Relationship Among Tasks in a Speaking Test:
In a Case of Eiken Speaking Test

Yukiko YABUTA　Atsushi IINO*　Youichi NAKAMURA

スピーキングテストにおけるタスク間の関係性:
英検スピーキングテストの場合

籔田由己子　飯野厚*　中村洋一

This study examined the following three relationships in Eiken Speaking Test: 1) Relationships between test tasks and total score of speaking test, 2) Relationships among test tasks, and 3) Relationships between test tasks and overall English proficiency. We found some relationships that leads us to the hypothetical model to show explain speaking ability as follows: speaking ability = listening comprehension + linguistic knowledge + reading aloud + idea expression.

key words: speaking test, task, relationship, speaking ability, English proficiency

1. Introduction

Since the introduction of “Five Proposals and Specific Measures for Developing Proficiency in English for International Communication” by the Ministry of Education, Culture, Sports, Science and Technology of Japan in 2011, measuring students’ English proficiency has become one of the central issues in English education in Japan. The government and educational boards encourage schools to actively use external certification tests to assess English proficiency of students, including speaking ability. In fact, Eiken and GTEC for STUDENTS are the most tests commonly-useded in junior high and high schools.

Speaking tests usually consist of several different tasks. Brown and Abeywickrama (2010) propose five assessment tasks for measuring speaking ability; imitative (imitate a word or phrase), intensive (reading aloud, sentence and dialogue completion), responsive (short conversation), interactive (longer conversation, discussion), and extensive (speech, oral presentation, storytelling). Weir (1990) presents eight types of speaking assessment; verbal essay, oral presentation, free interview, controlled interview, information transfer (picture description), information transfer (information gap), interaction task, and role play.

It is quite common that a test is made up of several tasks, each assessing a different aspect of speaking ability. Baba (1997) mentioned the importance of combining tasks to ensure the validity of measurement. This can be seen in commercial tests such as Eiken Speaking Test, TOEIC® SW Test, Versant™, GTEC, Cambridge ESOL speaking test, and Standard Speaking Test. Eiken Speaking Test consists of three sections: reading aloud, picture

* 清泉女学院短期大学・准教授（2005/04 - 2009/03）、法政大学・准教授（2009/04 - 現在）

description, and open-ended questions, whereas Versant™, a computer-assisted or telephone-mediated test, has additional sections for repetition, sentence building, and story retelling. Even though it is commonly believed that the assessment of speaking is quite complex and difficult, few studies have tried to explain the speaking ability using these tests.

2. Literature Review

Eiken Speaking Test for Grade 3 to pre-1 consists of three tasks: reading aloud, picture description and open-ended questions.

Firstly, reading aloud is used in some of the speaking tests. Scoring of reading aloud is relatively easy because the participants' oral production is pretty much controlled. (Brown & Abeywickrama, 2010). There are few studies that focus on reading aloud in English education. Miyasako (2002) investigated the relationship between reading aloud ability and overall English proficiency. In this study, 40 high school students were given a previous version of the Eiken Test to measure overall proficiency and at the same time, they were also given a reading aloud test. Two teachers evaluated the students reading aloud performance in terms of pronunciation, intonation, pause, conveyed meaning, and reading speed. Significant correlation was found between overall English proficiency and reading around ability ($r = .550$, $p < .01$). Regression analysis was also conducted with the same data to examine the degree of contribution of different factors to reading aloud ability. The results indicated that overall English proficiency contributed 30.2% and reading comprehension contributed 23.9% to reading aloud ability. From the study, it can be said that 1) there is a relationship between reading aloud ability and overall English proficiency, 2) if a test taker possesses a higher English ability and is good at reading comprehension, this person will perform well in the reading aloud test.

Iino, Yabuta & Thomas (2010) also investigated the relationship between reading aloud and English proficiency. In their study, CASEC, a computer adaptive test for English communication, The Edinburgh Project on Extensive Reading (EPER), a test to measure reading proficiency, and reading aloud tests were given to 80 junior college students. The results of the reading aloud test were evaluated by three teachers using three criteria; accuracy, intelligibility, and fluency. The study also found the significant relationship between reading aloud and overall English proficiency (CASEC score) ($r = .615$, $p < .01$), which confirmed the findings from Miyasako (2002). It also found that reading aloud and reading comprehension (EPER test score) had a significant correlation ($r = .529$, $p < .01$).

