Bulletin of Faculty of Education, Nagasaki University : Curriculum and Teaching No.50（2010）109-124

Whether Tis Nobler to Normalize Ⅱ

－ Increasing peer-evaluator consistency when evaluating presentations in the classroom －

Itsuro INAGE * Etsuko LAWN ** Murray LAWN ***

(Received October 30, 2009)

英語スピーキングの評価方法についての実践的研究

－評価数値の統計的標準化についての検討 (2) －

稲毛　逸郎*　ローン　悦子**　ローン　マリー***

（平成21年10月30日受理）

Abstract

In the evaluating of such as presentations in the classroom “peer evaluation” has become increasingly recognized as a potentially valuable part of the overall evaluation process. One of the challenges, however, in the evaluation of any activity involving speech is the aspect of achieving inter-evaluator consistency. In this paper the peer evaluated results of classroom presentations are normalized and the resulting increased inter peer-evaluator consistency is analyzed and discussed.

1．Introduction

Peer evaluation is commonly used in the classroom for a variety of reasons. Such reasons may include increasing student motivation on account of presenting to peers, motivation on account of an increased awareness of what they are being evaluated on, and giving the students a sense of both responsibility and accountability in the evaluation process (Evans, 2008). However, if the resultant data is simply averaged, it will quite possibly not represent a fair result. The inherent “variation of the individual” in their interpretation of the categories was evaluated, and actual allocation of a grade, even assuming each individual evaluator evaluates all peers consistently, will likely vary significantly between peer-evaluators. That is, inter-peer evaluator consistency is likely to vary significantly. The inherent process of averaging these grades will take care of

*長崎大学教育学部教授・国際文化講座
**長崎大学長大教育機能開発センター非常勤講師
***長崎純心大学人文学部英語情報学講師

absolute differences in grade averages but does not account for variation in standard deviation. This means that the standard deviation of the peer-evaluator in effect defines how much influence each evaluator will have on the final outcome.

In order to minimize this effect, normalization of the standard deviation has been used effectively (Inage, Lawn &Lawn, 2009).

Background

Previous papers by the authors have studied the aspect of increasing inter-evaluator consistency in the evaluation of speech, how it is dealt with by major testing agencies (Inage, Lawn, 2006) and they have proposed and tested the use of normalization of evaluator standard deviations to provide increased inter-evaluator consistency in the case on non-prior evaluator training (Inage, Lawn &Lawn, 2007).

Evaluation in the Classroom

Previous papers by the authors have focused on the evaluation of interviews or presentations by formally qualified “teachers of English.” In this study, however, the application of the same concept of normalization of evaluator standard deviations is applied to inter-peer evaluation results from the classroom. While peer-evaluation is generally agreed to as being potentially reliable, the specific degree of reliability will no doubt vary with the aggregate level of diligence and honesty exhibited by the students making up the class. In this regard the teacher will most likely be able to make a fair judegment.

１. Method

The Classes

This study uses data from two classes of students. The first class consisted of 17 students who gave presentations of whom 12 students provided evaluation data. The students in this class were majoring in English and/or education. The second class consisted of 52 students who gave presentations of whom 51 students provided evaluation data. The students in this class were majoring in Computer Science.

The Presentations

Presentation 1. Self Introduction – Self Evaluation Only

At the beginning of the term students were asked to prepare a short self-introduction in English. This short presentation was given in class and recorded on video. The video was placed online for the individual students to evaluate their own respective presentations. The evaluation criteria were explained prior to the evaluations and each evaluated item was modeled, both from poor (teacher’s best effort) to very good (again teacher’s best effort). The evaluation form (online) was as follows (blank and new lines abbreviated):

Presentation Evaluations, Self-Evaluation.

Your name, Grade each item out of 5 as follows: 1 = poor, 2 = below average, 3 = average, 4 = good, 5 = very good (outstanding)

1. Time management 1-2 minutes. (minus 1 mark per 10 seconds out of range min. = 1)

2. Loudness of voice 3. Posture and eye contact 4. Gestures 5. Correct pronunciation 6. Intonation 7. Stress and pauses 8. Fluency 9. Memorization (confidence) 10. Enthusiasm Total score /50

Comments: What I did well. What I need to improve. How can I improve my presentations?

Submission: Please copy and paste the above data into a word processor, complete the assignment and send it to [teacher's name] at [teacher's email address] by the [deadline, date and time].

In this initial exercise the goal was to help the students understand that they were at least aware of a number of the key aspects that go toward making an effective presentation in English. The students were given the internet URL (website), asked to watch their own presentation and evaluate it based on the above criterion. The final section was also considered important, namely "What I did well," and "What I need to improve." The data was explained in the form. It was to be copied and pasted into an email, completed and submitted by email. This exercise also provided the teacher with the student's preferred email address for future correspondence. Previously an auto-fill email form was used; however, it presumes the user's default email software is used, which is not often the case, so the students were asked to copy and paste the data from the webpage. Alternatively, a downloadable text (.txt) file could be used.

