Title: Development of a Practical Speaking Test With a Positive Impact on Learning Using a Story Retelling Technique

Authors: 1. Akiyo Hirai (University of Tsukuba)
2. Rie Koizumi (Tokiwa University)

Contact e-mail address: hirai.akiyo.ft@u.tsukuba.ac.jp

Phone: +81-29-853-4180

Address: Akiyo Hirai
University of Tsukuba
1-1-1 Ten'nodai, Tsukuba, Ibaraki 305-8571, Japan

Abstract

This paper presents a test development project for classroom speaking assessment. With the aim of enhancing and specifically easing the process of test preparation and administration and generating positive washback effects on learning, we developed a semi-direct speaking test called the Story Retelling Speaking Test (SRST). Although a story retelling technique has already been widely recognized as a teaching activity, its use for speaking assessment has not been fully studied. Thus, the paper discusses the potentiality of using this technique for the SRST and reports its pilot administration to 43 examinees. As a result, the high practicality of the test was confirmed at the test construction and implementation stages. In addition, the questionnaire distributed to the examinees yielded generally positive results regarding their perception toward the test usefulness and the appropriateness of the test procedures and task difficulty. With regard to the appropriateness of the texts, the examinees perceived that the retelling of stories was influenced most by text content and then by text length; however, these two factors appear to be interrelated. On the basis of these responses, we have suggested some revisions of the SRST and future validation and reliability studies.

Keywords:
speaking test development, classroom-based testing, story retelling

Acknowledgements

We are very grateful to the anonymous reviewers of this journal for their valuable comments on an earlier version of this paper. This study was partially supported by Grand-in-Aid for Scientific Research (KAKENHI) (C) (19520477).

Background

Communicative language teaching with an emphasis on speaking and listening has been widely encouraged in Japan (e.g., Ministry of Education, Science & Culture, 1999a, 1999b; Nishino & Watanabe, 2008). If the development of students' speaking skills is an instructional goal, their speaking ability needs to be evaluated with a careful deliberation of validity issues such as content coverage by the assessment and the way in which the assessment is conducted. Along with the trend of fostering communicative abilities, this paper reports on the efforts to develop a new speaking test suited to the needs of English teachers and students in Japan.

While several speaking assessment techniques have been developed thus far (Fulcher, 2003; Luoma, 2004), teachers who attempt to conduct speaking tests tend to face many difficulties such as the considerable amount of time necessary for administering and scoring the tests, the special techniques required for rating speaking performance, and the tremendous financial burden either on the students or the school if the available speaking tests in the market are used. These obstacles result in the low practicality of speaking tests, which seems to be of great concern, especially in classroom-based contexts. Owing to the low practicality of speaking tests, even when teachers organize speaking activities in class, they tend to conduct speaking assessments infrequently. Honda (2007) reported that the teachers at lower secondary schools in Japan evaluate the speaking performance of their students only 1.2 times per term/semester. A much lower frequency of speaking assessment can be predicted at upper secondary schools, mainly because of a greater emphasis on preparing students for the university entrance examinations, in which speaking performance tests are not included (Akiyama, 2004). When there is a wide gap between the frequency of opportunities for speaking activities and their assessment in English classes, the relationship between teaching and learning speaking skills could become fragile. As a consequence, the students may not take speaking activities seriously. Thus, generating a positive impact on learning by increasing the opportunities for speaking assessment in the classroom is a major concern in Japan.

To alleviate the issue of very little classroom speaking assessment in current English language education, we have been developing a classroom-based practical semi-direct speaking test, aiming to generate a positive impact on both teaching and learning. We adopted story retelling as the main speaking task and named the test the Story Retelling Speaking Test (SRST). If the SRST could achieve high practicality and strong positive impact as well as other test qualities, it could be widely used to enhance and assess students' speaking ability. Since, to our knowledge, there have been no studies using a story retelling task as a classroom speaking test, this study aims to report on the development of the SRST and examine whether or not the new test is perceived by the test takers as we have intended at

the test design stage.

Since our primary goal is to develop a practical speaking test with a positive impact, we will first review the test practicality and impact based on Bachman and Palmer's (1996) test usefulness framework and discuss how these concepts can be integrated in the designing of the SRST. We will then examine whether our intentions can be successfully perceived by students by administering the test and investigating the test responses.

Practicality and Test Impact

Bachman and Palmer's (1996) test usefulness framework has been influential in language testing. The framework may be suited to the evaluation of the SRST because practicality and impact—the main focus of our test development—are treated equally along with other qualities such as reliability, construct validity, authenticity, and interactiveness. Bachman and Palmer argue that all these aspects are essential for determining test usefulness and that a test should be aimed at maximizing its usefulness while maintaining the balance of these qualities.

Among the six qualities, practicality is the most difficult for speaking tests to achieve. The degree of test practicality can be determined by "the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities" (Bachman & Palmer, 1996, p. 36). Test practicality can be considered in the following four stages: (1) design and operationalization, (2) administration, (3) rating, and (4) analysis. At each stage, the test practicality is influenced by human resources (i.e., the number of people and the amount of burden entailed), material resources, and time as well as money.

