Test specification

KIT Speaking Test: English for the 21st Century

Last updated 15 April 2016

1. **Test purpose**
    a. To assess to what extent each student has acquired the speaking ability in English required to actively participate in 21st century global society
    b. To re-balance the teaching, learning and testing of speaking skills by sending these positive washback messages:
        - Speaking is as important as the other core language skills. It is not limited to daily conversation but is a serious communication mode essential to succeed in the 21st century.
        - Students are learning English as a lingua franca for communication between non-native speakers as well as with native speakers.
        - Students must learn to be confident users of their existing language resources at any stage of their language development, free from an excessive concern to conform to NS norms.
        - Students must be prepared to speak without preparation, as spoken interaction is normally spontaneous.
    c. To evaluate the feasibility of incorporating a speaking test into the current English education programme, and developing a speaking component for an English language admissions test to Japanese university postgraduate courses, with a possible future application to undergraduate entrance examinations

2. **Target Language Use domain**

    English as a Lingua Franca (ELF), reflecting the reality of language use which learners in the 'expanding circle' need to engage in. In this area English is not used on a daily basis and learners are learning the language for communication with other non-native speakers as well as native speakers.

3. **Definition of constructs to be measured**
    a. The ability to achieve a given task utilising spoken language proficiency and 21st century skills (creativity and innovation, critical thinking, problem solving, global awareness)
    b. Confident and fluent use of language

4. **Characteristics of test takers**
    a. Age/background:
    The students are 1st year undergraduates aged 19-21 in the Faculty of Science and Technology at Kyoto Institute of Technology. Most are Japanese with a few from other countries. Mixed male and female, but the majority are male.

b. Language learning background:
At school, Japanese students have studied English for six years: three years in Junior High School and three years in Senior High School. At university, they have completed one year of undergraduate study. Their level on the CEFR scale will be approximately A2-B1 with a bias towards reading and writing. Those who have prepared for the TOEIC test will also have enhanced listening skills. Some students have studied English additionally at private schools.

c. 21st century skills background:
Students have not been explicitly taught 21st century skills but courses they have taken may have included some exposure to these issues.

5. Test structure and sequence

The test consists of a series of tasks which are delivered by the computer. The candidate sees or hears a prompt, which may be one or more photographs, an audio soundtrack and/or a text message.

Future possibilities include the use of videos, graphs or other graphics to give instructions or deliver a prompt or dialogue.

The candidate responds by speaking into the microphone for up to 60 seconds per task, and the computer records the responses in digital sound files, omitting all the instructions except the question number. A total speech sample of 7 minutes is elicited from each student.

There are some introductory questions such as name and student number to check the volume level and recording function but these are not recorded or scored. Students see a volume indicator that tells them whether their voice is loud enough.

A microphone symbol on the screen indicates when the student's speech is being recorded. For most tasks, no preparation time is given, in order to encourage students to speak spontaneously. However, for some tasks, there is an opportunity for the student to rehearse their speech, and this rehearsal is not recorded.

Students see a time display that tells them how much time remains for their response in each question.

6. Task types

Q no.	Name	Speech sample	Prompt	21st century skills: Learners should…	Spoken language proficiency: Learners should …	Sample task
Part 1						
1 & 2	Imagine	45 seconds	Photo	think creatively and demonstrate originality	speak coherently and clearly	*Imagine who the owner of these shoes is and why they are here.*
3	Compare	45 seconds	Two photos	analyse alternatives and draw reasonable conclusion	compare, decide and justify	*Which of these cars would you prefer to have? Compare the cars and explain your reasons.*
Part 2						
4	Identify different values	45 seconds	Audio dialogue & photos	understand diverse values and perspectives	summarise and contrast different points of view	*How are Susan's and Kenji's opinions different? [join big company/security vs starting own business/exciting life]*
5	Take position	60 seconds	Same as task 4	evaluate arguments and decide own position	state and justify own position	*Which way of thinking do you support? Explain your position, giving examples.*
6	Identify problem	45 seconds	Audio dialogue & photos	interpret information to Identify problem	describe problem	*What is the problem Bill is facing? [One member of the lab team isn't collaborating]*
7	Problem solving	60 seconds	Same as task 6	find solution to problem	propose solution to problem	*If you were Bill, what would you do to solve the problem?*
Part 3						
8	Plan and organise	60 seconds, plus one rehearsal		identify and organize component parts to make a plan	suggest a plan and series of steps to achieve goal	*You want to organize a volunteer group on campus to help homeless people. Identify the different steps you would take and explain how you would organize them.*
9	Persuade	60 seconds, plus one rehearsal		promote and influence	persuade by presenting a positive image and message	*In an interview for a scholarship program, you are asked to explain why you should be selected. Talk about your personal achievements and strong points.*

