Validation of the EBB scale: A Case of the Story Retelling Speaking Test

HIRAI, Akiyo (University of Tsukuba)
KOIZUMI, Rie (Tokiwa University)

Abstract

Among different types of rating scales in scoring speaking performance, the EBB (Empirically derived, Binary-choice, Boundary-definition) scale is claimed to be easy to use and highly reliable (Turner & Upshur, 1996; 2002). However, it has been questioned whether the EBB scale can be applied to other tasks. Thus, in this study, an EBB scale was compared with an analytic scale in terms of validity, reliability, and practicality. Fifty-two EFL learners were asked to read and retell four stories in a semi-direct Story Retelling Speaking Test (SRST). Their performances were scored using these two rating scales, and then the scores were compared by using generalizability theory, a multitrait-multimethod approach, and a questionnaire delivered to the raters. As a result, the EBB scale, which consists of four criteria, was found to be more generalizable (i.e., reliable) than those of the analytic scale and generally assessed the intended constructs. However, the present EBB scale turned out to be less practical than the analytic scale due to its binary format and because it had more levels in each criterion. Further revisions seeking a better scale for the SRST are suggested.

Introduction

Increasing practicality while maintaining high validity and reliability has always been a challenge for speaking tests. This tends to lead to the infrequent use of speaking assessment both inside and outside the classroom (Akiyama, 2004; Honda, 2007). In addition, low reliability and validity for ratings are expected, particularly when raters are not trained and rating descriptors are vague (e.g., Brindley, 1998; Fulcher, 1996).

In order to tackle these problems, various measures have been taken to increase test practicality at the stages of (a) design and operationalization, (b) administration, (c) rating, and (d) analysis (Bachman & Palmer, 1996). At each stage, test practicality is influenced by human resources, material resources, time, and money. Of the four stages, the rating stage of speaking tests tends to require the most resources, for example, for rater training and the rating process itself. In order to overcome the difficulty of maintaining high practicality in this stage, one of the methods proposed was to use the Empirically derived, Binary-choice, Boundary-definition (EBB) technique (Upshur & Turner, 1995). The present study examines the characteristics of the EBB scale in the context of a Story Retelling Speaking Test (SRST), which is a tape-based test for
classroom use that was developed by the authors (see Appendix A).

As with all widely used rating scales such as holistic and analytic scales, the EBB scale also has several levels of the trait or traits being measured. It consists of a hierarchical set of explicit binary questions or descriptions. Then, listening to the student performance, the rater makes a series of yes-or-no choices (see examples in Appendix B) and reaches a final score. These descriptors, which are determined based on a small sample of actual task performances, define boundaries between score levels. This scale is therefore called an Empirically derived Binary-choice Boundary definition (EBB) scale (Upshur & Turner, 1995).

According to Turner and Upshur (1996, 2002), an EBB scale is easy to use and can obtain high intra- and inter-rater reliabilities. It is also claimed that the scale can assess constructs more precisely since it is made based on the actual performance of a test (Turner & Upshur, 1996). However, because of the scale’s empirically-derived and task-specific nature, it has been questioned whether results obtained by an EBB scale that was made specifically for a certain task can be generalized to other tasks (e.g., Bachman & Savignon, 1986; Fulcher, 2003). Fulcher (2003) claims that “the meaning of the [EBB] score relates only to the specific task that was used in the test, and its meaning cannot be generalized to any other test task, or any task in the real world, unless the criterion task is identical in every way to the test task” (p. 107), but to our knowledge, this claim has not been tested empirically. We see this problem as a reliability issue of the EBB scale and will investigate the consistency of EBB ratings from multiple tasks because if the results are consistent across tasks, the result of a task can be considered generalizable to that of another task.

Other than the practicality and reliability issue, an important research area for all testing contexts is validity. One useful way to demonstrate the validity of the EBB scale is to compare the scale with other existing scales, which has been attempted by Nakai (1997) and Koizumi and Kurizaki (2002). Nakai (1997) reported high correlations ($r = .726$ to $.958$) between three rating scales: analytic, holistic, and EBB scales, which suggests that the constructs these scales assess are similar. The evidence that he presented is important, but stronger evidence can be presented using a multitrait-multimethod (MTMM) approach. This simplified version was utilized by Koizumi and Kurizaki (2002) and showed that the EBB and analytic scales were correlated more highly ($r_s = .756$ to $.786$) than the EBB (or analytic) scale and non-speaking tests ($r_s = .142$ to $.439$). However, a more detailed MTMM approach needs to be taken to investigate whether each criterion such as communicative efficiency, grammar, and pronunciation of the EBB scale measures traits as intended.