The results of this initial short presentation prepared the students regarding expectations for the end of term presentation. The results of the initial presentations were generally very good and reflected levels perhaps typical for public university students in Japan and varied reflecting the respective majors. However, as is often the case, even among some English majors, content was simply "read." As a result, many students scored poorly in 2. Loudness of voice, 3. Posture and eye contact, 4. Gestures, 8. Fluency, 9. Memorization (confidence) and 10. Enthusiasm. For this reason the following presentation's criterion was simplified to focus on overcoming the tendency to just "read."

Presentation 2. Open Subject - Peer and Self Evaluated

At the end of the term students were asked to prepare a two to three minute presentation in English. The subject for this presentation was "open"; however, choice of a subject that was related to "what they wished to do or study in the future" was strongly encouraged. To that end "How to be a System Engineer" proved popular among Computer majors and "How to be a lovely wife" proved to be a popular topic for female students. For this presentation, the students were asked to use Microsoft PowerPoint or OpenOffice.

org presentation software media to compliment the presentation. The effective usage of “bullet points” and “supporting images” was taught during the term. Students were asked to write their intended script under the slides (visible on the presenter’s screen only in MS PowerPoint). The prepared file was submitted to the teacher two weeks or so, prior to the presentation for correction and grading. In the case of the final presentation, half of the final grade was given, based on the submitted slides and script. The other half of the grade was provided through peer-evaluation. This short presentation was given in class and also recorded on video. The students peer-evaluated the presentations in real time, that is, while listening to the presentations, and the video was placed online for the individual students to evaluate their own respective presentations. In the case of students being absent, they watched and evaluated other student’s presentations online. The evaluation criteria was explained prior to the evaluations and each item was modeled, both from poor (teacher’s best effort) to very good (again teacher’s best effort). The items were optimized to encourage the students to speak to the audience rather than read from their script. The concept of speaking about each of the key points, that is, bullet points, was encouraged, and simply reading a script was discouraged. The resultant modified evaluation items were as follows:

1. Loud voice? 2. Eye contact 3. Gestures 4. Fluent? 5. Memorized 6. Enthusiasm
7. Media

Item No. 7 “Media” was added to include evaluation of the presentation media software used. The remaining items, except for the additional “media” item, were chosen so as to be to some degree negatively impacted by a “script reader.” Items notably dropped from the first presentation included “5. Correct pronunciation,” “6. Intonation,” “7. Stress and pauses.” While these items are very important, they were for convenience summarized under “4. Fluent?” so as to focus on getting students eyes off their scripts.

For the purposes of convenience, a spreadsheet was used to prepare for data input. The spreadsheet was printed out and provided for students to fill in directly while listening to the presentations. Later the students were asked to access the same webpage previously mentioned and download a Microsoft Excel file. The Excel file was the same as the printed version provided in class. Access to free software namely “OpenOffice.org” was also explained for students who did not have access to the Excel software. Students were required to input their own data and submit the file as an attached email file. The students were also asked to watch their own presentation online and as with the first presentation to evaluate their 2nd presentation. The reasons for asking students to enter their own data were that particularly in the case of the larger class, entering 52 sets of data is time consuming, and that it was estimated to take 10 to 20 minutes per student. Initially the University’s e-learning software was considered but proved impractical at that time. By using a spreadsheet each student’s submitted file (excel file) could be opened and the data directly copied into a master spreadsheet for analysis. The file submission ratio reflected the student's majors, English/Education 12/17 and Computing 51/52. The classes were taught in regular classrooms, thus students needed to enter the data elsewhere in their own time.

３．Results – Class 1

Figure 1 shows the results of the second presentation discussed previously – class 1, a small class (17) of English major students. The presentation was evaluated by class peers in real time, and the student assessed or peer grades are labeled E1 to E12. The *x* axis shows grade scores. The grade scores are the accumulated totals of 7 items each with a possible score of 1 to 5. Thus the possible score range is 5 to 35. The *y* axis S1 to S17 indicates each student's score received. The significance of the S-number is that it represents the raw ranking of the presentations. That is, S1 is the student that received the highest average score. The relationship between student numbers (S1 to 17) and evaluator numbers (E1 to E12) is random and as is evident, 5 students did not submit grades. The presentations were videoed and graded later by the classes' teacher (a native speaker of English) and a Japanese teacher of English, these results are labeled NTE (native teacher of English) and JTE (Japanese teacher of English), respectively. Finally the raw average is also shown with a thicker and darker line.

Grading Characteristics – Class 1

As can be seen clearly in Figure 1, the spread of individual grading tendencies is very wide. With reference to Figures 1 and 2, the resulting improved inter-evaluator grading consistency can be clearly seen in Figure 2 after normalization. For the purposes of comparison the vertical *x* axes are identical so that the extent of the improvement can be seen and compared visually. Specifically looking at the main central area, which is excluding the very high and very low ranging grades, the raw data or original data shown in Figure 1 ranges between about 15 to 30.

