Assessing Speaking in a University General English Course

(総合英語クラスにおけるスピーキングの評価について)

Mark Brierley, Mary Aruga, Fred Carruth

Goals relating to communication, and specifically speaking, have been clearly foregrounded in recent education policies of the Japanese Ministry of Education (MEXT) and educational institutions in Japan. If such goals are set, how do we know whether they have been met? This paper will describe attempts to assess the speaking ability of a group of first-year university students enrolled in a general English course. We will give a review of the literature surrounding the testing of speaking, cover the design and implementation of a speaking interview test, and explore implications. A survey was conducted on students who had taken the test, and we report the results.

1. Introduction

For at least ten years now, MEXT has set communicative competence as a goal in the Japanese education system.

> Much value will be set on the improvement of fundamental and practical communicative competence in foreign languages and the subject of "Foreign Language" will be a required one at lower and upper secondary schools. (1998)
>
> English-language abilities demanded of all Japanese nationals on graduation from senior high school: Ability to hold normal conversations … on everyday topics. (2002)

However, all teachers are aware of how difficult it has been to make progress in this area. A central problem in the teaching of communicative competence is in its assessment, for which there are no established procedures at any level of the Japanese education system. The assessment of communicative ability or speaking in general would provide greater incentive for teachers to meet the ministry's goals, especially if they are teaching toward tests. In addition, such testing would provide a positive "washback" effect on students. (See 2.3 below.)

1.1. Interview tests

Interview tests were given by three teachers to nine *Sogo Eigo* classes of first-year students in three faculties. Candidates were paired, where possible with a student from a different class,

and where possible interviewed by a different teacher to their regular one. With a ten-minute slot for each interview, one class could be interviewed over two lessons. A range of topics was chosen from among the topics covered in the course, so that different students could be given different tasks and so would not be able to prepare answers to specific questions.

1.2. Questionnaire

At the beginning of the following semester, approximately two months later, students were asked to complete an online questionnaire. The questionnaire was in Japanese, the first language of most students, in order to increase reliability of responses. The questionnaire asked multiple-choice questions with answers on a Likert scale and open questions. Throughout this paper, we will refer to responses from 153 students in six classes. Three each of these classes were from the Education faculty and the Engineering faculty. Three were taught by teacher A and three by teacher B and the interview tests for three classes were conducted by a different teacher while the other three were conducted by the regular teacher. Results were compared in order to isolate significant differences between different teachers, different students or different test procedures.

While there are clear problems getting students to speak English in class, a majority of students believe speaking English is important in their lives. Overall, 63% of students questioned indicated that being able to speak was important or very important, with a further 31% suggesting it was somewhat important. The reasons students gave for these answers showed both instrumental and integrative motivation. There were clear differences between students of different faculties. While only 6 or 7% of students in each faculty felt speaking English to be “not very” or “not at all” important, 73% of Engineering students felt it important or very important compared with only 52% of Education students. Around 60% of all students felt their speaking ability should be assessed.

Over 70% of students thought it a good or very good thing that the speaking test had taken place, while only 6% felt it a bad thing or a total waste of time. Forty percent of students thought another interview test should take place at the end of the current semester.

2. Assessing Speaking

Key concepts in all language assessment are validity, backwash and reliability. Assessment tools and strategies invariably represent a compromise between these ideals, rather than optimisation of all three. Another practical consideration in our case is efficiency. Assessing speaking requires an examiner to listen to each candidate; we must ensure that the examiner’s time is spent effectively by assessing as many candidates as possible.

2.1. Validity

Spolsky (1975, cited by Kim, 2003) asserts that validity is the central problem in foreign language testing. If a test has construct validity, then it can be said to measure what it seeks to measure, which is called the construct. The speaking construct depends on the level and purpose of the speaker. Any construct of communicative speaking will include the ability to understand and be understood, and to process and respond to stimuli in real time.

