Articles

Assessing Speaking: Issues in School-Based Assessment and the Introduction of Speaking Tests into the Japanese Senior High School Entrance Examination[1]

Tomoyasu Akiyama
University of Melbourne

This paper[2] discusses ways of bridging a gap between teaching and assessment practice, focusing on the assessment of speaking skills in Japanese junior high school contexts. Through discussion of the assessment of speaking skills and based on a questionnaire survey, this paper identifies issues pertaining to the assessment methods of speaking skills employed by junior high school teachers. Based on the results of the survey, and on the concept of a task bank proposed by Brindley (2001), trial speaking tests were developed and piloted with 219 junior high school students. Results were analysed using Rasch techniques, and indicated that, although items across four speaking tasks fit Rasch measurement, differences of task difficulty between combinations of tasks might have an impact on student performance. The paper argues for the need to build up the task bank with relatively consistent tasks and discusses issues of the introduction of a formal speaking test in the senior high school entrance examination.

本研究は理論的枠組みをusefulness (Bachman & Palmer, 1996) をよりどころとし日本の中学英語教育の授業内容、教師による評価、入試問題の連携の欠如を「話す能力」に焦点を絞り考察した。その考察から高校入試にスピーキングテストを導入することは理論的に正当性があるということが判明した。また中学の英語教師（199名）へのアンケートの結果より中学教師の話す能力を評価する問題点、及び、高校入試にスピーキングテストを導入する必要性を論じた。また中学生（219名）に実施されたスピーキングテストのデータをラッシュ手法で分析した結果により、スピーキングテストを高校入試に導入する場合にはBrindley (2001)の提案した‘task bank’の概念が必要であることを論じた。最後に授業内容、評価、入試問題を意味のある連携にするためにはどのようにすればよいかを提案した。

Decisions regarding admission to Japanese senior high schools are usually made based on both school-based assessments implemented by junior high school teachers and test scores of the senior high school's particular entrance examination. In general, the weight given to test scores in proportion to school-based assessment ranges between 50/50 and 60/40. English is one of the core subjects for both assessments.

The Course of Study Guidelines (hereafter, the guidelines) for teaching English to junior high school students published by the Japanese Ministry of Education, Culture, Science and Technology (hereinafter the Ministry of Education) (1999) state that speaking is one of the most important skills junior high school students need to develop.

In the last two decades, the Ministry of Education has employed many Assistant Language Teachers (ALTs), native speakers of English, to assist junior high school students and Japanese English Teachers (JETs) in the improvement of their communicative skills. Despite the emphasis on the development of speaking skills evident in the guidelines and in the introduction of ALTs, few senior high school entrance examinations have included a means to assess speaking skills. Thus, there is a large discrepancy between the aims of the guidelines and the skills tested in senior high school entrance examinations.

This paper has three purposes. First, it discusses three assessment contexts (a) the 2001 English test in Tokyo senior high school entrance examination, (b) the inclusion of speaking tests in the senior high school entrance examination, and (c) the assessment of speaking skills in junior high schools in relation to the notion of "usefulness" (Bachman & Palmer, 1996). Second, it identifies the issues relevant to school-based assessment by junior high school English teachers in Tokyo based on a questionnaire survey while also reporting the results of a Rasch analysis of empirical data derived from test trials undertaken by junior high school students. Finally, in discussing the results of the questionnaire survey and the Rasch analysis, this paper argues for the need to build a "task bank," as suggested by Brindley (2001), to support the introduction of speaking tests in senior high school entrance examinations.

Evaluations of usefulness of three assessment contexts

Context A: The 2001 Tokyo Metropolitan Senior High School Entrance Examination

The notion of "usefulness" established by Bachman and Palmer (1996) provides a comprehensive and practical framework to investigate

test qualities. Usefulness consists of six aspects: reliability, construct validity, authenticity, interactiveness, impact and practicality. One of the principles underlying usefulness is that an evaluation of test quality needs to be made in a specific setting for an applied purpose. In using the notion of usefulness, I evaluated the 2001 English test in a Tokyo senior high school entrance examination (hereinafter "the English test"), the main purpose of which is to select students who wish to enter public senior high schools in Tokyo.

Reliability refers to consistency of test scores. Inconsistent test scores should not be used to make important decisions. Bachman and Palmer (1996) note that test scores tend to be reliable when the construct is defined relatively narrowly and test formats are uniform. As the English test primarily focuses on reading skills and grammatical knowledge and approximately 70 to 80 % of the test is allocated to a multiple-choice format (see Figure 1), the test scores of the English test are likely to be reliable. As the senior high school entrance examination is a high-stakes test, reliability in the entrance examination needs to be set as high as possible, yet not at the expense of construct validity.

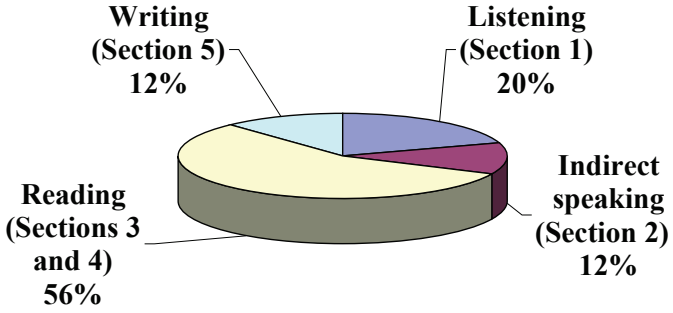


Figure 1: The proportion of skills tested in the Tokyo senior high school entrance examination in 2001

Construct validity refers to meaningfulness and appropriateness of the interpretations of test scores for an applied purpose in an applied setting. Given that the English test assesses a junior high school student's English language ability for the purpose of deciding entry to senior high schools, an entrance examination that does not include the assessment of speaking skills could be said not to have sufficient construct validity. In other words, it can be considered to be what Messick (1996, p. 252) calls "construct under-representation" of the focal construct.