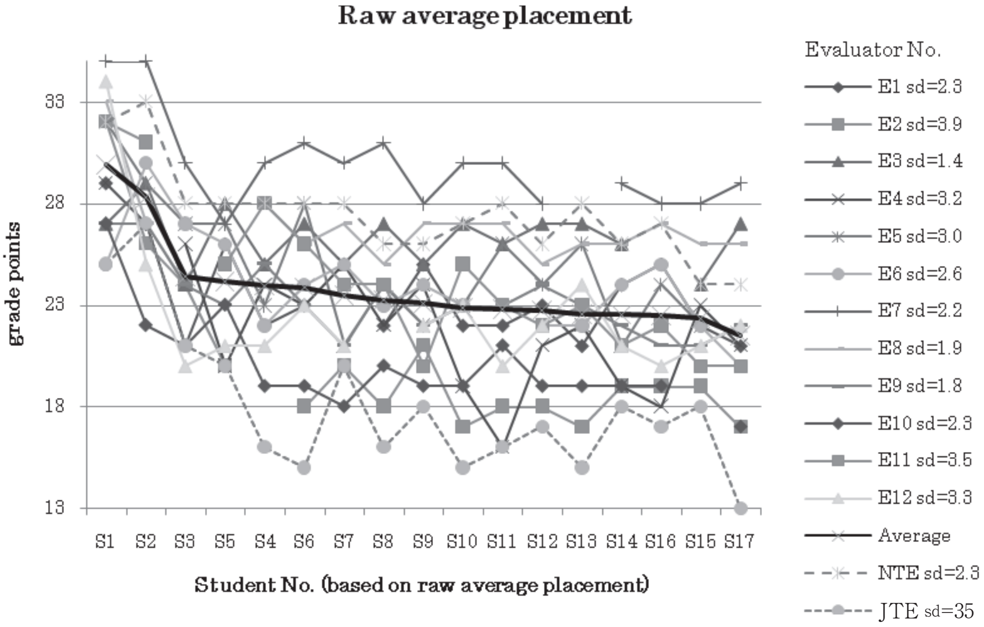

Figure 1. Grading results based on simple averaging - presentation 2 Class 1

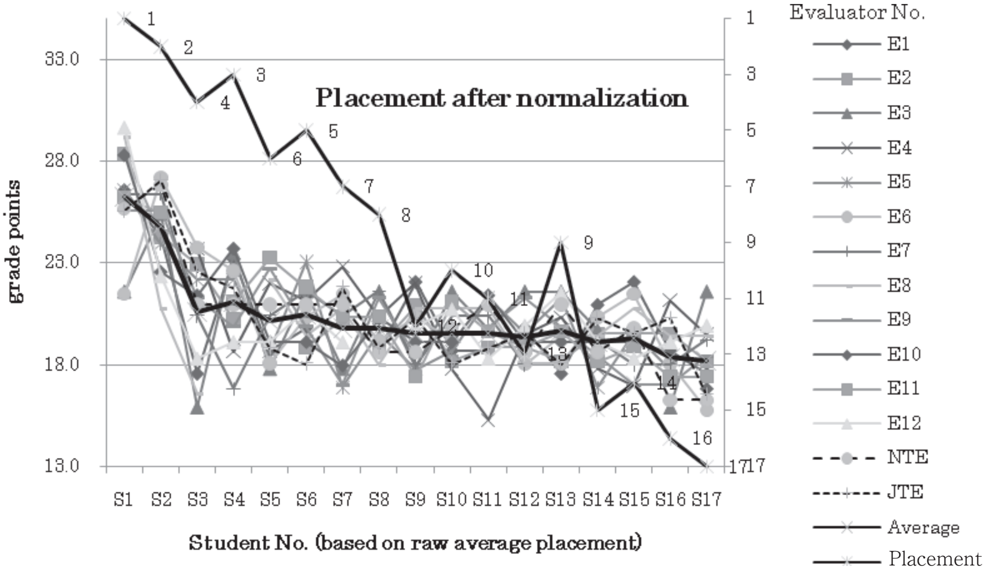

Figure 2. Grading results and placement based on normalization of peer-evaluator standard deviations presentation 2 Class 1

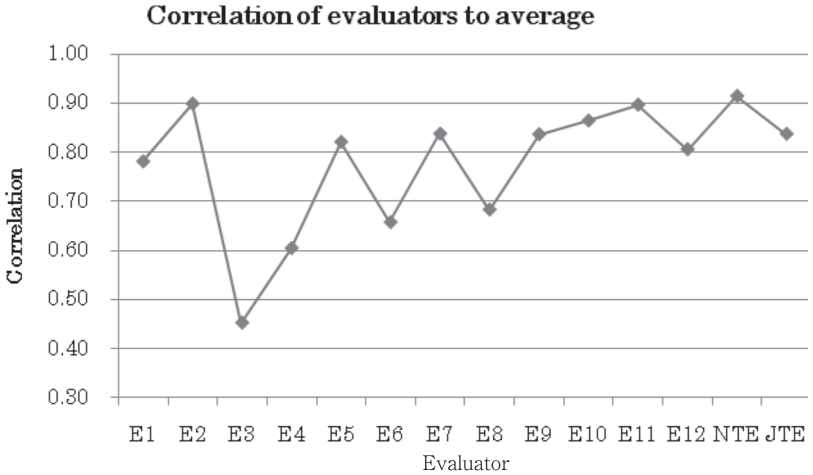

Figure 3. Correlation of evaluator grading profiles to the average grading profile presentation 2 Class 1

