IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.004, LIIR178
DOI: https://doi.org/10.52731/liir.v004.178

Analysis of Mutual Evaluation Activity and Student Performance in Creative Project-Based Learning

Motoki Miura *

Abstract

In this study, we analyzed mutual evaluation activities in a creative group learning lectures with some perspectives. Firstly, we analyzed correlations between personal achievements (mini-test, exercise) and group achievements (group works and presentations) for six years of lectures with Lego Mindstorms. From the result, we confirmed the significant correlations between oral presentations and demonstration. We also found the positive correlations between personal achievements and presentations. Secondly, we assessed correlations of students' mutual evaluation stats, rank correlations, and total scores. We confirmed the significance of correlation between average points and spearman rank correlations. A correlation between the standard deviation and the spearman rank correlation could be utilized to check the validity of mutual evaluations by students. Thirdly, we compared two types of peer-assessment activities, one in which learners were forced to enter comments and the other in which comments were optional. We confirmed that students with a larger standard deviation of the comment length are more serious about mutual evaluation.

Keywords: Peer review/assessment, group work, achievements, collaborative learning

1 Introduction

In creative group learning, peer review/assessment is popular, and commonly used to evaluate achievements and outcomes of students in many fields[2]. We also carried out mutual evaluations/peer review by students for evaluating group work achievements. However, fair and accurate assessment of group work by students can not always be guaranteed. Shiba et al. proposed a mutual evaluation technique for the fair assessment of individual students[3] but it requires several regroupings to generate a trust network. We consider that the regrouping is not applicable for longer-term projects. To address the fair and accurate assessment of group work by students without regrouping, we start gathering and analyzing our data of mutual evaluations of previous lecture courses. We discuss the results based on the correlations of student scores and stats of mutual evaluations.

* Chiba Institute of Technology, Chiba, Japan
Note: this article is a revised version of a manuscript [1] presented at an annual meeting of Japan Creativity Society.

In this paper, we analyzed student peer evaluation activities from three perspectives. The first is the correlation between the results of individual activities and the results of group activities. We investigated the relationship between individual outcomes related with programming skills, such as mini-tests and exercises, and peer review scores of presentations produced and presented by groups with creative projects. The reason why we investigated the first perspective is because we thought that by clarifying the relationship between individual performance and group performance would confirm the characteristics and reliability of the mutual evaluation itself.

The second is the relationship between the average and standard deviation of the scores given by the students in the student peer evaluation, the value of the correlation with the final group ranking, and the overall evaluation score of the students. The second perspective is based on the hypothesis that students with higher overall evaluation scores tend to evaluate each other more seriously.

The third analyzed the relationship between the amount of feedback comments in peer evaluations, the scores given by students, and the students' overall evaluation scores. Writing feedback comments puts a greater burden on students than giving marks, but students are expected to make a thorough evaluation accordingly. Therefore, we investigated the difference in mutual evaluation activities between cases where comment writing was compulsory and cases where it was not compulsory.

2 Analysis of Peer-Assessment in the 1st and 2nd Perspectives

The evaluations in the first and second perspectives were analyzed based on mutual evaluations conducted in PBL lectures, which are described in detail below. The lectures surveyed under the third perspective are described in detail in the next section.

Figure 1: Scene of the group work in LEGO lecture

2.1 Target Lecture: Practical Programming PBL

We explain the first target lecture *Practical programming PBL* for the first and second perspectives in detail. In this lecture, students were expected to learn the basics of programming using Lego Mindstroms, and then complete and present their work in groups. Since educational effectiveness of using robotics is high[4], Lego Mindstorms have been introduced for engineering education [5, 6] as well as peer learning[7] and project-based