Thus, the purpose of this study is to examine the reliability, validity, and
practicality of the EBB scale in scoring the SRST, in comparison with an existing analytic scale. We chose an analytic scale to compare with the EBB scale because both scales have multiple criteria to assess speech performance. This means that the scales can provide students with their speaking profiles, such as the strong and weak points of their speaking ability (e.g., Luoma, 2004). From the practicality point of view, a holistic scale would be the best but, since it has only one criterion, the validity of the scale is difficult to achieve unless raters are sufficiently trained (e.g., Upshur & Turner, 1995). Therefore, with the use of the EBB and analytic scales, this study attempts to answer the following three research questions (RQs):

RQ1: How many stories are needed to achieve high reliability in the two scales?
RQ2: Is the EBB scale related to the analytic scale as intended?
RQ3: Which scale is more practical, the EBB scale or the analytic scale?

To examine RQ1, the multivariate generalizability theory (Brennan, 2001) is adopted by utilizing the mGENOVA program (Center for Advanced Studies in Measurement and Assessment, 2007). With this program, we can analyze all the criteria of a rating scale together and identify how many tasks (i.e., stories) are necessary to obtain a reliable score for each criterion. For RQ2, a multitrait-multimethod (MTMM) approach is taken because it can indicate the adequacy of tests as measures of constructs (Campbell & Fiske, 1959). Creating a MTMM matrix of Pearson correlation coefficients between the criteria of the EBB and Analytic scales, we can compare whether correlations between the same or similar traits of the EBB and analytic scales are higher than between different traits. The MTMM method has advantages over an independent comparison of correlations between criteria of similar constructs in that the former can observe both trait and method effects together. Regarding RQ3, the practicality of the EBB scale is compared descriptively with that of the analytic scale by conducting a questionnaire with raters using both scales.

Method

Participants
Fifty-two EFL learners from two Japanese universities took the SRST. Among them, 48 learners were university students at either a beginning or intermediate level of English proficiency, majoring in English or engineering. The remaining four were graduate students who were majoring in English and who were relatively high in proficiency. They were asked to participate in this study in order to examine the ceiling performance of the test.
Materials

*Story Retelling Speaking Test (SRST).* A classroom-based SRST was developed that used a story retelling (or reproduction) technique as its main speaking task. The test consists of two sections: reading a story and retelling it (see Appendix A). In the reading section, reading ability is mainly measured. If test takers cannot answer the comprehension questions, they are assessed as being at the pre-speaking level, lacking basic linguistic knowledge. For the SRST, two 100-word and two 150-word stories were adopted from past Eiken (Society for Testing English Proficiency) Grades 3 and 4 tests. The text difficulty was nearly the same for the four stories (Flesch-Kincaid Grade Level: 4.1, 4.5, 4.6, and 5.3, respectively). Each story had three comprehension questions.

For the retelling section, four words, either proper nouns or keywords, were prepared from each story as prompts in order to help learners recall the content of the story when retelling. Examinees were tested on whether they could convey the information they had just received as clearly as possible, and on how much of the story they could recall. However, this was not a memory test as many studies have demonstrated that such elicited imitation tasks are not related to learners’ short-term memory capacity but rather to language abilities such as syntactic knowledge and sentence construction ability (Versant™ with Ordinate® Technology, 2007; Caplan & Waters, 1999).

The SRST was designed to enhance the practicality of classroom speaking assessment and to produce positive washback effects on teaching and learning (see Hirai & Koizumi, 2008, for details). It can be administered using tape-based instruction in a language laboratory or a CALL room to a large number of students at the same time. In addition, reproduction or retelling tasks have been widely used for classroom activities. They can encourage learners to pay more attention to not only the content of the story but also language form (e.g., Gambrell, Kapinus, & Koskinen, 1991; Yoshimura, 2006). Since these tasks demand deeper cognitive processing, more effective learning is thought to take place (e.g., Craik & Tulving, 1975; Hirai, 2005, Swain, 1985; Wittrock, 1990). In fact, according to examinees’ perceptions of the SRST, the connection from input to output enhanced students’ motivation to learn how to speak, which is positive evidence for washback (Hirai & Koizumi, 2008).

*EBB and analytic scales for SRST.* The EBB scale has the following four criteria: (a) Communicative Efficiency, (b) Grammar & Vocabulary, (c) Content, and (d) Pronunciation. These criteria were created in order to cover the intended speaking ability sufficiently. As shown in Appendix B, each criterion has five hierarchical levels,
which distinctly separates the performances between the upper and lower levels. The five levels were set up because they were the maximum that we could distinguish out of these performances. These levels and descriptors were made based on the construction procedures in the previous literature (e.g., Turner & Upshur, 1996; Koizumi & Kurizaki, 2002) and students' actual speech performances. We used two types of samples derived from two universities and each of us constructed an EBB scale separately to avoid constructor and sample effects as far as possible. Then, after thorough discussion, we made a final version of the EBB scale. Next, listening to the performances of three examinees who retold all the four stories, totaling 12 performances as benchmark speeches, and examining their transcriptions, we separately rated the performances with the pilot EBB scale to examine whether the scale could discriminate between the test takers, and discussed this matter until we reached agreement about the score.