This compares with the normalized results range excluding the extremes of about 16 to 23. In very approximate terms the inter-evaluator variation is halved. In regard to ranking as seen in Figure 2, some changes have also occurred. Placement is indicated on the right hand side axis, and generally varies from top left to bottom right, and any variation indicates change of placement due to the normalization process. Furthermore, the extent of placement variation can be seen by the rate of change either in line angle or by reading the rank number indicated at each point on the line. It is noted that most changes in rank order occur in areas where the average variation is low, that is, where there is only a very small margin in variation between the results. Possible causes for the grading characteristics observed in these graphs are discussed in later sections.

Figure 3 shows the correlation of individual evaluator grading profiles to the average grading profile - presentation 2 class 1 displayed in student No. order (random order). This is followed by the grading profiles of both a native speaker of English (Teacher) and a Japanese English Teacher. Noted in this graph is the fact that most of the peer evaluator's evaluations correlated closely with each other and with both of the English teacher's evaluations. This is evident by nine of the thirteen evaluator's correlations being over 0.8, three of the evaluations falling between 0.6 to 0.7 and finally one evaluator indicating a fairly low correlation of 0.45. This data shows that for this particular class most of the students graded their peer students in much the same manner as the teachers graded.

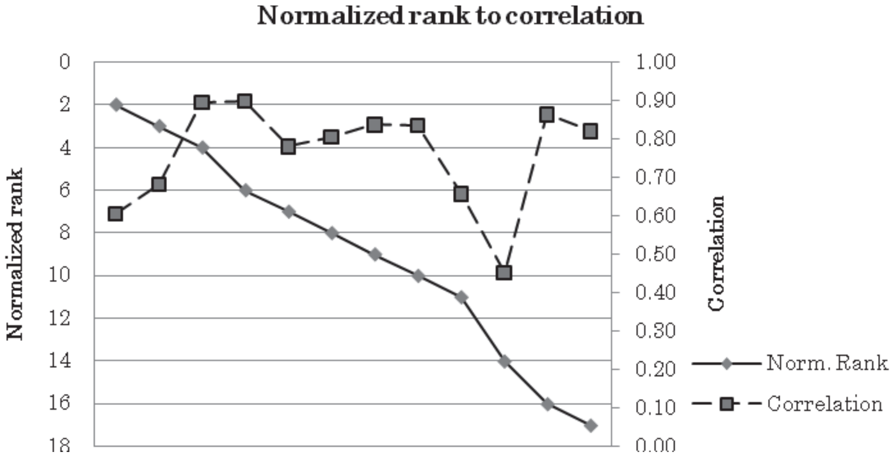

Figure 4. Correlation of evaluator grading profiles to the average grading profile - presentation 2 Class 1

Figure 4 shows the correlation of evaluator grading profiles to the average grading profile - presentation 2 Class 1 displayed in order of the respective evaluator's ranking received (actual ranking also shown). The purpose of this re-ordering was to see if there was any relationship between the evaluator's personal performance as indicated by their received rank. It would seem possible that higher performing students would also be better at grading; however, as can be seen from Figure 4, no such conclusion can be made in the case of this class. In other words, the relationship between the individual student's correlation and grade received appears quite random.

Re-evaluation based on the removal of a low correlating evaluator

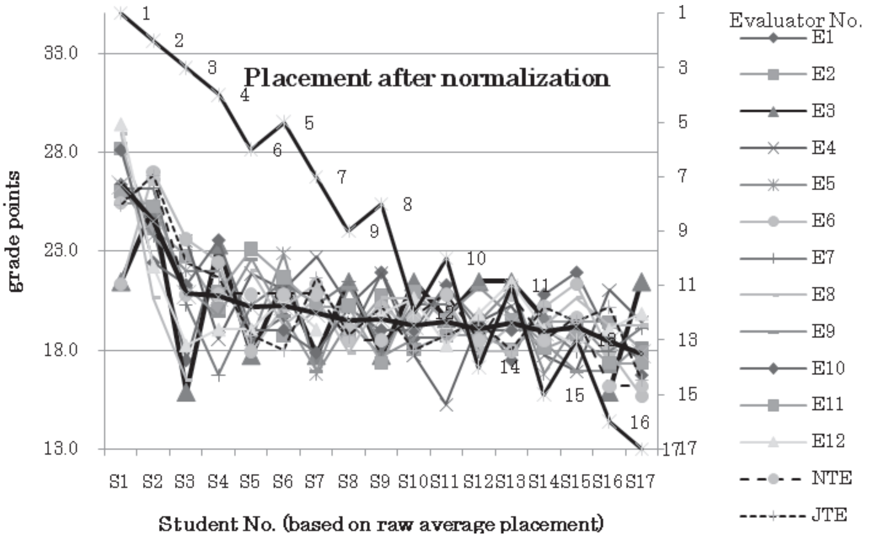

Figure5. Re-grading results and placement based on normalization of peer-evaluator standard deviations presentation 2 Class 1, less a slightly irregular evaluator

