroom-based teaching of speaking in connection with them, so that by 2017 KIT was in a position to incorporate its speaking test into the institute's admission office (AO) entrance examination. (For details of the development and administration of the KIT Speaking Test, see Hato & Kanzawa (2015), and Kanzawa *et al.* (2019).)

As part of its test development program, KIT collaborated on a video-conferencing-based test of spoken English with KGH and QQE (the KGH Speaking Test). This test was administered to students at Kyoto Kogakuin High School, and focused on English as a lingua franca, which has been defined as 'a "contact language" between persons who share neither a common native tongue nor a common (national) culture, and for whom English is the chosen *foreign* language of communication' (Firth 1996: 240). In the 'expanding circle' English is not used on a daily basis, and students are learning the language for communication with other non-native speakers as well as native speakers. As Jenkins *et al*. (2011) demonstrated, interlocutors in ELF situations use a variety of Englishes away from native-speaker norms in terms of phonology, lexicogrammar, and pragmatics, which undermines the assumption that conforming to these norms is essential for successful communication.

With the use of English as a lingua franca in mind, we opted for non-native speakers as interviewers and raters for the test. These interviewer-raters were instructors at QQE, a business based in Cebu, the Philippines, that provides English lessons via Skype, the video-conferencing system powered by Microsoft. The KGH test itself was also conducted via Skype. The rationale for choosing the video-conferencing mode is discussed in Section 2, and the details of the test which measures students' ability to use ELF is described in Section 3.

In addition to testing ELF, with the KGH Speaking Test we also attended to the effects of washback. 'Washback' is defined as the influence of a test on students' learning and teachers' instruction. These effects could be positive or negative (Alderson *et al*. 1993). For example, if the measurement of writing ability through multiple-choice questions led to lessons where practicing multiple choice questions predominated at the expense of actual writing, that would be considered a negative effect (Davis *et al*. 1999). On the other hand, if the introduction of a speaking test led to more oral communication in the classroom, that would be considered a positive effect (Taylor 2005).

As Hughes (2003) observes, the purpose of progress-achievement tests is not only to assess the progress of language learners toward the achievement of course objectives, but also to facilitate that progress through various washback effects. Accordingly, we were keen to design the KGH test in a way that would have a positive washback effect on the attitudes toward English language learning and testing of Japanese junior high and high school students. This implied constructing a test that

was not only a pleasurable and confidence-building experience for teenage students to participate in, but also integrated well into the existing school syllabus. We sought to ascertain whether the KGH Speaking Test had a positive washback effect on regular classroom instruction and activity by conducting a post-questionnaire conducted on the students. See Section 4 for the analysis.

The KGH Speaking Test was first administered in the academic year starting April 2016 as part of the Frontier Science and Mathematics Course at KGH. It was conducted in the school's computer suite, where students were connected online to interview-raters in QQE's Cebu center. Five tests were administered to approximately 50 ~ 60 students in two classes as part of the end-of-term examinations of the English Expressions I (EEI) and English Expressions II (EEII) courses, as shown in the following table.

Table 1: Summary of test administrations

	1st year 1st term	**1st year 2nd term**	**1st year 3rd term**	**2nd year 1st term**	**2nd year 3rd term**
Course	EEI	EEI	EEI	EEII	EEII
Test date	4th July 2016	28th Nov. 2016	6th Feb. 2017	26th June 2017	13th Feb. 2018
No. of examinees	58	58	54	54	53

2. Previous studies on video-conferencing speaking tests

Direct, face-to-face speaking tests have been conducted for over a century (Weir, Vidakovic, & Galaczi, 2013), but in the 1970s, semi-direct tests also started to appear. In these tests, stimuli were presented and examinees' audio responses recorded on tape recorders for raters to later listen to and evaluate (Clark, 1979). Early semi-direct tests included the Test of Spoken English (TSE) (Clark and Swinton, 1979) and the Recorded Oral Proficiency Examination (ROPE) (Lowe and Clifford, 1980). Subsequently, around the 1990s and concomitant with the evolution of technology and the spread of the Internet, speaking tests exploiting video-conferencing appeared. These allowed for the conduct of real-time tests even when interviewers and examinees were not in the same physical location.

Ever since their introduction, however, questions about examinees' performance on semi-direct tests in comparison to direct tests have been raised. In an early study addressing such concerns with respect to the video-conferencing mode, Clark (1992) sought to find if examinees performed differently in this kind of speaking test compared with face-to-face ones. To do so, he and his team administered tests in both formats to 32 students studying Arabic and Russian at the Defense Language Institute Foreign Language Center, USA. Regarding the Arabic test, the score of the face-to-face test (1.72) was slightly higher than that of the video-conferencing test (1.64), but this difference was not statistically significant. Concerning the Russian test, the score of the face-to-face test and video-conferencing test were almost identical (1.84 and 1.83 respectively). However, follow-up questionnaires showed that 57% of Arabic examinees and 74% of Russian examinees

preferred the face-to-face mode. Clark (1992) attributed this preference to technical issues, specifically motion discontinuity and audio dropout, and concluded that video-conferencing tests would become accepted if the technical issues were resolved.

Craig and Kim (2010) also carried out research into video-conferencing speaking tests, investigating differences in student performance on face-to-face and video-conferencing tests undertaken by 40 students studying English as a second language at a university in Korea. Regarding examinee performance, they found no statistically significant differences between modes with respect to overall scores or the analytic scores (i.e., fluency, functional competence, accuracy, coherence, and interactivity). This indicated that differences in test mode did not lead to differences in examinee performance. In the same study, Craig and Kim also assessed the levels of pre- and post-test anxiety experienced by examinees, and found the video-conferencing mode to create lower levels (i.e. 2.75 before the test and 3.0 after, as opposed to 3.45 before and 3.15 after for the face-to-face mode). In summary, Kim and Craig suggested that 'the video-conferenced interview was comparable to the face-to-face interview in terms of reliability, construct validity, authenticity, interactivity, impact, and practicality (2012: 257).'