2.2. Face Validity

Another kind of validity is face validity. This is a subjective judgment on "the degree to which a test appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of an observer" (Richards et al, 1992). As the real implications of tests are often different to those perceived by examinees or administrative personnel, face validity is often disregarded in testing. However, face validity is very important in a classroom speaking test, because the students are more motivated to speak if a test has good face validity (Hughes, 1989). Such a test is usually an achievement test, and as such should have face validity as well as construct validity (Davies, 1983 cited by Kim, 2003).

2.3. Backwash (or washback)

If a test has positive backwash, then it encourages students to participate in activities that will bring them closer to their learning targets. For example, students preparing for a test comprising only short written texts and multiple choice answers may not read long texts, listen to English, write English or speak English as these activities may be seen as irrelevant to passing the test. One may conclude that such a test had negative backwash.

Hughes (1989) recommends: "Test the abilities whose development you want to encourage. For example, if you want to encourage oral ability, then test oral ability" (p. 44). For beneficial backwash, he advises that we should sample widely and unpredictably and use direct testing. In other words, we should measure our students' abilities, rather than infer their abilities from their performance on other tasks. He warns that "Immediately we begin to test indirectly, we are removing an incentive for students to practice in the way we want them to" (p. 45). One of our student's comments was that they wished they had known about the interview test at the beginning of the semester.

Hughes goes on to recommend criterion-referenced testing based on objectives. "If test specifications make clear just what candidates have to be able to do, and with what degree of success, then students will have a clear picture of what they have to achieve. What is more, they know that if they do perform the tasks at the criterial level, then they will be successful on the test, regardless of how other students perform. Both these things will help to motivate students" (p. 45). He advises that the rationale for the test and its specifications should be known and understood by students and teachers, and that samples should be provided,

especially if the test is new (p. 46). This would produce a clear benefit in our classes by moving us nearer to the general goal of increasing students' use of English in class.

Luoma (2004) cites a study on two groups of advanced learners of English done by House (1996) on "pragmatic fluency", a combination of pragmatic appropriateness of utterances and smooth continuity in ongoing talk. She reports: "The group that received explicit instruction ... used a more varied range of gambits and strategies at the end of the course. However, even this was much narrower than native English speakers' repertoire, indicating that pragmatic fluency is a difficult skill to acquire" (p. 90, 91). This illustrates the relationship between implicit and explicit instruction and speaking test results.

Kim (2003) cites Bachman (1990) and Hartley & Sporing (1999) that positive backwash will come from assessing the skills and abilities taught in the class, and that students who have been taught communicatively should be assessed communicatively. In our context a communicate test is therefore likely to help teachers teach communicatively. "This conscious feedback loop between teaching and testing, in terms of content and of approach, is a vital mechanism for educational development" (Kim, 2003). In a future questionnaire, we should ask whether our students felt prepared for the test and whether we offered enough practice in class for the skills covered in the test.

2.4. Reliability

If a test is reliable, the result will not change for the same candidate on a different day, at a different time, in a different place or with a different examiner. This is clearly an important issue in a speaking test as the variables include the interview and other candidates who are taking the test. Although raters need to aim for reliability, it may not be the prime consideration all the time. There is reliability-validity tension; it is sometimes essential to sacrifice a degree of reliability to enhance validity and in certain circumstances, reliability and validity are mutually exclusive. However, if a choice has to be made, validity is more important for speaking assessment (Davies, 1990 and Bachman, 1990 cited in Kim, 2003).

2.5. The speaking construct

Before we can judge a test on any of these criteria, we need to know what we are trying to assess, a task which entails considering some of the most vexing questions in linguistics. According to Ellis (1994:696 cited in Kim, 2003), communicative competence is "the knowledge that users of a language have internalized to enable them to understand and produce messages in the language." Kim (2003) gives an excellent summary of the development of the speaking construct. Hymes (1971) broadened Chomsky's (1965) linguistic competence to include communicative competence. Canale & Swain (1980) identified grammatical competence, sociolinguistic competence and strategic competence within this,

Canale (1983) later adding discourse. Bachman (1990) addressed the question of communicative language ability, separating language competence from strategic competence. Luoma (2004) shows the inter-relationships between hypothesized components of language use such as language knowledge, topical knowledge and personal characteristics as they are mediated by strategic competence and affective factors (p. 98, after Bachman and Palmer, 1996, p. 63).