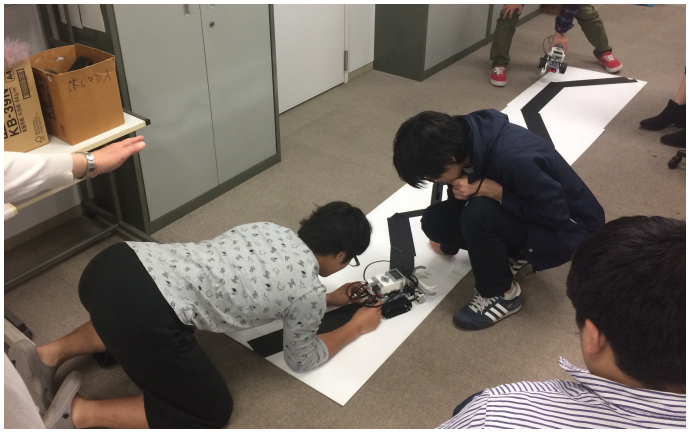

Figure 2: Scene of the group work in LEGO lecture: Learners passionate about line tracing

learning[8]. Our institute also introduced Lego Mindstorms for creative project-based learning at an engineering course and conducted lectures over six years.

The course term was 15 weeks. About fifty-five students (90% male, 10% female) attended the course every year. All participated students were in the second grade of university. The students worked in groups of 4 to 5 people through the course. There was no group replacement during the class. The students were expected to carry out a project with their original thoughts and motivations by a group (Figure 1). Except the project, we performed a line trace time trial in the fourth week to get familiar with LEGO Mindstorms (Figure 2). After that, the students mainly carried out the project by the group. In addition, a 10-minute mini quiz at beginning of each lecture time for confirmation of basic programming skills (Mini-test), and individual and group activity report (Exercise) were assigned.

The student groups were asked to orally present their achievement twice at the mid-term and the final. The mid-term presentation basically consisted of the purpose and the plan of the group project. The final presentation included demonstration and result of their works/products. Mutual evaluations (peer reviewing) were performed by a web interface (see Figure 3). All students were asked to evaluate individuals of other teams as well as other groups with the criteria. All points were 5-point Likert scale, and default value was 3. Averaged points were reflected to 20% of the total score for each presentation (40% in total). The students could check all comments and average/standard deviation of points just after the presentation.

We also conducted an open presentation for people outside the university and elementary school students to participate. The students were expected to explain their project to the outside/younger participants with demonstration. The outside/younger participants voted each project by seals, and the number of seals were reflected to the group score of open presentation as 10% of the total score.

Firstly, we analyzed score correlations between personal achievements (mini-test, exercise) and group achievements (mid-term, final, and open presentations) from 2013 to 2018. After that, we assessed correlation of total score and characteristics of mutual evaluation (average and standard deviation) by students.

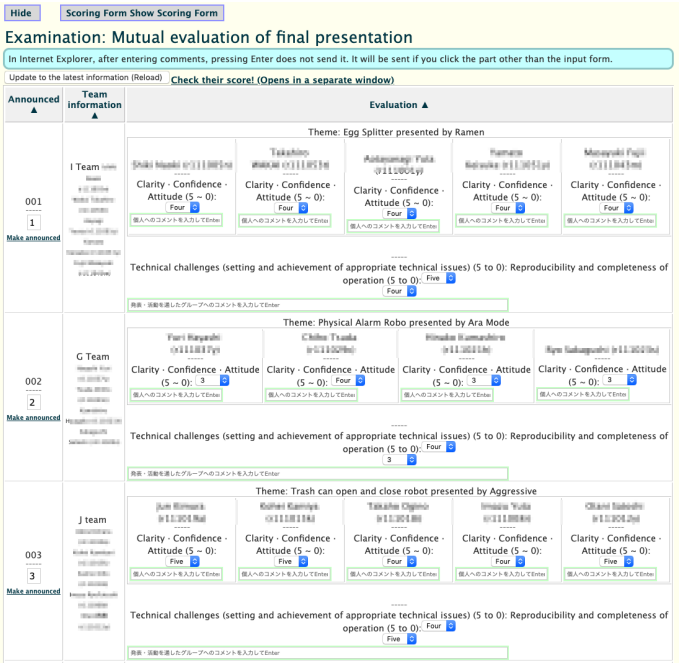

Figure 3: Web interface for mutual evaluation (Translated by Google Translate Extension)

