

A Pilot Study of Self-Evaluation and Peer Evaluation

Yoko Suganuma Oi

Graduate School of Education, Waseda University

yokosuganuma@suou.waseda.jp

Abstract

This study aimed to investigate the relationship between student evaluation and teacher evaluation of oral speech by addressing three research questions: 1) How different self-evaluation is from teacher evaluation when administered at three separate times within one month? 2) How does peer evaluation correlate with teacher evaluation when administered at three separate times within one month? 3) How did the students react self-evaluation and peer evaluation? The data of the present research for the analysis were the speeches of 26 Japanese senior high school students. The researcher decided to divide one group for self-evaluation group and the other group for peer evaluation group. Each group was composed of 13. One group was asked to evaluate their own speeches in relation to other students just after the speech. The other group was asked to evaluate their peers' speeches. At the same time one American English teacher and one Japanese teacher also evaluated students' speeches. The agreement between self-evaluation and teacher evaluation was low but there might not be difference in Language Use. On the other hand, peer evaluation did not correlate with teacher evaluation and the development of correlation during three sessions was not found. Though the reliability of self-evaluation was different among students, students positively recognized self-evaluation as the tool to improve English proficiency. Peer evaluation was not reliable even for high school students. In addition, students felt pressured because of their relationship as friends and the lack of confidence in terms of English proficiency. However, they also felt the necessity of competitiveness with peers. Therefore, it is important to implement both types of evaluations in class in good balance.

Keywords

Self-evaluation, peer evaluation, teacher evaluation, speech evaluation

1 Introduction

It is assumed that English proficiency would be improved using self-evaluation and peer evaluation (Oi, 2012) because the ability to judge themselves and peers helps students find their problems and solutions by themselves. The activities such as self-evaluation and peer evaluation encourage learners to be metalinguistically cognitive in terms of learner autonomy and learner responsibility. It is fundamental for learners to develop questioning attitudes, and to learn how to become independent and more self-aware learners. Learners can make effective decisions for themselves, can articulate their own language learning needs, and can work effectively with a "facilitating" teacher through self-evaluation and peer evaluation (Skehan, 1998). Breen (1987) showed five stages of self-evaluation, whose assumption is that learners are able to participate in meaningful discussion about goal setting, role allocation, the planning of learning, the activities to be used, and the forms of evaluation. Thus self-evaluation and peer evaluation help language education class to become "learner-centered."

Oi (2010) researched the correlation between peer evaluation of Japanese university students and teacher evaluation. Two groups consisting of three to six university students who scored 530 or more in TOEFL (PBT) or returnees from the US, Canada and Ireland, and two American English teachers participated in the survey. They evaluated oral discussion in English based on two main competences: language competence and communicative performance ability. The result showed that the correlation between teacher evaluation and peer evaluation was stronger over the course of five sessions.

On the other hand, the research (Oi, 2012) which focused on Japanese high school students showed a different result. It investigated the reliability of self-evaluation and peer evaluation of Japanese high school students. Ninety-two high school students, one American English teacher, and two Japanese English teachers participated in the survey. Students evaluated their own and peer's speeches on familiar topics. Teachers also evaluated students' speeches using the same evaluation sheet which was composed of General Evaluation,

Delivery, Language Use, and Topic Development. The consistency between student self-evaluation and teacher evaluation was high in Language Use. On the other hand, the correlation between teacher evaluation and peer evaluation was very low even with the students who had high proficiency in English. However, the research was conducted only once, so it should be discussed if the correlation between peer evaluation and teacher evaluation is affected by the age factor and the rating experience. This is the reason why the present research was carried out to confirm the relationship between student evaluation and teacher evaluation through plural researches. That is to say, more sustained observation is necessary to confirm the effectiveness of the evaluation by senior high school students in terms of oral speeches. Therefore the goal of the present study was to prove the reliability and usefulness of self-evaluation and peer evaluation of oral speeches in English.

2 Literature Review

Students' self-evaluation and peer evaluation has been implemented in English class, because teachers believe that reflection on one's proficiency and insight into evaluation criteria will stimulate students' motivation (Oi, 2012). Since the evaluation criteria of teachers' is the core in class, the reliability of students' evaluation should be investigated referring to teacher evaluation. So this section shows the previous studies on the relationship between student evaluation and teacher evaluation and the characteristics of student evaluation.

