Shiken: JALT Testing & Evaluation SIG Newsletter, 7 (2). June 2003. (p. 2 - 11)

Assessing speaking in Japanese junior high schools: Issues for the senior high school entrance examinations

by Tomoyasu Akiyama

(Dept. of Linguistics & Applied Linguistics, The University of Melbourne)

This paper has three purposes. First, it discusses three assessment contexts in relation to the notion of "usefulness" by Bachman and Palmer (1996). Those contexts are (1) the 2001 Tokyo senior high school entrance examination, (2) a proposal to include of speaking tests in that examination, and (3) a proposal to assess speaking skills in Tokyo junior high schools. This work also identifies some concerns by Japanese junior high school EFL teachers and students through various statistical procedures. Finally, it argues for the need to build up a "task bank," as suggested by Brindley (2001), for the speaking components used in senior high school entrance examinations.

> *"any high school entrance examination that does not include the assessment of speaking skills could be said to lack construct validity in light of the Ministry of Education, Culture, Sports, Science and Technology's 1998 revised guidelines"*

Evaluation of Usefulness of 3 Assessment Contexts

Let us begin by consider three assessment contexts.

1) The 2001 Tokyo Metropolitan Senior High School Entrance Examination

The 2001 Tokyo Metropolitan Senior High School Entrance Examination [*Toukyou-tou Koutou Gakkou Nyuugaku Shiken*] focused on reading skills and grammar knowledge and nearly 80% of the test had a multiple-choice format. Figure 1 indicates the way that the four skills are covered in this test A point of concern is that any high school entrance examination that does not include the assessment of speaking skills could be said to lack construct validity in light of the Ministry of Education's 1998 revised guidelines. The current entrance examination also appears to lack authenticity, since recent high school English curriculum guidelines by the Ministry of Education seek to develop speaking and writing skills as well as reading and grammar.

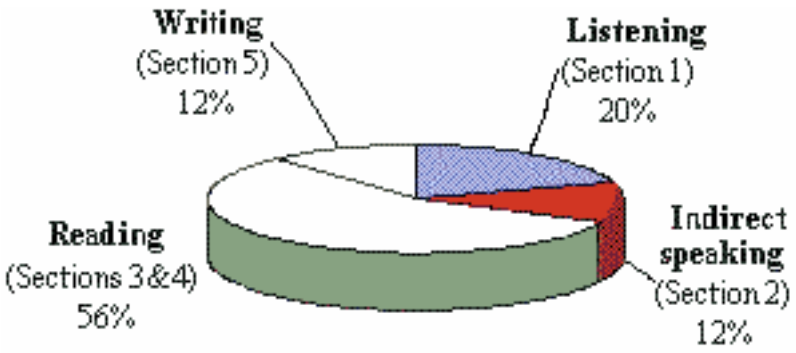

Figure 1. The proportion of skills tested in the 2001 Tokyo Senior High School English Entrance Examination

For the same reason, the current English test (which does not assess speaking skills) could be said to lack authenticity." Indirect" speaking tests are low on interactiveness because examinees are only required to select the English sentence which fits a given scenario most appropriately.

This paper reports how the inclusion of the speaking tests in the entrance examination may have some positive influence in junior high schools according to a survey of junior high school teachers. In terms of practicality, the current English examination test rates well. The English section of the 2001 Tokyo Metropolitan Senior High School Entrance Examination also rates well in well in terms of reliability and practicality. Its main problems involve construct validity, impact, authenticity and lack of interactiveness.

2) What impact would the introduction of speaking tests in entrance examinations have on teaching?

If speaking tests became a component of high school entrance examinations in Tokyo what would happen? Such a move might result in less reliability. The reason is that speaking tests have inherently many variables, such as rater behavior and interlocutors' variations (McNamara, 1996). The inclusion of speaking tests would represent a positive increase in authenticity, however, because the test would better reflect the curriculum content. Moreover, including speaking tests could engage students to complete tasks interactively, and such tests would be more interactive than the current examination. Introducing speaking tests in the entrance examination would also have great impact on teachers and students, as several studies (e.g. Shohamy, Donitsa-Schmidt, and Ferman, 1996; Cheng, 1997) suggest. As speaking tests require many resources such as administrators and raters, the inclusion of speaking tests might present problems in terms of practicality.

