Validity of Self-Assessment of Speech Performance:  
A Case of Japanese High School Students

Makoto FUKAZAWA  
Ibaraki Prefectural Takezono High School

Abstract

The main purpose of this study is to examine how valid self-assessment (SA) of speech performance can be in a high school setting in Japan. Fukazawa (2009) studied the validity of peer assessment (PA) and found that PA has reasonable degree of validity to support teacher assessment (TA). In the current study, another alternative assessment, SA, is the focus. Fifty-two students and three teachers participated in this study. The validity of SA was investigated from five aspects—the content, substantial, structural, external, and consequential aspect—according to Messick’s (1996) concept of validity. The validity of SA in this study means, to be more precise, the validity of interpretation and use based on self-assessment scores. The content and substantial aspects of validity were examined through analyzing a questionnaire, and both aspects were considered to have sufficient validity. The structural aspect was analyzed using Multi-faceted Rasch Measurement, and this aspect of validity was partially supported. In order to examine the external aspect of validity, the scores of SA were compared with that of TA in terms of correlations and mean scores. The result showed insufficient validity from the external aspect. The consequent aspect of validity was investigated analyzing students’ comments on SA, and some positive effects were found but validity of this aspect was not sufficient. According to these results, it was concluded that the validity of SA was shown to some extent but not sufficient to use as a part of formal assessment. Pedagogical implications are also suggested.

1. Introduction

Enhancement of practical communication ability has been emphasized in secondary school English education in Japan. In addition, the Ministry of Education, Culture, Sports, Science, & Technology (MEXT) introduced the new Course of Study (MEXT, 2009), the focus of which is to foster not only lessons conducted in English but also student-centered learning and integration of the four skills: reading, listening, writing, and speaking.

Although communicative language instruction has been improving, assessment for communicative language abilities, especially speaking skills, has not been focused
upon, even though such assessment is essential for effective language education (Akiyama, 2000; McNamara, 2000). This is because assessing speaking is one of the most challenging fields of language testing (Madsen, 1983) as speaking assessment is considered to be subjective (Brown, 1998; Hamaoka, 1995). In order to make speaking tests more reliable, ideally multiple raters are required (Hughes, 1989), but, in reality, it is difficult to find raters other than the instructor in charge of the class. A large number of students in a language classroom are another problem that makes conducting speaking tests difficult in Japan (Heaton, 1975). It is not practical to conduct speaking tests individually with forty students, which is the average number of students in a high school setting in Japan. Test administration and scoring require much time and work (Heaton, 1975), and as a result, teachers grow reluctant to conduct speaking tests.

In order to reduce these problems, making use of alternative assessments such as self-assessment (SA) and peer assessment (PA) has been considered. These assessments may provide new styles of assessing speaking so as to increase the practicality of speaking assessment. Fukazawa (2009) examined the validity of PA, and the study suggested positive outcomes for using PA for assessing speech performance. In contrast, previous studies on SA have shown various and sometimes contradictory results. Therefore, this study investigates the validity and the characteristics of SA, and suggests pedagogical implications for practical speaking assessment.

2. Previous Studies

2.1 Validity

From a traditional point of view, validity of a test is defined in such a way that “a test is said to be valid if it measures accurately what it is intended to measure” (Hughes, 1989, p. 26). According to this traditional view point, it was considered that there were five different categories of validity: content validity, construct validity, concurrent validity, predictive validity, and face validity. Essentially, Hughes describes content validity as the degree to which a test represents what it is meant to measure. Concurrent validity is the degree to which a test measures sub-categories of language skills. Predictive validity is the degree to which a test is related to another highly dependable test. Predictive validity concerns the degree to which a test predicts a test taker’s future performance. Face validity is the degree to which a test looks as if it measures what it is intended to measure.

Gradually, however, construct validity came to be considered the most important among these categories, and contains content and criterion-related validities as its sub-categories. Currently, validity is considered to be a unitary concept with construct validity as central, and it must be examined from various perspectives (Bachman, 1990). Following this concept, Messick (1996) categorized six different aspects of
validity: (a) the content aspect, (b) the substantive aspect, (c) the structural aspect, (d) the generalizability aspect, (e) the external aspect, (f) the consequential aspect. The content aspect of validity is described as the relevance between assessed content and assessment tasks. The substantive aspect examines theoretical processes are actually observed in assessment tasks. The structural aspect examines whether the observed dimensionality of response data matches the hypothesized construct domain at issue. The generalizability aspect refers to the degree to which observed scores and interpretations can be generalized to population groups, settings, and tasks. The external aspect includes the extent to which test scores and the results of other assessment are related. The consequential aspect refers to the implications of score interpretation in both positive and negative ways. The present study is conducted following Messick’s interpretation of validity.