Cheng and Warren (2005) indicated that students and teachers were different in their respective evaluating behaviors and the ways oral and written language proficiency were assessed. Cheng (2008) commented that students' personal traits or psychological characteristics such as confidence, and nervousness might affect their evaluation behavior. The students in Cheng's study tended to evaluate themselves lower and it was caused by their affective factors. With regard to peer evaluation, previous studies by Rolfe (1990), and Hughes and Large (1993) showed high agreement between teacher evaluation and peer evaluation.

Cheng (2008) also showed three implications and suggestions on self-evaluation. The first one was the development of self-reflection on their performance and learning process. The second suggestion was the help to train students to become better raters and learners. The final implication was students' psychological factors.

Patri (2002) investigated the effectiveness of the self-evaluation and peer evaluation of oral presentation skills of first year undergraduate students of ethnic Chinese background, and found that peer feedback enabled students to judge the performance of their peers in a manner similar to that of the teachers, when evaluation criteria were firmly set and students understood them well before evaluation. Patri concluded that peer feedback was supposed to help in achieving a higher correlation between teacher and peer assessment. In addition, teacher evaluation could be supplemented with peer evaluation at a lower cost especially in oral skills.

These previous studies focused only on undergraduate students and adult learners, not high school students. Therefore, it is meaningful to investigate the relationship between the evaluation of high school students and teacher evaluation to prove the hypothesis that student evaluation might be reliable as well as teacher evaluation.

3 Research Questions

In order to investigate the relationship between student evaluation and teacher evaluation of oral speeches, the following three research questions are addressed: 1) How different self-evaluation is from teacher evaluation when measured at three separate times within one month? 2) How peer evaluation correlates with teacher evaluation when measured at three separate times within one month? 3) How did the students react self-evaluation and peer evaluation? By investigating the above questions, it is assumed that respective characteristics and the relationship between self-evaluation and peer evaluation are found. The knowledge of student evaluation helps high school teachers to implement it more effectively in class.

4 Method

4.1 Participants

Participants were 26 Japanese senior high school students. They were comprised of 24 females and 2 males. The researcher decided to divide them into two groups: one group for the self-evaluation group, Group A, and the other group for the peer evaluation group, Group B. The students took Grade Pre-2 of the EIKEN Test in Practical English Proficiency and the students were divided into two groups so that the average score was the same, 30.4 out of 60 points. Standard deviation (SD) of Group A was 10.31, on the other hand, SD of

Group B was 6.23. This test was composed of listening, grammatical and reading sections. It was given to examine the English proficiency of individual students. The English proficiency of students was A1-A2 level of Common European Frameworks of Reference for Languages (CEFR). Each group was made up of 13 students. Group A for self-evaluation was composed of 13 female students. On the other hand, Group B for peer evaluation was composed of 2 male and 11 female students. There was no student who had lived in any English speaking countries. One American English teacher and one Japanese teacher participated in the research. Both of them have taught English in public senior high schools more than ten years in Japan. The researches were conducted three times on May 11th, June 1st, and June 8th.

Table 1: The SD and Mean of the Results of Pre-2 of English Proficiency Test

| | <i>n</i> | <i>MAX</i> | <i>MIN</i> | <i>M</i> | <i>S.D.</i> |
|---------|----------|------------|------------|----------|-------------|
| Group A | 13 | 48 | 18 | 30.4 | 10.3 |
| Group B | 13 | 51 | 22 | 30.4 | 6.2 |

Note: The maximum score is 60 points.

4.2 Data collection

4.2.1 Evaluation criteria and background questionnaire

Two kinds of background questionnaires were given to the students to ask them about their English educational backgrounds and the perceptions about self-evaluation and peer evaluation. The first one was distributed before the research to ask their backgrounds of English education (Appendix A). The second one was conducted to ask the perceptions about self-evaluation and peer evaluation at the end of one-month course. It was constructed using a 5 point Likert scale (1=very effective, 2=effective, 3=not effective, 4=not effective at all, 5=no idea) and open-ended question: what do you think of self/peer evaluation (Appendix B)?