3) Assessment of speaking skills in junior high schools

How should speaking skills be assessed in Japanese junior high schools? Studies by Brindley (1999) point out how the reliability of school-based assessments tends to be low. The construct validity could potentially be high, as Hamp-Lyons (1996) claims. Hamp-Lyons (1996) argues that portfolio assessment is much more valid than traditional tests. The reason that authenticity and interactiveness could be high is because school-based assessment provides ample opportunity to conduct speaking tests. However, these judgments need to be made with caution because they also involve issues about preferred teaching styles. Since entrance exams significantly determine how and what many junior high school students study, the impact of in-school speaking assessments would probably be lower than having speaking tests in the current junior high school entrance

examinations. Practicality may also be a problem, because the revised curriculum has decreased English instruction time from 4 to 3 hours per week.

While discussing these three assessment contexts in detail, many issues need to be considered to maximize the usefulness of any proposed speaking tests.

Research questions

Based on discussions for the three assessment contexts above, five questions are addressed in this paper. The first two involve a standard survey analysis and the remaining three questions involve Rasch analyses.

1. How do public junior high school teachers in Tokyo assess their students' speaking skills?
2. What impact would the introduction of speaking tests in entrance examinations have on teaching?
3. To what extent do tasks (speech, role-play, description and interview) differ in terms of perceived difficulty?
4. To what extent do the previous items fit Rasch measurement?
5. To what extent do students' performances as measured by four tasks fit Rasch measurement?

Methodology

Instrument 1

Please refer to Appendix 1 for an abridged copy of the questionnaire survey. This survey was designed to address research questions 1 and 2. Approximately 600 questionnaires were distributed to the public junior high school English teachers in Tokyo. The questionnaire was completed by 199 junior high school teachers (a response rate of 33%).

Instrument 2

Four of the five the most popular tasks according to the survey in Appendix 1 were used for a test trial (speech, role-play, description, and oral interview). Information gap tasks were not used because of difficulty in administration. All tasks had a duration of 5 minutes, including explanations of the test procedures.

Test-takers and interlocutors

The test-takers were all Japanese junior high school students and they ranged in age from 14 (second year students) to 15 (third year students) years. 219 students at twelve schools participated in the test trial. All students at each school undertook two of the four tasks (in total 438 students' performances).

The 13 interlocutors (12 Japanese English teachers at participants' school and the researcher) administered different tasks to the students.

Raters and scoring criteria

Five independent Japanese English senior high school teachers, with more than 10 years' teaching experience, rated students' performances from tape recordings. Scoring criteria consisted of 5 items (fluency, vocabulary, grammar, intelligibility and overall task fulfillment). The items were rated on a 0 to 5 points scale according to different levels of performance described for each item.

Results

Questionnaire survey

Research Question 1 ascertained how English teachers assessed students' speaking ability using direct speaking tests. Those who said they conducted direct speaking tests amounted to 57.3% of the same (n = 114). 42.7% (n = 85) of teachers said they did not administer speaking tests. However, further analysis shows that the combination of other assessment methods, such as class observation and pencil-and-paper tests were frequently used. Results revealed that the majority of English teachers assessed students' speaking skills based on classroom observation with a combination of pencil-and-paper tests and speaking tests.

Research Question 2investigated what impact the introduction of speaking tests would have on Japanese English teachers. Figure 2 indicates that more than 75% of the teachers reported that speaking tests would impact them, while 20% stated that little impact or no impact would occur to their teaching. Responses to this question showed that the introduction of speaking tests in entrance examinations would have a positive impact on teachers and their teaching activities, in that the majority of teachers would change their teaching styles towards improvement of students' communicative skills.

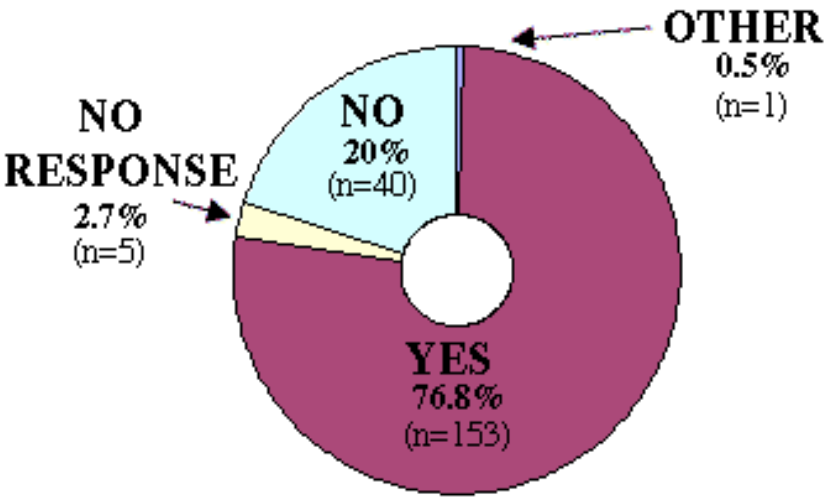

Figure 2. Teacher responses to Survey Question #9. (N=199)

Rasch analysis of the student test scores

Difficulty of items and tasks

Research question 3 investigated the difficulty of tasks (items) on each task. As indicated in the fifth column in Table 1, the description task is the most difficult and the interview task the easiest. The difference between the most difficult and the easiest tasks is approximately 1.5 logits.