Picture description tasks are popular ways to elicit oral performance. Types of pictures vary depending on the level. It can be very simple, designed to elicit specific words or phrases, or require the participants to tell a story or describe the incident illustrated in the pictures. This format is widely used in the major commercial tests such as Eiken Speaking Test, TOEIC® SW, Cambridge ESOL Speaking Test, and Standard Speaking Test.

Uenishi (2004) researched the factors contributing to the English proficiency of high school students. The students in his study were asked to take two types of speaking tests: picture description test and topic-based speech test. They also took a listening test, vocabulary test, grammar test, and cloze test. One of the results from the multiple regression analysis indicated a relationship between cloze tests and picture description task.

Yamakawa (2003) investigated the effects of repetition and planning on oral performance using picture description tasks. 108 university students were divided into two groups according to their level of proficiency and asked to complete two speaking tasks; picture description and topical speech. In each level, students were divided into planning group and repetition group. The students in the planning group had five minutes of planning time before the task, and the students in the repetition group repeated the same task three times after one minute of preparation. The result from analysis of variance revealed that the repetition group outperformed the planning group in both tasks in terms of fluency and the amount of production, but there was no significant difference between the two groups in accuracy and complexity.

Question-and-answer or dialogue interview tasks usually consist of one or two questions from an interviewer. The questions can vary from simple comprehension check questions to open-ended question asking interviewee's idea.

Nakamura (2004) compared the students' oral performance in two types of speaking tests, a dialogue interview test and a 'multilogue' discussion. 46 university students took both tests and in the multilogue discussion, students were divided into groups of three to four. Both data were evaluated in five categories: grammar, fluency, vocabulary, conversation strategies, and pronunciation for dialogue or content for multilogue. He concluded that as a whole, the multilogue test was more difficult than the dialogue test. Among the evaluation items, grammar is the most difficult item for students to score well on both tests followed by vocabulary in multilogue and fluency in dialogue.

A number of studies have been conducted for assessing speaking ability, few studies focus on the relationships among tasks in the test and how these tasks contribute to the total score of the test. The present study addresses this issue by examining the results of Eiken Speaking test. This study examines the following three research areas:

1) Relationships between test tasks and total score of speaking test
2) Relationships among test tasks
3) Relationships between test tasks and overall English proficiency

3. Method

3.1 Participants

We received cooperation from 41 Japanese students. 21 were university students, consisting of 13 sophomores and 8 juniors. The remaining 20 students were college students, consisting of 10 freshmen and 10 sophomores. They were asked to take CASEC test[1] to measure their comprehensive English proficiency level. Their mean score was $M = 555$ out of 1000 points with $SD = 82.0$, which can be equated to somewhere between Grade Pre-2 and Grade 2 in Eiken Test.

3.2 Data collection procedure

We used a set of Grade 2 Eiken Speaking Tests. The tasks in the test were reading aloud (RA, henceforth), a

[1] CASEC Test is the abbreviation of Computer Assessment System for English Communication, and a computer-based adaptive test made of four sections: vocabulary, reading and expression, listening comprehension, and dictation/listening. Each section is worth 250 points, for a total of 1000 point.

question about the passage read in reading aloud, description of three serial pictures, and two open-ended questions asking opinions of the participant. RA was conducted at the beginning of the test; the participants were provided with a passage of about 60 words, and given 20 seconds to read silently to comprehend the meaning of it. They then read it aloud and the examiner evaluated their performance on a five-point scale. After RA, one question, Question 1, was given to measure their comprehension. Often the question focuses on pronouns in the passage and participants are required to identify the referent located somewhere in the passage.

Picture description task, Question 2, followed after Q1. The participants were asked to plan their story in 20 seconds and describe the pictures. They were supposed to describe five important points in three pictures with proper vocabulary use, grammatical accuracy and proper usage of English in order to get a full score. The examiner evaluated the content and linguistic quality individually in a five-point scale.