More recently, Nakatsuhara *et al*. (2017) compared the speaking section of the International English Language Testing System (IELTS) test in a face-to-face test and video-conferencing mode. When the test was administered to 32 students participating in an IELTS preparation course at a college in London, no statistically significant differences between modes were found in the overall scores, or any of the four analytic scores (fluency, lexis, grammar, pronunciation). Moreover, analysis of the scores using the many-faceted Rasch model showed no difference in the difficulty of the tests between the two modes. There were, however, several differences in the language output of examinees. Among the language functions examined, *comparing* and *suggesting* were found more frequently used in the face-to-face test, while *clarifying* was more widely used in the video-conferencing test. Also, the researchers reported that, for interviewers, technical problems such as poor audio quality and video delay sometimes caused difficulties with the management of the test (e.g. by impeding intervention), as well as with the evaluation of pronunciation and grammar.

To examine the effect of technical problems in the video-conferencing mode, Davis *et al.* (2018) undertook research on a Skype-based test he and his team developed. This test comprised four speaking tasks which were delivered on a Skype-based platform developed by the Educational Testing Service (ETS) for testing purposes. This platform enabled a 'moderator' (i.e. interviewer-tester) and several examinees in different locations to talk together at the same time in a group. In the administration of trial tests with 72 participants in the U.S., 18 out of 28 sessions (64%) did not experience any technical problems. The situation was different in China, however, where 22 out of 24 sessions (92%) with 74 participants suffered mid-test video link drop out. This notwithstanding, 69% of examinees in China stated in the post-test questionnaire that they preferred the online speaking test over a face-to-face one, with only 18% responding that they favored the latter mode. Most examinees said that not having to meet a moderator and other test takers in person made the online test less stressful. Some also said they liked not having to go to a test center to take

the exam, and that they felt the controlled online environment made the test fairer to all examinees.

Although when designing the KGH Speaking Test we were keen to take advantage of various benefits of the video-conferencing mode, the literature raised awareness of challenges associated with it. The advantages were: (1) the ability to test a large number of students simultaneously, (2) the opportunity to connect students with non-native teachers of English whose first language was not Japanese, were based in a different country, and who had a different cultural background, (3) the lower anxiety levels associated with video-conferencing, and (4) the possible propensity of this mode to encourage student participation in the repair of communication breakdowns by, for instance, asking for clarification. Challengers included the need to: (1) thoroughly prepare a network environment that minimized sound quality deterioration and image delay, (2) allow for opportunities to retest examinees in the event of system failures, and (3) design evaluation criteria which took account of difficulties raters might have assessing pronunciation and grammar.

3. Steps from Test Development to Administration

3.1 Test purposes

The primary purpose of this test is to assess students' progress towards achieving the language goals of a specific course syllabus in a manner that current commercial standardized tests for high school students (e.g. GTEC and EIKEN) are unable to do.

As with the KIT Speaking Test (Hato *et al*. 2018), the secondary goal of the test is to assess students' ability to use ELF, which reflects the reality of contemporary language use that learners in the 'expanding circle' need to engage in.

A third aim relates to washback. With respect to this, the test is intended to (1) improve students' motivation to learn English, (2) develop among students a positive attitude toward speaking English, (3) raise both students and teachers' awareness of ELF, and (4) highlight for teachers alternatives to form-focused instruction.

3.2 Constructs

In consideration of the purposes outlined above, the scope of the test was defined as follows:

1) To measure achievement against the English Expression I and English Expression II course syllabi each term;
2) To measure ELF speaking ability; and
3) To foster a positive attitude toward communicating in English evidenced, for example, in attempts to employ various conversation strategies to keep channels of communication open, avoid breakdowns in communication, and repair misunderstandings when they occur.

As such, the test is intended to engender in students an attitude favorable to second language acquisition; one in which learners make efforts to exploit to maximum effect the language resources introduced to them through the syllabi, imperfect though their command of these may be. To this

end, it focuses less on measuring students' output against the yardstick of native speaker norms than on communicative efficacy in non-native speaker interactions.

3.3 Task types

For their English Expression I & II courses, KGH uses the course book *Departure*. Accordingly, test items are constructed to relate to the topics included in this core text. Items are designed through the collaborative efforts of teachers at KGH, instructors at QQE, and academics at KIT. All three groups continue to provide input into their creation.

The test has either three or four parts, depending upon the school grade of the students taking it. Each part includes a 'prepared speech' and/or a 'spontaneous interaction'. In the prepared speech, students deliver a planned talk on a given topic, while in the spontaneous interaction they engage in an unprepared dialogue with the interviewer. The prepared speech is designed to assess achievement of the term's learning objectives, and the spontaneous interaction to evaluate students' ability to use ELF. Tables 2 and 3 show the structure of the two versions of the test. As shown, a debate section is incorporated into the final exam for 2nd year students as well. This is because conducting a debate is part of the course syllabus for these students, and something they practice in their English Expression II class. Test items include propositions like 'The sale of junk food should be banned by law'.

