Comparing the Story Retelling Speaking Test With Other Speaking Tests

Rie Koizumi
Tokiwa University

Akiyo Hirai
University of Tsukuba

This study examines the validity of score-based interpretation of the Story Retelling Speaking Test (SRST) in comparison with the Versant (Pearson Education, 2008) and Standard Speaking Test (SST; ALC Press, 2010). In total, 64 participants took the three tests; their speaking functions, scores, and utterances were analyzed to probe the shared and varied aspects of the tests. The results showed that the SRST elicited more functions than the Versant but fewer than the SST, that it was moderately related to the latter two tests, and that it more successfully discriminated among a group of beginner and intermediate level learners. Additionally, the results suggested that (a) the tasks and speaking functions and (b) the aspects emphasized while rating may explain the differences in the test scores for the three tests. Based on the results, comparative advantages of each test were summarized, which may be useful for selecting appropriate speaking tests according to assessment purposes and situations.

本研究では、Story Retelling Speaking Test (SRST) を、Versant (Pearson Education, 2008) と Standard Speaking Test (SST; ALC Press, 2010) と比較することで 、SRSTの得点に基づく解釈の妥当性を吟味する。64名の受験者に3つのテストを受けてもらい、テストの共通点と相違点を調べるために、その言語機能と得点、発話を分析した。その結果、SRSTはVersantよりは多いがSSTよりは少ない言語機能を引き出すこと、SRSTは他の2つのテストと中程度の相

関を持ち、初級者・中級者グループで弁別力を発揮すること、(a)タスクと言語機能と(b)評価時に重きを置く要素の相違により、テスト得点の違いが説明されうることが分かった。結果に基づき、評価の目的と状況に沿って適切なスピーキングテストを選ぶ際に有益となる、各テストの相対的な利点を示した。

One difficulty related to speaking tests is ensuring that their administration and scoring are sufficiently practical. This seems especially true when tests are undertaken for formative and summative classroom assessment. While teachers can observe students' class performance in pair and group interactive activities as well as speech, discussion, and other presentation activities, speaking tests are needed to grasp students' achievement and proficiency in relation to speaking ability (Genesee & Upshur, 1996). A classroom speaking assessment can take a direct (or live) test format, such as one-on-one interviews with a teacher and interactions with a partner or group members; however, difficulties may arise as to securing interviewers and having time for such direct testing. When equipment for recording students' voices is available, employing a semi-direct (or tape-mediated) format becomes a viable alternative to direct speaking assessment in which "the stimulus is pre-recorded or text based, and the response by the candidate is recorded for distance rating" (Davies et al., 1999, p. 178). One example of semi-direct tests is the Telephone Standard Speaking Test (TSST), in which test-takers talk about their experiences, describe objects, and compare two objects through telephone (ALC Press, 2008). Another example is the speaking component of the TOEIC® (Test of English for International Communication) Speaking and Writing Test; test-takers read a text aloud into a computer microphone, describe pictures, answer questions, propose solutions, and express opinions (Educational Testing Service, 2011). These examples illustrate that semi-direct speaking tests adopt several tasks to elicit various types of performance from test-takers. However, a semi-direct task that has hitherto been underutilized is story retelling, in which test-takers retell a passage that they have just read or heard. This integrated speaking activity simulates natural speech in real-life situations.

A tape-mediated Story Retelling Speaking Test (SRST) was developed for Japanese learners of English as a practical resource for classroom use to assess speaking ability, especially the ability to produce extended spoken monologues (Hirai & Koizumi, 2009). In the test, students read a story silently, then retell the story, and express their opinions about it while looking only at keywords (see Figure 1 for the procedure and Appendix A for instructions and a story sample). The administration of the SRST with

the test instructions and one story takes about 8 minutes. Utterances are recorded and rated using an empirically derived, binary-choice, boundary-definition (EBB) rating scale to assess three criteria with five levels each: Communicative Efficiency (CE; including fluency, coherency, elaboration, adequacy of story-telling capability, and aptness of test-takers' opinion of the story), Grammar & Vocabulary (G&V), and Pronunciation (see Appendix B for the EBB rating scale). The descriptors of the scale were empirically derived on the basis of previous literature (e.g., Upshur & Turner, 1995), and included the salient aspects of students' actual speech delivery that separate higher and lower levels of the EBB scale (Hirai & Koizumi, 2011). The use of three rating criteria on the EBB scale is intended to increase the diagnostic value of the score report.

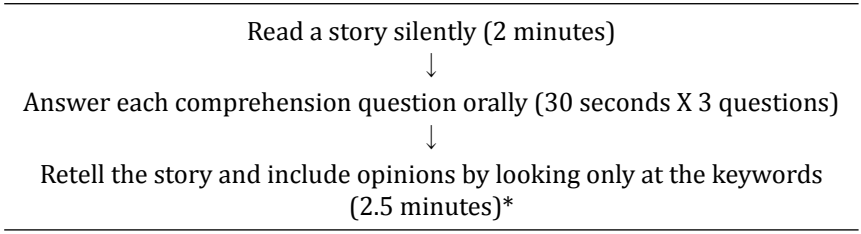


Figure 1. SRST Administration Procedure for One Story

*A beep sound is inserted after 2 minutes to inform test-takers of the time remaining (30 seconds) and when they should start expressing opinions.

The SRST is intended to have high practicality for test construction and administration: Teachers can convert lesson material into a test passage and the test procedure can be standardized with the recorded instructions. Although the test task is limited to story retelling and opinion statement, these skills are worth teaching and testing since L2 learners, especially Japanese learners of English, often lack skills in expressing their knowledge and opinions (National Institute for Educational Policy Research of Japan, 2007).

Previous studies (Hirai & Koizumi, 2009, 2011; Koizumi & Hirai, 2010) have examined test qualities of the SRST and shown evidence of its validity and usefulness. Hirai and Koizumi (2009) conducted a survey analysis and confirmed that test-takers generally felt that the test procedures and task difficulty were appropriate. In another study (Koizumi & Hirai, 2010), the effectiveness of the SRST components (e.g., keywords and opinions) was demonstrated by scrutinizing examinees' performances. For example, the

effect of text length on volume produced was found to be inconsistent and small, which suggests that memory has only a slight impact on SRST performance. Hirai and Koizumi (2011) compared two empirically developed rating scales (i.e., EBB vs. multiple trait) and demonstrated that the EBB scale has the more desirable characteristics of requiring fewer stories to maintain sufficient reliability (.70 or above) and of showing stronger discrimination. However, concerns regarding the validity of interpretation and use of the SRST scores remain. For example, does the SRST measure speaking ability similar to the range of skills assessed by other speaking tests? Although each speaking test is designed to meet purposes and situations in local contexts with varying operationalization of constructs and task characteristics, it is reasonable to assume that some aspects and constructs are commonly measured across tests and thus correlate between them. We will examine this question in this paper.