Figure 5 shows the re-evaluation of ranking after the removal of an evaluator's results that appeared slightly irregular. In Figure 3. "Correlations of evaluators to average," it was noted that one evaluator (Evaluator 3) exhibited a grading profile that differed from other evaluators. This could perhaps have been due to such as input error or inattentiveness and is discussed more in regard to class 2 later in this paper. The re-evaluation of ranking is shown in Figure 5 and is noted to result in an overall more linear placement. The profile of Evaluator 3 has been removed from the normalization calculation but the actual grading profile is shown as a bold black line. While the line is not clearly visible at all points, the notably low grading of Student 3 is seen clearly as being the cause of swapping 3rd and 4th places when compared to Figure 2. It must, however, be noted that Evaluator 3 was not alone in giving Student 3 a low grade.

Results – Class 2

Figure 6 shows the results of the second presentation discussed previously – class 2 a large class (52) of students majoring in computing (non-English). The presentation was evaluated by class peers in real time. The student assessed or peer grades are not labeled individually as in class 1 on account of the large numbers. The x axis shows grade scores. The grade scores are the accumulated totals of 7 items each with a possible score of 1 to 5. Thus the possible score range is 5 to 35. The y axis S1 to S52 indicates each student's score received. The significance of the S-number is that it represents the raw ranking of the presentations. That is, S1 is the student that received the highest average score. 51 of the 52 evaluators submitted grades. The presentations were videoed and made available online for the students' self evaluation, for absent students to use for evaluation of other students and as a final check for overall grading consistency. Finally, the overall raw average is shown with a darker and thicker line.

Grading Characteristics

In Figure 6, it can be clearly seen that the spread of individual grading tendencies is very wide. With reference to Figures 6 and 7, the resulting improved inter-evaluator grading consistency after normalization can be clearly seen in Figure 7. For the purposes of comparison, the vertical x axes are identical so that the extent of the improvement can be seen and compared visually. Specifically looking at the main central area, which is excluding the very high and very low ranging grades, the raw data or original data shown in Figure 6 ranges mainly between about 17 to 31. This compares with the normalized

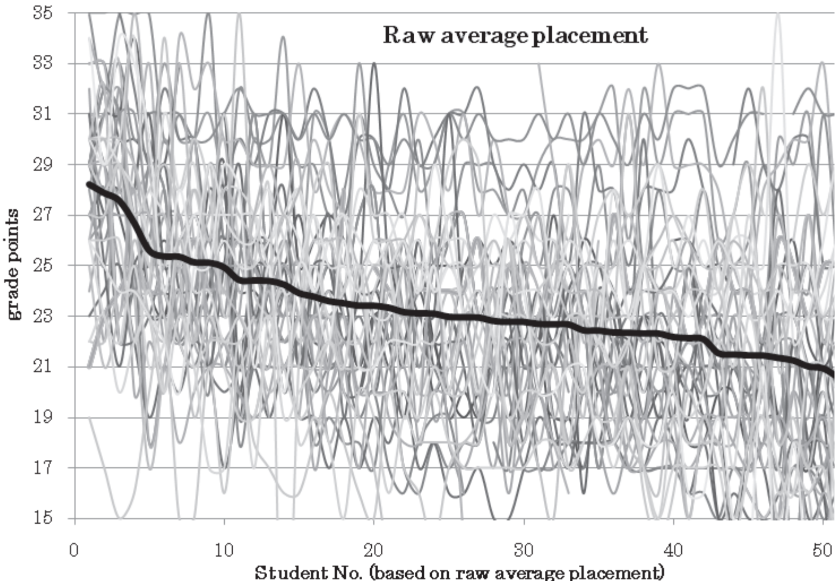

Figure 6. Grading results based on simple averaging - presentation 2 Class 2

results range excluding the extremes of about 19 to 27. In very approximate terms the inter-evaluator variation is halved although reflecting slightly less of a reduction compared to class 1. In regard to ranking as seen in Figure 7, the average line shows a negative angle in places. This indicates placement changes as was explained in the previous section in regard to class 1; however, on account of the large student numbers it is not practical to display on the same graph.

