

Self, Peer and Teacher Evaluation in English Oral Production

Yoko Suganuma Oi

School of Education, Waseda University,

yokosuganuma@suou.waseda.jp

Abstract

The paper discusses peer evaluation between students, and the correlation between teacher evaluation and peer evaluation in oral production in English. It also discusses the relationship of the criteria, delivery, language use, and topic development in self, peer, and teacher evaluation. The participants in this study were 92 Japanese public senior high school students, two Japanese English teachers and one native English teacher. The students were given a task to evaluate one minute speeches on the topic such as “Things that interest me”. After each speech, students and teachers evaluated the speaker’s speech. Speakers were also asked to evaluate themselves (self-evaluation). The result showed that there was not a significant correlation between teacher evaluation and peer evaluation. On the other hand, the self-evaluation showed a high consistency with teacher evaluation. The objective of the study is to explore an effective way to use peer evaluation and self-evaluation in school.

Keywords

Oral production, self-evaluation, peer evaluation, Japanese English teacher evaluation, native English teacher evaluation

Introduction

In this paper oral production stands for presentation and discussion. Peer evaluation is defined as the activity that students rate other classmate’s oral production. As Cheng et al. (2005) points out, peer evaluation in English oral production class is helpful for students because it provides learners with the opportunity to take responsibility for analyzing, monitoring and evaluating aspects of both the learning process and product of their peers. I have also given my students a chance to evaluate themselves in class. However, there are often discrepancy between teacher evaluation and peer evaluation. For this reason the present paper investigates whether there is inconsistency between teacher evaluation and peer evaluation.

Previous studies

As Patri (2002) discusses that self-evaluation and peer evaluation has been seen as having significant pedagogic relations with learner independence and learner autonomy. Furthermore self and peer evaluation are becoming more important as an alternative assessment method since they could encourage learners to be involved in the assessment process (Boud, 1999, pp.121-133). The involvement in evaluation is expected to increase students’ motivation and autonomy in language learning (Blanche & Merino, 1989, pp. 313-340; Brown & Hudson, 1998, pp.653-675). However, Cheng et al. (2005) argued that students and teachers were different in their respective evaluating behaviors and the ways oral and written language proficiency were assessed. Therefore self and peer evaluation of learners’ performance or ability have been considered to be sub-assessment for teachers (Brown & Hudson, 1998, pp. 653-675). Sub-assessment means that peer evaluation supports teachers’ final decision of their evaluation. The next section discusses evaluation in terms of native English teacher vs. non-native English teacher.

Traditionally, a native English teacher has been defined as a speaker of Standard English (Davies, 1991) and as such has long served as the norms for language teaching and testing (Zhang & Elder, 2011, p.32). Likewise, native speaking raters are commonly employed to mark tests, and native speaker standards are invoked in drawing up assessment criteria (Davies, 2004, pp. 431-450). However, in the wake of the rapid spread of English as a lingua franca in world communication, the status of the native speaker norm has been widely challenged. Zhang and Elder investigate the possibility of differing orientations to the oral proficiency construct of Chinese undergraduates by native and non-native English speaking teacher assessors. The results yielded by both quantitative and qualitative analyses of teacher data revealed no significant difference in raters’ holistic judgments of the speech samples and a broad level of agreement between groups on the construct components of oral English proficiency. Zhang and

Elder (2011) conclude that the analysis of raters' comments revealed both quantitative and qualitative differences in the way native English speaker and non-native English speaker teachers weighed various features of the oral proficiency construct in justifying the decisions made (2011, p. 31).

1 The study

1.1 Research question

The data presented in this paper is designed to explore the following questions:

- 1) To what extent does peer evaluation of English speech correspond to teacher evaluation?
- 2) To what extent does self-evaluation of English speech correspond to teacher evaluation?
- 3) Does native English teachers' evaluation correspond to Japanese teacher's evaluation?