Relationships between new tests and fairly well-established tests (external criteria) have been examined as part of validation processes (e.g., Messick, 1996). When tests are thought to assess similar abilities, moderate or high correlations are considered concurrent evidence for validity concerning new tests. One such investigation was done by Bernstein, Van Moere, and Cheng (2010), who reported strong relationships between the Versant™ tests and oral interview tests in Spanish, Dutch, Arabic, and English as a second language (L2). For instance, among 130 L2 English learners in Iran, correlations were high between the Versant English and the International English Language Testing System (IELTS; r = .77), the Versant English and the Test of English as a Foreign Language Internet-based Test (TOEFL iBT; r = .75), and the IELTS and the TOEFL iBT (r = .73). They argued that these strong relationships between the Versant™ tests and other tests suggest high validity of interpretation based on the Versant™ test scores. Concurrent validation often attracts criticism: The external criteria tests are often presumed to have perfect validity, which of course they do not (e.g., Bachman, 1990). However, this method is considered appropriate when the result is regarded as just one example of validity evidence, and when this method is used together with other methods for accumulating validity evidence.

Comprehensive test validation requires the demonstration of theoretical and empirical evidence (e.g., Messick, 1996). According to Chapelle, Enright, and Jamieson (2008), while theoretical evidence is usually obtained by describing (a) the importance of the target domain and the relevance and representativeness of the tasks, empirical evidence can be collected by investigating the following: (b1) the appropriateness of the rating scale and

the statistical properties of tasks and ratings; (b2) the reliability of the test and usefulness of the test specifications; (b3) consistency between actual test-taking processes and test developers' intentions, agreement between the difficulty order and the predicted order of the test tasks, and reasonable correlations between the target test and other tests assessing similar or different constructs; (b4) sound relationships between the target test and indicators of ability or real-life performance that the test scores are intended to predict (e.g., speaking ability or real-life performance); and (b5) the meaningfulness of test scores, the feedback for test users (e.g., teachers and test-takers), and the test's beneficial washback on intended aspects such as learning and teaching. Chapelle et al. demonstrated how evidence regarding (a) to (b5) was gathered for their validity argument for the TOEFL iBT using the argument-based approach to validity. Previous studies of the SRST (Hirai & Koizumi, 2009, 2011; Koizumi & Hirai, 2010) covered (a) to (b2). Further, the current study contributes to (a) by comparing the speaking functions (e.g., expressing opinions) elicited from the SRST versus those from other tests (Versant™ English Test and Standard Speaking Test; hereinafter, Versant and SST) in the discussion of the first research question (RQ1, below). Additionally, it aims to contribute to (b3) through comparison with the Versant and to (b4) through comparison with the SST (see RQ2 to RQ4, below).

Current Study

This study compares the SRST with two other speaking tests, the Versant and SST (see the Method section for details) to examine the validity of score-based interpretation of the SRST. The Versant and SST were selected because they have been thoroughly investigated with multiple sources of validity evidence reported (e.g., Nakano, 2002; Pearson Education, 2008) and are now used fairly widely in Japan. Moreover, the SRST, Versant, and SST seem to measure similar aspects of speaking ability.

This study investigates the similarities and differences in the three tests using multiple analytical methods. It should enable test users to grasp the strengths and weaknesses of each speaking test and to select appropriate speaking tests that are relevant to their purpose or situation. Four research questions are addressed:

RQ1: How do the speaking functions elicited by the SRST compare with those elicited by the Versant and the SST?

RQ2: To what extent are SRST scores related to Versant and SST scores?

RQ3: Are there differences in score distributions of the three tests between two groups: beginner and intermediate level learners combined versus higher proficiency level learners?

RQ4: What factors contribute to differences in the scores of the three tests?

For RQ2, given similar and differing test constructs and formats, correlations between the three speaking tests are expected to be moderate.

Method

Participants

Participants were 64 L2 learners of English, consisting of 40 undergraduates and 24 postgraduates from three universities in Japan (28 males, 36 females). Most were between 18 and 24 years of age and were majoring in English, art, culture, or physical education. The participants included 62 students from Japan and one each from China and France.

To investigate RQ2, the 64 test-takers were divided into either (a) a beginner and an intermediate or (b) a higher proficiency-level group, having regard to their majors, educational qualifications, and self-reported proficiency scores. All students who were specializing in English at graduate school (n = 23) and undergraduate students who had self-reported scores of 860 or higher on the TOEIC® (n = 3) were assigned to (b). Hence, 26 test-takers were assigned to (b). The remaining 38 students were assigned to (a). Although stratifying students on the basis of scores achieved in the same test would have been better, we failed to obtain such scores for all test-takers. When we compared students who reported their TOEIC® scores, we found that group (a) had a higher mean (M = 826.92; Median = 890.00; SD = 142.27; n = 13) than group (b) (M = 595.70; Median = 570.00; SD = 124.77; n = 10), Mann-Whitney U = 14.00, Z = -3.16, exact p < .001, effect size r = -.66 (a large effect size).

Tests Used

The procedure for administering the SRST has been described. We employed three stories of similar difficulty (Flesch-Kincaid Grade Level of 4.1 to 4.6), with the story lengths being short to relatively long (94 to 153 words). They were derived from past administrations of the EIKEN (Test in Practical English Proficiency), Grades 3 and 4, and were relatively easy to comprehend. One short story was used for practice, and the other two were

used for the analysis (see Appendix A for a longer story). This number was considered acceptable because Hirai and Koizumi (2011) showed that two stories can sustain sufficient reliability. The order of the two main stories was counterbalanced. It took approximately 22 minutes for SRST test-takers to finish the retelling of the three texts.

The Versant aims to assess "facility in spoken English—that is, the ability to understand spoken English on everyday topics and to respond appropriately at a native-like conversational pace in intelligible English" (Pearson Education, 2008, p. 7). Although this test is intended to assess "the core skills that are building blocks of speaking proficiency" (Bernstein et al., 2010, p. 371), which include both listening and speaking, we focus on speaking assessment and refer to it as a type of speaking test. The Versant is a semi-direct test conducted over the telephone or computer for about 15 minutes and consists of six tasks, including answering questions and retelling stories (see Table 1). Test-takers listen, then start speaking with virtually no planning time. Their utterances are recorded and scored by a fully automated scoring system in which human rating patterns are incorporated, and test results become accessible within minutes. An overall score is derived along with subscores for Sentence Mastery, Vocabulary, Fluency, and Pronunciation. These are reported on a scale of 20 to 80.