Regarding a scale to compare with this EBB scale, we used Nakai's (1997) analytic scale, which is a modified FSI (Foreign Service Institute) scale (see Hughes, 2003), because it was made for Japanese EFL learners and seemed best suited to score the SRST with only minor modifications. The analytic scale is composed of the following four criteria: (a) Communicative Efficiency, (b) Grammar, (c) Fluency, and (d) Pronunciation. The criteria are not exactly the same as the EBB criteria. In addition, each of the analytic criteria has only four hierarchical levels while each EBB criterion has five levels (see Appendix C). Although the same criteria and the number of levels might be better for the sake of comparison, we did not modify Nakai's (1997) scale much because our intention was to investigate whether the EBB scale could be used better than an existing scale to score the SRST.

**Questionnaire.** A questionnaire was prepared for the raters who evaluated the examinees using both the EBB and analytic scales. It consisted of five questions concerning the practicality of the two scales. These questions are shown in Table 6 in the Results and Discussion section.

**Procedure**

All the students took the SRST. As shown in Figure 1, first, they read a story silently for two minutes and then read aloud each of the three reading comprehension questions and answered them orally within a period of forty seconds per question (Appendix A). Next, they turned to the back of the sheet and retold as much of the story as possible while looking only at the four keywords to help them recall the story. They were also told to include their opinions about the story within two minutes. This procedure was repeated four times for the four stories, and the order of administering
stories was varied for all the examinees to counterbalance the order effect. The examinees' utterances were recorded and later transcribed; however, the transcription was not used at the rating stage but only for the interpretation of the results.

<table>
<thead>
<tr>
<th>2 minutes</th>
<th>40 seconds x 3 Questions</th>
<th>2 minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read a story silently</td>
<td>Answer each comprehension question orally</td>
<td>Retell the story and include the opinion</td>
</tr>
</tbody>
</table>

*Figure I. SRST procedure for one story.*

Regarding the scoring, six raters - four graduate students and two university teachers - who were majoring in English language education, had a training session for approximately three hours, using the benchmark performances. After training, each rater was assigned to rate approximately 10 examinees' performances. In order to avoid the order effect of the EBB and analytic scales, three raters evaluated their examinees using the EBB scale first, and two weeks later, they rated the same examinees with the analytic scale, without looking at the previously rated scores. The other raters did the same but in the reverse order. After scoring, five raters, excluding one rater who was an author of this study, answered the questionnaire.

**Table 1**

*Agreement Ratio of Each Criterion of the EBB and Analytic Scales*

<table>
<thead>
<tr>
<th>EBB scale</th>
<th>CE</th>
<th>G&amp;V</th>
<th>Content</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact agreement</td>
<td>.682a, .636d,</td>
<td>.636, .500,</td>
<td>.636, .500,</td>
<td>.455, .500,</td>
</tr>
<tr>
<td>ratio</td>
<td>.636a, .636d,</td>
<td>.727, .500,</td>
<td>.591, .591</td>
<td>.455, .591</td>
</tr>
<tr>
<td>Exact and one point difference ratio</td>
<td>.864, .909</td>
<td>.864, .909</td>
<td>.955, .818</td>
<td>.909, .864</td>
</tr>
<tr>
<td>Analytic scale</td>
<td>CE Grammar</td>
<td>Fluency</td>
<td>Pronunciation</td>
<td></td>
</tr>
<tr>
<td>Exact agreement</td>
<td>.727, .591</td>
<td>.545, .500,</td>
<td>.455, .591</td>
<td>.545, .545,</td>
</tr>
<tr>
<td>ratio</td>
<td>.591, .773</td>
<td>.500, .545,</td>
<td>.545, .455</td>
<td>.682, .636</td>
</tr>
<tr>
<td>Exact and one point difference ratio</td>
<td>.955, .909</td>
<td>1.000, .955</td>
<td>1.000, .955</td>
<td>.864, .909</td>
</tr>
</tbody>
</table>

*Note.* aNumber of ratings about which two raters agreed/22; bNumber of ratings about which two raters agreed or disagreed on only one point/22; cStory 1, dStory 2, eStory 3, and fStory 4. CE = Communicative efficiency; G&V = Grammar & Vocabulary.

To obtain the inter-rater reliability of the ratings, about 40% (i.e., 22/52) of the performances were rated by two different raters. As seen in Table 1, the agreement on each criterion was mostly moderate (.455 to .727 in the EBB scale; .455 to .773 in the analytic scale). When the ratio of the two raters who agreed with only one score difference was added to the exact agreement ratio, this agreement ratio was high (.773
to 1.000 in the EBB scale; .864 to 1.000 in the analytic scale). The results suggest that in most cases, two raters agreed on the rating of the same performance exactly, or disagreed with only one score difference. Since we considered this inter-rater reliability acceptable, the rest of the performances were rated by a single rater.