The evaluation sheet had five measures. It consisted of three components, Content, Delivery, and Language Use. It was devised by the researcher based on CEFR, TOEFL®, and STEP. Group A was asked to evaluate their own speeches in relation to other students just after the speech. Group B was asked to evaluate themselves and their peers. At the same time one American English teacher and one Japanese teacher also evaluated students' speeches. All of the participants learned how to evaluate speeches before the speech. Two teachers held a norm session to teach students how to evaluate speeches (Appendix C).

The topics of speeches were announced just before the speeches. Students made one-minute speeches for three times within one month on familiar topics: town, holiday, and future.

The types of measurement were analytical rating from one to five. In this case "5" indicated the highest evaluation, and "1" indicated the lowest evaluation. All of the participants rated the English oral speeches. The speeches were recorded and transcribed.

4.2.2 Methods of analysis

The following measurements to investigate the relationship between self-evaluation and teacher evaluation, and between peer evaluation and teacher evaluation were used in this research. Firstly, Wilcoxon signed-rank test was conducted to look for differences between self-evaluation and teacher evaluation. Kendall's tau was used to see the inter-rater reliability between the evaluations of two teachers. It was also used to examine the relationship between teacher evaluation and peer evaluation.

5 Results and findings

5.1 The inter-rater reliability between the evaluations of two teachers

There was a correlation between two teachers' evaluations in both groups. The correlation coefficients of the self-evaluation group presented, $\tau = .648$ for May 11th, $\tau = .804$ for June 1st, $\tau = .861$ for June 8th, $p < .05$. On the other hand, the correlation coefficients of the peer evaluation group presented, $\tau = .904$ for May 11th, $\tau = .682$ for June 1st, $\tau = .798$ for June 8th, $p < .05$.

The correlation coefficient between two teachers' evaluations was gradually getting stable. So it was assumed that teachers found the common and stable standard of evaluation for three times.

5.2 The Wilcoxon signed-rank test based on self-evaluation and teacher evaluation

Before conducting the Wilcoxon signed-rank test, the scores of student evaluation was subtracted from the scores of teacher evaluation. The sum of three evaluation components, Content, Delivery, and Language Use was 40 points. In terms of the sum, Content and Delivery, students tended to evaluate themselves lower than

teachers did.

However, the difference of Language Use is very small. The difference “0” indicates the perfect agreement between self-evaluation and teacher evaluation. In this research most of the students tended to evaluate their own speeches much lower than teachers did except Language Use.

Based on the difference between teacher evaluation and self-evaluation, the Wilcoxon signed-rank test was conducted to see differences between two data. Table 2 shows the results of three sessions between the native English teacher and Group A (self-evaluation group).

If the absolute value of the z score was higher than 1.96, we could say that there was a significant difference between two data. Concerning the May 11th data, the sum of 3 evaluation components showed -2.24. It meant that student self-evaluation was different from teacher evaluation in this case. On the other hand, the absolute values of the Content and Language Use z scores were less than 1.96. There might not be a difference between self-evaluation and teacher evaluation in the evaluation components of Content and Language Use.

Next is the result of the June 1st data. On this day Delivery and Language Use showed 1.93 and 1.95, which were less than 1.96. So self-evaluation might not be different from teacher evaluation in Delivery and Language Use.

The result of June 8th presented that only Language Use component showed less than 1.96. So self-evaluation of Language Use evaluation might not be different from teacher evaluation.

In every session, self-evaluation of Language Use might not be different from teacher evaluation.

Table 2: The Wilcoxon signed-rank test between native English teacher and Group A (self-evaluation group)

| | Sum of three components | | Content | | Delivery | | Language Use | |
|---------------------------|-------------------------|-------|---------|-------|----------|-------|--------------|-------|
| | T | z | T | z | T | z | T | z |
| May 11th NET & Group A | 13.50 | -2.24 | 1.50 | -1.55 | 7.50 | 4.56 | 1.50 | -0.52 |
| June 1st NET & Group A | 1.00 | -2.98 | 0.00 | -3.05 | 1.00 | -1.93 | 4.50 | 1.95 |
| June 8th NET & Group A | 9.50 | -2.51 | 6.50 | 3.56 | 12.00 | 8.94 | 1.50 | -0.70 |

Note: $p < .05$ T (the sum of the ranks of the smaller of the two ranks); z (Wilcoxon signed-rank test z -score)

Table 3 presents the results of three sessions between the Japanese English teacher and Group A (self-evaluation group). It also shows that student self-evaluation was different from teacher evaluation. However, self-evaluation might not be different from teacher evaluation in Language Use.