Table 1. Structure of the Versant

Part	Task	Number of items
A	Reading: Read a sentence aloud	8
B	Repeat: Listen to a sentence and repeat it	16
C	Short Answer Questions: Listen to a general knowledge question and answer it	24
D	Sentence Builds: Listen to three groups of phrases and reorder them into an understandable sentence	10
E	Story Retelling: Listen to a story and retell it	3
F*	Open Questions: Listen to a question eliciting an opinion and state an answer	2

*Not included in the final score but the sound files are accessible to test users.

The SST is a modified version of the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI), adjusted for Japanese learners of English (ACTFL–ALC Press, 1996; ALC Press, 2010).

Compared with the OPI, the SST has tasks that are more structured and more intermediate levels (three levels for Novice, five levels for Intermediate, and one level for Advanced) based on the ACTFL Proficiency Guidelines (ALC Press, 2010). The aim of the SST is to assess "functional speaking ability" (oral proficiency) and elicit face-to-face interaction that "simulates authentic conversation" (ACTFL–ALC Press, 1996, pp. 1-3) between a certified interviewer and an interviewee. The SST is "adaptive to the perceived level of the examinee as well as his/her personal and professional interests" (ALC Press, 2010). In other words, during simulated conversation that fits the test-taker's interests, the interviewer informally evaluates the test-taker's level based on his/her responses and selects tasks appropriate to the level. For this purpose, the interviewer finds a level at which the test-taker can consistently perform well and identifies "a ceiling of proficiency through prompts designed to elicit from the candidate speech acts at a level higher than s/he has thus far demonstrated" (ACTFL–ALC Press, 1996, p. 7).

According to ALC Press, (2010), the SST is completed in 10 to 15 minutes and comprises five stages: Warm-up questions, Single picture, Role-play with the interviewer, Picture sequences, and Wind-down questions. The recorded conversation is scored by at least two qualified raters. Test-takers receive a holistic score from Levels 1 to 9 with feedback in terms of Global tasks/functions, Communication with interviewer, Text type, Accuracy, Pronunciation, and Comment from interviewer.

Procedures

The participants took three speaking tests (SRST, Versant, and SST) on the same day or on separate days within 2 weeks. We counterbalanced the order of the three tests as far as the schedule allowed, but this was only possible for some candidates.

The Versant was administered using a telephone or computer, depending on the university settings. Before the test, each examinee received an individualized test sheet with test instructions, examples of tasks, and sentences for reading aloud. They had time to read through the sheet and prepare for the test, practicing tasks alone or using examples on the website. For the SST, test-takers met individually in a room with an interviewer and took the test.

Analyses

For RQ1, a checklist of speaking functions was made on the basis of O'Sullivan, Weir, and Saville (2002). Since their final checklist (Appendix 3 of their paper) contained only language functions elicited by the UCLES (University of Cambridge Local Examinations Syndicate) Main Suite examinations, other functions that O'Sullivan et al. omitted but which are observed in real life were included and used for the present analysis. The first author judged whether each function is elicited by all the tests. As a second rater conducting an independent evaluation, the second author, who is well acquainted with the SRST, judged the SRST functions, while another rater familiar with the Versant and SST judged their functions. Inter-rater reliability of all the judgments was high (Agreement ratio = .88; Kendall's tau-b = .82, $p < .001$). After examining the reasons for divergent points, we decided to use our more reasoned judgments as final ones. The open question section of the Versant, whose performance is not scored, was not included in the judgments.

For RQ2, the scores of two stories of the SRST were averaged. To calculate the inter-rater reliability of the SRST, four raters (two English teachers at secondary and tertiary levels and two TESOL graduate students) underwent a one-hour rating training in the use of the EBB scale and benchmark performances. After the training, 16 test-takers out of the 64 (25%) were evaluated by two raters independently. Pearson product-moment correlation coefficients between the two raters' ratings were found to be relatively high ($r = .81$ for CE; $r = .78$ for G&V; $r = .74$ for Pronunciation). Thus, the rest of the test-takers' responses were scored by only one of the raters, and these scores were used for analysis (our limited resources prevented us from asking two raters to evaluate all the students). The reliability of the three SRST criteria (e.g., CE) was found to be high ($\alpha = .84$). Then, all the three rating criteria were summed to produce the total SRST scores. Two examinees failed to complete one of the two stories in the SRST; their scores were imputed using the mean values of all the rest of the examinees.

A sequential multiple regression analysis was conducted using SPSS (Version 12.0.1) to examine the proportion of variance in the SRST scores (dependent variable) explained by the other test scores (independent variables). The sample size of 64 was not very large for multivariate analyses, but it was considered acceptable to use multiple regression analysis since it exceeded the minimum sample size required ($n = 63$) when a study has two independent variables with a medium effect size of R^2, a power of .80, and an alpha level of .05 (Green, 1991).

With regard to RQ3, we made histograms of the two proficiency groups and compared the score distributions to scrutinize each test's capability of discriminating between group members.

For RQ4, we took the following three steps. First, we converted the raw scores of the three tests to standard scores to enable direct score comparisons. Second, in order to examine the test performances of participants who showed large discrepancies among the three test scores, we calculated three types of subtractions in the standard scores by calculating (a) the SRST scores minus the Versant scores, (b) the SRST scores minus the SST scores, and (c) the Versant scores minus the SST scores. While a large number of cases (94%, 180/[64*3]) had similar standard scores (within the value of -1.50 to 1.50), 12 cases (6%; n = 10) showed different standard scores (with the absolute value being more than 1.50). Lastly, we transcribed the utterances of these 12 cases when the recordings were accessible. The performances that were accessible and analyzed were those of the two stories of the SRST, the Story Retelling Task of the Versant, and the overall interview of the SST. Interpretable differences are presented in the Results section.

Results

Comparison of Functions Elicited Using the Checklist

Table 2 shows that although the tests intend to assess aspects of speaking ability, overlapping functions were limited. For example, "describing" and "paraphrasing" were the only functions constantly (as indicated by O) or mostly (as indicated by Δ) elicited by the tests; "elaborating" was elicited by the SRST and mostly by the SST but not by the Versant. No functions were elicited only by the SRST.