Hughes (1989) reminds us that "the objective of teaching spoken language is the development of the ability to interact successfully in that language, and ... this involves comprehension as well as production" (p. 101).

2.6. Priorities in communicative testing

In view of the challenges inherent in encouraging students to speak English, we consider the positive backwash and face-validity of our interview test to outweigh the very real issue of reliability. Morrow (1979) notes the following characteristics of a test of communicative ability:

> 1. It will be criterion-referenced against the operational performance of a set of authentic language tasks…
>
> 2. It will be crucially concerned to establish its own validity as a measure of those operations it claims to measure…
>
> 3. It will rely on modes of assessment which are not directly quantitative, but which are instead qualitative…
>
> 4. Reliability, while clearly important, will be subordinate to face validity. Spurious objectivity will no longer be a prime consideration… (pp. 150-151)

Weir (1988) adds:

> Testing speaking ability offers plenty of scope for meeting the criteria for communicative testing, namely that: tasks developed within this paradigm should be purposive, interesting and motivating, with a positive washback effect on teaching that precedes the test; interaction should be a key feature; there should be a degree of intersubjectivity among participants; the output should be to a certain extent unpredictable; a realistic context should be provided and processing should be done in real time. Perhaps more than in any other skill there is the possibility of building into a test a number of the dynamic characteristics of actual communication. (p. 73)

2.7. Strategies so far: testing by wandering around; assessing presentations

A variety of speaking assessment strategies have been adopted by teachers at this institution. Informal discussion suggests that the three teachers in this study are not alone in the strategy of making notes next to each student's name throughout the semester as we observe them in classroom activities. However, we are also not alone when we find, at the end of the semester,

that we have insufficient data for a proper assessment.

Students show some sympathy for this approach, with around one-third of students feeling that teachers should assess their speaking in class activities, with over half ambivalent. Interview tests and presentations seem more popular, each supported by around 40% of students. Although they got more support, fifteen to twenty percent of all students felt that presentations and interview tests should not be used for assessing speaking, with a quarter of education faculty students opposed to these techniques.

The assessment of presentation ability has been more systematically approached by teachers, both in general English courses and dedicated presentation classes. This is possible because the speakers speak in a clearly defined time at a designated place with a specified purpose. A variety of rubrics exist, covering specific teaching points such as posture, eye contact and organisation. Unfortunately, vital elements of communicative English are missing from presentations, specifically the need to react in real-time to unexpected comments. A question-and-answer session at the end of a presentation would allow spontaneous interaction, although this is difficult to generate and even more difficult to assess than interviews.

3. An Interview Test

While the term ‘language assessment’ usually conjures up the image of paper tests, and spoken tests may seem like a modern innovation, oral examination has a long history. As far back as 1913, the University of Cambridge Local Examinations Syndicate (UCLES) included a compulsory oral component in the Certificate of Proficiency in English, and in 1945 John Roach published a paper looking into the many issues involved in oral testing, entitled *Some Problems of Oral Examinations in Modern Languages* (Taylor, 2003). Following the trend towards communicative teaching methodology, attempts to assess communicative competence are increasing.

According to Weir (1988), the controlled interview, with a set of procedures determined in advance for eliciting performance has the following advantages:

> 1. There is a greater possibility in this approach of candidates being asked the same questions and thus it is easier to make comparisons across performances.
>
> 2. The procedure has a higher degree of content and face validity than most other techniques apart from … role play and information gap exercises….
>
> 3. It has been shown elsewhere (Clark and Swinton, 1979) that with sufficient training and standardization of examiners to the procedures and scales employed, reasonable reliability figures can be reached with this technique.