Two open-ended questions followed the picture description task. The first one, Question 3, was a question ending with "What do you think about it?". That required the participant to express his/her opinion and their reasoning. The second one, Question 4, asked if the participant agrees or disagrees with a certain opinion. The topics of questions were different.

The participants took the speaking test individually in a face-to-face situation and the voice during the test was fully recorded. The data collection period was within a month of taking the CASEC Test.

4. Results

4.1 Evaluation of performance in the speaking test

Three examiners individually evaluated student performance using a five-point scale for each question during the interview. After the interview test, the three examiners listened to the recorded voices of participants whom they did not interview directly, adding their own evaluation. Later, the examiners had meetings to listen again to the voices of some outlying scores and adjusted each score depending on the evaluation policy. The rating criteria for each question were the ones made for the Eiken Speaking Test. Concerning Q2: the picture description task, proper use of vocabulary, grammar, and usage (VGU, henceforth) was evaluated in addition to the number of points described.

The score averages for the questions are indicated in Table 1. The inter-rater reliability coefficients among the three examiners for the questions were between .72 and .84, which meant the ratings were within a reliable band. As for the total score average (SUM), it was 17.15 out of 30, which was 57% score proportion. Therefore, compared to the case of real Eiken Speaking Test, the participants were on average somewhere around the borderline for passing Grade 2.

The highest score was 3.44 in Q2, the picture description task. The result indicated that many of the participants could describe three or more points from the pictures. On the contrary, the average VGU score in Q2 was 2.92 and it was not as high as Q2, which showed the difference between what was described and how the description was made.

The next highest score was observed in RA (M = 3.24), which indicated there were some participants who could read aloud the passage with fairly accurate pronunciation, proper intonation and pauses between sense groups. However, according to the result of Q1 (M = 2.1) the comprehension score was not high.

Table1

Descriptive statistics and inter-rater reliability coefficients for each question (N = 41)

		Examiner 1	Examiner 2	Examiner 3	Average among Examiners	Inter-rater reliability α
RA	*M*	3.37	3.12	3.22	3.24	.84
	SD	.70	.56	.88	.63	
Q1	*M*	2.17	1.78	2.34	2.10	.80
	SD	.74	.57	.76	.59	
Q2	*M*	3.56	2.80	3.95	3.44	.84
	SD	.90	.84	.97	.79	
VGU	*M*	3.05	2.39	3.32	2.92	.72
	SD	.71	.59	.88	.59	
Q3	*M*	2.71	2.22	3.07	2.67	.83
	SD	1.21	.82	1.21	.95	
Q4	*M*	2.88	2.29	3.20	2.79	.78
	SD	1.19	.72	1.14	.87	
SUM	*M*	17.73	14.61	19.10	17.15	.88
	SD	3.83	2.20	4.06	3.10	

Regarding the results of Q3 and Q4, which displayed how well the participants could express their own ideas, both showed almost the same score average, namely 2.67 in Q3 and 2.79 in Q4. Compared with RA and Q2, they were lower. These results may have come from the question type which requires more cognitive load due to the impromptu situation.

4.2 Relationship between the total score and the questions

The relationships between the total score average (SUM) and the questions as subcategories are shown in Table 2. Spearman's rank correlation coefficient (ρ) was adopted because some results did not have a normal distribution. As indicated in the table, all the questions are naturally related to their sum. All of them had significant correlations, and except for Q1 and Q2, all of the rest showed strong relationships with more than .70.

Table 2

Correlation coefficients between the total score average (SUM) and the questions (Spearman's ρ)

	RA	Q1	Q2	VGU	Q3	Q4
SUM	.766**	.477**	.530**	.803**	.719**	.775**

** $p < .01$

In order to find the contribution rato of each question to the total score, we conducted a regression analysis with all possible models. Although it was obvious that the total of the contribution ratio would be 100% because each question converge on the total score, it seemed worth investigating which questions contributed more or less to the total.

The results of the analysis are shown in Table 3 and visually displayed in Figure 1. Q3 (23%) and Q4 (22%) were higher than other questions. RA and VGU contributed 15%, followed by Q2 (13%) and Q1 (12%). From the viewpoint of question types, the sum of the two open-ended questions, Q3 and Q4, contributed the most, 45 %, to the total.