2.2 Self-assessment

SA is defined as “any assessments that require students to judge their own language abilities or language performance” (Brown, 1998, p. 53). The present study assumes this definition of SA. Brown offers advantages and disadvantages which are common to students’ assessments such as SA and PA. They “require little extra time or resources” (pp. 53-54). On the other hand, their disadvantage is that both are relatively subjective methods of scoring and may be unreliable in high-stakes situations. In addition, a unique advantage of SA is encouragement of learner autonomy, but scores seem to be less accurate.

Some studies support the validity and reliability of SA as a formal assessment. Bachman and Palmer (1989) investigated the construct validity of self-ratings of communicative language ability with 116 nonnative English speakers of various ages, using reliability coefficient alpha, the multitrait multimethod, and confirmatory factor analysis. The results indicated that self-ratings could be reliable and valid for measuring communicative language abilities. Shore, Shore, and Thornton (1992) also compared the construct validity of SA and PA. Their research revealed that the construct validity of PA was higher than that of SA. The results also showed that observable judgments were more valid than inferential judgments. Stefani (1994) examined the validity of SA and PA, comparing with teacher assessment (TA). The correlation between SA and TA was calculated and the results show strong correlation \( r = .93 \), which is stronger than the PA-TA correlation \( r = .89 \) from the same research. Furthermore, Yamashiro (2002) supported the validity of teacher and self ratings for public speech, including SA, through structural equation modeling. However, the tendency that the reliabilities of SA \( (\alpha = .67-.85) \) for public speaking traits were lower than those of TA \( (\alpha = .82-.93) \) and PA \( (\alpha = .79-.94) \) was observed.

On the other hand, there are other studies which do not support the validity of SA. Harata (2002) studied the validity of SA compared with TA. The results indicated
mostly moderate correlations between SA and TA, and she suggested the use of SA to facilitate learner autonomy. The results also show the tendency that analytic rating helps provide a more objective judgment of students’ own language ability than holistic rating. Boud and Falchikov (1989) also threw doubt on the validity of SA because students with high English proficiency tended to underestimate their language ability, while low proficiency students had a tendency to overestimate theirs. Stefani (1994) also examined such tendencies for both SA and PA. However, contrary to Boud and Falchikov, such tendencies were not observed in Stefani’s experiment.

Although previous studies have offered valuable information on the nature of SA, the results of the previous studies are mixed and inconclusive; moreover, they tend to use different types of analyses, which makes results of previous studies difficult to compare. Therefore, the objective of this study is to examine the validity of SA from the five aspects of validity, employing Messick’s (1996) validation framework. Five aspects of validity investigated are content aspect, substantive aspect, structural aspect, external aspect, and consequential aspect. The generalizability aspect is not included because the number of participants is limited in this study.

Chapelle (1999) suggested the following way of examining validity; (a) make hypothesis, (b) show relevant evidence to examine validity, and (c) argue validity with evidence and reasoning. Based on this procedure, the following hypotheses are set for each aspect of validity.

Hypothesis 1 (for the content aspect): Students’ speeches are related to what they have studied in the class.
Hypothesis 2 (for the substantive aspect): Students think that SA assesses their English speech.
Hypothesis 3 (for the structural aspect): The result of SA of student’s speech is unidimensional.
Hypothesis 4 (for the external aspect): There is a high correlation between SA and TA.
Hypothesis 5 (for the external aspect): There is no significant difference in mean scores between SA and TA.
Hypothesis 6 (for the consequential aspect): SA has positive effects on students’ speech.

3. Method

3.1 Participants

The participants in this study were 52 Japanese public high school students in the 11th grade, and all of them belonged to a science track. Eleven students were eliminated because of their inability to attend the required number of class meetings, and the valid data became 41. This experiment was conducted in a high school where
the researcher works and the school is considered as an academically oriented high school, which emphasizes science and English education. Two Japanese teachers of English (JTEs), including the researcher, and one assistant language teacher (ALT) also participated in this study. Both JTEs have more than 10 years of high school English teaching experience, and the ALT has taught English in Japan for almost two years.

3.2 Materials

Materials used for this research were (a) a proficiency test, (b) a rater training video, (c) assessment criteria, (d) an evaluation sheet, and (e) a questionnaire. The proficiency test (α = .75) was made by the researcher, mainly adapting the Eiken English Proficiency Tests (grade 2-3). It consisted of three parts—listening comprehension, grammar, and vocabulary sections—and contained 50 questions in total. The rater training video made from the previous speech contest was used following Freeman’s (1995) suggestion that the use of a video is useful training for reliable student assessment. The assessment criteria were composed of pronunciation, accuracy, delivery, fluency, and overall evaluation (see Appendix A), which were the points emphasized in communicative activities. According to these criteria, an evaluation sheet was made, which uses 6-point Likert scale (see Appendix B), adopting the style of the Common European Framework (Council of Europe, 2001). The questionnaire was written in Japanese, and mainly asked students’ attitude toward SA.