He also mentions disadvantages:

> 1. In interviews it is difficult to replicate all the features of real life communication such as reciprocity, motivation, purpose and role appropriacy.
>
> 2. Even when the procedures for eliciting performance are specified in advance there is still

> no guarantee that candidates will be asked the same questions in the same manner, even by the same examiner. (p. 76)

3.1. Format

Hughes (1989) calls for a long test, saying, “It is unlikely that much reliable information can be obtained in less than about 15 minutes, while 30 minutes can probably provide all the information necessary for most purposes” (p. 105). In the context of a high-stakes test this may be useful; however, in our fifteen-week-course context, such lengthy tests would leave little time for anything else. We note Hughes’ advice to “give the candidate as many ‘fresh starts’ as possible.” He advises more than one format, more than one tester, and many ‘items’ within each format. In the implementation of a testing regime over a one- or two-year curriculum, this may be possible, and it may be helpful for us to see each interview test as one component of a long-term assessment strategy. He goes on: “Particularly if a candidate gets into difficulty, not too much time should be spent on one particular function or topic. At the same time, candidates should not be discouraged from making a second attempt to express what they want to say, possibly in different words” (p. 106). This is useful advice to remind the examiner that an effective interview encourages the candidate to speak as much and as freely as possible.

The ideal method, according to Kim (2003), is to have an examiner and a scorer present during the test. While the examiner administers the test, the scorer, visible to the candidates, can record the information for the score. Hughes (1989) also recommends that a second tester be present for an interview because of the difficulty of conducting an interview and keeping track of the candidates’ performance. Cambridge ESOL exams follow this instruction with an interlocutor administering the test to paired candidates while an assessor listens and concentrates only on scoring.

Bound by practical constraints, and fully aware of the limitations, we settled upon single examiners interviewing pairs of students, with at least five minutes for each interview, within a ten-minute slot.

3.2. The examiner: own teacher or a stranger?

In deciding whether the regular teacher or a stranger should assess students, it is interesting to note that students who had been interviewed by their own teacher were more positive about the interviews, with 78% thinking the tests a good or very good thing against 64% of those who had been interviewed by a stranger. In addition, more of those interviewed by a stranger felt they had scored badly in the test (48% against 34%). Perhaps unsurprisingly, more could understand their regular teacher easily or very easily (61% against 39%). However, we were surprised to note that fewer of the students who had been interviewed by a stranger were

nervous (58% compared to 71%).

If we wish to accurately assess the candidate's ability, a stranger is likely to assess the performance more objectively, while a familiar teacher may take into consideration any previous encounters with the candidate. This may be more appropriate at the end of a course. However, if we see the interview as part of the course, and an opportunity to develop the teacher-student relationship and provide information that will help the teacher, then clearly we should be interviewing our own students. In the end, as a result of practical constraints, some students were interviewed by a different teacher, while others were interviewed by their own teacher.

3.3. Paired candidates

There are clear advantages to having pairs of candidates in an interview. Hughes (1989) comments that the traditional one-on-one interview "has at least one potentially serious drawback. The relationship between the tester and the candidate is usually such that the candidate speaks as to a superior and is unwilling to take the initiative. As a result, only one style of speech is elicited, and many functions (such as asking for information) are not represented in the candidate's performance" (p. 104). As well as allowing a variety of more authentic interactions, an examiner observing two candidates will be more efficient, allowing more time to listen to each candidate per minute of interview time.

There are of course challenges, primarily in the area of fairness. "The examinees' talk is almost inevitably influenced by the other participant's personality, communication style and possibly also language level. The concern is that all test takers may not get an equal opportunity to show their speaking skills at their best" (Luoma, 2004, p. 37). She suggests that scores from paired interactions must be interpreted carefully, and that candidates should have a different partner in the desirable event of another test at another time. Hughes (1989) suggests that "candidates should be carefully matched whenever possible" (p. 105).

With our goal of face validity in mind, we must be wary of students perceiving the test as being unfair. The questionnaire responses reflected the pros and cons of paired interviews, with around 40% of students in favour of paired interviews and 40% in favour of individual interviews. Most of the students' comments related to affective issues, although slightly more felt that there would be less tension in an interview with two candidates than with one. Of around 60 comments supporting paired interviews, nine mentioned cooperation, six that it would be closer to a real situation, three that they would have more time to think and three that there would be better conversation. On the other hand, eight students suggested the interviews would be fairer with one, and others commented that the teacher would be able to concentrate better on assessment with two candidates speaking to each other. One student commented that individual interviews would take too much time. The issue of personality also came up in many comments, some mentioning that they had performed well because they

knew their partners, some commenting how helpful their partners had been, and others complaining about their partners. Some commented that the test was not just of speaking ability, but also of character. This is clearly an important issue, although there is an argument that those whose character leads them to speak more are better at speaking, in much the same way that people who enjoy running are more likely to win in a race. In the context of paired interviews, assessors must be wary of one candidate's performance affecting the score of the other.