Results and Discussion

In order to determine how many stories are needed to achieve a high reliability in each scale (RQ1), the multivariate generalizability theory was used. The specific design was a fully crossed one facet (i.e., story) design with four dependent variables (i.e., four criteria). The EBB and analytic scales were analyzed separately, with missing data excluded. Rater facet was not included because the number of performances rated by two raters was small.

Table 2
Estimated Variance Components and Proportion of Variance Explained in G study

<table>
<thead>
<tr>
<th>Source of variability</th>
<th>Communicative Efficiency</th>
<th>Grammar &amp; Vocabulary</th>
<th>Content</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons (p)</td>
<td>0.709 (70.842%)</td>
<td>0.423 (46.327%)</td>
<td>0.616 (57.071%)</td>
<td>0.788 (77.381%)</td>
</tr>
<tr>
<td>Stories (s)</td>
<td>0.012 (1.242%)</td>
<td>0.000* (0.000%)</td>
<td>0.075 (6.946%)</td>
<td>0.000* (0.000%)</td>
</tr>
<tr>
<td>Residual (ps, e)</td>
<td>0.279 (27.917%)</td>
<td>0.491 (53.673%)</td>
<td>0.388 (35.982%)</td>
<td>0.230 (22.619%)</td>
</tr>
</tbody>
</table>

Note. *Negative variance was set to zero.

Analytic Scale (n = 49)

<table>
<thead>
<tr>
<th>Source of variability</th>
<th>Communicative Efficiency</th>
<th>Grammar</th>
<th>Fluency</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons (p)</td>
<td>0.237 (58.009%)</td>
<td>0.196 (47.870%)</td>
<td>0.268 (60.709%)</td>
<td>0.274 (69.811%)</td>
</tr>
<tr>
<td>Stories (s)</td>
<td>0.014 (3.421%)</td>
<td>0.005 (1.175%)</td>
<td>0.000* (0.000%)</td>
<td>0.005 (1.181%)</td>
</tr>
<tr>
<td>Residual (ps, e)</td>
<td>0.157 (38.570 %)</td>
<td>0.208 (50.956%)</td>
<td>0.173 (39.291%)</td>
<td>0.114 (29.008 %)</td>
</tr>
</tbody>
</table>

First, the sources of variability with the one facet design were examined based on the generalizability (G) studies for one story. This story had four sources of variability: (a) differences among objects of measurement (i.e., persons), (b) differences in item difficulty (i.e., stories), (c) the person-by-story interaction, and (d) random or unidentified events. The third and fourth sources of variability cannot be disentangled and are shown together (i.e., ps, e). As shown in Table 2, the substantial variability of the scores measured by both the EBB and analytic scales was explained by the variance component for persons (p; e.g., 70.842% for EBB Communicative Efficiency and
69.811% for Analytic Pronunciation). To a lesser extent, the variability of the score was explained by the variance components for residual (ps, e), except for EBB Grammar & Vocabulary (53.673%) and Analytic Grammar (50.956%), and only slightly by the variance components for stories (s; 0.000% to 6.946%). This is rather a good result because the variability in the measurement largely reflects the differences in the examinees' intended abilities, except for both grammar scales. It also indicates that the stories were almost the same in terms of difficulty and that the relative standing of examinees did not differ much from story to story.

Next, using D studies, we examined how the reliability coefficient of each criterion changed depending on the number of stories. Since the nature of rating scales is criterion-referenced rather than norm-referenced, the absolute values of phi coefficients (Φ), were used for the interpretation of reliability instead of the relative values of generalizability (G) coefficients. We considered a phi coefficient of above .70 to be high in reliability as a speaking test (Lado, 1964).

![Figure 2. Changes in phi coefficients in the decision studies (p x s Design).](image)

Figure 2 illustrates the change in phi coefficients (Y-axis) depending on the number of stories (X-axis; see also Table 3). When we looked at the cut-off point of .70, EBB Communicative Efficiency and EBB Pronunciation were highly reliable and only one story was enough to achieve .70, while most of the Analytic criteria required two stories.
(see the underlined values in Table 3). However, both EBB Grammar & Vocabulary and Analytic Grammar needed three stories to obtain reliable ratings. Thus, overall, to assess examinees' performances in the SRST reliably, the EBB scale required fewer stories than the analytic scale. These results suggest that scores from the EBB scale tend to be slightly more generalizable than scores from the analytic scale, which contradicts the claims in a previous study (Fulcher, 2003) stating that results in the EBB scale were hard to generalize across tasks. In addition, we found that two stories would be sufficient to obtain a reliable score in both scales if we could successfully revise both grammar criteria.