Table 3

Results of regression analysis of the speaking test

	Correlation coefficients with SUM	Unstandardized coefficients	Standardized coefficients	Contribution ratio
RA	.723	.984	.201	15%
Q1	.621	1.000	.192	12%
Q2	.518	1.007	.256	13%
VGU	.795	.988	.186	15%
Q3	.732	1.033	.314	22%
Q4	.833	.975	.270	23%
Total contribution ratio	1			100%

Figure 1. Contribution ratio in the interview test

4.3 Relationships among test tasks

Relationships among the questions in the speaking test are shown in Table 4 with Spearman's correlational coefficients, ρ.

First of all, the most prominent result was observed in VGU, which had a medium level of significant correlation coefficients with all questions. It was naturally and inevitably most strongly related with Q2 because this item was evaluated in parallel with Q2. However, this result suggested VGU is one of the fundamental elements of speaking ability.

Next, reading aloud (RA) also showed a medium level of significant correlation coefficients with all the questions except for Q2. The ability of reading aloud could also be one of the elements in speaking ability.

Question 3 and 4 showed strong significant relationship, .753, since both of them were a similar style of question. These open-ended questions also seemed to play a major role in speaking performance.

Table 4

Relationships among the questions in the speaking test (Spearman's ρ)

	RA	Q1	Q2	VGU	Q3	Q4
RA	---	.450**	.213	.558**	.527**	.537**
Q1	.450**	---	.19	.441**	.173	.310*
Q2	.213	.190	---	.680**	.074	.145
VGU	.558**	.441**	.680**	---	.336*	.474**
Q3	.527**	.173	.074	.336*	---	.753**
Q4	.537**	.310*	.145	.474**	.753**	---

* $p < .05$, ** $p < .01$

4.4 Relationship between CASEC and Eiken Speaking Test

Details of the relationship between CASEC scores and the speaking test are indicated in Table 5. Between the total of the speaking test (SUM) and CASEC total score, there was a medium strength of significant correlation coefficient, .648. In the CASEC Total column, it was found that RA, Q4, VGU, and Q3 showed significant relationships beyond .50 level in correlation coefficients.

Table 5

Relationship between CASEC and the speaking test (ρ)

Speaking test	CASEC				
	Total	Vocabulary	Reading/ Expressions	Listening Comprehension	Dictation
SUM	.648**	.467**	.531**	.712**	.489**
RA	.668**	.418**	.501**	.624**	.591**
Q1	.409**	.302	.364*	.370*	.291
Q2	.133	.097	.009	.304	-.026
VGU	.564**	.401**	.428**	.718**	.414**
Q3	.525**	.417**	.475**	.404**	.520**
Q4	.614**	.490**	.585**	.589**	.486**

* $p < .05$, ** $p < .01$

In the sum row of the speaking test, it can be seen that all the sections were significantly related to CASEC total and its components. The listening comprehension section showed the strongest relationship, .712. In other rows except for Q2 and some parts in Q1, significant positive relationships were observed. Question 2 in the speaking test seemed to be a unique item since it did not show any strong relationship to any items in CASEC.

To sum up the results so far, the followings are found:

1) Except Q1 and Q2, other tasks showed significant relationship with the total of speaking test.
2) Two open-ended questions, Q3 and Q4, contributed most to the total of speaking test.
3) VGU was related to all the test tasks of the speaking test
4) RA showed a correlation with VGU, Q3, and &Q4
5) Q2 showed no significant relationship with other tasks except VGU
6) RA, VGU, Q3 and Q4 also showed strong relationship with CASEC total score.
7) The total score of speaking test correlated most with listening comprehension in CASEC test.

5. Discussion

5.1 Relationships between test tasks and total score of speaking test

Regarding the relationship between test tasks and total score of speaking test, two things were observed.