In a related question in which students were asked whether they had held back from speaking out of consideration for their partners, over 40% said they had not, and another 20% could not remember. In a general question, of all students, 37% felt the test was fair, while 14% felt it was unfair and 47% could not say. Of thirty comments on fairness, six students commented on partners, while most commented on the topics being different for each student.

3.4. Tasks and topics

In order for a test to be reliable, the questions would have to be identical. With no provision of ushers, and no possibility for isolating students who had completed the test from those who had not, we decided to prepare four different sets of two topics, one for each student in each interview. The test would be very unreliable if the first students to take the test told later candidates all the questions! With the hope that the more the students knew about the test, the better, students were told all the topics before the test. In the questionnaire, we asked students what they did in the classroom while others were taking the test. Six of the students confessed that they had listened to other students talk about what had happened in the interview test. It is unclear how much difference this made, although students are unlikely to have been able to suddenly improve their communicative ability.

Hughes (1989, p. 106) recommends that we choose topics that would cause candidates no problems in their own language. We chose four sets of two or three topics all of which related to course content. The most difficult topics seem to have been surveys and quizzes: two class projects. Although all students had been living in Matsumoto for at least three months, some commented that was a difficult topic. In this kind of test, the topics act as catalysts; we are not interested in how much they know about their hometown, but in how they can express that knowledge in English.

The questionnaire results suggest that even easier topics may have been better, with around 30% finding the topics difficult, and less than 20% finding them easy. None commented that the topics were too easy, although one student commented that were it not for the easy topic he had been given, he would have been at a loss. Another claimed the test was difficult as she had thought about all the topics except the one allocated.

3.5. Rubric and instructions

Luoma (2004) advises assessment designers to consider how much explanation is needed for each part of the test. Brief, clear instructions are desirable; however, for the test to be fair candidates should know enough that they do not need to guess what is expected of them. She gives a range of options including providing information before the assessment, creating or agreeing on material with the learners. “The instructions in particular are important for task development because they set the scene for how the participants will perceive the task and their own performance on it” (p. 51). On a different note, Hughes (1989) advises us to avoid yes/no questions (p. 107); a candidate replying “yes” or “no” will not give us much data to assess their ability.

The test was split into three parts. In the first part, each candidate was asked standard questions: “What’s your name?” “How do you spell that?” “What do you study?” Next, a few introductory questions were chosen from a list. Where possible, follow-up questions were given, both to show interest in what the candidate was saying, and to model follow-up questions to encourage candidates to use them. This introductory part was also designed to put candidates at their ease.

In the second part, each student was given a topic in turn, (for example “Activities in Shinshu” or “Part-Time Work”) and directed to ask his or her partner questions about it. The candidate was given a card with the topic and a few prompts of information about which they could ask (see appendix 8.2).

In the third part, students were asked to discuss a topic connecting what each had been talking about (for example “What do you do when you are not studying?”) Following Luoma’s (2004, p. 38) advice to have a spare task or a set of additional questions available in case a task fails, a number of alternative questions and back-up questions were added to the interviewer’s script.

3.6. Speaking scales and scoring

Hughes (1989) warns that we will get valid and reliable scoring only if:

> Clearly recognisable and appropriate descriptions of criterial levels are written and scorers are trained to use them.
>
> Irrelevant features of performance are ignored.
>
> There is more than one scorer for each performance. (p. 110)

While time constraints prevented us from having more than one scorer, we responed to Hughes’s first suggestion by setting three criteria and five levels for assessors to allocate to each student (see appendix). It must be pointed out that these assessments are all subjective rather than objective, which brings reliability into question. It must also, however, be pointed out that speaking itself is also usually subjective, and any purely objective assessment risks missing the point, which is usually for one person to be understood by another.