3.3 Procedures

This research was conducted as a speech project, using six class periods of Scientific English. About two weeks before the actual speech, the topic “The thing that interests me most in science” was announced so that students would have enough time to prepare for the speech at home. In the first class, students had rater training with the training video. The students watched other high school students giving public speeches and then they had assessment practice three times. Rater training for the teachers had also been done in advance, following the same procedures. The students were told that the results of their assessments were included as a part of their grades as well as the teachers’ evaluations. In the second period, the students had a 30 minute English proficiency test. In the following classes, students made two-minute prepared speeches (13 speeches in each period). After each speech, one-minute assessment time was given. In addition to a speaker’s self-assessment, the rest of the class evaluated his/her speech as PA, and the teachers evaluated it as TA. Within a week from the last speech class, a questionnaire on the self-assessment was conducted.

3.4 Analysis

In this study, only the results of the overall assessment are analyzed of the five assessment criteria because the focus of this study is the validity of SA in general. The
five subordinate aspects of validity and the methods of analysis are summarized based on Koizumi (2005) in Table 1.

Table 1. Five Aspects of Validity and the Method of Analysis

<table>
<thead>
<tr>
<th>Aspects</th>
<th>What is examined in this study</th>
<th>Method of analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Relevance between test task and contents of lessons</td>
<td>Analysis of the questionnaire</td>
</tr>
<tr>
<td>Substantive</td>
<td>Rater’s perception of test as the assessment of their English speaking ability.</td>
<td>Analysis of the questionnaire</td>
</tr>
<tr>
<td>Structural</td>
<td>Fitness of observed scores into the theoretical model of the item response theory</td>
<td>Rasch analysis &amp; analysis of unexpected answers</td>
</tr>
<tr>
<td>External</td>
<td>Relationship between SA and TA</td>
<td>Correlation coefficient &amp; t test</td>
</tr>
<tr>
<td>Consequential</td>
<td>Positive effects of SA</td>
<td>Analysis of the questionnaire</td>
</tr>
</tbody>
</table>

In order to examine the content and substantive aspects of the validity of SA, the students’ questionnaire taken after the experiment was analyzed. For the questionnaire, a 4-point scale is used (1 = strongly agree; 2 = agree; 3 = disagree; 4 = strongly disagree; 5 = I don’t know), excluding number 5 for analysis. If the mean score of students’ answer is less than 2.5, which is the center value from 1 to 4, it is considered that students have a positive attitude toward the aspects in question.

The structural aspect is analyzed using FACETS (Linacre, 2008), the computer program of Multi-faceted Rasch Measurement (MFRM). Infit mean square value calculated by FACETS indicates the degree to which observed scores match the expected scores by the theoretical model. The expected value of the infit mean square is 1.00. As values tend toward 1.00, observed ratings are closer to expected ratings from the Rasch model. McNamara (1996) stated “more accurately, for n sizes of 30 or more, the range is the mean ± twice the standard deviation of the mean square statistic” (p. 181). This study takes this standard. Raters and speakers who do not fit into this theoretical model are considered as misfit raters and speakers. If the rate of misfit raters is less than 10% of all the raters (Stansfield & Kenyon, 1996) and the rate of misfit speakers is less than 2% of all the speakers (McNamara, 1996), the assessment is supposed to have unidimensionality, which indicates that the assessment measures one specific domain. In that case, enough validity of SA in terms of the construct aspect is implied. In this study, however, unidimensionality of SA cannot be investigated with MFRM, separating from the other assessments, because SA has only one score. Therefore, to be more specific, the rates of unexpected responses, which are observed scores with standard residuals greater than ±2, and the infit mean square values of each assessment as a whole, are also examined. In addition, the tendency of unexpected answers is investigated, dividing the students into high and low proficiency.
groups according to the result of the proficiency test for further information.

The external aspect is investigated using Pearson’s correlation coefficient and t-test. The correlation coefficient is used to determine the level of agreement between SA and TA. The interpretation of the correlation coefficient $|r|$ is as follows: less than ±.20 is considered as almost no correlation; from ±.20 to ±.40 is low correlation; from ±.40 to ±.70 is moderate correlation; more than ±.70 is high correlation (Kiyokawa, 1990). A t test is also conducted to compare the mean scores statistically. It examines whether or not the observed scores of SA and the average scores of TA are statistically different. In the case that both the Pearson’s correlation coefficient is .70 or higher and there is no significant difference in the t test, sufficient validity in terms of external aspect is confirmed.

The consequential aspect of validity is investigated by checking students’ positive comments on SA on their questionnaire. If more than half of the students write positive comments, SA is considered to have sufficient validity from the consequential aspect.

4. Results

4.1 Content Aspect

The content aspect of validity of SA was investigated based on the result of the students’ questionnaire (see Appendix C). This aspect refers to the relevance between class contents and students’ task for assessment. The following is the hypothesis to examine this aspect.

Hypothesis 1: Students’ speeches are related to what they have studied in the class.

In the questionnaire, Question 12 asked the relevance between the contents of the English class and their speeches, and Question 13 asked the relevance between assessment criteria and speaking activities done in the English class. Table 2 summarizes the results of the students’ answers.