1.2 Participants

The participants in this study were 92 second year (16-17 years old) senior high school students of a prefectural senior high school in Kanagawa in Japan. The participants comprised of 41 female (44.6%) and 51 male (55.4%). They met twice a week for 50 minutes for English lessons. Two Japanese English teachers, one is a regular teacher and the other is a guest teacher, and one native English teacher also participated in the survey. The survey was conducted in the school's compulsory English writing classes. The students were classified into five groups which composed of 19 to 20 students. The class is not related to their English proficiency. Table 1 shows the length of their stay in foreign countries.

Table 1: The length of stay in foreign countries

Years	8	7	6	5	4	3	2	1	0
Persons	1	0	1	0	1	3	2	0	84
	U.S.		U.S.		China	India	U.S.		
						Germany			
						Hong Kong			

1.3 Task

The students were asked to evaluate one minute speeches of students on the three familiar topics, "Things that interest me", "What I like", and "In my future". They were not allowed to make a speech with notes. After each speech, the students and teachers evaluated the speaker's speech. Speakers were also asked to evaluate themselves (self-evaluation). The Participants were encouraged to ask a speaker to say a question or opinion about the content of speech. They were also required to take the Measure of English Grammar Test (Shimizu, et al., 2006) to assess their English grammar level before the experiment. Table 2 indicates the result of MEG test of the students.

Table 2: The SD and Mean of the students

N	MAX	MIN	MEAN	SD
92	33	3	18.2	5.69

Note: The maximum scale of 35 points.

1.4 Rating scale

An analytic rating scale is made by a researcher, referring to the EIKEN (the Society for Testing English Proficiency) Can-do List (2006), TOEFL Test Integrated Speaking Rubrics (Educational Testing Service, 2004), and Common European Framework of Reference for Languages (Council of Europe, 2001). It incorporates four categories: General Description, Delivery, Language Use, and Topic Development. The rating scales are the five categories for bands 5-1. "5" indicates the highest evaluation, and "1" indicates the lowest evaluation. All participants in the survey learned how to use this scale.

1.5 Data collection

1.5.1 Evaluation sheet

Teachers and students received evaluation sheets to evaluate other participants and themselves after each speech. The evaluation sheets were collected at the end of the class. The evaluation time was 1-2 minutes. The evaluation sheet for students were translated from English into Japanese and modified for the Japanese level of senior high school students (see Appendix A). Japanese English teachers and a native English teacher used the same evaluation sheet written in English (see Appendix B).

1.5.2 Questionnaires

Two questionnaires were given to students. The first one asked them about their English education backgrounds. The second one asked them their opinions and impressions on peer evaluation and self-evaluation. Teachers were also given the questionnaire to ask how to use peer evaluation and self-evaluation in their English class. The result of the questionnaires was used to analyze the relationships of consistency between teacher evaluation, self, and peer evaluation.

1.5.3 Tools

Pearson product-moment correlation coefficient (Pearson's *r*) which is the statistic used for determining the correlation of two sets of continuous scale data was used in the analysis. It analyzes the correlation of evaluations between: 1) The first Japanese English teacher (JET1) and peer evaluation 2) The second Japanese English teacher (JET2) and peer evaluation 3) native English teacher (NET) and peer evaluation 4) JET1 and JET2 5) JET1 and NET 6) JET2 and NET. The individual evaluation of students is deducted from

teacher evaluation to see the difference of agreement. 7) JET1 and self-evaluation 8) JET2 and self-evaluation 9) NET and self-evaluation One minute speeches of 92 students and teachers' comments on speeches were recorded and transcribed.

2 Results

2.1 The correlation between peer evaluation and teacher evaluation

Tables 3, 4, 5, and 6 show the number of the students who score Pearson's r is more than 0.7, more than 0.4 and less than 0.7, and less than 0.4. If Pearson's r is higher than 0.7, the correlation between the two values could be said to be very high. However, higher correlation values are not found in all four categories. Moreover most of the values are less than 0.4. It means that students' evaluation does not correlate with teachers' evaluation.