Teachers with little specialized training can use objectified, or analytical scoring, which is consistent and easy to use (Bachman, 1990). However, it is possible to lose a sense of the candidate's overall performance, and even on a speaking test with objectified scoring, it is a good idea to indicate a very general impression of a student's performance. This is called "holistic scoring" and can be done simply by indicating whether the person is "high", "mid", or "low" (Bachman, 1990). Hughes (1989) cites research showing "a very high level of agreement between holistic and analytic scoring" (p. 110). Our scoring would have been improved by such a holistic score in addition to the three analytical scores: pronunciation, interaction and range. (See appendix 8.3.)

3.7. Affect

With 39% of all students saying they were nervous when they took the test, and 25% saying they were very nervous, stress is clearly an important issue for our students, and it is therefore of great concern as we design and implement these tests. Considering Krashen's "affective filter" (2003), it is incumbent upon us to reduce anxiety in our classrooms so that students may acquire more language. While Luoma (2004) comments that the proficiency of the interviewer is often not an issue, "personality and communication style certainly are" (p. 38). There are a number of things we can do to put our students at ease:

> Interviewer should be pleasant and reassuring throughout.
>
> The initial stages of the test should be very easy for all reasonable candidates.
>
> Testers should not make notes on the candidates' performance during the interview or remind candidates that they are being assessed.
>
> The interview should end at a level where the candidate feels comfortable to leave him or her with a feeling of accomplishment. (Hughes, 1989, p. 106)

4. Discussion

4.1. Improvements

Luoma (2004) advises that if test developers "actually analyse some performances and use the results to revise their scale … this may lead to the production of more concrete and user-friendly rating scales for speaking in the future" (p. 95). Such regular revision takes place in the case of Cambridge ESOL's tests (see, for example DeVelle, 2008).

If seen as part of students' education, it is important for feedback on performance to reach the students. Feedback can serve three functions: giving students specific information about their weaknesses; giving students general recommendations for their improvement; and giving teachers information on their students' abilities. The most straightforward way would be to return the grading slips to students after recording the necessary data. Feedback should

be positive, encouraging and helpful in showing students where and how they could improve. It may also be possible for online feedback to be given to students. We feel that in general, not just for this test, feedback to students could be improved.

4.2. Examiners

Although the sample was small, the average scores given differed between the three examiners. There are differences between the strictness and generosity of teachers (Brierley & Orlandini, 2007). If grading is standardised, as well as being fairer for students, it will allow us to compare students in different classes over the course of their learning careers. The process of reaching consensus over speaking assessment among the university's sixty or so English teachers may also be useful in clarifying the skills and abilities we wish to teach our students.

Our experience matches that of Kim (2003): "It seems that teachers need to have assistance and encouragement in trying communicative assessment. The accurate measurement of oral ability takes considerable time and effort to obtain valid and reliable results. Nevertheless, where backwash is an important consideration, the investment of such time and effort may be considered necessary." The use of video recordings of candidates and examiners may prove useful in: showing teachers good interviewing practice; giving standard examples of candidate performance to aid fair assessment; and analysing the abilities of our students to better focus our teaching efforts.

In this study, we took into account whether students were examined by their own teacher or a different teacher. In a future, larger study, we could also compare comments from students who had different examiners, in order to identify good practices among examiners. It would also be interesting to compare students' attitudes towards native and non-native teachers in assessing speaking.

4.3. A waste of time?

On the face of it, spending two whole lessons on a test seems extravagant, despite the fact that some researchers recommend spending at least three times as much. However, let us look carefully at what the teacher is doing and what the students are doing.

The teacher is spending most of these classes speaking or listening directly to students. We do not expect language teachers to stand at the front of the class lecturing students, and some personal interaction is desirable. If this is to happen by teachers patrolling the classroom, it may be less systematic and is likely to be biased towards more vocal students. While a teacher may get around a whole classroom in one lesson, it is unlikely that he or she will speak to every member of the class. As far as the teacher is concerned, therefore, the testing time is being well spent. It may be ideal for such tests to take place outside regular class-time, with students coming for an interview at a designated time and place. The

practicality of this may vary depending on students' and teacher's circumstances.