<table>
<thead>
<tr>
<th>Table 2. Students’ Reaction to the Questions on the Content Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
</tr>
<tr>
<td>(12) relevance between contents of the class and students’ speeches</td>
</tr>
<tr>
<td>(13) relevance between assessment criteria and speaking activities in class</td>
</tr>
</tbody>
</table>

*Note. $N = 41; M = mean; SD = standard deviation.*

According to the standard mentioned in Section 3.4, if the mean score is less than 2.5, it is assumed that students thought their speeches are related to what they have
studied. The mean scores of both of the questions were below the standard. Therefore, the Hypothesis 1 was supported and it is considered that SA has the content aspect of validity sufficiently.

4.2 Substantive Aspect

This aspect of validity examines whether theoretical processes are actually conducted in assessment tasks. It is investigated by analyzing the students’ answer of the questionnaire on SA. The following hypothesis is set to examine the substantive aspect of validity.

Hypothesis 2: Students think that SA assesses their English speech.

Question 2 in the questionnaire asked “I think SA assesses my own speech.” (see Appendix C). According to the students’ answers, 68.29% of the students thought they were somewhat able to assess their own speeches by themselves, while 14.63% answered that they did not think that they could assess their own speeches more or less.

Table 3. Students’ Reaction to the Questions on the Substantive Aspect

<table>
<thead>
<tr>
<th>Question</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2) SA can assess student’s own speech.</td>
<td>2.03</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Note. N = 41; M = mean; SD = standard deviation.

Table 3 indicates the mean score of this answer \( M = 2.03 \) was less than 2.5, the central value. Therefore, Hypothesis 2 was supported, which suggests sufficient validity of SA in terms of the substantive aspect. However, it should be mentioned that 17.07% of the students answered “I don’t know”, which was a rather high rate.

4.3 Structural Aspect

For the inquiry into the structural aspect of validity, unidimensionality was examined using MFRM. If SA fits into the theoretical model, unidimensionality of assessment is considered to be shown (McNamara, 1996). Therefore, the following hypothesis was made to examine this aspect.

Hypothesis 3: The result of SA of student’s speech is unidimensional.

In order to run FACETS (Linacre, 2008), however, all of the observed scores including TA, PA and SA were needed because of two reasons. First, each SA had only one observed score and it was not enough to run the computer program. Second, the other assessments were necessary to compare with SA. Therefore, Rasch Analysis...
including all three assessments was conducted. Table 4 shows the result of the analysis. Based on McNamara's (1996) formula, the acceptable range of the infit mean square value for the raters was from .22 to 1.70, while that for the speakers was from .42 to 1.50. There were two misfitting raters among 44 raters including three teachers, and the rate of the misfitting raters was 4.55%. On the other hand, there were three misfitting speakers out of 41 speakers, and the rate of misfitting speakers was 7.32%, which was over 2% and did not match the standard. Therefore, the requirements for unidimensionality of the assessment were not fully satisfied. This result partially supported the Hypothesis 3, which suggests the structural aspect of validity was somewhat limited.

Table 4. Infit Mean Squares and Acceptable Ranges

<table>
<thead>
<tr>
<th>Raters</th>
<th>M of Infit</th>
<th>SD of Infit</th>
<th>Acceptable Range</th>
<th>Numbers of Misfit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers</td>
<td>.96</td>
<td>.37</td>
<td>.22 - 1.70</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>.96</td>
<td>.27</td>
<td>.42 - 1.50</td>
<td>3</td>
</tr>
</tbody>
</table>

Note. N = 44(Rater); N = 41(Speaker); M = mean; SD = standard deviation

In addition, in order to take a closer look at SA, the rates of unexpected responses of SA, PA, and TA were also investigated. In this study, unexpected responses are defined as observed scores with standardized residuals greater than ±2, which are considered to have significant differences from expected scores given by the Rasch model. Table 5 shows the proportion of unexpected responses of each assessment. SA apparently had higher rate of unexpected responses (12.20%), while both TA and PA had similar rates of those responses (4.96%, 4.23%; respectively).

Table 5. Rate of Unexpected Responses of Each Assessment

<table>
<thead>
<tr>
<th>Status</th>
<th>Total Count</th>
<th>Unexpected Response</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA (n=41)</td>
<td>41</td>
<td>5</td>
<td>12.20%</td>
</tr>
<tr>
<td>PA (n=41)</td>
<td>1630</td>
<td>69</td>
<td>4.23%</td>
</tr>
<tr>
<td>TA (n=3)</td>
<td>121</td>
<td>6</td>
<td>4.96%</td>
</tr>
</tbody>
</table>

For additional information, the relationship between the unexpected responses and students' proficiency level was investigated since Boud and Falchikov (1989) mentioned the tendency of high proficiency students' underestimation and low proficiency students' overestimation as the reason of insufficient validity of SA. Table 6 indicates five students' unexpected responses of SA and their English proficiency levels. According to the result of the proficiency test which all the students took at the beginning of this experiment, they were divided into two groups, high and low proficiency groups. Then, t-test was performed to check the statistical differences in proficiency. The result of the t-test confirmed the significant differences between high
and low proficiency group ($t (39) = 9.81, p = .00$). As Table 6 shows, the students with high English proficiency gave themselves lower scores than expected, while those who belonged to low proficiency group gave themselves higher scores than the model expected, which supported Boud and Falchikov.