Firstly, RA, VGU, Q3 and Q4 showed significant relationship with the total score of the speaking test, though Q1 and Q2 indicated moderate correlation (Q1: $r = .444$, $p < .01$, Q2: $r = .530$, $p < .01$). Among the test tasks, VGU showed the strongest correlation with the total score ($r = .803$, $p < .01$). It can be said that vocabulary knowledge and grammar use are two of the main aspects of speaking ability as Bachman and Palmer (2010) mentioned grammatical knowledge can be described as the main resource of formulating utterances, therefore it is understandable that VGU is closely related to the total score of the speaking test.

Secondly, the total score also have a strong correlation with question-and-answer sections (Q3 and Q4), and these sections explain 46% of the test. It is quite obvious because speaking activities in the real world usually include some kind of interaction between people. One concern about open-ended questions like Q3 and Q4 is the topic of the question. Especially in Eiken Speaking Test, although it is limited to this study, whether the participant can perform well or not depend on the topic. If the participant is familiar with the question topic, it is easier for them to answer it. However, in the opposite case, the participant could not perform well even if she or he possessed the speaking ability. It is known that topic familiarity necessary to answer the question can impact participant's performance, especially fluency of the response (Taylor, 2011, Foster & Skehan, 1999). This suggests that test developers should carefully chose the topic in order to assess participants' ability accurately. As for participants, it is clear that they need to develop a wide range of knowledge about the topic and also an ability to describe their own idea on the topic.

5.2 Relationships among test tasks

Firstly, VGU is related to all the test tasks of the speaking test. It was indicated that vocabulary and grammar usage are foundations of speaking performance as mentioned in 5.1. VGU showed the strongest correlation coefficient with Q2 since this item was evaluated in parallel with Q2.

Secondly, RA had the second strongest correlation with VGU. The participants who can perform well in the reading aloud task have a tendency to possess a rich vocabulary and grammatical knowledge. This could be one

of the explanations of the relationship of VGU and RA. RA also correlates with Q3 and Q4. This implies that the relationship between reading aloud ability and speaking skills.

Thirdly, Q2 showed no significant relationship with other tasks except for VGU. It can be said that the picture description task is a unique task in Eiken Speaking Test. When Q2, a picture description task, and Q3 and Q4, open-ended tasks, are compared, the difference can be found in the conceptualization of the message. As Levelt's language production models (1989) stated, we plan the message content first and formulate it into the utterance. The content of Q2 is pretty much controlled since the story is already set. As a result, the conceptualization is very restricted compared to other tasks. Participants select the words and phrases for description of the pictures but it is limited to the ones which are related to the story. On the other hand, in Q3 and Q4, participants have a less controlled situation so that they can conceptualize their idea more freely based on their knowledge, including their linguistic knowledge. Q3 and Q4 are the questions which ask for the participants' own idea, so participants can plan the message freely without restriction.

5.3 Relationships between test tasks and overall English proficiency

When we observe the relationship between CASEC total score, an index of the participant's English proficiency, and the test tasks of the speaking test, relatively strong correlation coefficients were found with RA, VGU, Q3 and Q4. Among them, RA showed the strongest relationship with CASEC total score (r = .668, p < .01), and this result confirmed the finding from Miyasako (2002) and Iino et. al. (2010). RA also showed the second strongest relationship with the total score of the speaking test (r = .766, p < .01), therefore, it can be said that RA skills has something to do with English speaking ability.

The total score of the speaking test was most correlated with listening comprehension in CASEC test (r = .712, p < .01). This is understandable because the speaking test also required a participant to listen to the examiner's questions during the test. In the dialogue tasks like Q1, Q3 and Q4, participants have to carefully listen to the questions to produce the answer. Hence, the dialogue task does not purely assess the speaking skills since the participants need to use their listening skill to complete the task. However, it is obvious that listening skills plays an important role in the speaking test.

6. Conclusion

This study investigated Eiken Speaking Test in three areas: 1) relationships between the test tasks and the total score of the speaking test, 2) relationships among the test tasks, 3) relationships between the test tasks and overall English proficiency.