Table 2. Functions Elicited by the Three Tests

	Descriptions	SRST	Versant	SST
Informational functions				
Providing personal information	Give information on present circumstances, past experiences, and future plans	Δ	X	O
Expressing opinions	Express opinions	O	X	Δ

	Descriptions	SRST	Versant	SST
Elaborating	Elaborate on, or modify an opinion	O	X	Δ
Justifying opinions	Express reasons for assertions s/he had made	Δ	X	O
Comparing	Compare things/people/events	X	X	Δ
Complaining	Complain about something	X	X	Δ
Speculating	Speculate	X	X	Δ
Staging	Separate out or interpret the parts of an issue	X	X	X
Making excuses	Make excuses	X	X	Δ
Describing	Describe a sequence of events and a scene	O	O	O
Paraphrasing	Paraphrase something	O	O	Δ
Summarizing	Summarize what s/she has said	X	X	X
Suggesting	Suggest a particular idea	X	X	Δ
Expressing preferences	Express preferences	Δ	X	O
Interactional functions				
Agreeing	Agree with an assertion made by another speaker (apart from 'yeah' or nonverbal)	X	X	Δ
Disagreeing	Disagree with what another speaker says (apart from 'no' or nonverbal)	X	X	Δ
Justifying/ Providing support	Offer justification or support for a comment made by another speaker	X	X	X
Modifying	Modify arguments or comments made by other speaker or by the test-taker in response to another speaker	X	X	X
Asking for opinions	Ask for opinions	X	X	X
Persuading	Attempt to persuade another person	X	X	Δ

	Descriptions	SRST	Versant	SST
Asking for information	Ask for information	X	X	Δ
Conversational repair	Repair breakdowns in interaction	X	X	Δ
Negotiating meaning	E.g., check understanding and ask for clarification when an utterance is misheard or misinterpreted	X	X	Δ
Managing interaction functions				
Initiating	Start any interactions	X	X	Δ
Changing	Take the opportunity to change the topic	X	X	X
Reciprocating	Share the responsibility for developing the interaction	X	X	X
Deciding	Come to a decision	X	X	Δ
Terminating	Decide when the discussion should stop	X	X	X

Note. Functions and expressions used here are based on O'Sullivan et al. (2002). O = intended to elicit from all test-takers; Δ = intended to elicit from most test-takers or test-takers at higher levels; X = intended to elicit from a very limited number of or no test-takers.

Correlation and Multiple Regression Analyses

Table 3 shows that the three test scores were normally distributed. Correlations were moderate between the SRST and Versant (r = .64, p < .01) and between the SRST and SST (r = .66, p < .01). A high correlation between the Versant and SST (r = .79, p < .01) accords with Bernstein et al. (2010), who documented strong correlations between the Versant™ tests and various oral interviews in four languages (e.g., r = .77 to .92).

Next, all the assumptions for the multiple regression analysis were checked and confirmed to have been met. Table 4 shows that 43% (adjusted R^2) of the SRST scores were predicted by the SST scores alone and an additional 3% by the Versant scores. Similarly, 41% of the SRST scores were explained by the Versant scores solely, with an additional 5% explained by the SST scores. Overall, the SRST scores were substantially (46%) predicted by the scores of the other two tests. In other words, it was found that there is a general tendency that a candidate scoring high on the Versant and SST is

also likely to have a high SRST score. The finding that 43% of the SRST scores were predicted by the SST scores also means that 43% of the SST scores were predicted by the SRST scores, which suggests that the SRST scores can predict 43% of the SST scores.

Table 3. Descriptive Statistics of the Three Test Scores (*N* = 64)

	Mean	*SD*	Minimum	Maximum	Skewness	Kurtosis	Possible score range
SRST	9.83	2.47	3.00	14.50	-0.81	1.17	3-15
Versant	39.94	10.70	22.00	75.00	1.01	1.52	20-80
SST	4.59	1.61	2.00	9.00	0.88	0.69	1-9

Table 4. Regression Analyses for Predicting the SRST Scores

Variable	R^2	Adjusted R^2	SEE	*F* Change	Change *p*	*F*	*p*
SST only	.44	.43	1.87	47.87[a]	<.01	47.87[a]	<.01
SST + Versant	.48	.46	1.81	4.81[b]	.03	27.81[c]	<.01
Versant only	.41	.41	1.90	43.95[a]	<.01	43.95[a]	<.01
Versant + SST	.48	.46	1.81	7.24[b]	.01	27.81[c]	<.01

Note. SEE = Standard error of estimate. [a] (1, 62), [b] (1, 61), [c] (2, 61).

The finding that more than 40% of the SRST score variance can be explained by the other two tests suggests that the speaking ability assessed by the SRST may be similar to that assessed by the Versant and SST. Additionally, it indicates that more than half of the SRST variance is unexplained, suggesting each test measures distinctive test constructs. While the measurement error (e.g., test-takers' different conditions while taking the tests) could explain this difference, other factors, in addition to the abovementioned elicited different functions, are explored below.

Differences in Score Distributions

Table 5 shows score distributions for each test. For example, on the SRST, three test-takers received scores ranging from 3.00 to 3.99, whereas one had scores between 4.00 and 5.99. On the Versant, seven test-takers obtained scores ranging from 20 to 29. On the SST, three test-takers received Level 2 scores. Patterns become conspicuous in Figure 2, wherein the same

information as in Table 5 is displayed. For the beginner and intermediate level group, the Versant and SST had similar distributions, in which most students obtained lower scores. By contrast, the SRST overall dispersed students of the same ability group across the whole score range. For the higher proficiency group, scores of the Versant and SST were generally distributed within the whole score range, but the SRST scores were not observed in the lower end of the score range (i.e., 3.00 to 5.99). Overall, while the SRST seems to identify differences in the speaking abilities of students up to the intermediate level, the Versant and SST seem to better discriminate between highly proficient students.

Table 5. Score Range and Number of Students

SRST	3	4-5	6-7	8-9	10-11	12-13	14-15	
Beginner/ Intermediate[a]	3	1	4	15	11	4	0	
Higher[b]	0	0	1	5	10	8	2	
All[c]	3	1	5	20	21	12	2	
Versant	20-	30-	40-	50-	60-	70-		
Beginner/ Intermediate[a]	7	22	9	0	0	0		
Higher[b]	1	5	9	9	0	2		
All[c]	8	27	18	9	0	2		
SST	2	3	4	5	6	7	8	9
Beginner/ Intermediate[a]	3	13	12	8	2	0	0	0
Higher[b]	0	1	5	8	5	2	3	2
All[c]	3	14	17	16	7	2	3	2

Notes. SRST: 3 = 3.00-3.99; 4-5 = 4.00-5.99. Versant: 20- = 20-29; 70- = 70-80. SST: 2 = Level 2. [a]*n* = 38; [b]*n* = 26; [c]*N* = 64.