Table 1. *A Rasch analysis of some English classroom activities considered together*

No	Item name	Difficulty	Error	Task difficulty	IMS	Infit t
1	Speech /Fluency	-0.24	0.13		1.17	1.3
2	S /Vocabulary	0.06	0.13		1.01	0.2
3	S / Grammar	0.1	0.14		1.09	0.7
4	S / Intelligibility	**1.91**	0.16		1.04	0.4
5	S /Task fulfilment	-0.12	0.13	0.342	0.78	-1.9
6	Role-play /F	-0.35	0.18		0.92	-0.5
7	R /V	-0.11	0.19		1.08	0.6
8	R / G	0.19	0.18		1.09	0.6
9	R / I	0.59	0.21		0.88	-0.8
10	R /TF	-0.84	0.17	-0.104	1.14	0.9
11	Description /F	-0.30	0.15		0.93	-0.5
12	D /V	0.78	0.17		0.80	-1.5
13	D / G	0.99	0.17		1.12	0.9
14	D / I	**1.70**	0.19		0.99	0.0
15	D /TF	0.05	0.16	**0.644**	**1.38**	**2.5**
16	Interview/ F	**-1.34**	0.17		0.89	-0.8
17	I / V	-1.01	0.15		1.11	0.9
18	I / G	-0.94	0.17		0.87	-1.1
19	I / I	0.38	0.19		0.82	-1.4
20	I / TF	**-1.52**	0.15	**-0.886**	0.99	0.0
Mean		0.00	0.16		1.00	0.0
S.D.		0.91	0.02		0.15	1.1

F= Fluency, V= Vocab, G= Gram , I= Intelligibility, TF= Task Fulfillment

Research question 4 examined the quality of items, and the extent to which data patterns derived from the Rasch model differ from those of the actual data. Unexpected items that the Rasch model identifies are called either "misfit" or "overfit" items. The acceptable range of IMS here is from 0.70 to 1.30. As can be seen, only Item 15 is identified as a "misfit," indicating a larger than the acceptable range of IMS in the sixth and seventh columns. This shows that the actual data patterns from item 15 (Description: task fulfillment) varied unacceptably in comparison with data patterns estimated by Rasch measurement. Thus the items on four tasks appeared to produce relatively similar response patterns, suggesting that the items across tasks assessed the similar construct.

Person fit indexes

Table 2. *Overall misfit statistics for the student sample* (N=219)

Infit Mean square (IMS)	S.D.	The acceptable range Mean ± 2 S.D.	**Number of misfit** Students (%)
0.99	0.58	- 0.17 to 2.16	12 (5.4 %)

The last question focuses on students' scores across the four tasks. This is particularly important, since this question leads to issues of accountability for students. As can be seen in Table 2, 5.4% of the students were identified as misfit students. This indicates that the percentage of misfit students exceeds the limit of the acceptable percentages of misfit students. It is important to investigate why this happened.