Regarding the relationship between the test tasks and the total score of the speaking test, the significant relationship was found between test tasks and the total score of the speaking test, except for Q1 and Q2. In addition, two open-ended questions, Q3 and Q4, contributed most to the total score of the speaking test. It is important to express one's own idea in a speaking test. As for the relationships among the test tasks, we found the followings: 1) VGU is related to the rest of test tasks of the speaking test, 2) RA is fairly correlated with VGU, Q3, and Q4, 3) Q2 showed no significant relationships with other tasks, except for VGU. From these results, it can be said that vocabulary and grammar usage are the foundation of language production, and picture

description task is a unique task in Eiken Speaking Test.

Two things were found from the analysis of the relationship between the test tasks and overall English proficiency. All test tasks, except for Q2, showed strong relationship with CASEC total score, and the total score of the speaking test was most correlated with listening comprehension in CASEC test. These results imply the relationship between Eiken Speaking Test and CASEC, and show the importance of listening skills in the speaking test.

From these findings we constructed the hypothetical model or formula to explain speaking ability as follows: Speaking ability = listening comprehension + linguistic knowledge + reading aloud + idea expression.

The listening comprehension factor is drawn from the result of research question 3: the relationship between CASEC listening section and the speaking test. Linguistic knowledge, reading aloud ability and idea expression are all from research question 1 and 2, summarized in the contribution ratio through regression analysis.

It is certain that further research would be needed in the following three areas. First, participants' attitude towards the test should be analyzed to understand the results from this study deeply. A survey was conducted right after the speaking test so participants could self-evaluate their performance but it was not mentioned in this study. Combining the survey results and this study would be the next step. Second, we would like to investigate the participant's utterance more closely by transcribing the performance. This investigation will show us the tendency of their errors and speaking strategies. This leads to the third area where we could apply the results to speaking training in our everyday classroom. If we use the findings from the assessment efficiently, it would clarify effective teaching methods for English education.

Acknowledgements

This research was supported by the Society for Testing English Proficiency, Inc.

References

馬場 哲生 編. (1997).『英語スピーキング論―話す力の育成と評価を科学する』(英語教育研究リサーチ・デザイン・シリーズ). 東京: 河源社. (Baba. S (Ed.) (1997). *Eigo speaking ron*. Tokyo: Kagensya.)

Bachman, L. and Palmer, A. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Brown. H. D., & Abeywickrama, P. (2010). *Language Assessment – Principles and Classroom Practices*. New York: Pearson Education.

Foster, P. & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance, *Language Teaching Research, 3 (3)*. 215-247.

Iino, A. Yabuta, Y. & Thomas, J. (2010). Relationship between criteria for reading aloud evaluation and English proficiency. *Journal of Chubu English Language Education Society,* 40. 159-166.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, Mass.: MIT Press.

Miyasako, Y. (2002). Are there any relationships between reading aloud ability and English proficiency in high school students? *STEP Bulletin, 14*. 14-25.

Nakamura, Y. (2004). Dialogue interview test and multilogue discussion test. *International Christian University publications. I-A, Educational Studies, 46.* 197-209.

Taylor. L (Ed.) (2011). *Examining Speaking*. Cambridge: Cambridge University Press.

Uenishi, K. (2004). An empirical study of Japanese high school students' English speaking proficiency: Introducing two kinds of speaking tests. *Bulletin of the Graduate School of Education, Hiroshima University. Part. II, Arts and Science Education, Vol.52*. 121-126.

Weir, C. J. (1990). *Communicative Language Testing.* London: Prentice Hall.

Yamakawa, K. (2003). The effects of repetition and planning on EFL learner's oral performance. *Language Education & Technology, 40*. 175-189.

SUMMARY

This study investigated Eiken Speaking Test and examined the following three relationships: 1) Relationships between test tasks and total score of speaking test, 2) Relationships among test tasks, and 3) Relationships between test tasks and overall English proficiency. Regarding the relationship between the test tasks and the total score of the speaking test, we found that reading aloud and open-ended questions are related to the total score of the speaking test. As for the relationships among the test tasks, it could be said that the picture description task is a unique one in the Eiken Speaking Test. On the relationship between the test tasks and overall English proficiency, the results indicate a relationship between Eiken Speaking Test and CASEC test, and importance of listening skills in the speaking test.

Based on these findings, we constructed the hypothetical model to show explain speaking ability as follows: speaking ability = listening comprehension + linguistic knowledge + reading aloud + idea expression.