In spite of the difficulty of attaining high practicality of speaking tests, many attempts have been made to reduce the time and energy required at any of these testing stages. For example, as compared to a face-to-face interview, a tape-mediated speaking test (e.g., O'Loughlin, 2001; Shohamy, 1994) can save human resources at the administration stage (i.e., the number of, and burden on, interviewers) and the time required by interviewers (i.e., the time it takes to learn how to administer and to actually administer the interview test). One of the final goals of the SRST is to reduce the resources required at all the four testing stages.

In addition to practicality, test impact is also emphasized in the SRST. The importance of school-based tests with positive effects on teaching and learning cannot be overemphasized, as they have been recognized as a social entity in language testing (McNamara & Roever, 2006). The impact of testing can be divided into two categories: (a) the impact on society and educational systems and (b) the impact on individuals (stakeholders, especially test takers and teachers; Bachman & Palmer, 1996, pp. 29–35). An aspect of the impact that focuses on processes (learning and instruction) is referred to as *washback*. It is

defined as “the effect of testing on teaching and learning,” which can be either beneficial or harmful (Bachman & Palmer, 1996, p. 30).

Tests tend to have positive washback if the assessment uses authentic and direct samples of the communicative behaviors and can reflect good classroom practice (Bailey, 1996; Hughes, 1989; Messick, 1996; Morrow, 1991). More specifically, Messick (1996) points out the following:

> Ideally, the move from learning exercises to test exercises should be seamless. As a consequence, for optimal positive washback there should be little if any difference between activities involved in learning the language and activities involved in preparing for the test. (p. 241)

In this manner, communicative testing can elicit the best possible performance from the test takers (e.g., Bailey, 1996; Morrow, 1991). Taking these points into consideration, we have designed the SRST in a way that enables teachers to easily link classroom activities with the test.

However, empirical research has shown that washback involves dynamic and complicated processes with many uncontrollable variables such as the perspectives of the teachers and other stakeholders toward the test and whether it is high stake or low stake (Cheng, 2004; Shohamy, Donitsa-Schmidt, & Ferman, 1996; Wall, 1996; Wall & Alderson, 1993; Watanabe, 2004). In this sense, it is essential to investigate the washback effects by taking their potentially relevant factors into consideration. However, at this preliminary stage, we were able to examine only the test takers’ perceived washback effects; thus, we need to explore test impact more thoroughly in the future.

The remaining test qualities mentioned in the usefulness framework are also essential and are conventionally examined when a new test is created and implemented. Thus, we examined these other qualities as well by administering a questionnaire to the test takers.

Designing a Classroom-based Test

Applying Story Retelling Tasks for Speaking Assessment

We applied a story retelling technique to the SRST because it enables teachers to easily and accurately connect input and output or learning and assessment, which is a vital aspect of classroom assessment. The advantages of using this technique are discussed in the next section.

The key term *story retelling* can be defined by explaining each word separately. First, *retelling* refers to reproducing a story orally in English. An examinee can retell a story either in a different way or in the same way as the original (adapted from Chaudron, 2003, pp. 779–780). A *story* is defined as any type of written description consisting of two or more sentences that are connected to one another. It includes a description, either true or imaginary.

In this study, the following terms were used interchangeably: *retell* and *reproduce*; *story*, *text*, and *passage*. The term *story* is used in the test name in order to make it analogous with terms such as *story telling* and *story retelling*, which seem to have already become established as teaching activities (e.g., Kosuge & Kosuge, 1995).

Story retelling techniques have been widely used not only as a teaching activity but also a speech elicitation tool in second language acquisition (SLA; Chaudron, 2003). When story retelling is used for reading comprehension, the language of reproduction seems to be the first language (L1) in order to avoid the underestimation of reading comprehension (e.g., Donin, & Silva, 1993; Lee, 1986). However, when the target of a story retelling task is the learners' speaking ability, they are asked to retell the content in L2, as in the case of the SRST.

In regular retelling activities, the learners are provided with either a reading or listening text or shown a story using the TV or other related devices and are then asked to retell the content, mostly without looking at the original source. The SRST can also take the form of either reading or listening. However, the present study focuses only on the former, because a reading mode of text presentation would be suitable for learners at the beginning and intermediate levels of proficiency, which are our target groups. In general, when the same text is provided, the learners tend to find it easier to comprehend through reading rather than listening, as they have better control over the pace of their comprehension (Hirai, 2001).

The SRST consists of two sections: reading a story and retelling it. In the reading section, we mainly seek to measure reading ability. We predict that if the test takers cannot answer the comprehension questions, they will be determined to be at the pre-speaking level, where they lack basic linguistic knowledge. In addition, this section may inform us whether the level of text difficulty was suited to their ability. In the retelling section, the main construct that we want to measure is speaking ability. This is specifically measured with two functions: retelling the story and stating opinions about it. In the story retelling task, the examinees convey the information they have just received as clearly as possible and narrate as much of the story as they can. This task seems authentic (Underhill, 1987) because there are real-world situations where the examinees tell listeners about what they have read or heard. On the test sheet, in order to ensure maximum authenticity, we described a test-taking situation in which examinees retell a story that they read to their teacher and classmates (see Appendix). As for the function of stating their opinions about the story, we seek to measure their ability to generate their opinions in their own words, which is also essential in real-world situations. These dual functions assigned in the test would produce positive washback effects, as the learners are encouraged to acquire the ability to convey the message clearly and express their thoughts in their own words (e.g., Wittrock, 1974, 1990).