Table 2: Structure of 1st year test

Part	Task	Time (sec.)
Part 1	Student's prepared speech	45
	Spontaneous interaction (Q&A)	60
Part 2	Student's prepared speech	45
	Spontaneous interaction (Q&A)	60
Part 3	Student's prepared speech	45
	Interviewer's feedback	—
Part 4	Interviewer's prepared speech	45
	Spontaneous interaction (Q&A)	60

Table 3: Structure of 2nd year test

Part	Task	Time (sec.)
Part 1	Student's prepared speech	60
	Spontaneous interaction (Q&A)	60
Part 2	Student's prepared speech	60
	Interviewer's feedback	—
Part 3	Interviewer's prepared speech	60
	Spontaneous interaction (Q&A in which students argue against the interviewer's opinion)	120

As an example of a typical test item, in part 1 of the 1st year term 3 test, students participate in a role play in which they give the interviewer some information or advice about Japanese climate or manners. This relates to lessons students have had on language and cultural diversity:

1) Japanese climate and weather (from Lesson 15: Find out more about the world)
2) Japanese manners and etiquette (from Lesson 17: When different cultures meet)

In this part of the test, students know that they will have to talk about one of the two topics, but do not get to choose which one, and are not told in advance which topic they will be required to talk about. Rather, they find this out in the test, and therefore need to prepare for both, as is the case with most other test items.

In Part 4 of the 1st year test, the interviewer gives a 45-second speech about photographs and other visuals shown onscreen (see figure 1). After listening to the interviewer's talk, the examinee is required to take the initiative in developing a conversation with the interviewer on the topic presented. Examples from the term 3 test are:

1) How English is learned and used in the Philippines
2) The language**s** I speak
3) The language**s** used in the Philippines (from Lesson 18: Speak with the world)

Figure 1: Examinee listening to interviewer's talk

In preparation for part 4, each interviewer scripts a speech, which in the interests of standardization is edited for length and level of difficulty by a panel comprising teachers from KGH, senior instructors at QQE, and academics at KIT. This part of the test is included on the grounds that exposure to the cultures and languages of the Philippines and other Asian countries might promote the development of students' awareness of ELF. (Before taking the test, the majority of students do not know that many languages are used in the Philippines, with English a common means of communication.)

After the test items are decided, guidelines are written for interviewer-raters to provide form to their interactions with examinees. These guidelines include assistance on wording, procedures, and timing, as well as example conversations for reference (see figure 2). In the interests of standardization, interviewer training based on these guidelines is provided to all raters and back-up raters two months prior to the administration of the test. In this way it is ensured that even if staff change, replacement interviewers perform to the standard set.

Interview Guidelines, KGH Videophone Speaking Test 2017/02/06

Question No.	Procedure	Sample Dialogues & Possible Questions	Time (sec)
Greetings & Icebreaking [40sec]			
	Sound check	Hello! Can you hear me well? (st's response) *Solve any problems.	
	Self-introductions	OK, let's start. My name is _____ . I'm an English teacher, and I am now at my school on Cebu Island, Philippines. Could I have your name, please? (st's response) OK, XXX. Nice to meet you. (st's response)	
	Icebreaking	XXX, how are you feeling? Are you feeling nervous? Don't worry! Just relax and talk with me.	40/**40**
Part 1: Giving information and offering advice (roleplaying) [130sec]			
	Instruction	OK, XXX, let's begin Part 1 of the test. All right? XXX, I'm going to visit Japan for the first time this year. Is there any information or advice about __(a) or (b)__ you can give me? Please give me any information or advice about__(a) or (b)__. (a) Japanese climate and weather (b) Japanese manners and etiquette * Prior to the test, each interviewer is assigned either of the above topics, and asks the same question of all the students s/he interviews. OK, XXX? Then, please begin.	25/65

Figure 2: Interview guidelines

3.4 Steps to maximize positive washback

In order for the positive washback effect of the test to be maximized, it is necessary for each student to understand the content and purpose of the test fully, and to prepare sufficiently. Therefore, about two weeks before the exam, teachers distribute a handout in Japanese about test content and preparation. They also give an oral explanation. Finally, to facilitate the smooth delivery of the test, they provide students with a written outline of the procedure, and things to keep in mind while doing the test.

3.5 Test system

Several months before the actual test, a pilot is administered to eight university students in a real PC room to assist with the discovery of potential network issues, like delayed login to Skype, or

loss of video-conferencing connection. Problems such as these have been found to have a variety of causes. For example, a connection might suddenly fail if the same Skype account is being used for the main PC and backup PC. It might also fail if someone is using the network at another site on the campus during the test. Solutions to such problems include logging on all PCs with different Skype accounts, and requesting school staff to avoid heavy use of the Internet in other locations on campus while the test is in progress. A connection failure can also be worked around by logging on with a new account, for which purpose a buffer of additional accounts that can be used if needed during the test is prepared. Double the number of PCs than strictly necessary are also set aside for use just in case spares are required, and interviews are audio-recorded for reference purposes in both the Philippines (with a PC application) and Japan (on a digital recorder) to ensure a record can be retrieved should a link abruptly terminate.

Finally, experience with such technical problems led to the writing of a troubleshooting manual, which provides direction to interviewers and staff in the event of such difficulties. For example, should a network connection problem occur mid-test, the established procedure is to suspend the test at that juncture, and restart it at the same point on a spare PC in the next available time slot.

3.6 Test administration

As testing is conducted during the regular school day, it is essential that the test of a whole class be completed within one teaching period. In practice, this means delivering the test to approximately 30 students within one 50-minute time slot. To ensure this happens, a minute-by-minute timetable was created. In summary, the 30-student class is divided into 4 groups for testing via 8 PCs. Each test lasts 9 minutes, and is preceded by a 1-minute explanation. After the test, 3 minutes are allowed for the current group of students to vacate the PC room, and the next group to take its place.