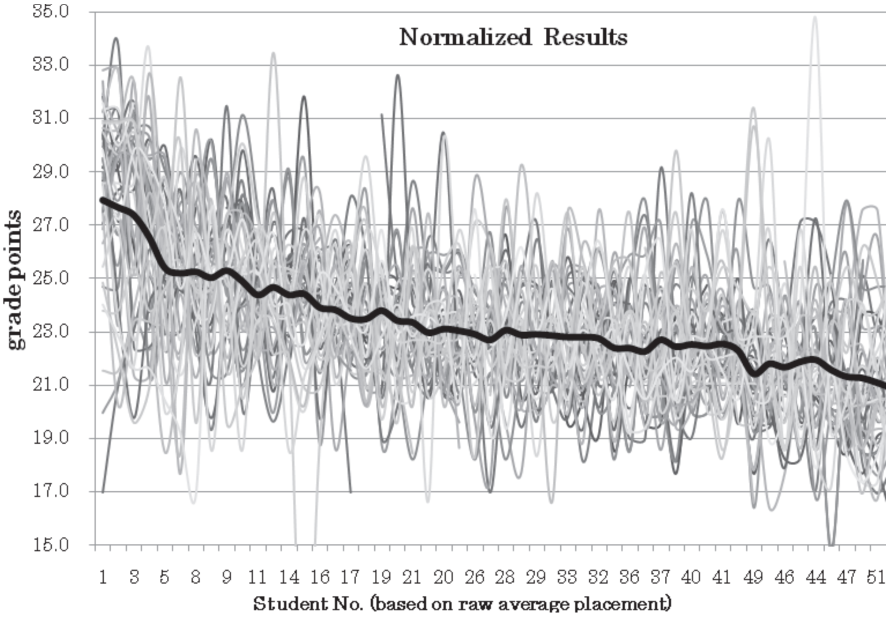

Figure 7. Grading results based on normalization of peer-evaluator standard deviations presentation 2 Class 2

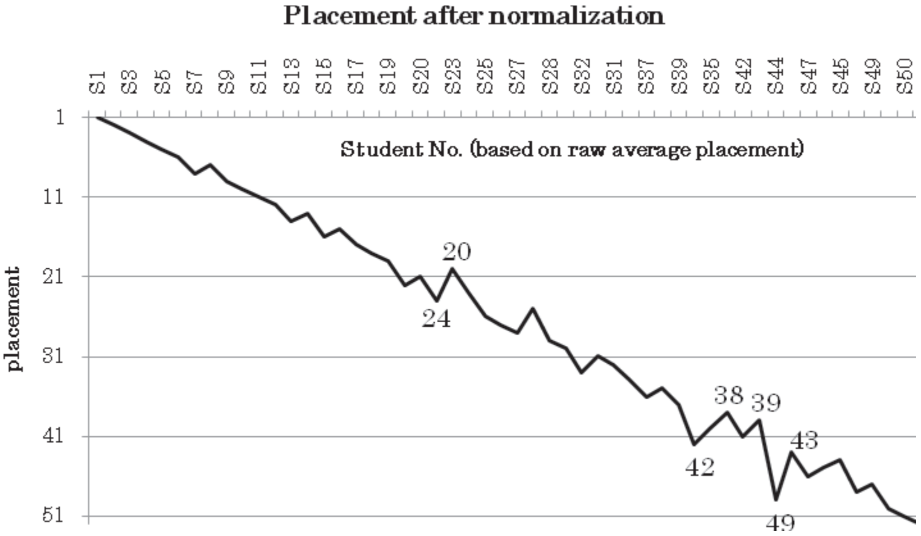

Figure 8. Placement based on normalization of peer-evaluator standard deviations presentation 2 Class 2

The resulting variation in placement on account of the normalization is shown in Figure 8. Placement is shown on the *x* axis, and generally varies from top left to bottom right, any variation indicates change of placement due to the normalization process. Furthermore, the extent of placement variation can be seen by the rate of change either in line angle or by reading the rank number indicated at points on the line which reflect significant change. It is noted that most changes in rank order occur in areas where the average variation is low, that is, where there is only a very small margin in variation between the results. Possible causes for the grading characteristics observed in these graphs are discussed in later sections.

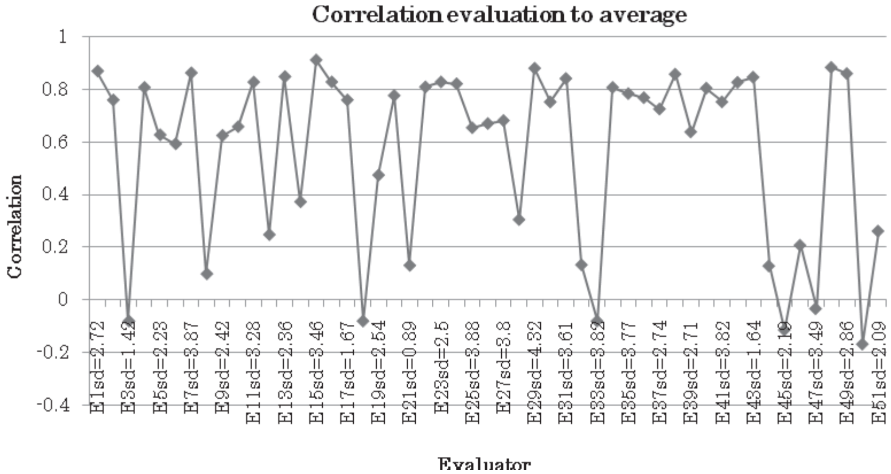

Figure 9. Correlation of evaluator grading profiles to the average grading profile - presentation 2 class 2

Figure 9 shows the correlation of individual evaluator grading profiles to the average grading profile - presentation 2 class 2 displayed in Evaluator No. order (random order). The correlations by evaluator reflect similar trends as those seen in Figure 3 in terms of percentiles above 0.5. However, the presence of some very low and negative correlations would most likely indicate a lack of attentiveness during one or more of the presentation classes or erroneous input of data into the spreadsheet by the students. Furthermore, 4 or so students produced identical data sets, indicating that the data was most likely just copied from another student. These very low correlations raise significant doubt over the data's validity; therefore, two sets of results are shown in Figure 10. One data set shows the placement with the low correlating data intact, and the second set shows re-placement with low correlating data removed (<0.5).