Strengths and Weaknesses of SRST

The SRST has several advantages as compared to face-to-face interview tests or other currently available speaking tests, the most important being its high practicality. At the design and operationalization stage, it is very easy for teachers to find a passage for a story retelling task and to put it into a test form. They need not create a passage or a task; rather, they only need to search for a passage in a book they have at hand. Thus, the time required to design and operationalize the test can be brief. In addition to the ease of test preparation, the difficulty of the SRST tasks can be more easily controlled by changing the level of difficulty of the text than it can through tasks that elicit a relatively natural discourse from examinees (e.g., a task asking them to talk about their favorite movies). Therefore, the SRST can be targeted at learners with various proficiency levels (Chaudron, 2003; Underhill, 1987), excluding, of course, those learners who cannot comprehend even a very basic text.

At the administration stage, the number of, and burden on, administrators is far less in the tape-mediated SRST (Underhill, 1987) than in a face-to-face interview test. Through the SRST, many examinees can be tested at the same time in a language laboratory and the same instruction tape can be used for any story. As a result, anyone who has facilities to record the examinees' performances may easily administer the test. In addition, administering the SRST normally does not cost money.

As for the rating stage, the scoring would also be simpler (Chaudron, 2003). Since the content and language of the examinee's production can be largely predicted by the story retelling task, the time required for both rater training and analyzing utterances would be shorter. With regard to the feedback, even non-native teachers may feel comfortable giving feedback on the examinee's story retelling performance by comparing it with the text. This is an important point taking into account that the majority of the teachers are non-native speakers of English, who may have to take the time to prepare a model answer if a task is open-ended, such as describing a picture or talking about a topic. Thus, the relative ease of scoring on SRST and giving feedback after the test may result in increased assessing opportunities in class and show students the importance of learning speaking ability.

The SRST also has the potential to produce some positive washback on teaching and learning. First, this text-provided task may elicit more utterances than tasks in which only a topic is provided. In particular, less proficient learners would be expected to speak for a longer time in the story retelling task (Bardovi-Harlig, 2000). Second, the teacher can include target learning points in the retelling material, such as grammatical points and key phrases (Chaudron, 2003). After the input of these points in the reading section, the students can practice using them in the retelling session. By receiving the text as a model answer immediately after such activities or the test, the students may easily find the points that they have not yet mastered and can practice by themselves. In other words, the task would show learners an achievable goal in terms of what to say or how to say. In this way, the teaching,

learning, and testing can all be closely linked, and learning can take place in a more efficient way, as Messick (1996) claims.

From the perspective of SLA, the reproduction can encourage learners to pay more attention to local and global information in the story (Kai, 2008), story structures (Gambrell, Kapinus, & Koskinen, 1991), and language form (Yoshimura, 2006). In addition, because reproduction opportunities strengthen meaning retrieval processes, they tend to enhance the ability to read fluently or comprehend speech delivered at a natural speed with sufficient accuracy of text comprehension (Hirai, 2001, 2005). The most powerful explanation of the effect of production opportunities has been provided by Swain (1985, 1995), who hypothesizes that learners who are pushed to produce concise and appropriate output are forced to move from "semantic processing," which is prevalent in comprehension, to the more "syntactic processing" needed for SLA. In other words, tasks such as "read and then recall or retell" demand "deeper" cognitive processing and integration of a learner's domain knowledge with the content of the text (e.g., Fincher-Kiefer, Post, Green, & Voss, 1988) and, thus, are thought to result in more effective learning (Craik & Tulving, 1975; Joe, 1995; Swain, 1985).

This deeper cognitive processing would be more actively elicited when the examinees express their opinions about the story at the end of the retelling session, because this generative processing, where target vocabulary items are likely to be used in new or similar contexts, requires such deeper processing (Wittrock, 1974, 1990). Activities that enhance generative processing are highly recommended to Japanese students, since the ability to develop this process has been found to be missing. The National Institute for Educational Policy Research of Japan (2007) conducted a wide-scale investigation of speaking abilities at randomly selected lower secondary schools. It was revealed that learners tend to have a low ability to produce extended talk and that they are relatively weak at a monologue task such as expressing what they know or stating their opinions in a limited amount of time. Moreover, it was reported that their speaking performances were correlated with the number of opportunities that they had practiced speaking in and outside of the classroom and had their speaking performances evaluated. This suggests the benefits of administering the SRST-type speaking test frequently in class.

Although the SRST has many advantages, we should also keep in mind its weaknesses. First, due to the passage dependency of the SRST, the test takers' memory capacity might affect their speaking performance. As a text gets longer, the amount of information to be processed (i.e., the memory load) is usually larger. However, Alderson (2000) states that the memory load is smaller if the reproduction is done immediately after the reading. Thus, at this preliminary administration stage, we used four stories of two different lengths and asked examinees to identify which of the four stories were easy to retell and why.

Second, since text comprehension may affect the speaking performance that follows (Underhill, 1987), the effects of reading texts in the SRST should be as small as possible. Therefore, reading comprehension questions that need to be answered are included after a silent reading of the text in order to see if learners understand the text sufficiently. If the text reading is fairly easy and its topic is easy to follow, the learners' reading ability and topic preference may have less effect on their speaking performance.