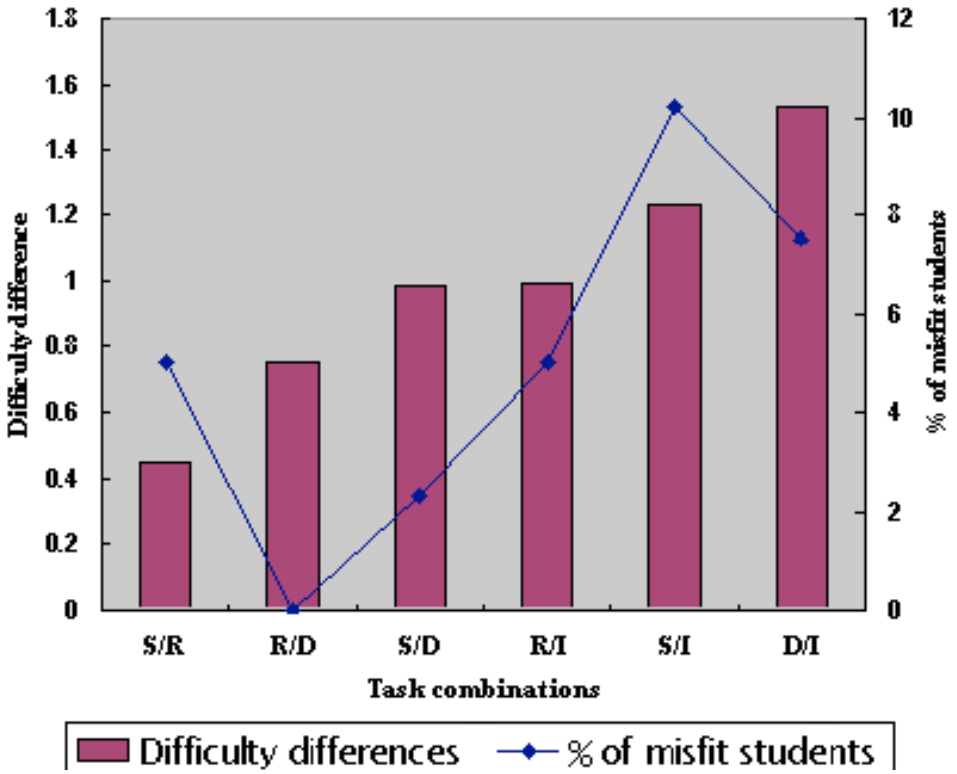

Figure 3 shows which combination of tasks produced misfit students. The combinations of tasks which produced misfit students the most frequently were speech and interview followed by the combination of description and interview. Other task combinations produced fewer misfit students than the above two combinations. One possible explanation for this is that differences of task difficulty in combinations might have an impact on increasing misfit students.

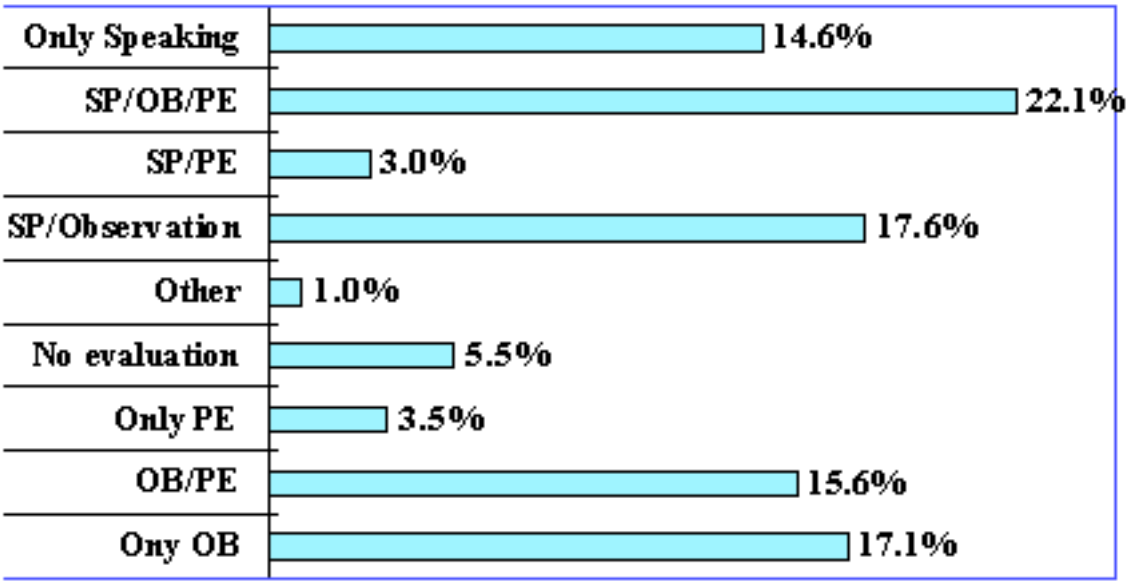

Figure 4. Ways that the teacher sample reported assessing their students' English skills (N=199)

If we look at Figure 4, we can see how speaking skills are assessed in considerably different ways by high school teachers in Japan. Over 22% of the nearly two hundred teachers responding to this survey indicated that they relied of a combination of speech analysis (SP), class observation (OB) and pencil-and-paper tests (PE) to assess speaking skills. However, it is worth noting that over 17% of the teachers relied solely on classroom observations to assess speaking skills.