The timing for each section of the test is provided to interviewers and students in the interview guidelines. In addition, interviewer training is carefully conducted to ensure competence at completing each section within the time allotted. Should any real-time advice on conducting the test be needed to be given to interviewers in the Philippines by staff in Japan (e.g., on delivery or time management), this is possible via Skype's chat function.

To further facilitate the most effective use of the time available, the topics for each question are provided on a list on each desk in the PC room as an aide-memoire for students.

3.7 Rating procedures

Interviewers also fulfill the role of test raters. Prior to each administration, in addition to the interviewer training mentioned above, rater training led by a senior rater is conducted. Post-test, interviewers listen to the sound files for all the students they interviewed and award scores. Table 4 shows the rating scales used. These were created with reference to the Cambridge Preliminary Test (PET) (University of Cambridge ESOL Examinations, 2012), since the target group is broadly equivalent to the A1 ~ B2 levels of the Common European Framework of Reference for Languages (CEFR).

Assessment of the prepared speech is divided into ‘content’ and ‘delivery’, and assessment of the spontaneous interaction into ‘response’ and ‘interaction’. Regarding weighting, 50% is awarded to the prepared speech, and 50% to the spontaneous interaction.

In addition to rating by interviewers, all responses are graded by the senior rater. If necessary, scores are corrected, and feedback provided in the next rater training session.

Table 4: Rating scales

Score	Prepared Speech (50% weighting)		Spontaneous interaction (50% weighting)	
	Content	**Delivery**	**Response**	**Interaction**
5	Produces fairly substantial and coherent speech, using a range of appropriate vocabulary and cohesive devices.	Speaks fluently enough to be comprehensible, and with some confidence.	Produces responses which are extended beyond short phrases, despite some hesitation.	Maintains the interaction with little prompting and support.
4	Between 3 and 5	Between 3 and 5	Between 3 and 5	Between 3 and 5
3	Produces meaningful speech, using a limited range of vocabulary and basic cohesive devices.	Just fluent enough to be comprehensible most of the time, but may lack confidence.	Produces responses which are characterized by short phrases and frequent hesitation.	Maintains simple exchanges with prompting and support.
2	Between 1 and 3	Between 1 and 3	Between 1 and 3	Between 1 and 3
1	Conveys very limited or unconnected information, using only simple words and basic phrases.	Is not comprehensible most of the time.	Only produces isolated words and memorized phrases.	Has considerable difficulty maintaining simple exchanges even with frequent prompting and support.
0	No relevant contribution.	Is not comprehensible at all.	No response at all.	No interaction possible.

3.8 Feedback

After the test, students’ scores are sent to their high school teachers for incorporation into their end of term grades for the course. Teachers are also given a disk containing sound files of all students’ tests to assist with the monitoring of progress and lesson planning. Students get to store the sound files of their own tests on their tablets so that they too may observe performance and monitor progress. In addition, each student receives a written report that along with the test score includes personalized feedback from the interviewer (see appendix 1).

To assist interviewers with writing reports, a wide range of sample report cards were created for reference purposes. These guide interviewers to focus feedback on the content of conversations, rather than on discrete language points (e.g., of pronunciation or grammar). As part of their comments, interviewers are asked to (1) remark upon what the student did well, (2) offer

suggestions about how to make improvements, and (3) provide words of encouragement. It was felt that compared to the scores provided by large commercialized tests, this kind of personalized feedback would be more likely to lead to an increase in student motivation.

4. Findings of Student Survey

4.1 Survey purpose

After each administration, students' opinions were surveyed. The main aims of these surveys were to determine whether the KGH speaking test has a positive washback effect upon students' learning of English conversation skills. (See Section 1 for a description of the washback effect). We hypothesize if students come to develop a positive attitude towards speaking English and the test itself, positive washback is in effect.

4.2 Questionnaire content

Two different questionnaires, A and B, were used to examine student attitudes toward English and the test itself. Questionnaire A contained 37 items which asked for feelings about the speaking test. Questionnaire B comprised 29 items asking for views about communicating with others in English. Both questionnaires also contained items asking about the content of the test itself, rather than for feelings about it. There were three such items in questionnaire A, and two in B. In addition, questionnaire B contained two questions about learning English, rather than communicating in it. The contents of the questionnaires are shown in appendices 2 and 3, here translated into English from the original Japanese.

4.3 Survey period

The tests and surveys were held along the timeline shown in figure 3 below, with English Expression courses ongoing between survey administrations. In figure 3, lessons are represented by grey rectangles, questionnaire administrations by downward triangles, and test administrations by upward triangles. Questionnaire A was distributed after the administration of the first and second tests of 2016. Questionnaire B was administered after the third of 2016, and the first of 2017. Questionnaire B was also administered before the first test of 2016 in order to enable the more accurate gauging of changes in attitudes toward English.