Table 6. Unexpected Responses of SA

<table>
<thead>
<tr>
<th>Student</th>
<th>Observed Score</th>
<th>Expected Score</th>
<th>Residual</th>
<th>Proficiency Test Score</th>
<th>Proficiency Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2.9</td>
<td>-1.9</td>
<td>39</td>
<td>High</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>5.3</td>
<td>-1.3</td>
<td>37</td>
<td>High</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>1.7</td>
<td>1.3</td>
<td>25</td>
<td>Low</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>2.7</td>
<td>1.3</td>
<td>23</td>
<td>Low</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>3.8</td>
<td>1.2</td>
<td>27</td>
<td>Low</td>
</tr>
</tbody>
</table>

Furthermore, the infit mean square values of each assessment as a whole were also examined. Table 7 indicates the result of the Rasch analysis. The mean score of the infit mean square value was 1.13 and the standard deviation was .27. According to McNamara’s (1996) standard, the acceptable range was 0.59 to 1.67 ($1.13 \pm [2 \times 0.27]$). The infit mean square value of SA was 1.44, that of TA was 1.01 and that of PA was 0.94. The infit mean square values of each assessment group fit into the acceptable range. However, the value of SA was apparently higher than 1.00 which shows expected rating by the Rasch model. This indicates distortion of the assessment to some extent; meanwhile the results of TA and PA were much closer to the ideal rating.

Table 7. Rater Measurement Report of SA, PA, and TA

<table>
<thead>
<tr>
<th></th>
<th>Total Count</th>
<th>Obsvd Avg</th>
<th>Measure</th>
<th>Infit Mnsq</th>
<th>ZStd</th>
<th>Outfit Mnsq</th>
<th>ZStd</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>41</td>
<td>3.29</td>
<td>1.04</td>
<td>1.44</td>
<td>1.8</td>
<td>1.45</td>
<td>1.8</td>
</tr>
<tr>
<td>PA</td>
<td>1630</td>
<td>3.81</td>
<td>-0.49</td>
<td>0.94</td>
<td>-1.5</td>
<td>0.95</td>
<td>-1.5</td>
</tr>
<tr>
<td>TA</td>
<td>121</td>
<td>3.86</td>
<td>-0.55</td>
<td>1.01</td>
<td>0.1</td>
<td>1.02</td>
<td>0.1</td>
</tr>
<tr>
<td>M</td>
<td>597.3</td>
<td>3.65</td>
<td>0.0</td>
<td>1.13</td>
<td>0.1</td>
<td>1.14</td>
<td>0.2</td>
</tr>
<tr>
<td>SD</td>
<td>895.2</td>
<td>0.31</td>
<td>0.9</td>
<td>0.27</td>
<td>1.7</td>
<td>0.27</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Note. $n=3$(TA), $n=41$(SA), $n=41$(PA). $M =$ mean. $SD =$ standard deviation.

Although the Hypothesis 3 was partially supported as overall assessment of speech performance, including SA, other results implied that there might be a tendency that SA does not fit into the theoretical model as much as PA or TA do.

4.4 External Aspect

The external aspect of validity of SA was investigated, comparing the result of TA as the external standard. In order to do so, not only correlation coefficients between SA and the average score of TA but also the mean scores of both assessments were examined. It is because, even if the correlation between the two assessments is high,
their scores might be different if their severities are different. Therefore, both correlation and the mean scores were compared to be more precise, and the following hypotheses were set to examine the external aspect.

Hypothesis 4: There is high correlation between SA and TA.
Hypothesis 5: There is no significant difference in mean scores between SA and TA.

Before discussing the above analyses, the descriptive analysis is shown in Table 8. The average score of SA was lower than that of TA, while the standard deviation of SA was higher than that of TA. It seems that SA was severer than TA and had wider range.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA (n=41)</td>
<td>3.29</td>
<td>.81</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>TA (n=3)</td>
<td>3.85</td>
<td>.75</td>
<td>2.33</td>
<td>5.67</td>
</tr>
</tbody>
</table>

Note. M = mean. SD = standard deviation. Min = minimum. Max = maximum.

First, the correlation between the observed score of SA and the average scores of TA was calculated. The correlation coefficient was moderate \( r = .46, p < .01 \), according to Kiyokawa (1990). Therefore, Hypothesis 4 was not supported.

Second, a t test between the observed score of SA and the average scores of TA was conducted. The results of the test indicated that the difference between the scores of SA and TA was statistically different \( t (82) = 2.33, p = .02 \). Therefore, Hypothesis 5 was not supported, either. According to these results, it is concluded that the validity of SA in terms of the external aspect is not sufficient.