While interviews were taking place, the other students were allocated tasks in the classroom. Just over half of all students felt they had used the time well, with a higher proportion of Engineering students (55%) and a lower proportion of Education students (48%). Only 4% felt it a waste of time. When asked what they actually did, most reported activities set by the teacher, such as group work or extensive reading, or activities related to English, such as "memorising vocabulary" or "studying for TOEIC". Others studied for the (non-English) test in the following lesson. A few confessed to having ignored the teacher's instructions, one enigmatically answering "did important things". In answering what students should do while others are being interviewed, out of four choices offered, extensive reading was the most popular (42%), more so among Engineers (49%, 70% in one class) than education students. "Self study" and "group activities" were more popular among Education students than Engineering students (35% and 18% compared with 15% and 6% respectively). Eighteen percent of students felt they should be free to do what they wanted. Few students took the option of "other"; those that did suggested allowing students to go home. In answer to another question, students suggested that rather than interviews taking place together, a few should be held each week. This approach has been used in a different context (see Ruzicka & Brierley, 2008), and it may be useful to start a rolling programme of student interviews from the beginning of the course.

4.4. Activity theory and sociocultural approaches to language learning

Luoma (2004) points out the anomaly that the interaction in the test is co-constructed, yet we score individuals (p. 103). She suggests a variety of "more socioculturally appropriate" approaches to testing that take into account the individual, for example offering a choice of tasks (also requested by students in our questionnaire), or encouraging candidates to bring a portfolio of their work. Again from a sociocultural perspective, she questions cultural appropriateness as a criteria when we assess our students, pointing out that student may be unaware of cultural norms beyond their classroom experience. In addition, she points to activity theory and the notion that the assessment is, and will be perceived by most candidates as an activity. However real-world the tasks are, candidates know that they are in a language test, with all the expectations and norms that that implies.

It seems from many of the comments that our students consider language to be a subject of study rather than a sociocultural activity, and that learning focuses on knowledge rather than skills. Although many thought they needed more practice, and appreciated the interview test as an activity, many perceived their shortcomings in terms of "not memorising enough words". This highlights an attitude towards language as a sum of knowledge rather than a set of skills, and once again casts doubt on the feasibility of MEXT's focus on communicative ability and the skills that implies. It also leaves unanswered the question of

how these attitudes can be changed.

5. Conclusion

Speaking interviews can play an important role in improving the speaking ability of students by giving credit and credibility to the importance of speaking ability, and through encouraging students to practice speaking for the test. A number of challenges must be met to implement a system that is both fair and practical. In addition, it is necessary to have an exact definition of the language we expect students to be capable of producing before tasks can be designed and performance criteria established.

6. Acknowledgements

The authors would like to thank Hisashi Miura for helping to translate the questionnaire into Japanese and Yasuna Horiuchi for the tedious tasks of transcribing recorded interviews and correlating survey responses. The development of these interviews would not be possible without the regular, but usually informal, interaction between teachers over what we should do in and out of our classrooms and how we should do it. Thanks are also due to all the students who participated in the tests and responded to the questionnaires.

7. References

Bachman, L. (1990) *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Brierley, M. and Orlandini, O. (2007) 'Standardised assessment of a general English course' Journal of Humanities, Shinshu University, 1, 164-184.

Brown, H. D. (2004) *Language assessment – principles and classroom practice.* White Plains NY: Longman.

Buck, G.. (2001) *Assessing listening.* Cambridge: Cambridge University Press.

DeVelle, S. (2008) 'The revised IELTS pronunciation scale' University of Cambridge ESOL Examinations Research Notes 34, 36-39.

Hughes, A. (1989) *Testing for language teachers.* Cambridge: Cambridge University Press.

Kim, H. S. (2003) *The types of speaking assessment tasks used by Korean junior secondary school English teachers*, Asian EFL Journal <http://www.asian-efl-journal.com/dec_03_gl.pdf>

Krashen, S. D. (2003) *Explorations in language acquisition and use.* Portsmouth, NH: Heinemann.