Some may wonder how the SRST could cover the broad score range for the beginner and intermediate learners, considering that it elicits utterances from a limited range of tasks. We believe that the SRST has this capability for two reasons. First, the SRST tends to elicit relatively long utterances, even from lower proficiency students, by presenting model language through the reading passage that can be used for production and by providing them with time to plan their speech; thus their speaking abilities can be well examined

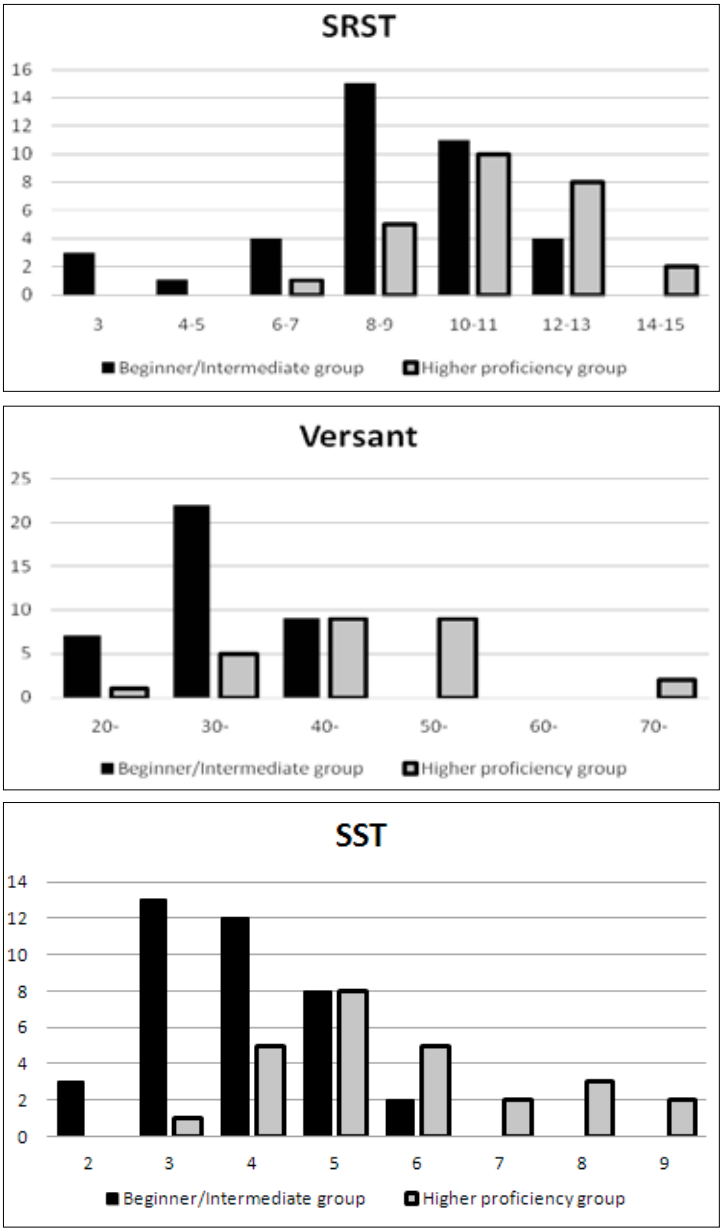


Figure 2. Score Distribution of the Three Tests

and discriminated across score levels. Second, the EBB rating scale for the SRST was empirically developed on the basis of utterances from novice and intermediate level learners, which maximized the discriminatory power of the SRST's EBB scale for such learners. However, this characteristic may vary depending on the difficulty of the stories that test-takers retell. The current study used relatively easy texts. Future studies should examine whether the use of more difficult texts leads to different discriminatory patterns.

Analysis of Transcribed Spoken Data

While the score distributions differ depending on the test-takers' proficiency levels, comparing the transcribed utterances of the three tests indicated two main factors that lead to score differences: (a) tasks and speaking functions and (b) aspects focused on while rating. The first factor, (a), was observed between the SRST and Versant, between the SRST and SST, and between the Versant and SST. First, the SRST and Versant seem to differ in terms of tasks, especially planning time. Examinees can take more time to plan future utterances in the SRST than in the Versant. The SRST does not have a specific time for planning, but candidates can think after they have finished reading a text. In contrast, the Versant gives virtually no planning time and requires quick responses. In one instance, a female student had a higher SRST score (Standardized = 0.44; Raw = 11.00; CE = 4.00; G&V = 4.00; Pronunciation = 3.00) than the Versant score (Standardized = -1.16; Raw = 31.00). She managed to explain the stories (with many pauses) within the specified time during the SRST; however, in the Story Retelling Task of the Versant, she could not finish the stories within the specified time. Her performances seemed to diverge substantially depending on the time allowed.

Between the SRST and SST, there was a case in which a task difference led to dissimilar performances, which resulted in different scores. The SRST has a story retelling and opinion stating task, whereas the SST calls for more varied and complex types of functions in response to an interviewer's prompts, especially for intermediate and advanced level learners. One female examinee achieved a higher SRST score (Standardized = 0.64; Raw = 11.50; CE = 4.00; G&V = 3.50; Pronunciation = 4.00) than SST score (Standardized = -0.99; Level 3, Novice High). She did fairly well in describing the stories she read during the SRST but produced fragmentary utterances and exhibited much difficulty in executing simple tasks, such as explaining her wish about an overseas tour to a tourist agency (in the role-play) during the SST. Given her successful performance in describing picture sequences during the SST, she seems to have the ability to express simple ideas in English when

the specific content to talk about is supplied; however, she is unlikely to have the ability to produce language while simultaneously considering the content.

Another task difference was noted between the Versant and the SST. The SST requires test-takers to use complex functions such as explaining details and giving solutions to problems by employing strategic skills, whereas the Versant's tasks are simpler. One female test-taker had a higher score on the Versant (Standardized = 2.76; Raw = 75.00) than the SST (Standardized = 0.87; Level 6, Intermediate Mid). She succeeded in retelling the gist in a story retelling task on the Versant. In contrast, during the SST, she could not execute her task in a role-play; she failed to convey to a shop clerk her request to exchange a product she had bought. In other SST tasks, she tended to stop when expressing details and complicated concepts. These divergent performances between the Versant and SST seem to suggest that she likely lacks strategic skills to manage and maintain interactions and the ability to describe details, which are elicited in the SST and led to a comparatively lower score on that test.