The English test could also be said to lack some authenticity, given that authenticity is defined as the degree of correspondence between the characteristics of test tasks and those of target language use (Bachman & Palmer, 1996). An authentic test ensures that 'nothing important' is omitted from the content of teaching (Messick, 1996, p. 243). This means that issues of authenticity are related to the content of the curriculum because the content of the curriculum draws upon the guidelines set by the Ministry of Education. As the aims of the English curriculum are to develop not only reading skills and knowledge of grammar but also to develop speaking and writing skills, an English test that omits the assessment of speaking skills could be said to lack authenticity.

Interactiveness is defined as the degree of interaction between test-takers and tasks. For example, if test tasks engage test-takers in using a range of strategies and knowledge of language, the tasks can be considered to be highly interactive. In terms of the 2001 English test, the "indirect speaking tests" in section 2 (see Appendix A) are low on interactiveness because students are only required to select that English sentence which captures a given scenario most appropriately.

Impact takes into consideration how test use has an impact on stakeholders such as test takers, teachers, and institutions. Bachman and Palmer (1996, p. 30) provide "micro" and "macro" aspects to be investigated in terms of the impact of tests. At the micro "washback effect" level (Alderson & Wall, 1993), the focus is on individuals such as students and their teachers, whereas at the macro level, the impact of a test on society and educational systems needs to be investigated. At the micro level, the results of a survey questionnaire suggest how the inclusion of speaking tests in the senior entrance examination would have an impact on junior high school teachers.

The final component of usefulness is practicality. Practicality takes into account the availability of time, space, equipment, and administrators, embracing all processes including test development, test administration, and scoring procedure. In terms of practicality, the current English examination test is highly practical.

Bachman and Palmer (1996) suggest that components of usefulness should make a relative evaluation, therefore each component was evaluated as high (3), moderate (2) and low (1). To sum up, the English test apparently has two high marks: reliability and practicality, and has four low marks: construct validity, impact, authenticity and interactiveness (see Figure 2).

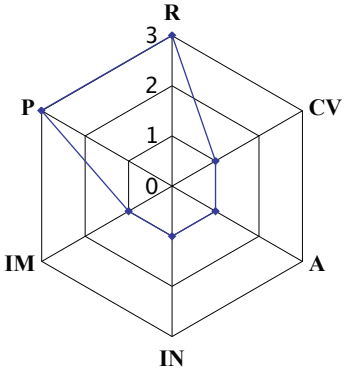


Figure 2: Usefulness of the senior high school entrance examination English

At least two options for assessing speaking skills can be considered under the current educational circumstances in the junior high school context: (1) the inclusion of speaking tests in the entrance examination and (2) assessment of speaking skills in junior high schools. Using the notion of usefulness, I evaluate the two assessment contexts with regards to the 2001 English test.

Context B: The introduction of speaking tests in senior high school entrance examinations

The second assessment context is the proposed introduction of a speaking test in the entrance examination for senior high schools (Figure 3). Although reliability has not yet been investigated, it is expected to achieve less reliability than the present English test. The reason for this is that speaking tests inherently have many variables which reduce reliability, such as rater behaviour and interlocutor variation (McNamara, 1996). However, the question is whether it is possible to maintain a minimal level of reliability in a high stakes test context. If the scores delivered by raters are not reliable, the inclusion of speaking tests is open to question. In terms of authenticity, the inclusion of the speaking tests could be regarded as authentic because the test would reflect the content of the curriculum. As the inclusion of speaking tests could engage students in completing tasks interactively, such tests could be more interactive than the current test. Introducing speaking tests in

the senior high school entrance examination would have great impact on teachers and students, as several other studies (Shohamy, Donitsa-Schmidt, & Ferman, 1996; Cheng, 1997) have attested. On the other hand, as speaking tests require many resources such as administrators and raters, the inclusion of the speaking tests can be low on practicality.

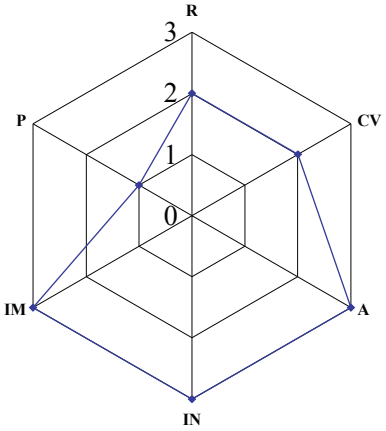


Figure 3: Usefulness of speaking tests included in the entrance examination

Context C: Assessment of speaking skills in junior high schools

The final assessment context is that of junior high school teachers assessing their students' speaking skills (Figure 4). In such a situation, speaking tests need not be administered in the entrance examination. As studies by Brindley (1989) and Rea-Dickins and Gardner (2000) showed, the reliability of teacher-implemented assessment tends to be low. As school-based assessment represents 40 % to 50 % of admission decisions, an important question is whether assessment implemented by teachers could enable senior high school teachers to make comparisons among students from various schools. On the other hand, the construct validity could potentially be high as Moss (1994) and Hamp-Lyons (1996) claim. Hamp-Lyons (1996) argues that portfolio assessment is much more valid than a traditional test, pointing out that portfolios allow teachers to take a closer look at their students' work over time and monitor their progress whereas the tests only cover a snapshot of student ability. However, as McNamara (2001) notes, little research into speaking