Table 3: The number of Pearson's r between three teachers and students in General Description

Pearson's r	JE1&students	JE2&students	NET&students
$r \geq 0.7$	0	1	2 persons
$0.7 > r > 0.4$	10	21	5
$0.4 > r$	82 (89.1%)	70 (76.1%)	85 (92.4%)

Table 4: The number of Pearson's r between three teachers and students in Delivery

Pearson's r	JE1&students	JE2&students	NET&students
$r \geq 0.7$	0	1	0 persons
$0.7 > r > 0.4$	14	17	20
$0.4 > r$	78 (84.8%)	74 (80.4%)	72 (78.3%)

Table 5: The number of Pearson's r between three teachers and students in Language Use

Pearson's r	JE1&students	JE2&students	NET&students
$r \geq 0.7$	0	0	1 persons
$0.7 > r > 0.4$	12	23	26
$0.4 > r$	80 (86.9%)	69 (75%)	65 (70.6%)

Table 6: The number of Pearson's r between three teachers and students in Topic Development

Pearson's r	JE1&students	JE2&students	NET&students
$r \geq 0.7$	0	0	5 persons
$0.7 > r > 0.4$	6	21	22
$0.4 > r$	86 (93.4%)	71 (77.1%)	65 (70.6%)

2.2 The correlation between Pearson's r of three teachers-peer evaluation and T-score of MEG test

Tables 7, 8 and 9 show the correlation between Pearson's r (teacher and peer evaluations) and T-score of MEG test. The result is that there is not positive correlation between Pearson's r and

T-score. Especially the correlation of NET indicates a strong inverse correlation. Only the category of Topic Development indicates slight positive correlations in all three teachers' correlations. The combination of JET1 and students tend to correlate slightly with T-score.

Table 7: The correlation between Pearson's r of JET1-peer evaluation and T-score of MEG test

General Description	Delivery	Language Use	Topic Development
0.14	0.15	0.02	0.003

Table 8: The correlation between Pearson's r of JET2-peer evaluation and T-score of MEG test

General Description	Delivery	Language Use	Topic Development
-0.02	0.11	0.08	0.27

Table 9: The correlation between Pearson's r of NET-peer evaluation and T-score of MEG test

General Description	Delivery	Language Use	Topic Development
-0.15	-0.05	-0.01	0.22

Tables 10, 11, and 12 show the correlation between Pearson's r (teachers and peer evaluations) and T-scores ≥ 60 points. The correlations between Pearson's r and T-scores ≥ 60 points do not present a strong correlation except the category of Topic Development of JET1 & peer evaluation. Compared to the other two teachers (JET2 and NET), Pearson's r between JET1 and students tends to correlate with the students' group with the higher T-score.

Table 10: The correlation between Pearson's r of JET1-peer evaluation and T-score ≥ 60 points of MEG test

General Description	Delivery	Language Use	Topic Development
0.11	-0.48	-0.68	0.47

Table 11: The correlation between Pearson's r of JET2-peer evaluation and T-score ≥ 60 points of MEG test

General Description	Delivery	Language Use	Topic Development
-0.52	-0.34	0.04	-0.07

Table 12: The correlation between Pearson's r of native English teacher-peer evaluation and T-score ≥ 60 points of MEG test

<i>General Description</i>	<i>Delivery</i>	<i>Language Use</i>	<i>Topic Development</i>
-0.31	-0.09	0.18	-0.14

2.3 The correlation between self-evaluation and teachers' evaluations

To see to what extent peer evaluation of English speech corresponds to teacher evaluation, the individual evaluation of students is deducted from teacher evaluation. Tables 13, 14, 15, and 16 present the difference between self-evaluation and teacher evaluation.