Third, after reading a text and answering the comprehension questions, the students are likely to use and learn the words and phrases in the text. This might be regarded as a weakness because during the speaking performance, the SRST may assess words and phrases that the test takers have newly learned during the reading section of the test along with those they already knew before the test. However, we regard this rather as one of the advantages we intended to have in the SRST. In a similar vein, Alderson (2004) states, in the context of the diagnostic test that he helped develop (DIALANG), that when examinees can take the test multiple times, the test results obtained later may be "a better indication, since it results from the users' having learned something about their first performance" (pp. 10–11).

Overall, when considering the nature of classroom-based testing, the advantages of the SRST seem to overweigh the disadvantages, and the present test development is significant because a speaking test with the use of story retelling tasks has not been fully developed, despite the many advantages mentioned above.

Administering SRST

For the purposes of (1) confirming the aforementioned intentions and (2) examining its test qualities, the SRST was administered to learners with low to intermediate levels of proficiency. After the test, a questionnaire was administered eliciting feedback from the examinees on the test. We believed that their responses toward the test should be taken into consideration at the initial test development stage, because they are one of the major stakeholders and tests are likely to have low validity if they do not take the test seriously (Alderson, Clapham, & Wall, 1995; Fulcher, 1999). The perception of the test takers was examined specifically from the following perspectives: (a) construct, (b) impact, (c) interactiveness, (d) appropriateness of procedures, and (e) appropriateness of task difficulty and test materials. Positive responses were considered to be evidence of some aspects of test usefulness.

Participants

Forty-three EFL learners from two Japanese universities took the SRST. We considered our participants' proficiency levels as low to intermediate based on their estimated productive vocabulary size, as calculated by two types of tests: an aural spelling and meaning

test (30 items; Hirai, 2005) and a written vocabulary test (78 items; Koizumi, 2005). It was found that their productive vocabulary size ranged from 694 to 2,417 words (see Table 1). Since the Course of Study (Ministry of Education, Science & Culture, 1999a, 1999b) stipulates that 900 words must be mastered in lower secondary school and 2,700 words—including the previous 900—in upper secondary school (Aizawa et al., 2007), the present participants were estimated to be within the secondary school levels and, thus, not at advanced levels.

Table 1 is here

Materials

Story Retelling Speaking Test (SRST). Four stories were used for the implementation of this pilot test. Two of the stories were approximately 100 words long, and the other two were 150 words long.[2] All four texts were selected from either the Reading or Interview Sections of past Eiken tests (Society for Testing English Proficiency, 2008; see Table 2). The Eiken test is the most widely used proficiency test in Japan, and both Grades 3 and 4 are used at the lower secondary school levels for those who have studied English for two to three years. The degree of text difficulty was nearly the same for the four stories.

In the reading section of the SRST, three comprehension questions were prepared for each story. In the retelling section, four words, either proper nouns or keywords, were selected from each story as prompts in order to help learners recall the content of the story when retelling. The English instruction read by a native speaker was recorded on tape, which was used in administering the test. In case the participants did not follow the tape-recorded instruction, both Japanese and English instructions were printed on the test booklet (see the Appendix for the English version).

Table 2 is here

Questionnaire. In order to examine the perceptions of examinees with regard to the SRST, a three-part questionnaire was prepared. Part 1 inquires about the examinees' experience of living abroad; it also asks them to list any English qualifications in which they may have gained certification. Parts 2 and 3 aim to gather evidence regarding the test qualities (see Table 3). In Part 2, the examinees answered each question using a five-point Likert scale ranging from strongly disagree (1) to strongly agree (5), while in Part 3, they could write their opinions freely.

Procedure

The experiment was conducted in language laboratories, with examinees wearing headsets with microphones to record their performance. Each examinee received a booklet with the instructions on the front page and the four stories on the following pages. The order of the stories was counterbalanced, so that examinees sitting next to each other started reading different stories. The instructor explained the purpose of the SRST and its procedures in Japanese.

The tape-recorded SRST then began. The examinees proceeded to the next page and silently read one of the stories within two minutes. Next, they answered the three comprehension questions orally while referring to the story. The examinees were instructed to first read the question aloud and answer it within 30 seconds; in order to include time for both reading the question and answering it, a 40-second pause was inserted for each question. After the reading section was over, the examinees were directed to turn the sheet over and retell as much of the story as possible in two minutes, including their opinions about the story at the end of their narration. During the retelling, the examinees were permitted to look at the four keywords. A stopwatch was displayed on the monitor in order to help them utilize the two-minute time limit optimally. This procedure was repeated for each story. After all the stories were retold, the questionnaire was administered to the examinees.

Results and Discussion

Analysis of Questionnaire

The questionnaire was analyzed and the test qualities of the SRST were examined one at a time. The passages had not been read by the participants prior to this experiment (Q2-23). Although one examinee had lived abroad for three years, his performance in the SRST was not exceptionally superior. Thus, we simultaneously analyzed the answers of all 43 examinees.