A second factor that seems to contribute to diverging scores is aspects focused on while rating, which was observed only between the SRST and SST. One female student who obtained a lower SRST score (Standardized = -0.74; Raw = 8.00; CE = 4.00, G&V = 1.00, Pronunciation = 3.00) than SST score (Standardized = 0.87; Level 6, Intermediate Mid) received a low score on grammar and vocabulary on the SRST because her performance contained relatively numerous minor errors. Minor errors were also obvious during the SST; however, her talk was intelligible and convincing with high fluency on the SST, which resulted in a higher SST score. The SRST uses three criteria; when one of the three yields a lower score, the total derived by adding the three scores results in a lower score. On the contrary, in the holistic rating system the SST employs, if test-takers make themselves understood very effectively and achieve the set tasks, they can gain higher scores despite some minor errors in utterances in terms of grammar and pronunciation. The SST holistic scale is weighted more towards communicatively effective performance than minor errors, while the SRST EBB scale gives equal weight to each of communicative efficiency, grammar and vocabulary, and pronunciation. Thus, the SST holistic scale may be able to compensate for minor errors with impressive holistic performance.

These two key factors, (a) tasks and speaking functions and (b) aspects focused on while rating, seem to lead to differences in the evaluation of test performance and the resulting scores, which could invite varied decisions

based on the scores. However, recall that as much as 46% of the SRST score variance was shared by the other two tests (see the *Correlation and Multiple Regression Analyses* subsection). Therefore, we can conclude that the three tests tend to produce close results overall, with some variation caused by the aforementioned factors.

Discussion and Conclusion

The SRST is a semi-direct speaking test devised for classroom use and for measuring the ability to produce extended spoken monologues. This study investigated the relationships between the SRST, Versant, and SST to probe the validity of score-based interpretation of the SRST. RQ1 was "How do functions elicited by the SRST compare with those elicited by the Versant and SST?" We found that when we considered both O and Δ, the SRST elicited more functions (k = 7) than the Versant (k = 2), but fewer than the SST (k = 20). Few functions were shared: The SRST shared two functions with the Versant and seven functions with the SST, while the Versant shared two functions with the SST. The SST was found to elicit more functions by providing several tasks (e.g., picture sequences, role-play) and chances for test-takers to interact with an interviewer. Although the functions elicited by the SRST were limited compared with the functions elicited by the SST, we intended to limit the functions and tasks to focus on areas that Japanese learners of English find difficult and to increase the practicality for administration. Similarly, the Versant elicited a limited number of functions, which corresponds with the developers' intentions to elicit "core skills that are building blocks of speaking proficiency" (Bernstein et al., 2010, p. 371) without using many real-life functions.

RQ2 asked to what degree the SRST scores are associated with the Versant and SST scores. The results showed that correlations of the SRST with the Versant and SST were moderate (r = .64 to .66), that a substantial proportion of the SRST score variance (46%) was predicted by the other two tests, and that the SST alone explained the score variance as much as the Versant did (43% vs. 41%). Such relationships were expected on the basis of intended test constructs and formats, and were empirically supported by the moderate to strong correlations; hence, it is concluded that the SRST likely assesses some of the "facility in spoken English," measured by the Versant and some of the "functional speaking ability," tested by the SST. Moreover, the result—43% of the SST scores was explained by the SRST—seems to show that the construct that the SRST measures is related to real-life interactive communication, since the SST aims to simulate natural conversation.

RQ3 examined differences in score distributions of the three tests between the beginner and intermediate-level learner group combined and the higher level learner group. Figure 2 showed that differences existed in score distributions of the three tests between the two proficiency groups: The SRST scores of the beginner and intermediate-level learners ranged widely, whereas their Versant and SST scores clustered at the lower end of the score range. Conversely, the Versant and SST differentiated higher level learners within the possible score range better than the SRST. These results suggest that the SRST can better differentiate speaking abilities in beginner and intermediate-level students, whereas the Versant and SST can better discriminate such abilities in students of higher proficiency.

RQ4 aimed to identify factors contributing to score differences in the three tests by analyzing the transcripts of test-takers' utterances. The analysis indicated that (a) tasks and speaking functions and (b) aspects emphasized while rating could cause score differences. As for task differences, the SRST allows story retelling and opinion statement after the possibility of some planning time, whereas the Versant asks test-takers to perform several simple tasks immediately after they are provided. The SST elicits fluent use of interactive functions using various tasks such as talking about familiar topics, stating opinions, negotiating, and elaborating on details and complex matters, depending on test-takers' levels. With respect to differences in scoring systems, the SST rating focuses more on fluent and effective communication than on errors that do not impede understanding, whereas the SRST concentrates equally on communicative efficiency and accuracy aspects.

Three implications are discussed. First, this study contributed to the accumulation of validity evidence for the SRST and demonstrated one instance of the validation process by providing multiple new strains of empirical evidence derived through comparison with the other tests. This study and previous ones (Hirai & Koizumi, 2009, 2011; Koizumi & Hirai, 2010) covered most critical analyses in the validation framework, as delineated in the introductory section (i.e., regarding the [a] to [b4] aspects). However, investigation into the meaningfulness of the test scores, the feedback to test users, and the beneficial washback of the test on intended aspects such as learning and teaching (i.e., [b5] in the framework above) remains to be done. The impact of the SRST on learning speaking skills, especially when used as a formative and summative assessment tool in language classrooms, should specifically be inspected.

The second implication is that this study explained one difference between the three tests (i.e., score distributions between the beginner/intermediate

group and the higher proficiency group) and two factors differentiating the scores (i.e., tasks and speaking functions, and rating method). This information may provide a useful basis for selecting one of the three speaking tests as appropriate for a given assessment purpose and testing situation. Although qualified interviewers are needed along with test budgets, the SST typically elicits interactive and complex functions by assigning test-takers various tasks that fit their proficiency levels, while focusing on effective communication. The Versant, while requiring monetary and equipment resources (telephones or computers), measures natural-paced listening along with the ability to react promptly. Further, the SST and Versant tend to discriminate between learners of a higher proficiency group. On the other hand, the SRST requires teachers or peers to evaluate performances using the EBB rating scale, and it uses a limited range of tasks (i.e., story retelling and opinion stating) and elicits a limited number of functions (i.e., providing personal information, expressing opinions, elaborating, justifying opinions, describing, paraphrasing, and expressing preferences). However, it has three chief advantages, particularly when used as a classroom test. First, it is free. Second, teachers can incorporate the test into classroom activities by using passages that students have already learned. Third, it is likely to discriminate between speaking performances, particularly in students at beginner and intermediate levels. Thus, the SRST might be capable of effectively discriminating between students who have achieved speaking goals in class from those who have not, and of demonstrating students' short-term speaking development. These three advantages indicate that, for the purpose of formative assessment, teachers can provide feedback regarding aspects that students have been taught (e.g., pronunciation) during the lessons, conduct remedial activities, and test the same aspects again to scrutinize the improvement, when their resources allow them to do so. These classroom uses of the SRST might encourage speaking activities comprising extended monologues and enhance the speaking ability of L2 learners, although this needs to be empirically tested. It may be worthwhile for teachers to use the SRST for their class, considering its advantages and its shared aspects with the Versant and SST.