Table 13: The difference between self-evaluation and teachers' evaluations in General Description

Difference	Student-JET1	Student-JET2	Student-NET
4	0 (0%)	0(0%)	0(0%)
3	1(1%)	1(1%)	2(2%)
2	10(11%)	12(13%)	20(22%)
1	25(27%)	25(27%)	28(30%)
0	34 (37%)	35(38%)	37(40%)
-1	20(22%)	16(17%)	3(3%)
-2	2(2%)	2(2%)	2(2%)
-3	0(0%)	1(0.01%)	0(0%)
-4	0(0%)	0(0%)	0(0%)

Table 14: The difference between self-evaluation and teachers' evaluations in Delivery

Difference	Student-JET1	Student-JET2	Student-NET
4	0(0%)	0(0%)	0(0%)
3	1(1%)	2(2%)	1(1%)
2	7(8%)	14(15%)	18(20%)
1	32(35%)	23(25%)	32(35%)
0	36(39%)	27(29%)	33(36%)
-1	14(15%)	20(22%)	7(8%)
-2	2(2%)	6(7%)	1(1%)
-3	0(0%)	0(0%)	0(0%)
-4	0(0%)	0(0%)	0(0%)

Table 15: The difference between self-evaluation and teachers' evaluations in Language Use

Difference	Student-JET1	Student-JET2	Student-NET
4	0(0%)	0(0%)	0(0%)
3	0(0%)	1(0.01%)	0(0%)
2	10(11%)	14(15%)	14(15%)
1	35(38%)	25(27%)	26(28%)
0	36(39%)	33(36%)	37(40%)
-1	11(12%)	12(13%)	12(13%)
-2	0(0%)	6(7%)	3(3%)
-3	0(0%)	1(1%)	0(0%)
-4	0(0%)	0(0%)	0(0%)

Table 16: The difference between self-evaluation and teachers' evaluations in Topic Development

Difference	Student-JET1	Student-JET2	Student-NET
4	0(0%)	0(0%)	0(0%)
3	0(0%)	0(0%)	2(2%)
2	10(11%)	9(10%)	9(10%)
1	33(36%)	26(28%)	28(30%)
0	43(47%)	38(41%)	38(41%)

-1	5(5%)	18(20%)	14(15%)
-2	0(0%)	1(1%)	1(1%)
-3	1(1%)	0(0%)	0(0%)
-4	0(0%)	0(0%)	0(0%)

"0" indicates the perfect agreement between students' self-evaluation and teachers' evaluation. From 29% to 47 % students evaluate themselves as the same as teachers evaluate students in every category. There is also a tendency that students evaluate themselves slightly higher than teachers' evaluations. The ratio of the differences, "1" and "0", accounts for about 60-70% of the whole. Especially Topic Development category shows the higher ratio of perfect agreement, compared to other three categories.

2.4 The correlation between teachers

Tables 17, 18, 19, and 20 show the correlation between teachers in four evaluation categories.

Table 17: The Pearson's r between teachers in General Description

JET1& ET2	JET1& NET	JET2& NET
-0.06	0.04	0.63

Table 18: The Pearson's r between teachers in Delivery

JET1&JET2	JET1& NET	JET2& NET
0.19	-0.06	0.28

Table 19: The Pearson's r between teachers in Language Use

JET1&JET2	JET1& NET	JET2& NET
-0.01	0.003	0.44

Table 20: The Pearson's r between teachers in Topic Development

JET1&JET2	JET1& NET	JET2& NET
0.07	0.04	0.73

The correlations between JET2 and NET in every category are comparably higher than the other two combinations, JET1 and JET2, JET1 and NET. Especially the Pearson's r of General Description and Topic Development are closer to the agreement. On the other hand, the combinations between JET1 & JET2 and JET1 & NET do not present any correlations. Yet the Pearson's r of Topic Development indicates positive values in all combinations though the values are much lower than JET2 & NET combinations.

2.5 The result of questionnaire of students

Tables 21 and 22 present the ideas of students about peer evaluation and self-evaluation. About 85-89% students answered that peer and self-evaluations are very effective.

Table 21: The result of questionnaires about peer evaluation of students

Very effective	Effective	Little effective	Not effective	Uncertain
24	58	1	1	8 persons

Table 22: The result of questionnaires about self-evaluation of students

Very effective	Effective	Little effective	Not effective	Uncertain
32	46	2	2	10persons

Table 23 shows the replies of students' free comments on peer and self-evaluations. 75 students wrote their ideas on both types of evaluations. The following comments are summarized from a broader view of the students' ideas.