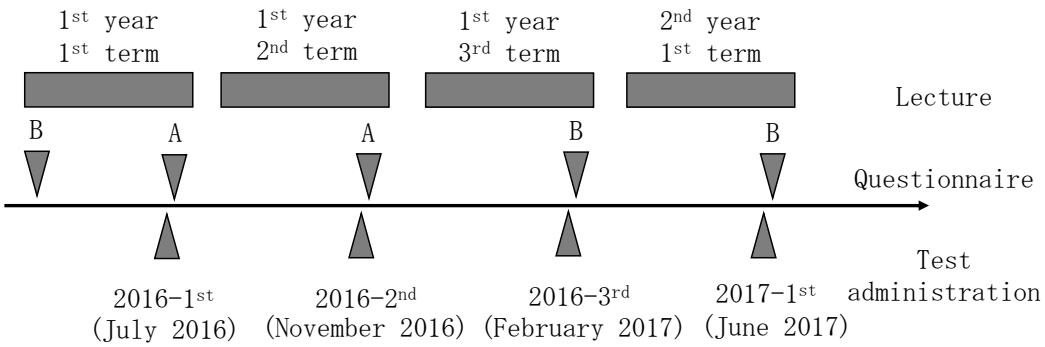

Figure 3: Lesson, test & questionnaire timeline[1]

4.4 Factor Analysis

A factor analysis of the data from questionnaires A and B was conducted, postulating latent factor structure behind the datasets (see appendices 2 and 3). SPSS 23.0 was used to estimate factor loadings. After multiple analyses varying the number of factors retained, three factor solutions for questionnaire A (N=107) and B (N=160) were obtained, making the structure of latent factors easier to interpret.

Appendices 2 and 3 show the factor loadings and communality h^2 for questionnaires A and B respectively. Regarding questionnaire A, factor 1 concerns positive impressions toward the speaking test, factor 2 self-confidence with respect to the test, and factor 3 negative feelings about the test, hesitancy in particular. Regarding questionnaire B, factor 1 shows willingness to communicate in English, factor 2 self-confidence in English, and factor 3 hesitancy about speaking English.

4.5 Explaining changes in test scores

Differences between tests scores were calculated by subtracting the score of the 1st test of 2016 from the 2nd test of 2016 for questionnaire A, and by subtracting the score of the 3rd test of 2016 from that of the 1st test of 2017 for questionnaire B[2]. Differences in factor scores were also calculated for both questionnaires. These differences indicated how the characteristics of each factor changed between questionnaires A and B.

Multiple regression analysis using the test score differences as the dependent variable, and factor scores as the independent variable, was also undertaken in order to ascertain which independent variance contributed to the explanation of the variance of the test score difference. (See figure 4). The values in figure 4 show partial regression coefficients. Large values signify that factors greatly contribute to the explanation of the variance of the dependent variable. The left half of figure 4 indicates that, with respect to questionnaire A, none of the factors was significant in changing the test score. The right half, however, shows that, with respect to questionnaire B, the variance of test scores difference can be explained by hesitancy before and after taking the test. In short, students who felt less hesitation about speaking English attained higher scores on the speaking test.

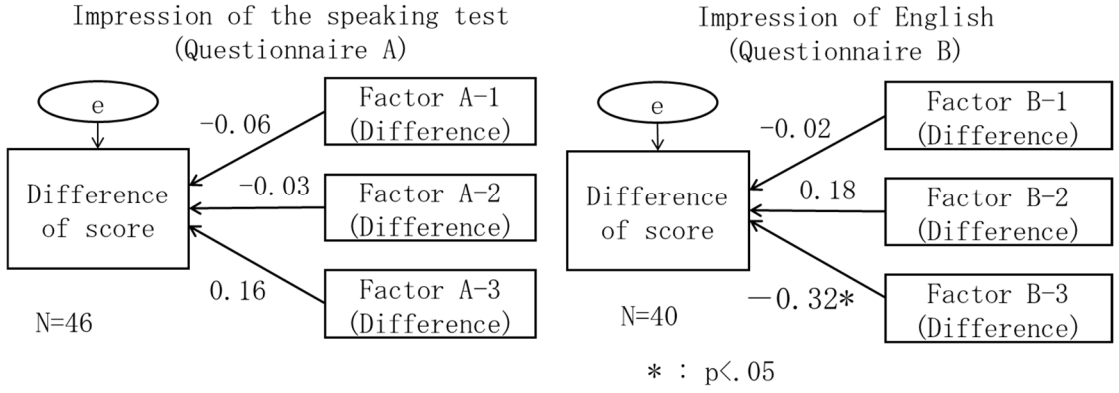

Figure 4: Results of Multiple regression Analysis

4.6 Relationship between impressions of the test

Table 5 and table 6 show the correlation coefficient between questionnaire items in questionnaires A and B respectively. Weak but significant correlations were found between some items in questionnaire A, while stronger, more significant correlations were found between items in questionnaire B. From these findings, we conclude that (1) students who wanted to learn English tended not to hesitate to speak English, (2) students who were eager to improve their conversation skills in English tended to want to take speaking tests periodically, and (3) students who thought the test worked well as a method of assessing achievement on the English Expression courses tended also to feel that it was effective at improving those skills.

Table 5: Correlation analysis of questionnaire A (N=107)

	(A)	(B)	(C)	(D)
Difference of test score (A)	1	0.166	0.200	-0.042
This test is suitable for use when awarding a grade for the English Expression I course. (Difference score, B)	0.166	1	.334*	0.160
This kind of test held at the end of every semester helps improve my English speaking skills. (Difference score, C)	0.200	.334*	1	.282*
Because of this test, I can identify what points to address to improve my English speaking skills. (Difference score, D)	-0.042	0.160	.282*	1

*: $p < .05$

Table 6: Correlation analysis of questionnaire B (N=107)

	(A)	(B)	(C)	(D)	(E)
Difference of test score (A)	1	-0.074	0.202	-0.128	-0.005
I am reluctant to speak in English. (Difference score, B)	-0.074	1	-0.188	-.519**	-.495**
I am good at the English Expression II class. (Difference score, C)	0.202	-0.188	1	0.253	0.286
I became motivated to improve my English speaking skills because I took the speaking test where I talked with an interviewer in the Philippines via Skype. (Difference score, D)	-0.128	-.519**	0.253	1	.648**
I want to take this kind of speaking test periodically. (Difference score, E)	-0.005	-.495**	0.286	.648**	1