Table 3: The Wilcoxon signed-rank test between Japanese English teacher and Group A (self-evaluation group)

| | Sum of three components | | Content | | Delivery | | Language Use | |
|---------------------------|-------------------------|-------|---------|-------|----------|-------|--------------|-------|
| | T | z | T | z | T | z | T | z |
| May 11th JET & Group A | 4.00 | -2.90 | 22.50 | -0.54 | 12.50 | -2.09 | 0.00 | -3.17 |
| June 1st JET & Group A | 0.00 | -3.06 | 17.50 | -0.71 | 6.00 | -2.41 | 12.0 | -1.64 |
| June 8th JET & Group A | 8.50 | -2.58 | 6.00 | -2.61 | 7.50 | -2.29 | 0.00 | 0.00 |

Note: $p < .05$ T (the sum of the ranks of the smaller of the two ranks); z (Wilcoxon signed-rank test z -score)

In this research, the agreement between self-evaluation and teacher evaluation was not found. Besides, students evaluated their speeches lower than teachers did. However, concerning Language Use, there might not be a difference between self-evaluation and teacher evaluation. The agreement between high school student's self-evaluation and teacher evaluation was not always consistent.

5.3 The correlation coefficient of Kendall tau based on peer evaluation of each individual student and teacher evaluation

To analyze the correlation between peer evaluation and teacher evaluation more precisely, the correlation

coefficient of Kendall tau for each individual student was calculated (Appendix D & E). The Delivery component showed a slightly weak correlation between student evaluation of each individual student and teacher evaluation. However, in every session, there was almost no correlation between peer evaluation and teacher evaluation. This results show the same tendency as in Oi (2012).

5.4 The responses of the questionnaire

Appendixes F and G present the responses of the self-evaluation Group A and the peer evaluation Group B about the effectiveness of self-evaluation and peer evaluation. Twelve of thirteen students in the self-evaluation group replied that a self-evaluation activity was very effective or effective to improve English proficiency (Appendix F).

On the other hand, the responses of Group B presented that only five of thirteen students in the peer evaluation group answered that a peer evaluation activity was effective to improve English proficiency. One student insisted on the importance of concrete feedback rather than analytical evaluation to develop the level of speech. Only two students recognized the “usefulness” to improve the level of speech by using peer evaluation. One student commented peer evaluation produced competitiveness, and it was important to stimulate their motivation (Appendix G).

6 Discussion

In this research, the agreement between self-evaluation and teacher evaluation was not found. Moreover, students evaluated their speeches lower than the teachers did. Those are differences from Oi (2012). The general academic level of the students in the present study was lower than that of the students in the previous research. One possible reason is that the lack of confidence caused the students in the present study to evaluate themselves low. It was assumed that the lower self-evaluation was caused by misunderstanding their level of performance because of the general academic level rather than evaluating their speeches. The students of the previous study had confidence in their whole school life because of their academic level. So one of the reasons was supposed that the lower self-evaluation of the present study was affected by the lack of self-confidence.

The agreement between high school student’s self-evaluation and teacher evaluation was not always consistent. However, concerning Language Use, there might not be a difference between self-evaluation and teacher evaluation. This is the same results as those of Oi (2012). So it supports the hypothesis that high school students might evaluate their language use in the same manner as teachers. One of the reasons is assumed that the consistency between teacher evaluation and self-evaluation was caused by the less number of subcategories of Language Use compared to the other categories. It is also assumed that students have stronger confidence of their knowledge on language use because they have spent more time studying language use than the other components such as delivery and content.

Peer evaluation was not correlated with teacher evaluation. However, the Delivery component presented weak correlation in several students’ data. The development of correlation during three sessions was not found, being different from the data of university students (Oi, 2010). Rolfe’s (1990) previous study presented the high agreement between teacher evaluation and peer evaluation. Yet the peer evaluation of high school students did not correlate with teacher evaluation at all. Therefore, the peer evaluation of high school students might not be consistent. The correlation between the peer evaluation and teacher evaluation should be investigated in terms of age as further study.