Table 3 is here

Construct. Four questions were prepared for gauging the examinees' responses to the construct of the SRST (i.e., face validity). As shown in Table 3, Q2-2 and Q2-3 concern the reading comprehension task, which was mainly intended to measure the ability to grasp both the main idea and details, while Q2-7 and Q2-8 were related to the story retelling task primarily constructed to measure speaking ability. If the test takers felt as we had expected, Q2-2 should receive a higher score than Q2-3, and Q2-8 should obtain a higher score than Q2-7.

A repeated-measures ANOVA ($F(2.285, 42) = 24.91$, $p < .001$, $\eta_p^2 = 0.37$) followed by the Bonferroni post hoc test among these four questions revealed that the scores on Q2-8

(M = 4.51) were significantly higher than those on Q2-7 (M = 2.67) at $p < .001$, and the effect size of the difference was large (g = 1.75)[3]. Thus, the fact that the test takers felt that the story retelling section measured their speaking ability more than their reading ability was positive evidence for the face validity of the SRST. On the other hand, the scores on Q2-2 (M = 3.56) and Q2-3 (M = 3.58) were nearly the same, and the difference was not significant at $p < .05$; further, the effect size of this difference was negligible (g = −0.02). Thus, the examinees felt that the reading section of the SRST measured both reading and speaking abilities, probably because the reading section requires them to read questions and answers aloud.

Impact. In order to observe whether the SRST had a positive impact on the examinees' learning, we prepared three questions: Q2-20 ("Was the test beneficial to your English study?"), Q2-21 ("Were you motivated to study English after taking the test?"), and Q2-22 ("Do you think you could improve your English if you frequently had these types of speaking activity?"). The mean responses to all these questions were relatively high at 3.51, 3.77, and 3.53, respectively.

Moreover, all the 27 examinees who agreed or strongly agreed on Q2-22 wrote that they could improve their speaking ability and/or summary skills if they frequently had such practice. Among them, three examinees mentioned that the test would help them increase their vocabulary in addition to their speaking ability. This is also a piece of evidence for our intended construct. Although the present SRST allowed students to choose between retelling the original story and paraphrasing it, the exact memorization of the passage was limited and after a few sentences at the most, the examinees were obliged to paraphrase or summarize the content that they comprehended. Thus, assigning them the task of reproducing the story seems to help them to pay close attention not only to the language (i.e., form) but also the content of the story, and to enhance their paraphrasing or summarizing ability.

Furthermore, four learners commented in Q3-4 or Q3-5 that they had not taken a speaking test of this kind and that it was very interesting and beneficial. One student wished to receive the feedback on his speaking performance and another suggested that it would be beneficial to include items in a text that would help students to learn. Thus, we have confirmed that, in terms of the test takers' perceptions, the SRST gave positive washback effects on their speaking ability and motivation to learn English, as was originally estimated.

Interactiveness. The evidence of interactiveness was obtained from the analysis of the examinees' affective conditions. Two questions relevant to this quality were Q2-17 ("Were you nervous or anxious during the test?") and Q2-19 ("Did you try to speak a lot during the test?"). It was observed that most examinees seem to have been somewhat tense during the test (M = 3.58) and tried to speak as much as possible (M = 3.95). Thus, the SRST can provide the evidence regarding interactiveness (Bachman & Palmer, 1996), namely, that the procedure allowed them to be moderately anxious and motivated.

Appropriateness of procedures. The test takers stated that the tape-recorded instruction was sufficiently loud (M = 4.53 for Q2-14) and clear (M = 4.30 for Q2-15) and that it was easy to record their voice on the tape (M = 4.30 for Q2-16). In addition, even though all examinees spoke simultaneously in the classroom, most of them were not disturbed by the surrounding noise as they had their headphones on (M = 3.26 for Q2-18). This suggests that the tape-mediated SRST can be administered to class members at the same time without interfering with their performance. From these results, the procedures of the SRST were found to be appropriate and practical, and seemed to elicit reliable performances from the examinees.

In addition, the procedure of the retelling section was confirmed to be reasonable (from Q2-9 to Q2-13); in particular, the responses to Q2-10 ("Was it difficult to remember the content of the story you have read?") seem to show that the effect of memory on retelling performance was not very strong (M = 3.56).

However, two points should be raised for reconsideration from the responses to Q3-2. One examinee pointed out that she could not do well on the very first story because she was not sure if what she was doing was right. Four examinees felt that they found the allotted time of 2 minutes to be insufficient for including their opinions. Therefore, we have decided to insert a practice story at the beginning of the test in order to prolong the retelling section from 2 minutes to 2 minutes and 30 seconds, and to insert a small sound after 2 minutes in order to let examinees know the time left for expressing their opinions. Another revision was suggested in the responses to Q2-4 and Q2-5 concerning the time allocation the reading section. The time allotted for answering comprehension questions was a little too long (M = 4.35 for Q2-5), and most of the examinees left more than 10 seconds unused in completing each question. Thus, we decided to shorten the response time by 10 seconds—from 40 to 30 seconds.