4.5 Consequential Aspect

The consequential aspect of validity refers to short and long-term effects of assessment. It was investigated through analyzing the students' comments on SA in the questionnaire. In order to achieve this goal, the following hypothesis was examined.

Hypothesis 6: SA has positive effects on students' speech.

This hypothesis was examined calculating the rate of students' positive comments on SA. Question 11 asked students to write what they learned through SA. Since this question was asking positive effects of SA, any answers were basically considered as positive responses. Eighteen out of 41 students wrote some comments on this question, which is 43.9 % of all the students. Although the rate was relatively high, it was less than 50%. According to the standard suggested in section 3.4, Hypothesis 6 was not fully supported.
For more detailed information on students' response toward SA, their answers were divided into three categories by the researcher; (a) direct positive impact, (b) enhancing motivation, and (c) confirmation.

Table 9. Categories of Students' Positive Responses

<table>
<thead>
<tr>
<th>Category</th>
<th>Response</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Positive Impact</td>
<td>5</td>
<td>27.8%</td>
</tr>
<tr>
<td>Enhancing Motivation</td>
<td>11</td>
<td>61.1%</td>
</tr>
<tr>
<td>Confirmation</td>
<td>2</td>
<td>11.1%</td>
</tr>
</tbody>
</table>

Note. N = 18.

The first category represents students’ responses which indicate students’ having direct benefits through SA. As Table 9 shows, 27.8% of responses belonged to this category. Some of their comments were, “I got a chance to be conscious about my own pronunciation and intonation,” “I came to think of how my messages are conveyed to the audience,” and “I can now judge my own speech objectively.”

Students’ responses that review their own speech and try to make use of it for their next speech fall into the second category. This was the most common category, to which more than 60% of the responses belong. One of the responses said, “Through SA, I could recognize my good and bad points, and I would like to make use of them for the next speech.” This was the most common response in this category, and there were eight other responses which had similar comments. The other comments were “Through reviewing my bad points, I am motivated to improve those points,” and “It would be better for me not to get nervous too much and to use more facial expressions next time.”

The last category, confirmation, represents the students’ responses which confirm their present ability or skills. Two responses fell into this category and they were “I became aware of my present English level,” or “It was a good opportunity for me to be aware of my English ability.”

Although Hypothesis 6 was not fully supported, many students still felt the positive effect of SA such as above. Therefore, it seems possible to conclude that SA still has the consequential aspect of validity to some extent.

5 Discussion

Validity of SA was examined from five aspects according to Messick’s (1996) concept of validity. First, the content aspect was investigated by analyzing the students’ answers on the questionnaire on SA. The students tend to think that their speech and what they had learned in the class were related, which supported the first hypothesis. According to the result, SA appeared to have the content aspect of validity.
sufficiently. Second, the substantial aspect was also checked analyzing the questionnaire. More than half of the students thought they evaluated their speeches in a valid manner, and Hypothesis 2 was also supported, which suggested reasonable validity in terms of the substantive aspect. Third, the structural aspect was examined using MFRM. Unidimensionality of the speech assessment including SA was partially seen, and the structural aspect of validity was also partially supported. The fourth aspect, the external aspect of validity was inquired, comparing to TA. The correlation between SA and TA was moderate \( r = .46, p < .01 \), which was the same result as Harata’s (2002) study on SA; furthermore, significant difference in mean scores between the two assessments was shown \( t (82) = 2.33, p = .02 \). Therefore, both Hypotheses 4 and 5 were not supported, and the external aspect of validity seemed insufficient. Fifth, the consequential aspect was investigated through analyzing students’ comments on the questionnaire. Nearly 44% of the students felt positive effect of SA, but Hypothesis 6 was not fully supported and the consequential aspect of validity seemed not sufficient.

According to these results, it is reasonable to conclude that SA seems to have validity to some extent but not sufficient to use it as a part of summative assessment such as term tests mainly because the apparent difference was observed between SA and TA. Further investigation on infit mean square value (1.44) of SA and the rate of unexpected answers (12.20%) also threw doubt on validity of SA. The analysis of the unexpected answers on SA showed the tendency that the students with high English proficiency underestimated themselves, while the students with low proficiency overestimated themselves. The same tendency was observed by Boud and Falchikov (1989). In addition, more than 17% of the students selected “I don’t know” when they were asked whether they could assess their own speeches, which might reflect their uncertainty of their SA as a part of formal assessment.

It is interesting that the results of this study on SA appear to be quite different from the result of Fukazawa (2009) on PA, which was conducted and analyzed with almost the same procedures using the same validity framework. According to that study, sufficient validity of PA was indicated in five aspects—the content, substantial, external, consequential, and generalizability aspects—and the structural aspects of validity was partially seen in PA. As a result, it was concluded that PA had comparable validity to TA. Compared to the results on PA, SA seems to have relatively lower validity, which is the similar tendency observed in Shore, Shore, and Thornton (1992) and Yamashiro (2002).