7. Rating scales

Score rating	Task achievement (80% weighting)	Task delivery (20% weighting)
5	-task is achieved with satisfactory supporting detail.	- speaks fluently enough to be comprehensible and with some confidence. - given time is well used despite some hesitation or repetition.
4	Between 3 and 5	Between 3 and 5
3	-task is partially achieved, or is achieved with minimal supporting detail.	- just fluent enough to be comprehensible most of the time but may lack confidence. - given time is not effectively used because of frequent hesitation or repetition
2	Between 1 and 3	Between 1 and 3
1	- some relevant words but task is not achieved.	- is not comprehensible most of the time.
0	- no relevant contribution.	- no comprehensible contribution.

8. Test usefulness qualities

a. Validity
   - Semi-direct, not interpersonally interactive
   - Elicits authentic and meaningful language
   - High face validity: 74% of students in the second administration said their speaking ability was properly assessed by this test, and 86% said the ability tested by this test was important to them (n=575).

b. Reliability is a focus for continuing research. This includes refinement of the rating scales, rater training and consistency of rater judgments, based on statistical analysis using Item Response Theory (IRT). Each student's response to each question is scored by two raters (100% double-marking).

c. Impact
   - Important washback effect (test purpose b).
   - The test is relevant to course objectives and will contribute to the overall score of the English programme.

d. Practicality
   - can be delivered to a large number of candidates (70-80) at the same time in the same computer room
   - The sound files of each student's responses are made available remotely to the raters, who rate each response online.

9. Test delivery

The test is delivered via a computer, a headphone with a microphone. The CBT system is designed so that all students begin the same response at the same time. This means the students are less likely to be influenced by the neighbors' responses. The student's spoken responses are temporarily saved on the drive of each student's university computer account and then retrieved through multiple servers to provide a secure data sharing system. In case of failure to retrieve the data through servers, a backup is stored on USB flash memory.

Students are informed in advance of:

- the test purpose
- the test format
- task types
- sample items, with additional guidance highlighted in Japanese
- rating scales

In particular, students are advised that they will get a score of up to five marks for task achievement and task delivery for each task, and it is emphasized that the intention of the mark scheme is to reward them for achieving the task confidently, making the best use of their linguistic resources.

10. Test marking

The sound file for each question is scored remotely by two raters, one native speaker of English and one non-native speaker based in the Philippines. They rate each response online,

using the rating scales. A senior rater adjudicates the discrepancy of the scores between the first two raters where there is a gap of two or more marks.

Mark scheme

Each of the nine tasks is marked 0-5 marks on two scales, 'Task Achievement' and 'Task Delivery'. To produce the final score, the marks on the scales are weighted 80% to 'Task Achievement' and 20% to 'Task Delivery'.

The design of the rating scales has been informed by, among other sources, the Cambridge English: Preliminary Speaking scales.

11. Score reporting and test results

Scores for the test are reported to each individual as

1) An overall final score out of 100, with the combined scale scores weighted as described above
2) A graphic display of the overall score distribution, with an indication of that student's overall score in comparison with his or her peers.
3) A breakdown of that student's scores out of five on each scale, shown against the rating scale descriptors.

The test is pitched at CEFR A2-B1. However, given the innovative nature of the test skills and tasks, it is not possible to pre-determine the equivalence of levels of performance to established tests or scales. Data on scores on other established tests (TOEIC, TOEFL) collected from the first cohorts will be used to explore whether it is possible to establish equivalences for each of the bands on the KIT test, and establishing comparability against the CEFR will be a goal for future research.

Impact hypothesis: a baseline study

If the test is successful in creating a strong positive washback, the overall results will gradually improve and be more widely distributed when successive cohorts of similar students have a greater awareness of the test and take the test after a period of tuition and self-study. The first cohorts of this test should therefore be seen as a 'baseline' study.