Table 23: The result of questionnaires about self & peer evaluation of students

Comments	Peer evaluation	Self-evaluation
The importance of metacognition	0	23
Effectiveness to improve English	7	1
Motivation to study English	3	1
Difficulty to evaluate	1	8
Opinions about the items and contents of evaluation sheet	4	3

Students recognized the importance of metacognition to improve their English ability. Many students stated that it was a good chance to be conscious of themselves and their ability. They also reflected on themselves by comparing their English with that of their classmates. Some students felt ashamed of their inferiority and also highly motivated to develop themselves next time. On the other hand, they confessed the difficulty to evaluate peers because of their insufficient ability of language use. Some students mentioned that it was difficult to evaluate others since they did not have confidence in their English. Some of them pointed out that their English ability was not enough to evaluate peers. They also claimed that the number of categories of evaluation was beyond their judgment. A few students commented that the description of the evaluation sheet looked complicated for them. Ten students also mentioned the presence of peer pressure. Peer pressure might have caused tension or stress.

2.6 The result of questionnaire of teachers

Table 24 presents the opinions about peer and self-evaluation. JET2 and NET have adopted peer evaluation in class, using written feedback or numerical evaluation, but both of them are biased

toward the accuracy of students' evaluation. Therefore, peer evaluation will not influence their final decision, i.e., school record. NET mentioned that he would have considered taking peer evaluation more seriously, if the students were unknown to each other. He thought that his students were biased against the proficiency of peers. He also recognized the effectiveness of peer evaluation because they would listen more carefully and be more serious if they believed that they were grading other students. JET1 applied students' self-evaluation in his final decision to evaluate a student's English oral production using written feedback, but he did not say that he would adopt peer evaluation in class.

Table 24: The result of teachers' questionnaires on self and peer evaluation

	JET1	JET2	NET
Peer evaluation	Effective to develop English	Very Effective to develop English	Very effective to develop English
Self-evaluation	Effective to develop English	Very Effective to develop English	Uncertain

3 Discussion

3.1 Comparison of self, peer and teacher evaluations

1) To what extent does peer evaluation of English speech correspond to teacher evaluation?

Only a few students present high correlation (≥ 0.7) in all of the evaluation categories.

The correlation between Pearson's r (teachers and peer evaluations) and T-score of MEG test also does not yield any significant values. The correlation between the group that had a T-score of more than 60 points and JET1 indicates higher values in Topic Development. Topic Development category judges whether a speaker presents a clear progression of ideas and conveys the relevant information required by the task. The correlation between Pearson's r (JET1 and peer evaluation) and T-score ≥ 60 points presents much higher correlation than the other teachers' correlations in Topic Development. So there is a possibility that the students who had more than 60 T-scores may share the same standard of evaluation with JET1 who is a regular teacher for the students in Topic Development. On the other hand, the other two teachers, JET2 and NET who are guest teachers may not have the same standard in the evaluation of Topic Development.

2) To what extent does self-evaluation of English

speech correspond to teacher evaluation?

Self-evaluation and teacher evaluation indicate higher agreement in all of the evaluation categories. Above all, the difference in Topic Development indicates the least of all categories. The perfect agreement ratio of JET1 is more prominent than those of the other two teachers. The difference between self-evaluation and JET1 evaluation in Topic Development indicates 47%, and both of the ratios of JET2 and NET are 41% in the same category. Whereas students evaluate themselves slightly higher than the teachers do. Self-evaluation for senior high school students could be stated to be more reliable. Moreover the evaluation of a regular English teacher tends to agree with self-evaluation by the students in Topic Development.

3) Does native English teacher's evaluation correspond to Japanese teachers' evaluation?