**: $p < .01$

4.7 Survey conclusions regarding washback

From the results of the survey, we can conclude that there is evidence of positive washback from the test on students' attitude toward speaking English, in particular regarding hesitancy and nervousness in conversation skills. Furthermore, comparison of the answers to the survey conducted prior to the first-term test and the one administered after the third-term test reveals the number of students who answered 'agree' or 'partially agree' to the question, 'Do you feel hesitation when speaking in English?' decreased over 10% from 61.3% to 50.9%. Furthermore, the number of students who answered 'agree' or 'partially agree' to the question, 'Do you feel nervous when speaking in English?' decreased over 20% from 91.9% to 69.9%. However, there is no statistical evidence of any washback, positive or negative, from the analysis regarding students' feelings about the test itself.

5. Project achievements to date, and challenges for the future

As mentioned in Section 4, the analysis of student questionnaires indicates that the scores of students who felt hesitant about speaking English rose when they overcame this hesitancy. Teachers have also provided positive feedback about the test with remarks like the following:

1) 'It has become easier to conduct day-to-day classes because of the clear goals set by the end-of-semester speaking test.'
2) 'Commercialized tests which assess students' general English skills do not work as well as this.'
3) 'We were able to evaluate students' speaking ability much more easily than by ourselves or by asking an ALT [assistant language teacher, native speaker or native speaker-like in proficiency].'

For the authors, perhaps the most gratifying part of the project has been receiving feedback from students like, 'It was fun,' 'The interviewer was very kind,' and, 'At first I was nervous, and it was hard for me to prepare for the test, but as my level improved, speaking English became more interesting.' It is certainly rewarding seeing students enjoying conversing with the Filipino interviewers, using gestures to facilitate communication when language resources prove insufficient.

Several important issues that need to be addressed have arisen, however. Firstly, the survey uncovered an increasing polarization in terms of learning motivation. For example, the questionnaire administered immediately after the 3rd-term, end-of-term test for grade 1 asked, 'Does taking such a speaking test in the final exam of each term help improve your English speaking ability?' To this, 9.4% of students answered 'completely disagree', while 19% answered 'completely agree'. Moreover, in response to the question, 'Do you want to stop taking such a speaking test from now on?' 20.8% of the students answered 'completely agree', while 15.1% of answered 'completely disagree'. Also, class teachers said they perceived increasing polarization with regard to speaking ability, although this is not shown by statistical analysis of the test scores.

Secondly, there are practical issues that need to be considered with respect to the school's English syllabi, and the demands of implementing such a test. For instance, as the teachers pointed out, most classes follow lexicogrammatical syllabi, and use textbooks with grammar as their organizing feature. As there are only two English Expression classes per week, there is insufficient time for speaking practice, and students themselves have made remarks like, 'It is difficult to speak English without practicing daily in class,' and, 'I do not know how to practice speaking.' The issue here is how to integrate preparation for the term-end tests into day-to-day classroom activities. The solution, we feel, lies in the direction of developing day-to-day class content in parallel with the term-end tests.

Finally, there is the matter of the demands that administering an online speaking test to a large number of students places on Information and Communication Technology (ICT) teachers. Delivering the test requires not only dedicated access to a PC room for a significant period of time, but also the dedication of a substantial number of work hours from ICT staff. How to reduce the workload for these teachers is another of our main practical considerations going forward, and the resolution of this issues will be the subject of further research.

The authors look forward to the day when the teaching and assessment of speaking skills, along with the skills of listening, reading, and writing, become a part of daily practice in Japanese high schools. We hope this paper functions as a step in that direction.

Notes

1. Test administrations are named after the Japanese school years in which they were held. Therefore, test names may differ from calendar years of administration.
2. The targets of multiple regression analysis are students who have taken both of the tests indicated, and answered all the questions in both questionnaires A and B.

References

J. C Alderson and D. Wall, Applied Linguistics. 14(2), 115-129, 1993.

J. L. Clark, Concepts in language testing: Some recent studies, TESOL, Washington D.C., 35-49, 1979.

J. L. Clark and D. Hooshmand, System. 20(3), 293-304, 1992.

J. L. Clark and S. S. Swinton, ETS Research Report Series. 1979(1), i-69, 1979.

D. A. Craig and J. Kim, Multimedia Assisted Language Learning. 13(3), 9-32, 2010.

A. Davies, A. Brown, C. Elder, K. Hill, T. Lumley, and T. McNamara. Dictionary of Language Testing, Studies in Language Testing, Cambridge University Press, Cambridge, 1999.

L. Davis, V. Timpe-Laughlin, L. Gu, and G. J. Ockey, Useful assessment and evaluation in language education, Georgetown University Press, Washington D.C., 115-130, 2018.