### 6. Conclusions and Implications

This study investigated the validity of SA in terms of the content, substantial, structural, external, and consequent aspects, based on Messick’s (1996) validation
framework. The validity of SA was argued according to the validation process. Chapelle (1999) suggested; (a) hypotheses were made in section 2.2, (b) relevant evidence were shown for validity examination in section 4, and (c) the validity was discussed with evidence and reasoning in section 5. The results showed that SA seemed to have sufficient validity from the content and substantial aspects. The structural and consequent aspects of validity was shown to some extent but was not sufficient. In terms of the external aspects, the validity of SA was considered to be not enough. Considering these results, it is suggested that SA has validity to some extent but it is not sufficient as summative assessment.

There were other findings in this study. First, the higher rate of unexpected answers and higher infit mean square values of SA from Rasch Analysis suggested that the score of SA might have a tendency to be distorted compared to other assessments, TA and PA. This result is quite similar to Yamashiro (2002) which supported the validity of TA, PA and SA for public speech, but less reliability for SA. Second, some positive effects of SA was found from the students’ comments on the questionnaire, such as giving positive feedback for students’ speeches, enhancing their motivation toward studying English, and making them realize their present English levels.

These findings suggested two pedagogical implications. One is that it may be difficult to use SA as a part of summative assessment, unlike PA whose validity as a summative assessment was reasonably supported (Fukazawa, 2009), although SA seems to have some validity as this study indicated. In addition to less validity of SA compared to PA or TA, the problem arises when students' observed scores are directly used in case of SA, which has more impact on their grades compared to PA where observed scores are averaged. Some positive washback of SA, however, lead to the other implication. It would be more effective to use SA for giving students opportunities to review their performance, making use of lessons they have, and motivating them for the next performance. Brown (1998) and Harata (2002) also point out the encouragement of learner autonomy as an advantage of SA. Using SA as formative assessment is not a new concept, but in reality SA has not been used effectively enough.

It is important to mention the limitations of this study. First, the amount of data is limited. In addition, speech is only one type of speaking activity. Using different speaking tasks might change the results. Therefore, it may be risky to generalize the results from only this study. Second, the evidence for each aspect of validity to prove is also limited. For example, Question 2 for investigating content validity also has an aspect of face validity. Therefore, more evidence for each validity aspect needs collecting. Third, quantitative analysis is primarily used this study. However, more qualitative analysis is essential in order to deepen the understanding of SA, for example finding out the reasons of students’ uneasiness toward SA. Therefore, further study is needed.
Today, practical communication skills, including speaking, are emphasized more than ever in English classes in high school according to the new Course of Study. At the same time, assessments of those skills must be more focused upon. For practical and effective assessment of students’ speaking skills, alternative assessments such as SA and PA have an important role to play. It is important to know the characteristics of SA and PA, and to make the most use of these assessments in order to improve students’ language skills, enhance their learner autonomy, and encourage teachers to provide students with more opportunities for communication in a language classroom.

Acknowledgements

This research was partially supported by the Grant-in-Aid for Scientific Research (KAKENHI; No. 22908023, Term 2010) by the Ministry of Education, Culture, Sports, Science and Technology. I would like to thank Akihiko Mochizuki, Rie Koizumi, Masatoshi Nagahashi, and Chris Hale for giving me precious advice.