versions of portfolio assessment has been reported. Authenticity and interactiveness could be potentially high because school-based assessment could provide ample opportunity to conduct speaking tests. However, these judgements need to be made with caution because they depend upon teachers, teaching styles and assessment criteria. If teachers assess only reading skills and the knowledge of grammar, and so transfer to their evaluation of speaking ability an overemphasis on accuracy, assessments implemented by junior high school teachers may prove less authentic and interactive. Therefore, it would be necessary to investigate exactly how junior high school teachers assess speaking skills. The impact of tests in schools would be lower in comparison with that of tests of speaking in entrance examinations. Practicality would also be low in the school situation because the revised curriculum has decreased English classes hours from 4 to 3 hours per week.

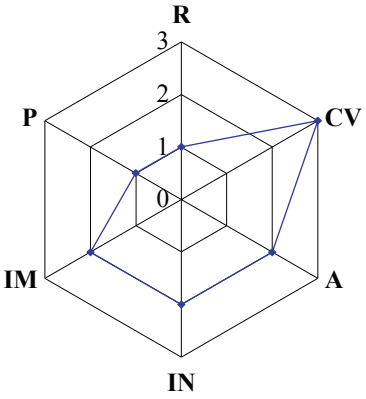


Figure 4: Usefulness of speaking skills assessed in junior high schools

As can be seen in Figures 2, 3 and 4, each assessment context has advantages and disadvantages. For example, the English test in the entrance examination has great advantages of reliability and practicality, but there are disadvantages in terms of construct validity, authenticity, and interactiveness and impact. The assessment of speaking tests in schools has the potential to become highly authentic and interactive. However, given the high stakes there may be reluctance to accept locally administered results as equally valid. On the other hand, the inclusion of speaking tests in senior high school entrance examinations

has the potential of engaging students in interactive speaking tasks and thus impacting on the teacher and students, although reliability and practicality might be problematic.

Through discussion of these three assessment contexts, key questions arise as to which aspects of usefulness should be prioritised and which assessment context could maximize the usefulness of speaking tests. As I propose to show, one way of addressing them is to strengthen the linkage between teaching and assessment practice based on the aims of the guidelines.

Research Questions

Based on the previous discussions of usefulness in the three assessment contexts, five questions are addressed in this paper. The first two questions follow analyses of a questionnaire survey of junior high school teachers in Tokyo. Questions 3, 4 and 5 arise from Rasch analysis.

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 senior high school entrance examinations in Tokyo have on teachers/teaching?
3. To what extent do tasks (speech, role-play, description and interview) differ in terms of perceived difficulty?
4. To what extent do items fit the Rasch model?
5. To what extent do students' performances as measured by the four tasks fit the Rasch model?

The first question focuses on current assessment methods of speaking skills. If such assessment is not sufficient to enable senior high school authorities to make admission decisions, it is important to seek an alternative to school-based assessment in order to assess speaking skills. What then (question 2) would be the impact on teachers/teaching if speaking tests were introduced in entrance examinations? The third question investigates difficulty of speaking tasks. Given that differential difficulty of tasks might have an influence on students' performances, it would be important to investigate task difficulty statistically. The fourth question examines speaking task items, investigating to what extent the items assess the focal construct. The last question investigates to

what extent scores derived from tests can be used to make important decisions. If significant numbers of students are not assessed appropriately, test scores cannot be interpretable. This suggests that tasks need to be revised.

Data Collection Methods

Data collection 1: A questionnaire survey [3]

A questionnaire survey was designed to address research questions 1 and 2. For research question 1, the teachers were presented with a range of assessment options and were asked to choose the two tasks most often used to assess students' speaking ability (see Appendix B). In order to answer question 2, junior high school teachers were required to make dichotomous responses and speculate on what impact the inclusion of speaking tests would have on teachers. Distributed to 600 junior high school English teachers in Tokyo, the questionnaire was completed by 199 (a response rate of 33 %).

Data collection 2: Test trials

Based on results from the questionnaire survey, four of the five most popular tasks with the exception of information gap tasks [4] (speech, role-play, description, and oral interview) were used for a test trial (see Appendix C). All test instructions were given orally in Japanese, and Japanese written cards were provided for the role-play, thus clarifying what students were required to do. Each task had a duration of 5 minutes, including explanations of the test procedures.

The first task was a speech task. After 30 seconds of planning time, each student was to speak on one topic from a choice of five; for example, a) things students want to do in their high school, b) students' best friends, c) students' favourite school events, d) students' club activities, and e) things students did during the winter vacation. The duration of the speech task was 90 seconds, excluding test instructions. After finishing their speeches, the students were each asked two questions based on the content of the speech by the interlocutors (the English teacher and the researcher).

The second task was a role-play. This task required students to buy presents at a shop in Sydney for their family and friends. Students were required to read a task card in Japanese, and were given only 50 Australian dollars. They were also required to ask a cashier (an interlocutor) where

a good restaurant was, after paying for the presents. The main reason this shopping situation was chosen was that a shopping dialogue was included in their texts, so students already had some background knowledge.

The third task was a description task. After 30 seconds of planning time, students were given 90 seconds to describe an illustrated scene in front of a station at 11:30 a.m., people were waiting, smoking, walking with a dog, and buying tickets. A couple was eating lunch in the restaurant near the station. A boy was also waiting for someone. A second illustration showed the young man getting angry and quarrelling with his (girl) friend. The clock at the station showed 1:00 p.m., indicating that he had been waiting for her for a long time. After describing this picture (90 seconds), students were asked a set of three questions about the scenes.