Table 3

<p>| Decision Studies of the EBB and Analytic Scales for Each Criterion (p x s Design) |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>1 story</th>
<th>2 stories</th>
<th>3 stories</th>
<th>4 stories</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EBB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicative Efficiency</td>
<td>.717, .708</td>
<td>.835, .829</td>
<td>.884, .879</td>
<td>.910, .907</td>
</tr>
<tr>
<td>Grammar &amp; Vocabulary</td>
<td>.463, .463</td>
<td>.633, .633</td>
<td>.721, .721</td>
<td>.775, .775</td>
</tr>
<tr>
<td>Content</td>
<td>.613, .571</td>
<td>.760, .727</td>
<td>.826, .800</td>
<td>.864, .842</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>.774, .774</td>
<td>.872, .872</td>
<td>.911, .911</td>
<td>.932, .932</td>
</tr>
<tr>
<td><strong>Analytic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicative Efficiency</td>
<td>.601, .580</td>
<td>.750, .734</td>
<td>.819, .806</td>
<td>.857, .847</td>
</tr>
<tr>
<td>Grammar</td>
<td>.484, .479</td>
<td>.653, .647</td>
<td>.738, .734</td>
<td>.790, .786</td>
</tr>
<tr>
<td>Fluency</td>
<td>.607, .607</td>
<td>.756, .756</td>
<td>.823, .823</td>
<td>.861, .861</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>.706, .698</td>
<td>.828, .822</td>
<td>.878, .874</td>
<td>.906, .902</td>
</tr>
</tbody>
</table>

*Note. *$\rho^2 =$ generalizability coefficient; $\Phi =$ phi coefficient. Underlined $= $ phi coefficient over .70 in the smallest number of stories.

Since the grammar aspect was likely to be less generalizable across stories, we examined the scores and the transcription of students' utterances of the grammar criterion. As a result, it seems that a different length of utterances influenced the occurrence of grammatical errors. For example, Student A, as shown in Table 4, obtained the best rating for EBB Grammar & Vocabulary in Story 3, in which she spoke very little, whereas she received the lowest rating for the same criterion in Story 4, in which she spoke more. In this manner, even such low-level students could memorize a few sentences of a story almost exactly, but when they spoke for longer, their grammatical ability was arguably reflected in their utterances.

Given this result, in order to assess students' grammatical knowledge accurately and consistently, a certain length of utterances may be needed. Thus, we will modify the grammar descriptors and give the lowest rating when the utterances contain only a few
sentences, assuming that the grammatical knowledge of students who utter little is so lacking that they cannot produce a series of monologic sentences.

Table 4

<table>
<thead>
<tr>
<th>Story</th>
<th>Number of words uttered</th>
<th>EBB</th>
<th>Analytic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CE</td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>31</td>
<td>1.0</td>
<td>2.5</td>
</tr>
<tr>
<td>S2</td>
<td>12</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>S3</td>
<td>5</td>
<td>1.0</td>
<td>5.0</td>
</tr>
<tr>
<td>S4</td>
<td>44</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Note. *pruned tokens; CE = Communicative Efficiency; G&V = Grammar & Vocabulary; Pro = Pronunciation. Rating scores were rated by two raters.

On the other hand, the pronunciation criterion was the most reliable among the four criteria (see Table 3) although it also assesses an aspect of errors. This may be because grammar-related ratings can be improved if students memorize the exact sentences written in the story, whereas pronunciation does not improve readily from the written source.

Next, we examined to what extent the criteria of the EBB scale is related to specific criteria of the analytic scale, showing a multitrait-multimethod (MTMM) matrix. The methods we put in the matrix were the EBB and analytic rating methods, and traits examined were the four criteria of each method. In this matrix, correlations are divided into two types: convergent and discriminant. Convergent validity coefficients are correlations between measures of the same construct using different measurement methods (i.e., monotrait-heteromethod) or correlations between measures of the same construct using the same measurement methods (i.e., monotrait-monomethod), which should be high. Discriminant (also called divergent) validity coefficients are correlations between measures of different constructs using the same method of measurement (i.e., heterotrait-monomethod) or correlations between different constructs using different measurement methods (i.e., heterotrait-heteromethod; Crocker & Algina, 1986). Ideally, these discriminant validity coefficients should be lower than the convergent validity coefficients.

Based on this theory, we identified which correlations would be convergent validity coefficients by comparing the descriptors of the EBB and analytic scales shown in Appendixes A and B. First of all, Communicative Efficiency in both the EBB and analytic scales measures not only communicative efficiency but also some aspects of fluency and content. For example, phrases such as "no long pauses," and "with some fluency" in the EBB Communicative Efficiency descriptor indicate fluency aspects,
while the phrases "completes the story," and "with few opinions" in the Analytic Communicative Efficiency descriptor are indications of measuring some aspects of content. Therefore, high convergent validity coefficients would be expected between EBB Communicative Efficiency, Analytic Communicative Efficiency, EBB Content, and Analytic Fluency. Second, EBB Grammar & Vocabulary and Analytic Grammar both evaluate grammatical accuracy. Third, EBB Pronunciation and Analytic Pronunciation measure almost the same trait. Thus, correlations between these criteria were regarded as convergent validity coefficients and were expected to be higher than the other coefficients.