References


## 評価基準確認表

<table>
<thead>
<tr>
<th>項目</th>
<th>基準</th>
<th>Basic（初級）</th>
<th>Intermediate（中級）</th>
<th>Advanced（上級）</th>
</tr>
</thead>
<tbody>
<tr>
<td>発音</td>
<td>声の大きさ</td>
<td>開き取るのが難しい。</td>
<td>開き取ることが可能である。</td>
<td>良く開き取ることができる。</td>
</tr>
<tr>
<td></td>
<td>発音</td>
<td>開き取取どのように難い。</td>
<td>開き取取どのように可能である。</td>
<td>良く開き取取どのように可能である。</td>
</tr>
<tr>
<td>正文</td>
<td>間違いが多い。</td>
<td>間違いがあるが、まずまず正確</td>
<td>間違いが少ない。</td>
<td></td>
</tr>
<tr>
<td></td>
<td>間違いが多く、文として高品質に含まれている場合もある。</td>
<td>間違いのない文</td>
<td>間違いほとんど無し</td>
<td></td>
</tr>
<tr>
<td>流暢性</td>
<td>ときめのなさ</td>
<td>話が中断してしまう。</td>
<td>時々、話が中止されることがある。</td>
<td>話が自然に流れ。</td>
</tr>
<tr>
<td></td>
<td>強弱のリズム</td>
<td>間違い大変多い/単語レベルの発話</td>
<td>間違い大変多い/音節レベルの発話</td>
<td>間違い大変多い/文章レベルの発話</td>
</tr>
<tr>
<td></td>
<td></td>
<td>間違い大変多い/単語レベルの発話</td>
<td>間違い大変多い/音節レベルの発話</td>
<td>間違い大変多い/文章レベルの発話</td>
</tr>
<tr>
<td></td>
<td></td>
<td>間違い大変多い/単語レベルの発話</td>
<td>間違い大変多い/音節レベルの発話</td>
<td>間違い大変多い/文章レベルの発話</td>
</tr>
<tr>
<td>態度</td>
<td>ディスクリプター</td>
<td>積極的、に話そうという姿勢が弱い。</td>
<td>積極的、に話そうとしている。</td>
<td>積極的、効果的、に話している。</td>
</tr>
<tr>
<td></td>
<td></td>
<td>アイ・コンタクト、ジェスチャー、表情等が弱い。</td>
<td>アイ・コンタクト、ジェスチャー、表情等を有している。</td>
<td>アイ・コンタクト、ジェスチャー、表情等が多く、効果的に使われている。</td>
</tr>
<tr>
<td>性</td>
<td>機械的ではない</td>
<td>少し積極的</td>
<td>おおむね機械的</td>
<td>積極的、非常機械的</td>
</tr>
<tr>
<td>性</td>
<td>難度：×</td>
<td>難度：△</td>
<td>難度：△</td>
<td>難度：△</td>
</tr>
<tr>
<td>総合評価</td>
<td>内容の理解が難しい。</td>
<td>意境に伝わる。</td>
<td>意境に伝わる。</td>
<td></td>
</tr>
<tr>
<td></td>
<td>内容の理解が難しい。</td>
<td>意境に伝わる。</td>
<td>意境に伝わる。</td>
<td></td>
</tr>
<tr>
<td></td>
<td>間違いが多く、文として高品質に含まれている場合もある。</td>
<td>間違いが多く、文として高品質に含まれている場合もある。</td>
<td>間違いが多く、文として高品質に含まれている場合もある。</td>
<td></td>
</tr>
<tr>
<td></td>
<td>間違いの大変多い/単語レベルの発話</td>
<td>間違いの大変多い/音節レベルの発話</td>
<td>間違いの大変多い/文章レベルの発話</td>
<td></td>
</tr>
<tr>
<td></td>
<td>間違いの大変多い/単語レベルの発話</td>
<td>間違いの大変多い/音節レベルの発話</td>
<td>間違いの大変多い/文章レベルの発話</td>
<td></td>
</tr>
<tr>
<td></td>
<td>間違いの大変多い/単語レベルの発話</td>
<td>間違いの大変多い/音節レベルの発話</td>
<td>間違いの大変多い/文章レベルの発話</td>
<td></td>
</tr>
<tr>
<td></td>
<td>間違いの大変多い/単語レベルの発話</td>
<td>間違いの大変多い/音節レベルの発話</td>
<td>間違いの大変多い/文章レベルの発話</td>
<td></td>
</tr>
</tbody>
</table>

---

Appendix A: Assessment Criteria
スピーーチ自己評価用紙

発表者（年組番氏名）

★適当であると思われるものに〇を付けなさい。

<table>
<thead>
<tr>
<th>評価項目</th>
<th>評価</th>
<th>良かった点に〇</th>
</tr>
</thead>
</table>
| (1) 発音
（声の大きさ・発音の明瞭さ） | Basic Intermediate Advanced | 1 声の大きさ 2 発音の明瞭さ |
| (2) 正確さ
（文法） | A1 A2 B1 B2 C1 C2 | 1 文法 |
| (3) 流暢さ
（とぎれのなさ・強弱のリズム） | A1 A2 B1 B2 C1 C2 | 1 とぎれのなさ 2 強弱のリズム |
| (4) 発表態度
（アイコンタクト・ジェスチャー・積極的態度） | A1 A2 B1 B2 C1 C2 | 1 アイコンタクト 2 ジェスチャー 3 積極性 |
| (5) 総合評価
（上記項目＋内容・構成等） | A1 A2 B1 B2 C1 C2 | 1 内容 2 構成 3 その他（ ） |
Appendix C  Examples of Questions in the Questionnaire

自己評価についてのアンケート

（2）自己評価は、発表者（自己）のスピーチの能力を評価していると思う。
①とてもそう思う  ②ややそう思う
③あまりそう思わない  ④全くそう思わない  ⑤わからない

（11）自己評価をすることで、英語学習や他の面で勉強になった点があれば記入してください。

（12）今回のスピーチは、科学英語の授業内容と関連していると思う。
①とてもそう思う  ②ややそう思う
③あまりそう思わない  ④全くそう思わない  ⑤わからない

（13）スピーチの評価項目（発音・正確さ・流暢さ・態度・総合）は、これまでの英語の授業で行ってきたスピーキングの活動と関連していると思う。
①とてもそう思う  ②ややそう思う
③あまりそう思わない  ④全くそう思わない  ⑤わからない