The last task was an oral interview, consisting of a set of four questions, the first asking the student's name. The next three questions were based on the results of the survey conducted by the study group of Tokyo metropolitan junior high schools (*Tokyo-to Chugako Eigo Kyoiku Kenkyukai,* 2000). The survey was conducted by distributing questionnaires to approximately 3,000 junior high school students in Tokyo to find out what topics students in Tokyo were interested in talking about in English. Favourite topics included 1) students' club activities, 2) their daily life 3) their plans during the holidays, and 4) their favourite types of music, singers, sports and athletes.

Research participants

Table 1 summarizes information about the participants, tasks, and raters for the test trial. Because of school events and time constraints, different numbers of students undertook each of the tasks due to school events and time constraints. This occurred because more than the anticipated number of students completed the speech and interview tasks. Due to technical problems with tape recorders, performances of some students were not recorded: 11 were not recorded in each of two speech and role-play tasks, and 3 performances were not recorded in each of two description and interview tasks.

Test-takers

The test-takers were 219 Japanese second year (age 14) and third year (age 15) junior high school students at 12 schools in Tokyo. All students at each school undertook two of the four tasks, totalling 438 student performances.

Table 1: The research participants: test-takers, tasks and raters

School ID (Year) (n)	Speech	Role-play	Description	Interview	Rater (ID)
2 (3rd) (20)	✓			✓	1, 2, 5
3 (2nd) (7)		✓	✓		1, 3, 4
4 (2nd) (20)			✓	✓	1, 2, 4
5 (3rd) (20)			✓	✓	1, 4, 5
6 (2nd) (20)		✓		✓	1, 4, 5
7 (3rd) (20)	✓	✓			1, 2, 4
8 (2nd) (22)	✓		✓		1, 3, 5
9 (3rd) (20)		✓		✓	1, 2, 3
10 (2nd) (20)	✓		✓		1, 2, 4
11 (3rd) (17)		✓	✓		1, 2, 5
12 (3rd) (19)	✓			✓	1, 4, 5
13 (2nd) (14)	✓	✓			1, 2, 5
Total (219)	115	98	106	119	

Interlocutors

Thirteen interlocutors (12 Japanese teachers of English at the participants' school and the researcher) administered different tasks to the students. In general, in order to minimize differences between interlocutor effects, the English teachers had undertaken interlocutor training with the researcher and the role-play task, which required more interactions with students was conducted by only the researcher. However, owing to time constraints and for practical reasons, the researcher also took part in other tasks.

Raters and scoring criteria

Five independent Japanese English senior high school teachers, with more than 10 years' teaching experience, rated students' performances from the tape recordings. Each task was rated by two of the four raters and Rater 1 (the researcher), who was an anchor rater. This was done to make a meaningful connection with facets of the speaking test for further study. Scoring criteria consisted of 5 items (fluency, vocabulary, grammar, intelligibility and overall task fulfilment). The items were rated on a 0 to 5 point scale according to different levels of performance described for each item.

Results

Questionnaire survey

Research question 1 ascertained what percentage of English teachers assessed students' speaking ability using direct speaking tests. Those who did amounted to 57.3 % (114 English teachers). However, further analysis shows that direct speaking tests were not the only methods of assessing students' speaking ability. The combination of other methods, such as class observation (OB) (frequency of students' utterances and evidence of a positive attitude towards speaking) and pencil-and-paper tests (PE) (testing accents and choosing appropriate words or phrases within conversations) were frequently used (see Figure 5).

Of the 57.3% (114) of teachers who conducted direct speaking tests, 42.7% (85) combined direct speaking tests with other methods, including observation and pencil-and-paper tests, while 14.6% of English teachers assessed speaking ability using only direct speaking tests (SP). On the other hand, 42.7% of teachers did not use direct speaking tests, 17.1% of the teachers (34) used only class observation, 3.5% (7 teachers) used only pencil-and-paper tests and 15.6% (31 teachers) combined observations with these two methods of assessment. Eleven teachers (5.5%) did not include assessments of speaking ability at all and 2 (1.0 %) teachers used other methods. Although this question showed that approximately 60% of English teachers sometimes employed direct speaking tests as an assessment method, only 15% used direct speaking tests as their only assessment. The most frequent assessment method was "only observation" and observation combined with other methods (72.4% in total). Results revealed that the majority of English teachers assessed students' speaking skills based on classroom observation with a combination of other methods.

Research question 2 investigated what impact the introduction of speaking tests would have on Japanese English teachers, which is closely related to the washback effect. Figure 6 indicates that more than 75 % of the teachers reported that speaking tests would have an impact on them, while 20 % expected little impact or no impact on their teaching. All comments have been translated into English by the researcher (see Appendix D). 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. Furthermore, most teachers who gave negative