Appropriateness of task difficulty and test materials. In order to gauge the perception of the test takers with regard to the task difficulty, the scores of Q2-1 and Q2-6 were compared. The results suggested that the test takers felt that they did much better in answering the comprehension questions on the text they read (M = 3.33) than in retelling it (M = 1.84; paired $t(42) = 7.13$, $p < .001$, $g = 1.27$). This result indicates that the difficulty of the speaking task may not directly arise from the passage difficulty but rather from the nature of the speaking task itself. Therefore, we concluded that the level of task difficulty was appropriate. In fact, in responses to Q3-3, in which test takers were asked to note down the difficulties of the task, most of the answers concerned the difficulties of the retelling section; and none of the examinees claimed that the texts were difficult to read. Thirteen examinees had difficulties retelling the story in English, not because they did not remember the content,

but because they could not immediately find the right words for what they wanted to say or because they had difficulty in constructing grammatically correct sentences. Another 10 examinees felt that stating their opinions was difficult because they were not used to expressing their opinions in English or because they could not find anything to say about "such an unimpressive short story." Regarding this point, we may have to prepare some questions on the test sheet that prompt the examinees' opinions in order to encourage opinions about such a short story.

Table 4 is here

Concerning the appropriateness of the texts for SRST (Q3-1), the examinees chose easy stories to retell and mentioned reasons for their choice (see Table 4). On the whole, 30 examinees (69.77%) found either Story 1 or 3 (both short) easier, whereas 11 examinees (25.58%) found either Story 2 or 4 (both long) easier, and seven examinees (16.28%) did not show any preferences about the stories. Among the 16 examinees who chose Story 1, the strongest reason was its content (n = 12), followed by its length (n = 4). A similar tendency was found among those who chose Story 3 (short). On the other hand, among those who chose a longer story—either Story 2 or 4—the strongest reason was the content of the story (n = 5), followed by the familiarity of its topic (n = 2). Thus, the factors that affect speech reproduction were mainly the content (or topic) of the story, followed by its length.
However, the factors of content and length seem to be interrelated on the basis of the fact that about 70% of the examinees chose one of the short stories, even though they chose its content, not length, as the reason for their choice. This may be because the plot of a short story tends to be simpler and contains less detailed descriptions; consequently, it is easier to retell. In this regard, both the content and length of a story would influence the examinees' perceived difficulty of story retelling. Taking these results into consideration, we may use texts of the current length (i.e., from 100 to 150 words) for beginning to intermediate level learners, but at least two texts—a short and a long one—may be necessary for the SRST in order to alleviate the content and length effects.

Conclusion and Plans for Future Research

This paper presented efforts to develop the SRST, a new test for classroom speaking assessment. Given most classroom situations in Japan, where input may not lead to intake owing to few opportunities that enable the learners to use what they have comprehended, the speaking test that employs a story retelling technique will be useful since it attempts to connect input to output directly in the form of speaking assessment. Through the test design and implementation stages, the SRST, so far, seems to have satisfied the two main qualities of practicality and positive impact on teaching and learning, as had been intended. The results

of the questionnaire also drew evidence of perceived construct, interactiveness, and positive impact on the examinees' learning and motivation to enhance speaking ability, especially oral paraphrase and summary abilities. Furthermore, it was confirmed that the test administration procedures and level of task difficulty were appropriate. However, some revisions were found to be necessary, such as (a) inserting a practice test at the beginning of the test, (b) adding some questions to prompt opinions of a story, (c) reducing the time allotted for answering each comprehension question, and (d) providing more time for retelling.

In terms of the text materials, it was found that the most influential factor was the content of a text, followed by the text length; however, the two factors seemed to be interrelated. In other words, the examinees had a tendency to perceive that the plot of a shorter text was simpler and was thus easier to retell. From these results, we reached the conclusion that using texts of different length for the SRST would lead to draw more reliable performances from the examinees.

Because the test development is still in the preliminary stage, further validation studies are necessary before it is put to use. A validation study we urgently need to do is to establish a practical and valid scoring system. One practical scoring system might be an EBB (Empirically derived, Binary-choice, Boundary-definition) scale (Turner & Upshur, 1996; 2002). Thus, we plan to develop an EBB scale which consists of four criteria and compare it with an existing analytic scale that has similar criteria. Then, we will ask raters which one is easier to use. In order to investigate intra- and inter-rater reliabilities and to determine how many stories are needed to achieve high reliability, we are planning to use a many-faceted Rasch measurement model (Linacre, 1989) and multivariate generalizability theory (Brennan, 2001). To collect evidence of the construct validity of the criteria, we may adopt a multitrait-multimethod (MTMM) approach and examine to what extent the EBB scale is related to the existing analytic scale.

In another validation study, we are planning to examine the usefulness of the SRST more thoroughly from various perspectives—considering particularly the aspects of impact, validity, and reliability. Although the SRST is a school-based monologic test, it will be necessary to compare it with other acknowledged speaking tests. These tests should include one that has two-way interaction type tasks, such as the Standard Speaking Test (SST; a modified Oral Proficiency Interview; ALC Press, 2008a) or the Eiken interview test (Society for Testing English Proficiency, 2008), and one that has similar one-way interaction type tasks, such as the Versant test (Pearson Education, 2008) or the Telephone Standard Speaking Test (TSST; ALC Press, 2008b). Comparing scores obtained by these different speaking tests using Pearson correlations, we may obtain evidence of the concurrent validity of the SRST. It would be interesting to interpret how much variance in the SRST can be explained by these already-established tests and how much unique variance is left. Such an analysis may suggest strengths, weaknesses, and characteristics of the SRST.