A. Firth, Journal of Pragmatics. 26(2), 237-259, 1996.

Y. Hato and K. Kanzawa, Official Bulletin of Center for Information Science. 34, 30-48, 2015.

Y. Hato, K. Kanzawa, H. Mitsunaga, and S. Healy, Waseda Working Papers in ELF. 7, 87-99, 2018.

A. Hughes, Testing for Language Teachers, Cambridge University Press, Cambridge, 2003.
J. Jenkins, A. Cogo, and M. Dewey, Language teaching. 44(3), 281-315, 2011.
K. Kanzawa, M. Mori, Y. Tsubota, and Y. Hato, Official Bulletin of Center for Information Science. 37, 22-36, 2019.
J. Kim and D. A. Craig, Computer Assisted Language Learning. 25(3), 257-275, 2012.
P. Lowe Jr, and R. T. Clifford, Measuring spoken language proficiency, 31-39, 1980.
Ministry of Education, Culture, Sports, Science and Technology, An Action Plan to Cultivate "Japanese with English Abilities," 2003.
F. Nakatsuhara, C. Inoue, V. Berry, and E. Galaczi, Language Assessment Quarterly. 14(1), 1-18, 2017.
L. Taylor, ELT Journal. 59(2), 154-155, 2005.
University of Cambridge ESOL Examinations, Cambridge English Preliminary: Handbook for Teachers, 2012.
C. J. Weir, I. Vidaković, and E. D Galaczi, Measured constructs: A history of Cambridge English examinations, 1913-2012 (Vol. 37), Cambridge University Press, Cambridge, 2013.

Appendix 1: Example of Student Report Card and Interviewer Feedback

Score Report Kyoto Kogakuin High School Videophone English Speaking Test

Class-Student No.:	[redacted]
Student's name:	[redacted]
Interviewer's name:	Chu
Date and time:	2016/11/28, 2-2 (15:34-15:43)

Your overall score (Full marks: 60) **39**

Interviewer's comment:

Hi! [redacted]

You did well in the test. Your speech in Part 1 was especially interesting. Shooting is not a very common sport a boy at your age would be interested in. I was a little surprised to hear that you enjoy shooting, but it sounds exciting. You corrected yourself in your speech in Part 3, which shows that you were careful of what you were saying.

I felt that you were genuinely interested in talking, and really tried to have a conversation with me. When I asked you where you lived, you had some trouble finding the right words in answering my question. I think you wanted to say that you live in the "outskirts" of the city. Perhaps, you could have said "I live in the suburb." Also, the word "funny" has the wrong feeling for what you were trying to say in your talk about the countryside. A better choice of word would be "fun" because it means enjoyable or amusing, but the word "funny" means something that makes us laugh. Therefore, it would be better to say, "The countryside is fun."

I advise you to try to have more opportunity to use English so you can become a better English speaker.

Good luck, [redacted] :)

Chu (QQ English, Cebu, Philippines)

Your score breakdown:

	Prepared Speech (50% weighting)		Q & A (50% weighting)	
	Content	Delivery	Discourse Management	Interactive Communication
Your scores (Full marks: 5)	3.0	3.3	3.3	3.3
5	- produces fairly substantial and coherent speech, using a range of appropriate vocabulary and cohesive devices.	- speaks fluently enough to be comprehensible and with some confidence.	- produces responses/questions which are extended beyond short phrases, despite some hesitation.	- maintains the interaction with little prompting and support.
4	Between 3 and 5	Between 3 and 5	Between 3 and 5	Between 3 and 5
3	- produces meaningful speech, using a limited range of vocabulary and basic cohesive devices.	- just fluent enough to be comprehensible most of the time but may lack confidence.	- produces responses/questions which are characterized by short phrases and frequent hesitation.	- maintains simple exchanges with prompting and support.
2	Between 1 and 3	Between 1 and 3	Between 1 and 3	Between 1 and 3
1	- conveys very limited or unconnected information, using only simple words and basic phrases.	- is not comprehensible most of the time.	- only produces isolated words and memorized phrases.	- has considerable difficulty maintaining simple exchanges even with frequent prompting and support.
0	- no relevant contribution.	- is not comprehensible at all.	- no response at all.	- no interaction possible.

Appendix 2: Questionnaire A Items, Factor Loadings and Communality h^2

Item	Factor1	Factor2	Factor3	h^2
I never want to take this exam again.	-0.967	0.188	0.299	0.775
I had fun during the exam.	0.865	-0.065	0.045	0.779
I felt the interviewer was easy to get on with.	0.773	-0.093	-0.239	0.676
I think it would reduce the anxiety about university entrance speaking tests if high schools administer this kind of test periodically.	0.741	-0.192	0.058	0.655
I am looking forward to the next time I take this exam.	0.726	0.111	-0.004	0.762
I wanted to talk to the interviewer much longer.	0.720	0.024	-0.067	0.699
I fully understood what the interviewer was saying.	0.655	-0.047	-0.219	0.628
I want to have more opportunities to speak English outside the classroom having taken the exam.	0.628	0.112	0.204	0.760
I am eager to improve my speaking ability in English having taken the exam.	0.622	-0.048	0.438	0.764
The interviewer's speech in Part 4 was interesting for me.	0.620	-0.026	-0.044	0.596
English spoken by the interviewer was easy to understand.	0.616	-0.074	-0.063	0.642
I want to be a good speaker of English having taken the exam.	0.616	-0.068	0.380	0.670
I want to work harder in 'English expression I' lessons having taken the exam.	0.608	-0.026	0.482	0.745
I want to take Skype lessons with a foreign language teacher having taken the exam.	0.607	0.053	0.122	0.600
I could speak without hesitation better than I expected.	0.583	0.040	-0.283	0.548
In the Prepared Speech section, I did not know how to prepare the task.	-0.523	0.297	-0.034	0.570
I fully committed to the conversation with the interviewer in order to communicate well.	0.466	0.301	-0.083	0.581
During the test, I think that my speech in English could be understood by others more than I expected.	0.441	0.283	-0.365	0.656
I want to have more opportunities to communicate with people in the non-English speaking world, especially Asia, having taken the exam.	0.428	0.315	0.161	0.748
I paid close attention to speaking English with more accurate pronunciation and syntax during the exam.	0.380	0.280	0.047	0.597