Discussion

> *"the inclusion of the speaking tests has the potential to assist in bridging the gap between skills taught in classes and skills tested in entrance examinations, and between goals of the guidelines and assessment policy."*

Results of the questionnaire survey revealed that teachers' assessment methods varied, suggesting that it would be difficult to compare students' speaking ability across schools. The introduction of speaking tests would have a positive impact on approximately 80% of public English junior high school teachers in Tokyo, and most teachers maintained that they would change to a more communicative style of teaching. Thus, it can be

argued that the inclusion of the speaking tests would have the potential to assist in bridging a gap between skills taught in classes and skills tested in entrance examinations, and between goals of the guidelines and assessment policy.

Results from test trials undertaken by junior high school students showed that all items except one fit Rasch measurement, indicating that items on each task were effective in assessing the target construct. However, results also showed that the four tasks frequently used by English teachers were different in terms of difficulty. This means that students who undertake a variety of difficulties of tasks might not be assessed appropriately. Given that variables, including rater behavior and interlocutors, are inherent in performance tests, difficulty of tasks needs to be relatively equal in order to reduce variables. The concept of task banks, presented by Brindley (2001), and item banks by Ikeda (2000) could have important implications for the introduction of formal speaking tests in entrance examinations:

Conclusion

Implications for this study are that speaking tasks used in a classroom need to be trialed, and also investigated with Rasch measurement, given that school-based assessment represents half of the selection procedures for students who wish to enter senior high schools. In junior high school contexts, a role play task bank, such as shopping situation, inviting friends to a party, or giving directions to a stranger could be developed. In order to not only administer speaking tests in a high stakes context, but also to enable teacher implemented assessment to be comparable across schools, it would be necessary to investigate tasks with Rasch techniques, based on empirical data, and to build up a task bank with a relatively consistent quality of tasks.

References

Akiyama, T. (2001). The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context. Melbourne Papers in Language Testing. 10 (1), 1 - 22.

Alderson, J. C., and Wall, D. (1993). Does washback exist? Applied Linguistics, 14, 115 - 129.

Bachman, L. F. (1990). Fundamental consideration language testing. Oxford University Press

Bachman, L. F., and Palmer, A. S. (1996). Language testing in practice. Oxford University Press.

Brindley, G. (2001). Outcome-based assessment in practice: some examples and emerging insights. Language Testing. 18, (4) 393-407.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11* (1), 38-54.

Ikeda, H. (1999). What we need for research on language testing in Japan – A psychometrician's view. *21st Century Language Testing Research Colloquium*. Plenary speech made at LTRC 99 Tsukuba in Japan.

Japanese Ministry of Education, Science and Culture. (1998). *Chugakuko Shidosho: Gaikokugo-Hen*. [Guidelines for Junior High Schools: Foreign Language Study Revisions]. Tokyo: Kairyudo.

McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Addison-Wesley Longman.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13* (3), 239 - 256.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: washback effect over time. *Language Testing, 13* (3), 298-317.

Appendix 1: Original Survey (Abridged)

Q1: 授業でどのようなタスク（オーラルコミュニケーション活動）を主に使っていますか。下記の選択群からもっとも使われるタスクから順番に２つ選んでください。選択群にない場合、その他のところに具体的にお書きください。

Q2: どのようにスピーキング能力を評価しますか。主なものひとつ、下の中から選んでください。

Q9: 仮にスピーキングテストが高校入試の英語の一部に導入された場合、先生ご自身の授業の仕方に影響があると思われますか。(はい、いいえ）。またその理由をお書きください。

質問２で、２と答えられた方は質問８にお答えください

Appendix 2: English Translation of Original Survey (Abridged)

A questionnaire survey of Japanese junior high school English teachers in Tokyo

The purpose of this questionnaire is to investigate speaking tasks which you conduct in assessing your students' speaking ability in the classroom.
Please answer questions below: Your cooperation will be highly appreciated.

Question 1. *What kind of tasks are used to facilitate oral communicative activities in your classes? Choose the two tasks from the most to the second tasks used below.*

Task numbers: the most often used task ( )

Choices of tasks:

(1) Oral interview (2) Information gap (3) Show and tell (4) Skit (5) Other

Question 2. *How do you evaluate your students' speaking ability? (Please choose the major one)*

Your answer Number ( )

If your answer is 2, please go to question 8

(1) speaking tests (2) speaking ability is not evaluated at all (3) classroom observation

(4) paper and pencil tests (5) the system entrance examinations

(6) other: _____

(Your reasons)

Question 9. *If speaking tests are introduced into entrance examinations, would the test affect you or your teaching? (Please give brief reasons for your answer).*

Your answer is (Yes / No)

(Your reasons)

Note that Questions 3, 4, 5, 6, and 7 were omitted due to space limitations.

HTML: www.jalt.org/test/aki_1.htm / PDF: www.jalt.org/test/Akiyama1.pdf