A set of correlations among the eight criteria was first obtained for each story separately. However, since all the four sets were similarly patterned, the mean scores of the four stories were used as a final matrix as shown in Table 5. The convergent validity correlations we predicted are indicated with underlines.

**Table 5**
*Multitrait-multimethod Matrix of the EBB and Analytic Scales*

<table>
<thead>
<tr>
<th></th>
<th>Method 1: EBB</th>
<th>Method 2: Analytic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CE</td>
<td>G&amp;V</td>
</tr>
<tr>
<td>EBB CE</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>EBB G&amp;V</td>
<td>.341*</td>
<td>--</td>
</tr>
<tr>
<td>EBB Content</td>
<td>.814**</td>
<td>.277</td>
</tr>
<tr>
<td>EBB Pro</td>
<td>.461**</td>
<td>.353*</td>
</tr>
<tr>
<td>Ana CE</td>
<td>.635**</td>
<td>.358*</td>
</tr>
<tr>
<td>Ana Grammar</td>
<td>.447**</td>
<td>.636**</td>
</tr>
<tr>
<td>Ana Fluency</td>
<td>.589**</td>
<td>.264</td>
</tr>
<tr>
<td>Ana Pro</td>
<td>.437**</td>
<td>.247</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>3.204</td>
<td>3.352</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.882</td>
<td>0.739</td>
</tr>
</tbody>
</table>

*Note. *p < .05  **p < .01. n = 49. Missing values were deleted listwise. CE = Communicative Efficiency; G&V = Grammar & Vocabulary; Pro = Pronunciation; Ana = Analytic scale. Underlined = convergent validity correlations; Italicized = unexpectedly high correlations.*

Overall, the convergent and discriminant validities seem to be successfully distinguished. All the underlined convergent validity correlations were moderately high at .561 to .821 and were mostly higher than heterotrait-monomethod validity coefficients such as .277 between EBB Grammar & Vocabulary and EBB Content, and heterotrait-heteromethod validity coefficients such as .247 between EBB Grammar & Vocabulary and Analytic Pronunciation. Thus, the results were considered to provide
strong evidence that the EBB scale can measure what we intended to measure, and that the constructs of the EBB scale are similar to those of the analytic scale. These results were also in line with Nakai (1997) and Koizumi and Kurizaki (2002).

However, there were three cases where discriminant validity correlations were higher than convergent ones (see italicized values in Table 5). These were correlations of Analytic Pronunciation with EBB Content, Analytic Communicative Efficiency, and Analytic Fluency (r = .627, .708, and .632, respectively), which were higher than correlations between Analytic Pronunciation and EBB Pronunciation (r = .561). In this regard, Analytic Pronunciation seems to be problematic in that it failed to discriminate pronunciation from other aspects of speaking ability accurately. In contrast, EBB Pronunciation did not have such patterns, indicating that the pronunciation criterion in the EBB scale could differentiate pronunciation from the other traits such as content and fluency.

Regarding convergent validity coefficients, the correlations between EBB Communicative Efficiency and EBB Content, between Analytic Communicative Efficiency and Analytic Fluency, and between EBB Content and Analytic Fluency were very high (r = .814, .778, and .758, respectively). The results indicate that in order to produce a “coherent story” as stated in the Communicative Efficiency criterion, both content and fluency aspects are necessary. However, though high convergent validity correlations were expected, such high relationships within the same scales (monotrait-monomethod validity correlations) may show a redundancy among the criteria. Thus, we may be able to combine EBB Communicative Efficiency and EBB Content into one. For the sake of practicality, by reducing the number of criteria from the current four criteria, raters may find it easier to give scores on all the three criteria while listening to a two-minute performance once.

Lastly, to answer RQ3, the practicality of the EBB and analytic scales were compared. As shown in Table 6, it took approximately 1.5 hours to train raters for each scale, which seemed sufficient for all of them. However, Questions (2) and (3) showed that only two raters were able to rate all four criteria in one listening (i.e. within two minutes) using the EBB scale, whereas four raters were able to do so with the analytic scale.

In addition, as indicated in Questions (3) and (4), the raters spent slightly less time on scoring performances using the analytic scale, and found it easier to use. As indicated in Question (5), there seem to be at least three reasons for this. One is the difference in the number of levels in a criterion; the analytic scale has only four levels in each criterion while the EBB has five levels. Another reason is the difference in the format of each scale. Rater 2 felt that it was sometimes harder to make a binary decision
in the EBB scale, while, as Rater 5 mentioned, the analytic scale was easier to score because she could simply move her eyes down to the descriptors from top to bottom until the performance fit the description. The third reason pointed out by some raters was that EBB Grammar & Vocabulary requires raters to count the number of grammatical and lexical errors, which means they have to listen to the whole performance just to satisfy this criterion.