Finally, although we developed the SRST for novice and intermediate learners in this study, we can alter and use it for more advanced learners by employing more difficult texts or changing the text modes from the present reading version to a listening version. To this end, we need to further investigate the influence of factors such as text difficulty and text mode on the speaking performance.

Note 1. JACET8000 is a word list compiled from textbooks used at many secondary schools, newspapers, and books (JACET Basic Words Revision Committee, 2003).

Note 2. The length of the text for this experiment (100 and 150 words) was estimated based on the previous literature. According to Hirai (2001), the average reading speed of low and intermediate level learners who read an easy text with sufficient comprehension was 85.21 words per minute with the standard deviation (*SD*) of 28.37 (p. 64). In this respect, a text with 113.68 words (= [$M - SD$] × 2 minutes = [85.21− 28.37] × 2) could reasonably be read by most of the examinees in two minutes. Thus, we prepared two 100- and two 150-word-length texts for this study in order to determine an appropriate length for the story retelling text.

Note 3. An effect size, Hedges' *g* (Hedges & Olkin, 1985), was calculated as $g = (M_1 - M_2)$ / *Pooled SD* = (4.51 − 2.67) / 1.05 = 1.75, using the Effect Size Calculator (Curriculum, Evaluation and Management Centre, 2006). Hedges' *g* was interpreted based on |0.2| as a small effect, |0.5| as medium, and |0.8| as large (Cohen, 1988, pp.24–27).

REFERENCES

Aizawa, K., Murata, M., Uemura, T., Mochizuki, M., Tono, Y., Sugimori, N., et al. (2007). *Construction of a vocabulary list for Japanese learners of English and development of a system for analyzing educational materials based on large-scale corpora.* Research Project, Grant-in-Aid for Scientific Research on Priority (B), 2004-2006, Project number: 16320076.

Akiyama, T. (2004). *Introducing EFL speaking tests into a Japanese senior high school entrance examination.* Unpublished Ph.D. dissertation, University of Melbourne.

ALC Press. (2008a). *Standard Speaking Test (SST).* Retrieved December 13, 2008, from http://www.alc.co.jp/edusys/sst/english.html

ALC Press. (2008b). *TSST.* Retrieved December 13, 2008, from http://tsst.alc.co.jp/tsst/e/index.html

Alderson, J. C. (2000). *Assessing reading.* Cambridge University Press.

Alderson, C. (2004). The shape of things to come: Will it be the normal distribution?. In M. Milanovic & C. Weir (Eds.), *Studies in language testing 18: European*

language testing in a global context: Proceedings of the ALTE Barcelona Conference July 2001 (pp. 1–26). Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

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

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing 13*, 257-279.

Bardovi-Harlig, K. (2000). Tense and aspect in second language acquisition: Form, meaning, and use. *Language Learning, 50* (supplement 1), 1–461.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Chaudron, C. (2003). Data collection in SLA research. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 762–828). Malden, MA: Blackwell.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng & Y.Watanabe with Curtis, A (Eds). (2004). *Washback in language testing: Research contexts and methods.* (pp. 147-179). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294.

Curriculum, Evaluation and Management Centre, Durham University. (2006). *Effect size calculator*. Retrieved March 6, 2008 from http://www.cemcentre.org/renderpage.asp?linkID=30325017

Donin, J., & Silva, M. (1993). The relationship between first- and second-language reading comprehension of occupation-specific text. *Language Learning, 43,* 373–401.

English Educational Foundation of Japan. (1992). *Jitsuyo eigo gino kentei 3 kyu zen mondai shu* [Test in Practical English Proficiency: All test items in Grade 3]. Tokyo: Author.

Fincher-Kiefer, R., Post, T. A., Green, T. R., & Voss, J. F. (1988). On the role of prior knowledge and task demands in the processing of text. *Journal of Memory and Language, 27,* 416–429.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics, 20,* 221–236.

Fulcher, G. (2003). *Testing second language speaking.* Essex, U.K.: Pearson Education

Limited.

Gambrell, L. B., Kapinus, B. A., & Koskinen, P. S. (1991). Retelling and reading comprehension of proficient and less-proficient readers. *Journal of Educational Research, 84*, 356–362.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hirai, A. (2001). *The relationship between listening and reading rates: A comparative study of Japanese and international EFL learners.* (Doctoral dissertation, Temple University, Japan, 2001). (UMI No.3031531)

Hirai, A. (2005). Factors predicting EFL learners' listening and reading fluency. *JACET Bulletin, 41*, 19–36.

Honda, T. (2007). *Assessment of speaking performance in Japanese junior high school EFL classes--Task types and good task combinations--*. Unpublished master's thesis, Tokyo Gakugei University, Japan.

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

Japan Association of College English Teachers (JACET) Basic Word Revision Committee (Ed.). (2003). *JACET List of 8000 Basic Words.* Tokyo: Author.

Joe, A. (1995). Text-based tasks and incidental vocabulary learning. *Second Language Research, 11*(2), 149–158.

Kai, A. (2008). The effects of retelling on narrative comprehension: Focusing on learners' L2 proficiency and the importance of text information. *ARELE (Annual Review of English Language Education in Japan), 19,* 21–30.