According to the results of questionnaires, the self-evaluating group recognized the effectiveness of self-evaluation to improve English proficiency because it helped students to reflect on themselves. This is the same results as those of Oi (2012). Oskarsson (1989) presented six advantages of using self-evaluation: 1) promotion of learning 2) raising level of awareness 3) improving goal-orientation 4) expansion of range assessment 5) sharing assessment burden 6) beneficial post-course effects. Therefore, the result of the present study supports the advantages Oskarsson pointed out.

On the other hand, the peer evaluation group felt more pressured towards peer evaluation because of the lack of confidence in English proficiency. According to the responses of Background Questionnaire, they also felt some doubt on whether they were qualified to evaluate others. Some of them mentioned that they did not have confidence in their ability to rate because of their English proficiency. They commented that concrete descriptive feedback was more useful to improve English proficiency rather than analytical rating evaluation.

The implementation of self-evaluation and peer evaluation in class has different advantages. Though the reliability of self-evaluation was different among students, students positively recognized self-evaluation as

the tool to improve English proficiency. Peer evaluation was not reliable for high school students. In addition, students felt pressured because of their relationship as friends. Students also commented that they did not have confidence in their ability to evaluate others in terms of English proficiency. However, some students commented that they felt the necessity of competitiveness with peers. Therefore, it is important to implement both types of evaluations in class, recognizing the advantages and disadvantages of two evaluations.

7 Conclusion

The implementation of self-evaluation and peer evaluation in class has different possibilities. Students positively recognized self-evaluation as the tool to improve English proficiency and bridge the gap between their present English proficiency and their goals, though the reliability was different among students. Besides students might judge their language use in the same manner as teachers. Therefore self-evaluation of language use might be positively implemented in an English class. On the other hand, it should be investigated whether the reliability of self-evaluation about Delivery and Content would be developed after the training of evaluation.

Statistically significant correlation was not found between teacher and peer evaluation, though there was a slight correlation between teacher evaluation and peer evaluation in the Delivery component. It was assumed that students shared similar evaluation criteria with teachers in the Delivery component. Some students commented that they were not confident in rating other students' speeches. They also mentioned that they felt peer pressure. On the other hand, some of them felt a necessity to be competitive with peers to motivate themselves.

Therefore adopting both types of evaluations in class has educational implications because it encourages students' participation and initiative in class. Adopting both types of evaluations may enhance learner autonomy because it is related to self-reflection and motivation.

As further study, it is needed to explore descriptive evaluation because some students mentioned the usefulness of descriptive evaluation about their performance. The present study did not show the correlation between high school students' peer evaluation and teacher evaluation though the previous study about university students (Oi, 2010) showed the consistency. Therefore, age factors should be investigated to explore the possibility of student evaluation in class.

References

- Breen, M. (1987). Contemporary paradigms in syllabus design: (Parts 1 and 2). *Language Teaching*, 20, 157-74.
- Cheng, W. & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing*, 22, 93-110.
- Cheng, Y. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research*, 12(2), 254-255.
- Hughes, I. E. & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education*, 18, 379-85.
- Oi, S. Y. (2010). A Pilot Study of Teacher Evaluation and Peer Evaluation in English Oral Discussion. *Proceedings of the 15th International Conference of Pan-Pacific Association of Applied Linguistics*, 135-142.
- Oi, S. Y. (2012). A Study of Student Evaluation and Teacher Evaluation (Master thesis). Waseda University.
- Oskarsson, M. (1989). Self-assessment of language proficiency: Rational and applications. *Language Testing*, 6 (1), 1-13.
- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, 19, 109.
- Rolfe, T. (1990). Self and peer-assessment in the ESL curriculum. In Brindley, G., editor, Vol.6: The second language curriculum in action (pp.163-86). Sydney: NCELTR, Macquarie University.
- Skehan, P. (1998). *A Cognitive Approach to Language Learning* (p.266). Oxford University Press.

Appendices

Appendix A

Background Questionnaire 1

1. If you have ever lived in a foreign country, how many years did you live? Please circle one item.

More than 10 years 9 years 8 years 7 years

6 years 5 years 4 years 3 years
2 years 1 year I have never lived abroad.