Table 6

<table>
<thead>
<tr>
<th>Rat-er</th>
<th>(1) Was the rater training sufficient? (hours)</th>
<th>(2) Were you able to score a speech while listening to it once?</th>
<th>(3) Which required more time?</th>
<th>(4) Which was easier to use?</th>
<th>(5) State reasons for your answers to (3) and (4).</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Sufficient (EBB 1.5h, Analytic 1h)</td>
<td>No</td>
<td>No</td>
<td>EBB Analytic</td>
<td>I had gotten used to analytic scales for other tests like the STEP Test.</td>
</tr>
<tr>
<td>R2</td>
<td>Sufficient (1.5h each)</td>
<td>No</td>
<td>Yes</td>
<td>EBB Analytic</td>
<td>I had to listen to the whole performance, counting the number of errors on the EBB G&amp;V criterion. It was hard to decide yes or no every time.</td>
</tr>
<tr>
<td>R3</td>
<td>Sufficient (1.5h each)</td>
<td>Yes</td>
<td>Yes</td>
<td>Same Analytic</td>
<td>I had to count the number of errors on the EBB G&amp;V criterion.</td>
</tr>
<tr>
<td>R4</td>
<td>Sufficient (1.5h each)</td>
<td>Yes</td>
<td>Yes</td>
<td>EBB Analytic</td>
<td>The descriptor of the analytic scales did not explain grammatical errors in detail, but once I had established some criteria for myself, the analytic scales were easier to use.</td>
</tr>
<tr>
<td>R5</td>
<td>Sufficient (1.5h each)</td>
<td>Sometimes, No</td>
<td>Yes</td>
<td>EBB Analytic</td>
<td>Each analytic scale had only four levels, and the format allowed me to simply go straight down to the descriptors which matched the performance.</td>
</tr>
</tbody>
</table>

*Note.* The questionnaire was answered in Japanese and translated later.

For all the above reasons, the current analytic scale with four levels in each criterion seems to be more practical than the EBB scale with five levels each. However, due to the vague band descriptors of the analytic grammar criterion, it also has a drawback. Rater 4 mentioned that due to its simple description, she had to establish some standards for herself in order to give scores, and said this might be a cause of inconsistency in rating among the raters. In fact, the lower reliability of the grammar
criterion was revealed in the analysis of the D studies mentioned above. If the descriptors allow much leeway for using one’s own interpretations, since judging performances involves different kinds of knowledge due to raters’ different backgrounds and characteristics, it is not expected that raters will agree with each other even though rater training is provided (e.g., Eckes, 2008; Lumley & McNamara, 1995).

The result that the EBB scale had lower practicality than the analytic scale seems to be somewhat different from the previous literature (Turner & Upshur, 1996, 2002). However, the practicality of rating scales may vary depending on other conditions and contexts such as the number of levels of each criterion. Therefore, in order to get more generalizable results, further comparisons are necessary.

Based on the results, three points need to be considered in order to create a better scale for the SRST. First, the Grammar & Vocabulary criterion needs to be revised. Counting grammatical and lexical errors seems to be difficult and inefficient since raters cannot concentrate on other criteria (i.e., Communicative Efficiency, Content, and Pronunciation). Second, as suggested above, if we stick to the current five levels in each criterion but still enhance the practicality of the EBB scale, reducing the number of criteria by combining the communicative efficiency and fluency criteria into one might be a reasonable solution. Considering the fact that three out of five raters could not rate examinees on all the four criteria by listening to a performance once, this revision is worth considering. Third, the binary format of the EBB scale may need to be considered. One idea is to adopt the format of an analytic scale but use the empirically derived descriptors as created in the present EBB scale.

Conclusion

This study attempted to show the reliability, validity, and practicality of the EBB (Empirically derived, Binary-choice, Boundary-definition) scale in comparison to the existing analytic scale, in scoring the Story Retelling Speaking Test (SRST). Three research questions (RQs) were posed. RQ1 investigated how many stories were needed to achieve high reliability in the two scales. The results of the generalizability theory suggest that in order to achieve high reliability (φ = .70 or more), two stories would be necessary in the SRST except for the grammar criterion, and that the results using the EBB scale are mostly consistent across stories and are slightly more generalizable than the results using the analytic scale. RQ2 examined whether the EBB scale was related to the analytic scale as intended. A multitrait-multimethod approach revealed that in general, correlations between the same or similar traits were higher than those between different traits. Thus, it was found that the EBB and analytic scales can generally assess the intended constructs. Lastly, to answer RQ3, practicality was compared based on the
feedback of the raters, and the EBB scale was found to be less practical than the analytic scale, mainly due to its binary format and the fact that it has more levels for each criterion.