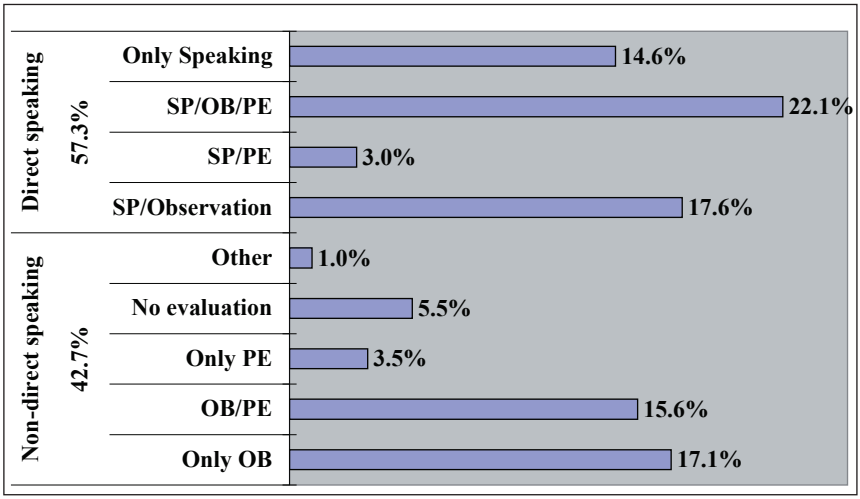


Figure 5: Teacher's assessment methods of speaking skills (n=199)

responses to this question indicated that it was not necessary to put greater emphasis on speaking skills because teachers were already placing emphasis on the development of speaking. While speaking tests have not been yet implemented in the senior high school entrance examination, the inclusion of these tests seemed to potentially engage junior high school teachers who favoured more communicative teaching and direct speaking tests. Thus the inclusion of speaking tests could be one of the ways to bridge the gap between aims of the guidelines and the content of teaching, and between the content of teaching and assessment practice.

Rasch analysis of the student test scores

Application software for Rasch measurement, known as Quest (Adams and Khoo, 1996), was used to address research questions 3, 4 and 5. One advantage of using Rasch measurement software, including Quest, is that item difficulty and person ability, based on responses to specific tasks, are estimated in terms of relative probabilities, so that items, tasks, and students' ability can be compared on the same scale of probabilities. Quest also provides fit indexes, indicating to what extent responses to items on tasks display a consistent pattern (McNamara, 1996). Fit indexes signal whether the necessary patterning is largely present or relatively absent. In the latter case, the item is said to display a

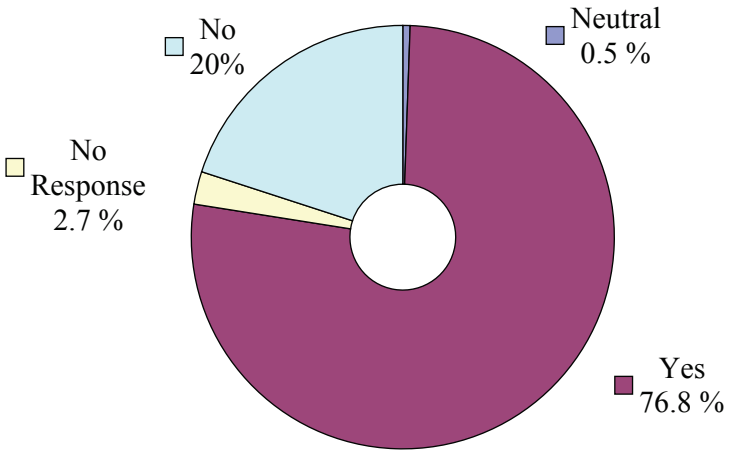


Figure 6: Responses to research Q2 (n=199)

misfit. We can also seek this kind of consistency of response in students' performances and then identify instances of misfit in relation to students, too. Table 2 shows the names of the four tasks used in the test trial, the item difficulty (the third column), task difficulty (the fifth column), and fit indexes (sixth and seventh columns).

Difficulty of items and tasks

Research question 3 investigates the difficulty of tasks (items) on each task. An item with a positive value indicates that the item is more difficult than the mean (logit), and a negative logit shows that the item is easier than the mean. In the third column in Table 2, item 4 (Speech / Intelligibility) is the largest value (1.91 logit), indicating that this item is the most difficult among all items, followed by item 14 (Description / Grammar: 1.7). On the other hand, the easiest item of the interview task is identified as item 20 (Interview /Task Fulfilment: -1.52), followed by item 16 (Interview / Fluency: -1.34). As indicated in the fifth column, 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 logit. This result will be discussed later.

Table 2: Rasch measurement report

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= Vocabulary, G= Grammar, I= Intelligibility, TF= Task Fulfillment

Fit indexes across four tasks

Research question 4 examines 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. Both infit mean square (IMS) and infit *t* in the sixth and seventh columns interpret the same information in different ways. The acceptable range for infit mean square (IMS), according to McNamara (1996, p. 181), is "the mean ± twice standard deviations of the IMS", and the infit *t* statistics -2 to 2. Thus, the acceptable range of IMS here is from 0.70 to 1.30. As can be seen in Table 2, only item 15

(IMS: 1.38; Infit t; 2.5) is identified as 'misfit', indicating a larger than the acceptable range of IMS in the sixth and seventh columns. This suggests that the actual data patterns from item 15 vary unacceptably in comparison with data patterns predicted by the Rasch model. Table 2 also shows that no overfit items (less than 0.7 on IMS or less than -2 on *t* statistic) were identified. This suggests that data patterns across items have some meaningful variations. In summary, the items on four tasks appeared to produce relatively similar response patterns, suggesting that the items across tasks are functioning to measure the similar construct.

Person fit indexes

The last question focuses on students' scores across the four tasks. Quest can also provide misfit persons, just as the misfit item which was identified in the previous analysis. This is particularly important, since this question leads to issues of accountability for students. For example, if the particular task combination includes misfit students, some students who undertake a task combination might be treated unfairly. McNamara (1996) states that the numbers of misfit persons should be within 2% of the total candidates. Tests with more than 2% of misfit students need to be amended. Table 3 presents the numbers of misfit students and their percentages of the total, including infit mean square statistics and standard deviation. As can be seen in Table 3, 5.4% of the students were identified as misfit students. This indicates that the percentage of misfit students exceeds the acceptable percentages of misfit students. It is important to investigate why this happened.