Luoma, S. (2004) *Assessing speaking*. Cambridge: Cambridge University Press.

MEXT (1998) <http://www.mext.go.jp/english/news/1998/07/980712.htm>

MEXT (2002) <http://www.mext.go.jp/english/news/2002/07/020901.htm>

Morrow, K. (1979) ‘Communicative language testing: revolution or evolution?’ in *The Communicative approach to language teaching*, Edited by C.J. Brumfit and K. Johnson, Oxford: Oxford University Press.

Richards, J. C., Platt, J. and Platt, H. (1992) *Longman dictionary of language teaching & applied linguistics*. Harlow: Longman group UK limited.

Ruzicka, D. and Brierley, M. (2008) ‘Selling ER: investigating factors in classroom management that affect reading performance’ Journal of Humanities, Shinshu University, 2, 223-238.

Taylor, L. (2003) ‘The Cambridge approach to speaking assessment’ University of Cambridge ESOL Examinations Research Notes 13, 2-4. <http://www.cambridgeesol.org/rs_notes/rs_nts13.pdf>

Weir, C. J. (1988) *Communicative language testing.* Hemel Hempstead: Prentice Hall.

8. Appendices – test materials

8.1. Teacher script

Part 1

(Questions from teacher to student A, then to student B)

All students:

“What’s your name?”

“How do you spell that?”

“What do you study?” “What is your major?”

“What is your teacher’s name?”

Ask a range of questions.

Do not ask both students the same questions.

...

“Do you like reading?”

“What book are you reading?”

“Do you like it?”

“Why?”

...

“What is your hobby?”

“What do you do in your free time?”

“Do you have a part time job?”

“What do you (want to) do?”

Part 2:

“[Student A], please ask [student B] some questions about **his/her hometown**. Please use this [paper] to help you.

[Student B], please answer the questions.”

(prompt if necessary)

“Where are you from?”

“Why?”

“[Student B], please ask [student A] some questions about **Matsumoto**. Please use this [paper] to help you.

[Student A], please answer the questions.”

(prompt if necessary)

“Do you like Matsumoto?”

“What is different to your hometown?”

Part 3:

“Please discuss **places in Japan**.”

4. Discussion: Places in Japan

“Talk about different places in Japan”

(Prompt: “Ask what is different between Matsumoto and his/her hometown...”

“That’s all! Thank you. The test is over”

8.2. Student prompt sheets

4A Hometowns
Ask your partner questions about his or her hometown.

Ask questions about:
- Famous places
- Famous people
- Food
- Nature

4B Matsumoto
Ask your partner questions about Matsumoto.

Ask questions about:
- Best place
- Why?
- Best food
- Local culture

8.3. Speaking assessment criteria

Score	Pronunciation	Interaction	Range
5	Always easy to understand for the teacher and the other candidate	Able to initiate, maintain and close conversation, quick to respond	Suitable vocabulary and grammar always available to complete the task
4	Can usually be understood	Some initiation, usually quick to respond	Vocabulary and grammar are usually suitable and available
3	Can be understood with some effort	Interaction continues, mostly reactive, response often slow	Vocabulary and grammar allow task to be completed
2	Much effort is required to understand	Interaction is limited, pauses cause strain	Lack of vocabulary and grammar make task difficult
1	Impossible to understand	Interaction impossible	Insufficient linguistic resources
0	No language		

Note: The student should be understood in the context of the questions and the audience. Students do not need to sound like the Queen of England or Walter Cronkite.

8.4. Sample assessment sheet

Name:	Date: 28/7/2008
Student number:	Class: TI (2)
Interview Room: 311	Time: 11:10 AM

For the teacher: S: 32 T: C I: A

Category	Pronunciation	Interaction	Range
Score			
Comments:			

(Associate Professor, School of General Education, Shinshu University)
(Part-time teacher, School of General Education, Shinshu University)
(Part-time teacher, School of General Education, Shinshu University)
30/ Dec./ 2008 Accepted