JET2 and NET constantly present higher correlations in all of the evaluation categories than other two combinations JET1&JET2 and JET1 & NET. Especially the values of General Description and Topic Development are closer to the perfect agreement. The correlation of Language Use category indicates 0.44. On the other hand, the correlation of Delivery category is 0.28.

It was assumed that the evaluations of Japanese teachers would have higher consistency, but the combination of a native English teacher and the other Japanese English teacher (JET2) presents higher correlations in all of the categories. So being native or non-native as evaluators does not have a significant factor in this research.

JET1 does not show any correlations with the other two teachers in all of the categories. The reason why JET1 presents different values from the other two teachers may be related to the understanding of evaluation categories or influenced by the regular English class evaluation. A regular teacher better knows English competence of his students. JET1 may use his regular class standard when he evaluates students' oral speeches. JET1 commented that it was difficult to evaluate many evaluation items of speech production at the same time or in the limited time. The other two teachers also stated similar opinions: 1) JET2 said that she did not have confidence in the evaluation of oral production because there are many elements. Moreover the evaluation time was limited. 2) NET also mentioned that it was rather difficult to evaluate many areas at the same time unless the speech was recorded. All of the three teachers tell the difficulty in judging speeches at the same time and in the limited time.

4 Conclusion

The correlation in Topic Development is higher than the other three categories in all of the correlations: 1) the correlation between teacher evaluation and student evaluation 2) the correlation between Pearson's r and T-scores of MEG test 3) the correlation between self-evaluation and teachers' evaluation. 4) the correlation between teachers, though this category is related more to the individual evaluator's interpretation of the content.

About 40% of the students evaluated themselves as teachers evaluate them. So it is suggested that self-evaluation should be more implemented in the senior high school classroom because it is useful for students to develop their self-consciousness and to motivate themselves. On the other hand, peer evaluation is not so reliable even if students have higher grammatical competence. So it is necessary to study the relationship between age and peer evaluation.

Both teachers and students recognized the effect of peer pressure as the tension or the stimulation to motivate students. This phenomenon is also expected to be more explored. Teacher evaluation is not related to being native or not being native teachers. Other factors seem to affect the consistency of evaluations. The result must not be generalized because of the limited number of participants. Future studies are necessary to confirm the findings.

Acknowledgements I am grateful to a prefectural school in Kanagawa for their cooperation with the research.

References

- Blanche, P., & Merino, B. J. (2001). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, 39, 313-340.
- Boud, D. (1999). Avoiding the traps: Seeking good practice in the use of self-assessment and reflection in professional courses. *Social Work Education*, 18(2), 121-133.
- Brown, J. d., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32, 653-675.
- Cheng, W. et al. (2005). Peer assessment of language proficiency. *Language Testing* 22, 93-94.
- Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press.
- Davies, A. (1991). The native speaker in applied linguistics. Edinburgh University Press.
- Davies, A. (2004). The native speaker in applied linguistics. In Alan Davies and C. Elder (Eds.).

- The handbook applied linguistics, 431-450. Malden, MA: Blackwell.
- Educational Testing Service. (2004). TOEFL Test Integrated Speaking Rubrics. Retrieved date (051511) from <http://www.ets.org/>
- Nakano, M. et al. (2003). Measure of English Grammar (MEG). *SEISAKU KAGAKU* 10(3), 59.
- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing* 19, 109.
- Shimizu, Y., Yamakawa, K., Sugino, No., Ohba, H., and Nakano, M. (2006). Developing a sample-free grammatical proficiency test for SLA research. *Proceedings of the 10th Conference of Pan-Pacific Association of Applied Linguistics*, 227-236.
- The Society for Testing English Proficiency. (2006). Eiken-Can-do List. The Society for Testing English Proficiency.
- Zhang, Y. & Elder, C. (2011). Judgments of oral proficiency by non-native English speaking teacher raters: Competing or complementary constructs?. *Language Testing*, 28(1), 31-34.