Table 3: The number of misfit students (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 %)

Table 4 shows that the combinations of tasks, which include misfit students the most frequently, were speech and interview followed by the combination of description and interview. Other task combinations produced fewer misfit students. One possible explanation for this is that differences of task difficulty in combinations might have the effect of increasing the number of misfit students. Figure 7 shows that when

a difference of task combination in terms of difficulty becomes larger, the difference affected student performance. However, given the small number of students examined, and the fact that rater behaviour is not considered here, this interpretation must be treated with caution.

Table 4: Relationships between differences of task difficulty combinations and percentage of misfit students

Task combinations (n)	S/R (n=34)	S/D (n=42)	S/I (n=39)	R/D (n=40)	R/I (n=40)	D/I (n=40)
Difference of task difficulty on each task combination	0.45 (logit)	0.98	1.23	0.75	0.99	1.53
Numbers of misfit students	2	1	4	0	2	3
(%)	(5 %)	(2.3 %)	(10.2%)	(0 %)	(5 %)	(7.5 %)

S= Speech, R= Role-play, D = Description, I = Interview

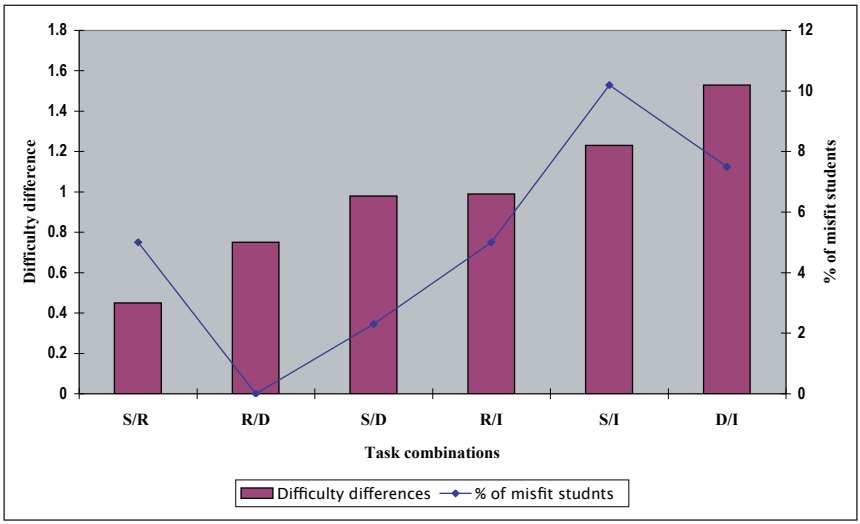


Figure 7: Relationship between difficulty difference of task combinations and % of misfit students (n=15)

It is clear that more comprehensive analyses, including rater behaviour analysis and differential item functioning (DIF) analysis, would be needed. In terms of DIF analysis, six specific schools (2, 5,9,10,

12 and 13) had misfit students, while the others (3, 4,6, 7, 8 and 11) did not. This suggests, as Brindley (2000) states, that not only differences of task difficulty, but also other factors irrelevant to students' performance, such as rater characteristics and interlocutor's behaviour, might have an undue impact on students' scores. These factors might pose threats to validity.

Discussion

Results of the questionnaire survey revealed that the majority of teachers assessed students' speaking skills mainly by observation, and by combining observation with other methods, such as direct speaking tests and pencil-paper tests. The results also showed that the teachers' assessment methods varied. Thus, it would be difficult to compare students' speaking ability across schools, even within the same school where there were more than two teachers, without someone to moderate the teacher-evaluators' efforts.

The introduction of speaking tests would have a positive impact, stated approximately 80% of teachers, and most of these maintained accordingly that they would change to a more communicative style of teaching. From a junior high school teacher's point of view, speaking skills need to be tested because English classes are designed to develop students' oral communicative ability based on the guidelines. As some teachers commented, "The high school entrance examinations should reflect the proportion of time we spend teaching conversation in English classes at junior high school level." The discrepancy between the lack of speaking tests at the entrance examination and the emphasis on the development of speaking ability in class might lower teachers' and students' motivation to speak English in class. Rea-Dickins and Rixon (1997) point out issues that reside in a disparity between the aims of teaching, which puts an emphasis on the skills of listening and speaking, and assessment practices implemented by teachers.

> There is often a major discrepancy between assessment and the underlying construct and content of YL [young learners] language learning programs. Much EFL primary practice emphases the oracy skills of listening and speaking.... Tests of this narrow content coverage and format, will give the 'wrong' message to both teacher and children about the nature of language learning. (p. 158)

Through the previous discussions, it can be argued that the inclusion of the speaking tests would have the potential to assist in bridging the gap between skills taught in classes and skills tested in entrance examinations, and the disparity between the aims of the guidelines and the skills tested in the senior high school entrance examination. In fact, the introduction of speaking tests in the entrance examination would link the aims of the Ministry of Education to the teaching and assessment practice.