Koizumi, R. (2005). *Relationships between productive vocabulary knowledge and speaking performance of Japanese learners of English at the novice level.* Unpublished Ph.D. dissertation, University of Tsukuba, Japan.

Kosuge, A., & Kosuge, K. (1995). *Speaking no shido* [Teaching speaking]. Tokyo: Kenkyusha.

Lee, J. F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition, 8,* 201–212.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

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

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA: Blackwell.

Messick, S. (1996). Validity and washback in language testing. *Language Testing 13*, 241-256.

Ministry of Education, Science & Culture. (1999a). *Chugakko gakushu shido yoryo (heisei 10 nen 12 gatsu) kaisetsu—gaikokugo hen--* [Explanation of the Course of Study for junior high school concerning foreign languages]. Tokyo shoseki.

Ministry of Education, Science & Culture. (1999b). *Kotogakko gakushu shido yoryo*

kaisetsu—gaikokugo hen—eigo hen [Explanation of the Course of Study for senior high school concerning foreign languages with a special focus on English]. Tokyo: Kairyudo.

Morrow, K. (1991). Evaluating communicative tests. In S. Anivan (Ed.), *Current developments in language testing. Anthology Series 25.* Singapore: Regional Language Centre (pp. 111-118).

National Institute for Educational Policy Research of Japan. (2007). *The investigation on the special project 'English speaking.'* Retrieved February 21, 2008 from http://www.nier.go.jp/kaihatsu/tokutei_eigo/05002051033004000.pdf

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly, 42*, 133-138.

O'Loughlin, K. (2001). *Studies in language testing 13: The equivalence of direct and semi-direct speaking tests.* Cambridge University Press.

Pearson Education. (2008). *Versant*™. Retrieved December 4, 2008, from http://www.ordinate.com/

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*, 99–123.

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

Society for Testing English Proficiency. (2008). *Eiken: Test in Practical English Proficiency.* Retrieved December 13, 2008, from http://stepeiken.org/

Swain, M. (1985). Communicative competence: Some roles of comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition*. Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honor of H .G. Widdowson* (pp. 125–144). Oxford University Press.

Turner, C. E., & Upshur, J. A. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth, & C. Elder (Eds.). *The language testing cycle: From inception to washback* (pp. 55-79). Australia: Applied Linguistics Association of Australia.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36,* 49-70.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques.* Cambridge University Press.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*, 334-354.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10,* 41-69.

Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng & Y.Watanabe with Curtis, A (Eds). (2004). *Washback in language testing: Research contexts and methods.* (pp. 129-143). Mahwah, NJ: Lawrence Erlbaum Associates.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11,* 87–95.

Wittrock, M. C. (1990). Generative process of comprehension. *Educational Psychologist, 24,* 345–376.

Yoshimura, F. (2006). Does manipulating foreknowledge of output tasks lead to differences in reading behaviour, text comprehension and noticing of language form? *Language Teaching Research, 10,* 419–434.

Zen mondai & kaito 2001 nendo dai 2 kai kentei: Ichiji shiken [All test items and answers at the second Eiken test in the academic year of 2001: First stage]. (2001). *STEP the Latest on English, 21*(Supplement), 1–74.

Zen mondai & kaito 2001 nendo dai 3 kai kentei: Ichiji shiken [All test items and answers at the third Eiken test in the academic year of 2001: First stage]. (2002). *STEP the Latest on English, 24*(Supplement), 1–74.

Appendix

English Version of Instructions for SRST and Story 1

Story Retelling Speaking Test

- Directions: This is a test to measure your speaking ability in English. The test has the following three steps. First, you will read a story silently for two minutes and be ready to answer three questions. Do not take notes. Next, you will read each question aloud and answer it within thirty seconds. Finally, you will retell as much of the story as possible and then include your opinions about the story within two minutes. You may use keywords written on the test sheet. Even if you have not finished within two minutes, you will not lose points.

- Situation and scoring:

  Imagine the following situation. In class, you are going to read a story written in English and retell its content to your classmates orally. The audience is your teacher and classmates. Your oral presentation will be evaluated based on (1) whether you can answer the comprehension questions correctly, (2) how much of the story you can convey clearly, and (3) whether you can add your opinions about the story at the end of your speech. Concerning correctly answering comprehension questions, as far as your

answer is appropriate and natural as a response to the question, it will be regarded as correct.

Now, let's begin. First, look at the next sheet.

----------<Next page>----------

Read the story silently within two minutes.

Story 1

Kenji goes to school by train. One morning he was very sleepy. After he left the station, he remembered that he left his bag in the train. Some textbooks, a box lunch and a dictionary were in the bag. At school he telephoned the lost-and-found office of the station to ask about the bag. But "We don't have your bag" was the answer. He was shocked. He returned home and told his mother about it. His mother said, "You are lucky. A kind man brought your bag to the house. He found your name and address on it."

After the signal, read each question aloud and answer it in English.

Q1: Where did Kenji leave his bag?

Q2: What was there in the bag?

Q3: Why was Kenji lucky?

----------<Next page>----------

Retell as much of the story as you can in English in two and half minutes. You can look at the keywords while you are retelling. At the end of your retelling, be sure to include your opinions about the story.

Keywords:

Kenji, train, bag, mother