2. If you have an English Proficiency Test Grade, what grade do you have? Please circle one item. If you have other English Proficiency Test Grades (eg. Kokuren Eiken, TOEIC etc.), please write the name and grade/scores.

1st grade Pre-1st grade 2nd grade Pre-2nd grade
3rd grade 4th grade 5th grade

3. What do you think about evaluating other participants' performance in English interaction? Please circle one item from 1) to 4).

- 1) Very effective to develop English
- 2) Effective to develop English
- 3) Not effective
- 4) Uncertain

Appendix B
Background Questionnaire 2

1. What do you think about evaluating English performance of other participants? Please circle one item from 1) to 5). If you have any opinions or ideas, please write it below.

- 1) Very effective to develop English
- 2) Effective to develop English
- 3) Not so effective
- 4) Not effective at all
- 5) Uncertain

Opinions / ideas

2. What do you think about evaluating your own English performance by yourself? Please circle one item from 1) to 5). If you have any opinions or ideas, please write it below.

- 1) Very effective to develop English
- 2) Effective to develop English
- 3) Not so effective
- 4) Not effective at all
- 5) Uncertain

Opinions / ideas

Appendix C

EVALUATION SHEET

年 組 番 氏名

自分と友達のスピーチを評価しよう！

スピーチをした人の名前→

| JUDGING TIMES | POINT VALUE | | | | | |
|--|--------------|-----------|-----------|------|------|----------|
| SPEECH DEVELOPMENT structure, organization, support material ● 構成がしっかりしているか ● 理由が述べられているか | CONTENT | EXCELLENT | VERY GOOD | GOOD | FAIR | NOT FAIR |
| EFFECTIVENESS achievement of purpose, interest, reception ● スピーチのテーマに話があっているか | | すばらしい | とても良い | 良い | ふつう | あまりよくない |
| SPEECH VALUE ideas, logic, original thought ● 聞いていて感動があったか ● オリジナリティがあったか | | EXCELLENT | VERY GOOD | GOOD | FAIR | NOT FAIR |
| PHYSICAL appearance, body language ● 身振りや手振り、アイ・コンタクトなどをして効果的にスピーチを盛り上げたか | DELIVERY | すばらしい | とても良い | 良い | ふつう | あまりよくない |
| PRONUNCIATION AND INTONATION ● 発音やイントネーションは良かったか | | EXCELLENT | VERY GOOD | GOOD | FAIR | NOT FAIR |
| VOICE flexibility, volume ● 速さはちょうど良かったか ● 声は十分に大きかったか | | すばらしい | とても良い | 良い | ふつう | あまりよくない |
| MANNER directness, assurance, enthusiasm ● 積極的な姿勢や熱心に取り組む態度が見られたか | | EXCELLENT | VERY GOOD | GOOD | FAIR | NOT FAIR |
| CORRECTNESS of LANGUAGE USE grammar, word selection ● 文法や単語の選択は正しかったか | LANGUAGE USE | すばらしい | とても良い | 良い | ふつう | あまりよくない |

Appendix D

The correlation coefficient between peer evaluation and teacher (NET) evaluation

| | | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B11 | B12 | B13 |
|----------|--------------------------------|-------|--------|-------|-------|--------|-------|-------|--------|--------|------|--------|------|--------|
| May 11th | Sum of 3 evaluation components | .312 | .608** | .425* | .395 | .355 | .349 | .167 | .298 | .193 | .06 | .253 | .296 | .347 |
| | CONTENT | -.24 | .350 | .125 | .019 | .262 | -.06 | -.13 | -.29 | -.06 | -.1 | .432 | -.14 | -.35 |
| | DELIVERY | .551* | .388 | .629* | .092 | .324 | .556* | .491* | .313 | .258 | .52* | .297 | .422 | .394 |
| | LANGUAGE USE | -.32 | -.13 | .132 | .054 | -.28 | -.30 | -.14 | -.21 | -.50 | -.3 | -.37 | -.04 | -.39 |
| June 1st | Sum of 3 evaluation components | .312 | .608* | .425* | .195 | .355 | .349 | .167 | .298 | .193 | .06 | .253 | .347 | -1.0** |
| | CONTENT | .026 | .313 | .225 | .215 | -.11 | -.12 | -.14 | .039 | -.09 | -.0 | .000 | -.02 | .311 |
| | DELIVERY | .522 | .549* | .422* | .491* | .700** | .454* | .236 | .644** | .664** | .26 | .398 | -.11 | .171 |
| | LANGUAGE USE | -.18 | .056 | .080 | .226 | .097 | .155 | -.14 | -.26 | -.27 | -.2 | -.04 | -.19 | -.36 |
| June 8th | Sum of 3 evaluation components | -.10 | .319 | .031 | .298 | -.20 | .154 | -.19 | .150 | .262 | .03 | .320 | .270 | .378 |
| | CONTENT | -.35 | -.02 | -.08 | .301 | -.15 | -.06 | .269 | -.03 | .088 | -.3 | .289 | .197 | .180 |
| | DELIVERY | .191 | .211 | .295 | .259 | .040 | .379 | .159 | .509* | .163 | .18 | .582** | .240 | .384 |
| | LANGUAGE USE | -.17 | .189 | .516* | .026 | .189 | .386 | .253 | .029 | .181 | .18 | .378 | -.36 | -.08 |