I prepared well for the Prepared Speech section once it was announced.	0.319	-0.009	0.119	0.531
I had self-confidence about my pronunciation of English after I took the exam.	0.161	0.735	-0.188	0.789
I wondered about the interviewer not being a native English speaker.	-0.378	0.720	0.115	0.503
I am self-confident in speaking English after the test.	0.265	0.613	-0.137	0.790
If this exam were held face-to-face rather than via Skype, I could have spoken better.	-0.115	0.495	0.061	0.503
I have courage to talk with foreigners in English having taken the exam.	0.409	0.463	-0.083	0.708
I am interested in the Philippines having taken the exam.	0.232	0.402	-0.039	0.581
When I could not make myself understood by the interviewer, I improved my communication strategies: e.g., by repeating the same sentence, or changing the vocabulary.	0.244	0.373	0.012	0.476
I felt awkward because I could only talk in Japanese-style English during the exam.	-0.042	0.159	0.491	0.522
I could not speak English as well as I expected.	-0.048	0.330	0.419	0.460
I felt humiliated.	-0.336	0.114	0.407	0.596
I want to speak English more confidently during classes at school having taken the exam.	0.393	0.245	0.394	0.679
I was nervous.	0.073	-0.145	0.361	0.438
During the exam, I often thought in Japanese.	-0.248	0.015	0.325	0.497
I paid little attention to pronunciation and grammar in the test.	-0.072	0.088	-0.194	0.267
I asked the interviewer to repeat, speak slowly or use different words when I had a lot of difficulty understanding what s/he was saying.	-0.080	0.056	0.182	0.297
I could understand the interviewer's speech in Part 4.	-0.060	0.156	-0.166	0.198

Appendix 3: Items Used in Questionnaire B, Factor Loadings and Communality h^2

Item	Factor1	Factor2	Factor3	h^2
I am interested in communicating with foreigners.	0.795	-0.031	-0.159	0.631
I want to have the chance to use English overseas.	0.787	-0.015	0.097	0.674
I want to have more opportunities to speak English outside the classroom.	0.780	-0.048	0.234	0.693
I want to talk with native speakers of English.	0.766	-0.067	0.020	0.571
I want to study or do a homestay abroad in the future.	0.740	-0.155	-0.047	0.548
I am interested in foreign countries.	0.664	-0.061	-0.188	0.530
I want to speak English like a native speaker.	0.578	-0.021	0.312	0.544
I have fun speaking in English.	0.511	0.387	-0.043	0.634
I am eager to improve my English conversational skills.	0.456	0.274	-0.013	0.556
I will not be anxious if I can't speak English.	-0.354	0.017	-0.017	0.255
I am interested in Asian countries.	0.353	0.068	-0.141	0.290
I am working harder in my 'English Expression I' high school class.	0.333	0.282	-0.081	0.524
I don't want to communicate with foreigners who speak with an unusual accent.	-0.267	-0.201	0.129	0.326
I am good at English.	-0.031	0.846	0.166	0.726
I am good at reading English.	-0.068	0.744	-0.047	0.557
I am good at the 'Communication English I' class in high school.	-0.111	0.677	0.122	0.515
I am good at speaking English.	0.003	0.627	-0.057	0.526
I have fun learning English.	0.288	0.563	-0.030	0.643
I am good at writing English.	-0.094	0.555	0.358	0.484
I am good at listening English.	0.010	0.490	-0.187	0.416
I have little confidence about English grammar.	0.000	-0.432	0.182	0.361
I have little confidence about English pronunciation.	0.032	-0.407	0.190	0.307
I have little idea of how to improve my English speaking skill.	0.016	-0.351	0.168	0.297
I always pay close attention to speaking English with more accurate pronunciation and grammar.	0.201	0.232	0.111	0.285
I feel awkward because I can only speak Japanese-style English.	0.179	-0.108	0.794	0.581
I feel awkward when I speak in English.	-0.097	-0.199	0.577	0.486
I think it is impossible to make myself understood in Japanese-style English.	-0.284	0.091	0.530	0.389

When I talk in English, I am not worried about my pronunciation and grammar if I can make myself understood.	-0.053	-0.057	-0.373	0.297
I am nervous when I speak in English.	0.028	-0.146	0.360	0.411

Skype 方式の英語スピーキングテストの開発と日本の高等学校における実施

要旨

本論文では、京都市立京都工学院高校で実施したビデオフォン (Skype) 方式英語スピーキングテストの詳細を述べるとともに、生徒を対象に複数回実施した質問紙調査の概要と結果について論じる。京都工芸繊維大学の教員を中心とする研究チームは、京都工学院高校、および、フィリピン・セブを拠点にオンライン英語レッスンを提供する株式会社 QQ English と共同でビデオフォン (Skype) 方式英語スピーキングテストを開発し、京都工学院高校フロンティア理数科の一学年を対象に 2 年間に渡ってテストを実施した。テストは学期末考査として実施し、生徒が使用する教科書に準拠した内容とした。テストの目的は、(1) 生徒の学期中の達成度を測定すること、(2) リンガフランカとしての英語能力を測定すること、(3) テストを通じて、生徒の異文化への関心・理解を深めるとともに、ポジティブな波及効果を得ることである。生徒を対象に実施した質問紙調査の分析から、このテストにはある程度のポジティブな波及効果があることが示唆された。具体的には、英語で話すことに対する恥ずかしさが低下した生徒はスコアが上昇する傾向があることが分かった。また、2 年間の実施の後、英語で話すことを恥ずかしいと感じる生徒の割合が減少した。

キーワード: ビデオ会議システム、スピーキングテスト、リンガフランカ、波及効果、質問紙調査