A third implication is that, although the SRST is primarily constructed as a test for classroom settings, the relatively large proportion of variance shared by the Versant and SST might indicate that retelling tasks are useful in other settings. The TOEFL iBT already has such speaking integrated tasks, in which examinees read or listen to texts and speak based on the information provided (Chapelle et al., 2008). The Versant also utilizes a story

retelling task based on a listening stimulus. While there are variations in text types (academic and nonacademic) and specific activities elicited (oral summary vs. retelling as much as possible; with or without adding opinions), generally retelling tasks could be useful with novices, intermediates, and advanced learners.

Finally, researchers should modify two aspects of the current procedures to obtain stronger evidence for validity in future studies. First, it is necessary to counterbalance the order in which the three speaking tests are taken to offset possible order effects. Second, the criterion for dividing test-takers into two proficiency groups should be improved. This study stratified students, primarily using information about students' majors and educational qualifications, with minor adjustments to accommodate their self-reported proficiency scores. However, separating the two groups on the basis of the same proficiency test scores would better clarify actual proficiency levels and provide more meaningful interpretations of the results. More rigidly generated evidence for validity would strengthen the SRST validity argument and enhance the usefulness of the SRST in the classroom context.

Acknowledgement

This research was partially supported by the Grant-in-Aid for Scientific Research (KAKENHI, C) [grant number 19520477]. We would like to thank the following for their contributions to our paper: Knowledge Technologies, Pearson, for letting us use the Versant and its sound files; ALC Press, for providing sound files of the SST; Yujia Zhou and Emiko Kaneko for their assistance in test administration and analysis; and Yo In'nami for his critical comments regarding our draft.

Rie Koizumi is an Assistant Professor at Tokiwa University. *Akiyo Hirai* is a Professor at the University of Tsukuba. They are interested in validating speaking tests.

References

ACTFL–ALC Press. (1996) *Standard Speaking Test manual.* Tokyo: Author.

ALC Press. (2008). *TSST: About the test format and assessment.* Retrieved from http://tsst.alc.co.jp/tsst/e/assessment.html

ALC Press. (2010). *The Standard Speaking Test (SST).* Retrieved from http://www.alc.co.jp/edusys/sst/english.html

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

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27,* 355-377. doi:10.1177/0265532210364404

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™.* New York: Routledge.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Educational Testing Service. (2011). *TOEIC® Speaking and Writing: About the tests: Test content.* Retrieved from http://www.ets.org/toeic/speaking_writing/about/content/

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499-510.

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly, 6,* 151-167. doi:10.1080/15434300902801925

Hirai, A., & Koizumi, R. (2011). *Validation of empirically-derived rating scales for the Story Retelling Speaking Test*. Unpublished manuscript.

Koizumi, R., & Hirai, A. (2010). Exploring the quality of the Story Retelling Speaking Test: Roles of story length, comprehension questions, keywords, and opinions. *ARELE (Annual Review of English Language Education in Japan), 21,* 211-220. Retrieved from http://ci.nii.ac.jp/els/110008512411.pdf?id=ART0009707205&type=pdf&lang=jp&host=cinii&order_no=&ppv_type=0&lang_sw=&no=1322450319&cp=

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241-256. doi:10.1177/026553229601300302

Nakano, M. (2002). Standard Speaking Test (SST) to TOEIC, TOEFL, EIKEN tono kaiki bunseki [Regression analysis of Standard Speaking Test (SST), TOEIC, TOEFL, and the EIKEN]. *Research report from Institute of Oral Communication, Waseda University* (pp. 23-50). Tokyo: Institute of Oral Communication, Waseda University. Retrieved from http://www.alc.co.jp/edusys/sst/pdf/article3.pdf

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

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19,* 33-56. doi:10.1191/0265532202lt219oa

Pearson Education. (2008). *Versant™ English Test: Test description and validation summary.* Retrieved from http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12. doi:10.1093/elt/49.1.3

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

Appendix A

Instructions and a Story Sample

Read the story silently within two minutes. 2分間で次の文章を黙読しなさい。

Story 2

> Last year, Bob and his sister Jean went to Florida for their summer vacation with their parents. They visited Florida for one week. The weather was very nice. So everyone was really happy.
>
> On the first day, Bob and his family went to the beach. It was beautiful. The sand was white and the water was very clean. Bob and Jean swam for over three hours. After that, they played volleyball with some other children on the beach. Their parents watched them and smiled.
>
> After playing volleyball, Bob and Jean felt tired. They sat down on the sand next to their parents and drank cold coconut water. Their father said, "It's getting late. Let's go back to the hotel." But Bob and Jean didn't want to leave the beach. Jean asked, "Can we come back again tomorrow?" Their mother said, "Of course we can." Bob and Jean were very happy to hear that.

(Zen mondai & kaito 2001 nendo dai 3 kai kentei: Ichiji shiken, 2002, p. 59; Copyright 2002 by the Society for Testing English Proficiency, Printed with permission)

After the signal, read each question aloud and answer them in English.
1問ずつ合図があってから、質問を読み上げて、英語で答えなさい。

Q1: What did Bob and his family do on the first day?

Q2: How long did Bob and Jean swim?
Q3: Why were Bob and Jean happy?