Appendix A Speaking Evaluation for students

Score 点数	General Description 全体評価	Delivery 発音・イントネーション・速さ	Language Use 文法・語句などの運用	Topic Development 話題の発展性
5	内容がわかりやすく、話しに筋（すじ）がとおっている。少しのまちがいはあるが印象的だ。	全体的に、はっきりと流れるように話している。また、落ち着いて話すのでわかりやすい。発音・イントネーションのまちがいは少しある。	文法をうまく使って理由や意見を表現している。単語（たんご）の選び方もよい。少しのまちがいはある。	自分の考えをはっきりと話している。くわしく表現している。
4	まちがいは少し目立つが、だいたいはわかりやすい。話しに筋（すじ）も通っているが、完ぺきではない。	発音・イントネーションのまちがいが少しあり、聞きとるのが難しい時がある。しかし、全体的にははっきりとしたスピーチである。	まちがいがあるが、だいたいは、文法や単語をうまく使っている。	まちがいは少しあるが、だいたいは、自分の考えを話している。
3	テーマに関連しないところや正確でない部分も少しあるので、ときどきわかりにくい。あいまいな部分もあるが、理解はできる。	ある程度は、はっきりと話しているが、そうでないときもある。発音・イントネーションのまちがいがある。聞きとるのは難しい（むずかしい）。	文法や単語をあまりうまく使っていないため、話しの内容があいまいになっている。	テーマにそった話しが少ししかない。大切な点を話していない。まちがいがある。
2	話しの筋（すじ）が最小限しかとおっていない。全体をとおしてわかりにくい。	発音・イントネーションに問題があるため、話しの内容があいまいになっている。とぎれとぎれで、ためらいがちである。	文法や単語をうまく使っていないため、考えがうまく表現できていない。	テーマにそった内容がほとんどない。あいまいで、まちがいがある。
1	話をしなかった。			

Appendix B Speaking Evaluation for Teachers

Score	General Description	Delivery	Language Use	Topic Development
5	It fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse.	It is generally clear, fluid, and sustained. It may include minor lapses or difficulties with pronunciation or intonation. Pace may vary at times as the speaker attempts to recall information. Overall intelligibility remains high.	It demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient expression of relevant ideas. Contains generally effective word choice. Though some minor errors or imprecise use may be noticeable, they do not require listener effort.	It presents a clear progression of ideas and conveys the relevant information required by the task. It includes appropriate details, though it may have minor errors or minor omissions.
4	It addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas.	It is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation, or pacing and may require some listener effort at times. Overall intelligibility remains good.	It demonstrates fairly effective use of grammar and vocabulary, and coherent expression of relevant ideas, though it may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the comprehension of the message.	It is sustained and conveys relevant information required by the task, though it exhibits some incompleteness, inaccuracy, lack of specificity with respect to content, or choppiness in the progression of ideas.
3	It is connected to the task, though some relevant information may be missing or contain inaccuracies. It contains some intelligible speech, but at times problems with intelligibility and/or overall coherence may have obscure meaning.	It is clear at times, though it exhibits problems with pronunciation, intonation, or pacing and so may require significant listener effort. It may not be sustained at a consistent level throughout. Problems with intelligibility may obscure meaning in places (but not throughout).	The range of vocabulary and grammar is limited. Some complex structures may be used, but typically contain errors. It results in limited or vague expression of relevant ideas and imprecise or inaccurate connections.	It conveys some relevant information but is clearly incomplete or inaccurate. It omits key ideas, or demonstrates limited development of important information, or misunderstanding of key ideas.
2	It is very limited in content or coherence or is only minimally connected to the task. It may be unintelligible.	Consistent pronunciation and intonation problems cause considerable listener effort and frequently obscure meaning. Delivery is choppy, fragmented, or telegraphic. It contains frequent pauses and hesitations.	Range and control of grammar and vocabulary severely prevent expression of ideas and connections among ideas.	It fails to provide much relevant content. Ideas that are expressed are often inaccurate, limited to vague utterances, or repetitions.
1	Speaker makes no attempts to respond OR response is unrelated to the topic.			