Figure 10 shows a comparison of presentation 2 class 2's results normalized. The *y* axis labeled S1 – 52 indicates the original student placement using simple averaging (ranked by all). The "Ranked by all - normalized" data set indicates the placement after the normalization process including all evaluator's data (see Fig. 8). The "Ranked by high correlating evaluators only–normalized," is the result after evaluators exhibiting correlations of less than 0.5 have been removed (see Fig. 9). Notable features of the re-grading are similar to those in class 1 shown in Figure. 5; most variations occur in the mid

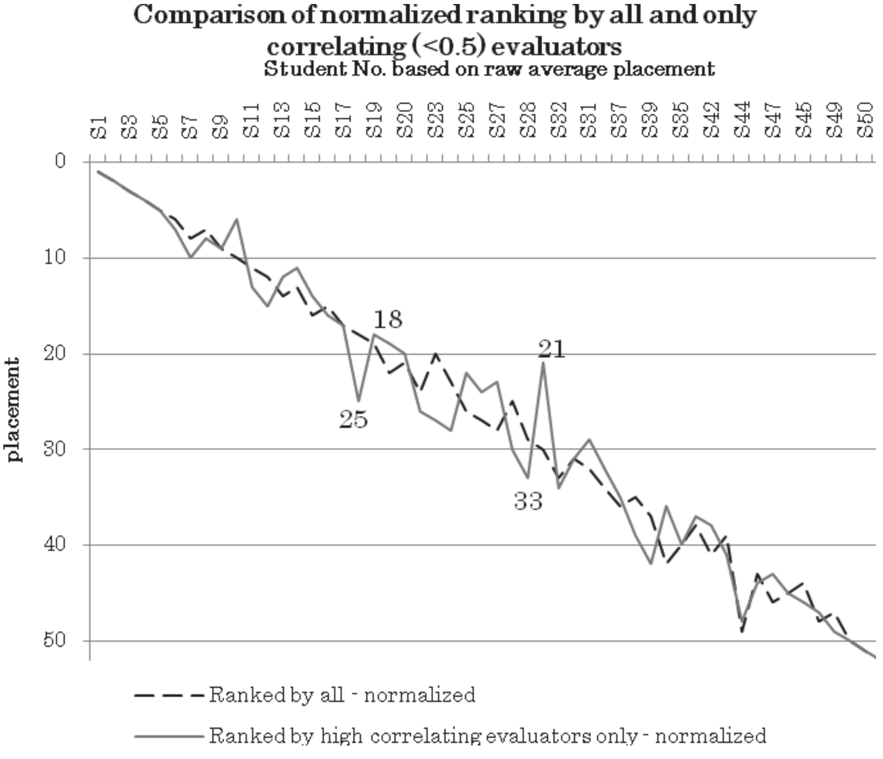

Figure 10. Re-grading results and placement based on normalization of peer-evaluator standard deviations presentation 2 class 2 – grading by all compared to grading by evaluators exhibiting a high correlation with other evaluators (>0.5)

section where actual grading variations are minimal (See Figs. 6-raw and 7 Normalized). Very significant changes occur in the change of placement of Student 18, whose "raw" and "normalized by all" placement is 18th . However, after low correlating evaluators' data is removed falls to 25th place. The reverse is Student 30, whose "raw" and "normalized by all" placement is 30th . After low correlating evaluators' data is removed rises to 21st place. It is difficult to ascertain as to why this may occur, except to question the possibility of personality or personal issues being involved, peer-friendship or otherwise issues. The noted favor or disfavor received by certain class members by the low correlating students as noted in Figure 10 would tend to indicate the possibility of favoritism or discrimination towards certain class members that extends beyond the presentation under evaluation. It is no doubt that this is to some degree an unavoidable issue to contend with in regard to peer-evaluation. In the case of major organizations that evaluate speech [2], no personal prior knowledge of, or relationship in any form with the person being evaluated is a prerequisite. However, in the case of peer-evaluation this is unavoidable, and perhaps an issue requiring further study towards minimization. It is perhaps unrealistic to ask student's closest friends in an equal light as those who are somewhat distant or even hostile toward them on a regular basis.

４. Discussion

Checking Student Papers

In considering the preparation and checking of a class of 52 students presentations (class 2), the time required to check the presentations word for word, is quite substantial. Likewise, the class time required to stage 52 presentations each of 2 to 3 minutes duration is bordering on the "impractical." However, for a class which focuses on presenting "Technical English" it seemed logical. The presentations took up a total of 3 full classes (class length 90 minutes) and required special arrangements for absent students. The checking process also doubled as grading of the content and provided the first half of the student's grade.