----------<Next Page>----------

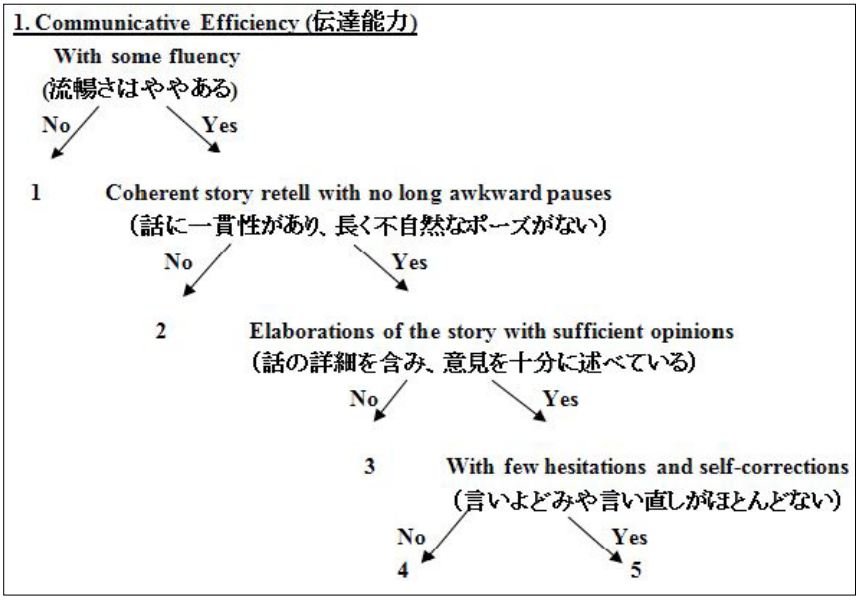
Retell as much of the story as you can in English in two and half minutes. You can look at the keywords while you are retelling. At the end of your retelling, be sure to include your opinions about the story. You will hear a signal 30 seconds before closing.

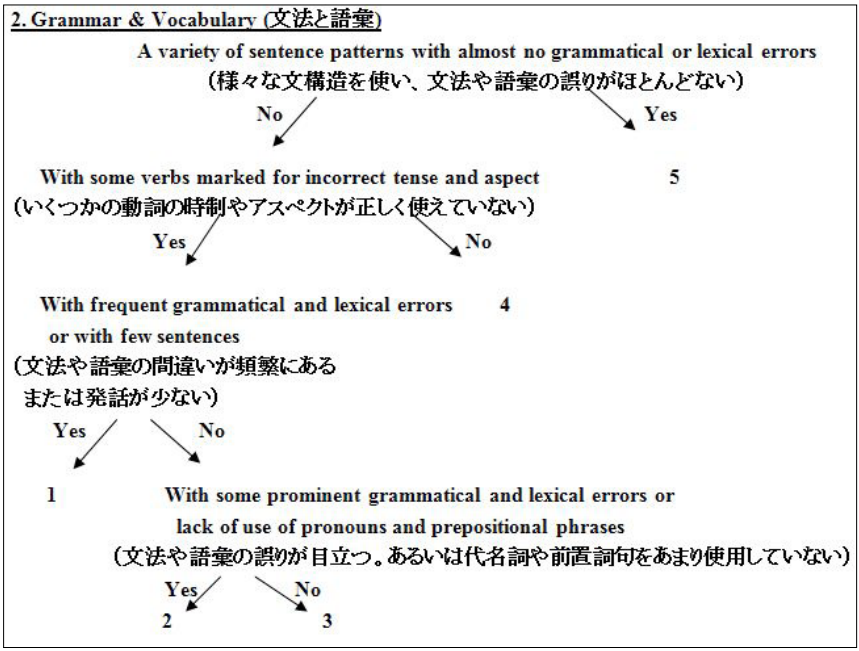
今読んだ内容をできるだけ詳しく、2分30秒間英語で話しなさい。話しながら、キーワードを見てもかまいません。読んだ内容を話し終えたら、必ず、その内容についての感想や意見も英語で述べなさい。終了30秒前にチャイムがなりますので、感想を始める目安にできます。

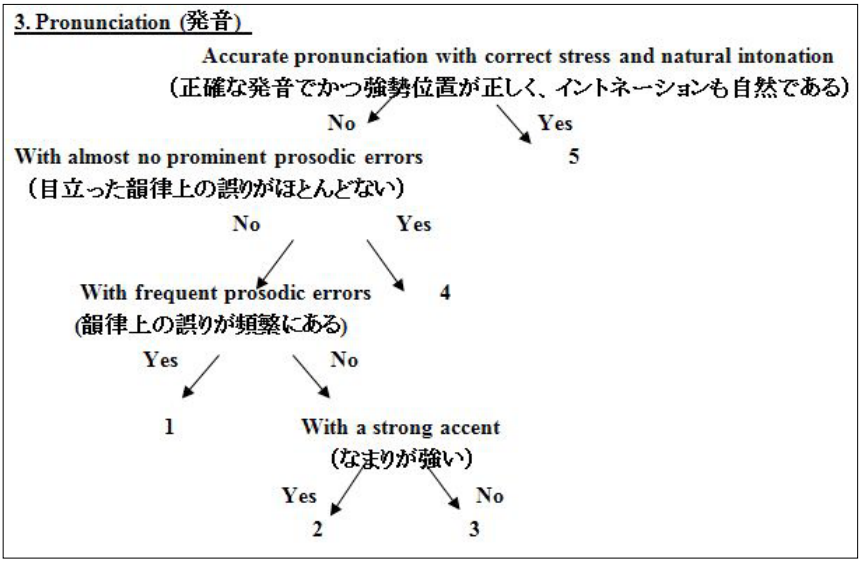
Keywords: Bob, Jean, Florida, beach

Appendix B

EBB rating scale for the SRST




2. Grammar & Vocabulary (文法と語彙)
A variety of sentence patterns with almost no grammatical or lexical errors
(様々な文構造を使い、文法や語彙の誤りがほとんどない)
No
Yes
With some verbs marked for incorrect tense and aspect
5
(いくつかの動詞の時制やアスペクトが正しく使えていない)
Yes
No
With frequent grammatical and lexical errors
4
or with few sentences
(文法や語彙の間違いが頻繁にある
または発話が少ない)
Yes
No
1
With some prominent grammatical and lexical errors or
lack of use of pronouns and prepositional phrases
(文法や語彙の誤りが目立つ。あるいは代名詞や前置詞句をあまり使用していない)
Yes
No
2
3


3. Pronunciation (発音)
Accurate pronunciation with correct stress and natural intonation
(正確な発音でかつ強勢位置が正しく、イントネーションも自然である)
No
Yes
With almost no prominent prosodic errors
5
(目立った韻律上の誤りがほとんどない)
No
Yes
With frequent prosodic errors
4
(韻律上の誤りが頻繁にある)
Yes
No
1
With a strong accent
(なまりが強い)
Yes
No
2
3