Results from test trials undertaken by junior high school students showed that all items except one fit the Rasch model, indicating that items on each task were effective in assessing the target construct. However, the results also showed that the four tasks frequently used by English teachers were different in terms of difficulty. This means that students who do not undertake all possible tasks might not be assessed appropriately. For example, scores from students who undertake two tasks, such as the most difficult and the easiest tasks, could be different from scores of those who undertake two task of similar difficulty. Given the variability inherent in performance tests, including rater behaviour and interlocutors, the difficulty of tasks needs to be relatively equal in order to reduce variability. The concept of "task bank" presented by Brindley (2001), could have important implications for school-based assessments and the assessment of speaking skills in the senior high school entrance examination:

> The first is to develop, in collaboration with practitioners, a bank of fully-piloted exemplar assessment tasks with known measurement properties that teachers can use either for specific assessment in their own classrooms or as models for writing their own tasks. This task bank will be continuously updated as new tasks are developed and piloted, using Rasch-calibrated tasks as 'anchors'. In this way tasks can be mapped on to different levels of achievement. (p. 401)

Implications for this study are that speaking tasks used in a classroom need to be trialled, and also investigated using the Rasch technique, given that school-based assessment represents approximately half of the selection criteria for students who wish to enter senior high school. In junior high school contexts, a role-play task bank, such as a shopping situation, inviting friends to a party, or giving directions to a stranger could be developed. Thus, the task bank is one way of facilitating systematic assessment of students' speaking skills. Collaboration

between researchers and English teachers would make a significant contribution to the task bank.

Another important implication for this study is a question raised by Shohamy (1995, p. 204): "How many performances are needed in order to arrive at valid conclusions?" In achieving more valid evaluations of students, given the time constraints in the senior high school entrance examination, school-based assessment has advantages over the inclusion of speaking tests in entrance examinations. More frequent short 'direct' speaking tests and systematic classroom observations need to be conducted by English teachers. As results of the questionnaire survey indicated, the classroom assessment of speaking skills in schools would have little impact on teachers or students. On the other hand, the inclusion of formal speaking tests would significantly affect junior high school teachers. Therefore, it is important to investigate ways of maximizing the advantages of both school-based assessment and the senior high school entrance examination.

Conclusion

This paper has identified issues of school-based assessment implemented by junior high school teachers, showing that assessment methods of speaking skills varied among junior high school teachers and that only a small number of teachers used only direct speaking tests, despite the emphasis on developing speaking skills in the guidelines. Therefore, the application of results derived from varied assessment methods in a high-stakes context is open to question. However, the above statements do not imply that school-based assessments are not necessary. Rather, school-based assessment has the potential of high construct validity and authenticity.

Through discussions of the three assessment contexts, and the results of the questionnaire survey, this paper has argued for the need to introduce speaking tests in senior high school entrance examinations in order to compensate for the inherent weakness of school-based assessment. The results also showed that tasks frequently used by junior high school teachers varied in terms of task difficulty and that differences of task difficulty had an impact on students' performances. Therefore, in order to not only administer speaking tests in senior high school entrance examinations, but also to enable school-based 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.

Acknowledgements

I would like to express my sincere gratitude to the junior high school teachers who completed the questionnaires. I am grateful to Prof. Tim McNamara at the University of Melbourne, an anonymous reviewer, and the editors of JALT Journal for their insightful comments and suggestions on earlier drafts.

Tomoyasu Akiyama is a Ph.D candidate at Department of Linguistics and Applied Linguistics, The University of Melbourne. His research interests include validity investigations in educational contexts and applications of Rasch measurement to large scale speaking tests.

Notes

1. A condensed summary of this research appeared in the June 2003 issue of the Testing and Evaluation Special Interest Group Newsletter *Shiken*, 7 (2): 2-8.
2. An earlier version of this paper was presented at the JALT conference at Kyoto Sangyo University in May 2002.
3. Questions 3, 4, 5, 6, and 7 were omitted due to space limitations.
4. Information gap tasks were omitted because at that time the researcher and junior high school teachers thought these tasks were not appropriate in testing contexts.

References

Adams, R. and Khoo, S. (1996). *Quest. ACER Quest. The Interactive Test Analysis System. Version 2.1*. The Australian Council for Educational Research.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brindley, G. (1989). *Assessing achievement in the learner-centered curriculum*. Sydney: National Center for English Teaching and Research.

Brindley, G. (2000). Task difficulty and task generalisability in competency-based writing assessment. In G. Brindley (Ed.), *Studies in immigrant English language assessment. Research series 11* (pp. 125-157). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

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.

Hamp-Lyons, L. (1996). Applying ethical standards to portfolio assessment of writing in English as a second language. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3. Performance testing, cognition and assessment* (pp. 151-164). Cambridge: University of Cambridge Local Examinations Syndicate, Cambridge University Press.

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

McNamara, T F. (2001). Language assessment as social practice: challenge for research. *Language Testing*, *18* (4), 333-349.

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

Ministry of Education (1999). *Chugaku Gakushu Shidouyourou*. [The Course of Study Guidelines]. Tokyo Shoseki.

Moss, P. A. (1994) Can there be validity without reliability? *Educational Researcher*, *23* (2), 5-12.

Rea-Dickins, P. & Gardner, S. (2000). Snares and silver bullets: disentangling the construct of formative assessment. *Language Testing*, *17* (2), 215-243.

Rea-Dickins, P. & Rixon, S. (1997). The assessment of young learners of English as a foreign language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education*. *Volume 7*: *Language Testing and Assessment* (pp 151-161). Dordrecht, Netherlands: Kluwer Academic Publishers.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188-211.

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

Tokyo-to Chugako Eigo Kyoiku Kenkyukai (2000). *Goi to Eigokyoiku [Vocabulary and teaching of English] (23): Student Talk (3)*. Kenkyu-Bu.

Appendix A

An example of Section 2 in the 2001 Tokyo Senior High School Entrance Examination

You want to know the English name of an animal that you saw on TV yesterday. You draw a picture of the animal in your notebook and show it to your English teacher, Ms. Smith.