The Modified Grading to Encourage Eye Contact

In regarding the modification of the grading criterion to emphasize the importance of "facing the audience," speaking based around the key points (bullet points) was provided on the screen (projector) rather than reading a script from start to finish. The reality was that most students did simply read from start to finish, which made the grading very difficult, as most items encouraged memorization or "ad lib," thus facilitating eye contact, gestures, etc. This is because the strict grader most students could have validly given a grade of "poor" for most items. The result made it very difficult to find any variation between most students. The students that excelled were those who succeeded in getting their eyes off the script to look at the audience or who excelled greatly on one of the evaluated items. Such examples were students that spoke with outstanding enthusiasm, or outstanding fluency (mainly students who had studied overseas), or students who diligently memorized their script or at least significant parts of the script. Perhaps it is unrealistic to expect Japanese university students to memorize or ad-lib a presentation provided with only key points; however, in a previous study [1], nearly all the private "high school" students memorized near word perfect their respective speeches with no prompts (Speech contest). Perhaps some practice would assist one computer major student when asked regarding a notable lack of presentation, fluency, she proudly said in Japanese "of course I didn't practice." The phrase "practice makes perfect," in English perhaps should be emphasized more in these classes.

Further Suggestions to Improve Eye Contact

Discussions with an experienced teacher of English provided suggestions of a compromise, specifically using a small "palm of the hand" sized card with just the main points and a requirement never to break eye contact with the audience for more than about 3 seconds. In the case of using such software as PowerPoint, simple use of the slide would suffice, and also requesting a second person to operate the computer was suggested to further improve the eye contact with the audience.

Realistic Concentration Spans

In comparing the above small class (17) of English/ Education major students with the large class (52) of computing major students, not only the aspect of "major field" but also the actual time the students spent evaluating other students must be considered. In the case of the small class, all presentations were held and graded within a single 90 minute class and moreover, the students in that particular class all seemed to work closely together as a class and were equally interested in each student's speech, as was noted by vibrant feedback throughout the presentations. The larger class may have had some vibrancy in regard to some of the presenters; however, the extended time span over 3 90 minute classes no doubt took its toll on perhaps even the most diligent of evaluators. For those students who had minimal interest to begin with it was perhaps unrealistic to expect fair results over such an extended period of time. This perhaps was the cause of near zero or negative correlations, indicating that more or less random grading data was entered for most students, while certain students seemed to have been singled out either positively or negatively.

The Value of Peer-Evaluation

In the case of the diligent evaluator, it is assumed that careful evaluation of other students in regard to given important aspects of presenting would raise the student's awareness of, and increase the student's motivation toward, working on the aspects being evaluated. However, this assumption should perhaps be tested more empirically.

5. Conclusion

This paper has discussed and analyzed on a small scale the validity of peer - evaluation. Based on the assumed validity of peer–evaluation, this paper focused on means to increase peer–evaluator consistency when evaluating presentations in the classroom. The use of "normalization," previously used in speech contests, was used and provided a notable increase in inter (peer) evaluator consistency. Furthermore, after analyzing inter–evaluator correlation the presence of low correlating (<0.5) evaluator grading profiles were noted. The low correlations were assumed to reflect either data entry errors or a lack of interest or diligence in the evaluating process by the respective students. This was particularly notable in the larger non-English major class. The low correlating data was removed from the evaluation process and the grading was re-graded, based on the modified data set. The result, particularly in the case of a large class, showed clear cases of students with low correlating evaluator profiles, exhibiting favoritism towards and discrimination against other specific students.

6. Future Issues

The overall value of peer evaluation in the classroom (of such as presentations) which has been beneficial for the student is largely presumed but should be more empirically tested. Furthermore, a means to counter or "normalize" the natural tendency in order to evaluate favorably student's peer friends and to evaluate more stringently those who are perhaps more distant or hostile peers needs to be addressed.

References

Evans, D. 2008. Reflections on Peer Evaluations in an English Language Course*, The Journal of Nursing Studies,* National College of Nursing, Japan, Vol. 7, No. 1. pp. 41-49.

Inage, I., Lawn, E. and Lawn, M. 2009. Whether tis nobler to normalize – increasing inter-evaluator consistency in the evaluation of oral communication based activities -, *Bulletin of Faculty of Education,* Nagasaki University: Curriculum and Teaching No.49 (March 2009), pp. 93-102

Inage, I. and Lawn, E. 2006. A preliminary Consideration of English Speaking Tests – Based on the Analysis of Current English Proficiency Tests -, *Bulletin of Faculty of Education,* Nagasaki University: Curriculum and Teaching No.46 (March 2006), pp. 137-151.

Inage, I., Lawn, E. and Lawn, M. 2007. A Study of Native and Japanese Speakers of English Grading Tendencies of Speaking Ability - Based on the Analysis of Interview Evaluations and Background Questionnaire -, *Bulletin of Faculty of Education,* Nagasaki University: Curriculum and Teaching No.47 (March 2007), pp. 129-143.