* $p < .05$ ** $p < .001$

Appendix E *The correlation coefficient between peer evaluation and teacher (JET) evaluation*

| | | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B11 | B1 | B1 |
|----------|--------------------------------|------|-----|------|------|------|------|------|------|------|------|------|------|-----|
| May 11th | Sum of 3 evaluation components | .05 | .40 | .24 | .39 | .18 | .37 | -.04 | .44 | .19 | .075 | .05 | .02 | .26 |
| | CONTENT | .32 | .32 | .24 | .01 | -.05 | .52 | .42 | .11 | -.06 | .157 | .49 | .04 | .03 |
| | DELIVERY | .66 | .52 | .56 | .09 | .25 | .50 | .51 | .32 | .25 | .246 | .31 | .59 | .30 |
| | LANGUAGE USE | .27 | .25 | .11 | .05 | -.34 | .22 | -.07 | .63 | -.50 | .462 | .10 | .34 | .29 |
| June 1st | Sum of 3 evaluation components | .41 | .71 | .52 | .33 | .50 | .46 | .37 | .61 | .19 | .400 | .20 | .45 | .05 |
| | CONTENT | .16 | .59 | .53 | .29 | .09 | .01 | -.02 | .34 | -.09 | .203 | .17 | -.02 | .33 |
| | DELIVERY | .48 | .82 | .50 | .59 | .48 | .55 | .47 | .57 | .66 | .664 | .57 | -.11 | .49 |
| | LANGUAGE USE | .39 | .21 | .74 | -.12 | .09 | .30 | .26 | .61 | -.27 | .150 | -.11 | -.19 | .38 |
| June 8th | Sum of 3 evaluation components | -.13 | .40 | .12 | .36 | .04 | .12 | -.06 | .15 | .21 | .415 | .16 | .37 | .33 |
| | CONTENT | -.38 | .05 | -.07 | .30 | -.24 | -.09 | .25 | -.03 | -.03 | .044 | -.29 | .31 | .24 |
| | DELIVERY | .32 | .32 | .35 | .37 | .12 | .51 | .12 | .45 | .54 | .380 | .28 | .65 | .37 |
| | LANGUAGE USE | -.17 | .18 | .51 | .02 | .11 | .18 | .25 | .38 | .25 | .029 | .18 | .18 | .37 |

* $p < .05$ ** $p < .001$

Appendix F

The responses of the self-evaluation Group A

| ID | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 |
|---------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|
| Group A | 2 | 2 | 2 | 1 | 1 | 5 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |

Note: 1 Very effective to improve English proficiency
 2 Effective to improve English proficiency
 3 Not so effective to improve English proficiency

- 4 Not effective to improve English proficiency at all
- 5 No idea

Appendix G

The responses of the peer evaluation Group B

| ID | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B11 | B12 | B13 |
|---------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|
| Group B | 2 | 2 | 5 | 5 | 3 | 5 | 2 | 3 | 1 | 5 | 2 | 3 | 5 |

- Note:* 1 Very effective to improve English proficiency
 2 Effective to improve English proficiency
 3 Not so effective to improve English proficiency
 4 Not effective to improve English proficiency at all
 5 No idea