At that time, what do you say to her?

1. Ms. Smith, why do you want to know the name of this animal in English?
2. Ms. Smith, why did you draw this animal in this notebook?
3. Ms. Smith, why do you want to know about this animal?
4. Ms. Smith, what do you call this animal in English?

Appendix B

A Questionnaire Survey to 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 the questions below: Your cooperation will be highly appreciated.

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

Task numbers: the most often used task ( ) →( )

Choice of tasks

(1) Oral interview (2) Information gap (3) Show and tell (4) Skit
(5) Role-play (6) Speech (7) Description
(8) Others

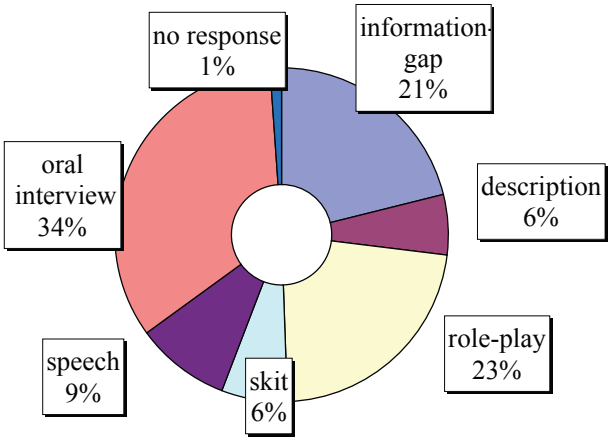
Question 2. How do you evaluate your students' speaking ability? (Please choose the primary method)

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

Question 8. Do you think speaking tests need to be introduced as a part of high school entrance examinations? (Please give brief explanations for your answer.) (Yes / No)
(Your explanations) _____

Question 9. If speaking tests are introduced into entrance examinations, would the test affect you or your teaching? (Please give brief explanations for your answer.) Your answer is (Yes / No)
(Your explanations)_____

Appendix C
Percentage of tasks used in English classes (N=199)



Appendix D

Junior High School Teachers' Responses to the Research Question 2

Tests would influence teachers and their activities because

1. I would be forced to put more emphasis on speaking activities in class (53 teachers).
2. I would have to increase the number of short speaking tests, which would be similar to the speaking tests because students and their parents require teachers to do so (25).
3. Tests would partially influence my teaching styles (23).
4. Students' and teachers' motivation would be directed towards more speaking skills (5).

Tests would not influence teachers or teaching activities because

1. I have already put emphasis on the development of speaking, so that it is not necessary to put greater emphasis than we already have present in the syllabus (28).
2. I don't feel it is necessary to organize classes for the test. If students participate in my class, why should I prepare for them? (4).
3. This is a students' issue, so that our teaching styles are not influenced by tests (2).
4. Introducing speaking tests would contaminate real conversations, which we are trying to achieve (2).

15 REASONS WHY YOU SHOULD JOIN THE JAPAN ASSOCIATION FOR LANGUAGE TEACHING

1 . . . . Leading authorities in language teaching regularly visit us: Henry Widdowson, David Nunan, Jane Willis, Bill Grabe, Kathleen Graves, Jack Richards…

2 . . . . Tips on the job market, introductions… JALT plugs you into a network of language teacher professionals across Japan.

3 . . . . Eighteen special interest groups and their publications: Bilingualism, Global Issues, College and University Educators, CALL, JSL, Teaching Children, Materials Writers, Teacher Education, Testing, Gender Awareness, Pragmatics, Other Language Educators, Junior and Senior High School, Learner Development, Pragmatics, and more.

4 . . . . JALT is a place to call your professional home. With 40 chapters across Japan, it also certain to be not far from the other place you call home.

5 . . . . Monthly chapter programs and regular regional conferences provide valuable workshops to share ideas and sharpen presentations skills.

6 . . . . Professional organizations look great on a resume. Volunteer for a position as a chapter executive, work in a conference, or edit for the publications. You gain organizational and management skills in the process.

7 . . . . JALT maintains links with other important language teaching organizations such as TESOL, IATEFL, AILA and BAAL. We have also formed partnerships with our counterparts in Korea, Russia, Taiwan and Thailand.

8 . . . . Do you have research ready for publication? Submit it to the internationally indexed *JALT Journal*, the world's fourth largest language teaching research journal.

9 . . . . Looking for a dependable resource for language teachers? Check out each month's issue of *The Language Teacher* or any of the many fine publications produced by our SIGs.

10 . . . . JALT produces Asia's largest language teaching conference with all the best publishers displaying the latest materials, hundreds of presentations by leading educators, and thousands of attendees.

11 . . . . JALT develops a strong contingent of domestic speakers: Marc Helgesen, Kenji Kitao, Chris Gallagher, David Paul, Tim Murphey, Kensaku Yoshida, David Martin, Michael Guest, and many others.

12 . . . . Conducting a research project? Apply for one of JALT's research grants. JALT annually offers partial funding for one or two projects.

13 . . . . Free admissions to monthly chapter meetings, discounted conference fees, subscriptions to *The Language Teacher* and *JALT Journal*, discounted subscriptions to *ELT Journal*, *EL Gazette*, and other journals. All for just ¥10,000 per year for individual membership, ¥8500 for joint (two people) membership, or ¥6500 if you can get a group of four to join with you.

14 . . . . Access to more information, application procedures, and the contact for the chapter nearest you.

15 . . . . You don't need a reason. Just do it!

Keep current at JALT2003 in Shizuoka