The results did not lead us to determine which scale was more useful for the SRST. The EBB scale was slightly superior in reliability and validity, whereas the analytic scale excelled in practicality. However, the results helped us find points for revision in the scales. First, the descriptors of EBB Grammar & Vocabulary criterion should be modified. Second, the Communicative Efficiency and Content criteria of the EBB scale can be combined to enhance its practicality. Third, the current EBB binary format might be changed into one similar to the analytic scale. Since both scales have strengths, combining the good aspects of these scales may enable us to create a better scale in the future.

Implications and Further Research

Even though more than ten years have passed since the EBB scale was proposed (Upshur & Turner, 1995), EBB scales have not been constructed and used much. There may be at least two reasons for this. First, there has been little empirical investigation of EBB scales, so their characteristics have not been well understood. Another reason may be that developing an EBB scale for each test task is laborious work. Some may feel that the benefits of doing this work for each task do not justify the effort involved. The present study, however, demonstrated that the same EBB scale can be used for different stories and it functions as well as, or better than an analytic scale in the context of the Story Retelling Speaking Test. In this regard, this study is significant because it may facilitate the construction and use of an EBB scale for classroom-based assessment. In addition, using an EBB scale with multiple criteria may help students recognize the importance of these aspects of speaking ability and provide diagnostic profiles of their speaking ability.

In the future, it will be necessary to conduct scale revisions and rescore the SRST performances in order to re-examine the reliability, validity, and practicality of the revised scale. In particular, the present study revealed that there is much to be done on practicality, and this seems to be a prerequisite for classroom-based speaking assessment. We also need to employ other qualitative and quantitative methods such as many-facet Rasch measurement using FACETS (e.g., McNamara, 1996), that can detect and measure rater and task effects and provide a more detailed profile of each criterion. We hope that through these revisions and analyses, a more practical and reliable scale can be made and that this may help teachers increase classroom assessment opportunities and consequently enhance Japanese students’ speaking ability.
Acknowledgement

We are grateful to the Japan Language Testing Association for giving us a research grant for this research. This study was also partially supported by Grand-in-Aid for Scientific Research (KAKENHI) (C) (19520477).

References


Appendix A  An example of Story Retelling Speaking Test

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?

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
Appendix B  The EBB scales for the SRST

1. Communicative Efficiency (伝達能力)
Coherent story retold with no long pauses (話に一貫性があり、長いポーズがない)
No With some fluency (流暢さはややある)  With little hesitation and with few self-corrections (言いようみや言い直しがほとんどない)
No Yes Yes
With seven or more sentences (7 文以上の発話がある) No Yes
1 2 3 4 5

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 With some verbs marked for incorrect tense and aspect (いくつかの動詞の時制やアスペクトが正しく使っていないう)
Yes Yes No
With frequent grammatical and lexical errors (文法や語彙の間違いが頻繁にある)
Yes Yes No
Use of pronouns and prepositional phrases (代名詞や前置詞句を使用している)
No Yes
1 2 3

3. Content (内容)
With most of the key storylines (話の筋をほとんどカバーしている)
No Yes
With more than a few key storylines (話の筋を3つ以上述べている) Elaborations of the story with few content errors (話の詳細を含み、その内容に誤りがほとんどない)
No Yes No Yes
1 2 3 4 5

With sufficient opinions (感想が十分で適切である)
No Yes
4 5
4. **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ある)

With almost no prominent prosodic errors
(目立った韻律上の誤りがほとんどない)

<table>
<thead>
<tr>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

With frequent prosodic errors
(韻律上の誤りが頻繁にある)

<table>
<thead>
<tr>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

With a strong accent
(なまりが強い)

<table>
<thead>
<tr>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Appendix C  The Analytic Scale for the SRST

1. **Communicative efficiency**
   - 4 Very coherent story with sufficient opinions.
   - 3 Completes the story using some discourse markers (e.g., well, then, after) with few opinions.
   - 2 Occasional connection between sentences or phrases.
   - 1 Just listing, no connection between phrases or words.

2. **Grammar**
   - 4 No more than two errors during the interview.
   - 3 A few errors which do not interfere with misunderstanding.
   - 2 Frequent errors causing occasional misunderstanding.
   - 1 Control of very few major patterns, which frequently prevents communication.

3. **Fluency**
   - 4 Speech is effortless and smooth.
   - 3 Speech is occasionally hesitant.
   - 2 Speech is very slow and uneven except for short or routine sentences.
   - 1 Speech is very halting and fragmentary.

4. **Pronunciation**
   - 4 No conspicuous mispronunciations.
   - 3 Occasional mispronunciations which do not interfere with understanding.
   - 2 Mispronunciations lead to occasional misunderstanding.
   - 1 Pronunciation frequently unintelligible.

*Note.* A modified analytic rating scale of FSI (based on Nakai, 1997, p. 126).