2.2 Result

Figure 4 shows one of the correlation matrices between personal achievements (mini-test, exercise) and group achievements (mid-term, final, and open presentations) in 2016. The diagonal elements of the matrix show the histogram, the upper right numbers are correlations, and the lower left cells indicate scatter plots. Number of asterisks after the correlation shows the p-value (*, **, and *** represent p-value less than 5%, 1%, 0.1%, respectively). Since the *Presen* score was the sum of *MidP* and *FinalP*, these correlations were obviously higher than others. We will omit them in the following analysis. Figure 5 shows the summary of all correlation matrices from 2013 to 2018. Since the mini-test had started in 2016, the correlations with the mini-test before 2016 were missing. The most reliable finding was that the correlation between P (*Presen*) and *OpenP* was high. Other correlations varied from year to year. One possible interpretation was that mini-test quizzes were more correlated with the presentation. The mini-test scores were more closely related to programming comprehension and ability than exercises. The high-skilled students tended to get high scores in presentations.

Figure 6 shows one of the correlation matrices between students' mutual evaluation stats (average and standard deviation of points), rank correlations with the averaged result by group (*Spearman*), and the total student score. Note that the rank correlations were calculated by group evaluation scores. Generally, the high rank correlations (*Spearman*) represent the student fairly and precisely evaluated groups. Figure 7 shows the summary of all correlation matrices from 2013 to 2018. These results also contained variations from year to year, but some tendency could be confirmed. The most significant tendency was the high correlation between average and spearman (*Avg-Spearman*). The initial point in the web interface was 3 of 5. Thus, the student who rated higher points in average could fairly evaluated the group achievement. The second major correlation was *Sdev-Spearman* (nega-

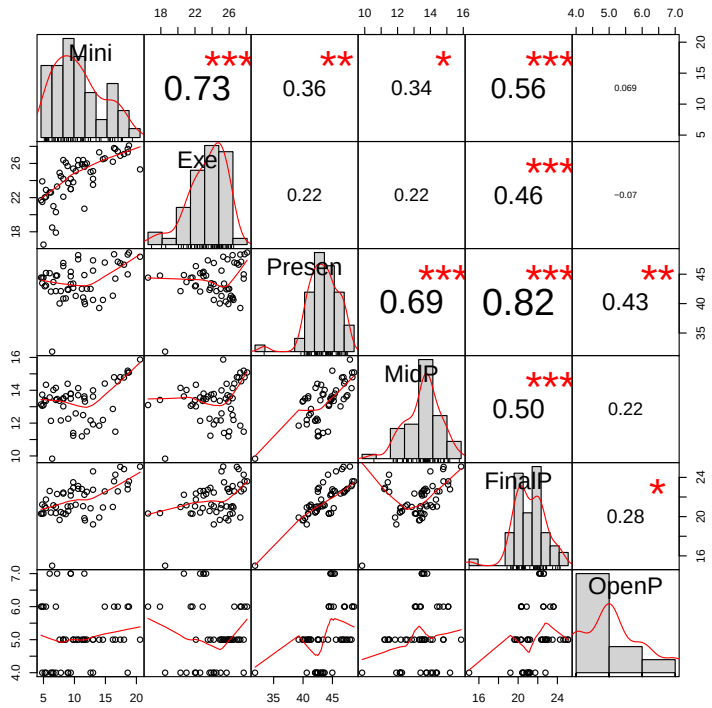

Figure 4: A correlation matrix between personal and group achievements in 2016

	Exe-P	Exe-MidP	Exe-FinalP	Exe-OpenP	P-OpenP	Mini-Exe	Mini-P	Mini-MidP	Mini-FinalP	Mini-OpenP
2013				0.25 +	0.97 ***					
2014					0.95 ***					
2015	0.47 ***	0.68 ***	0.41 **	0.32 *	0.72 ***					
2016			0.46 ***		0.43 **	0.73 ***	0.36 **	0.34 *	0.56 ***	
2017			0.25 +		0.83 ***	0.39 **	0.31 *	0.26 +	0.28 *	
2018					0.89 ***	0.25 +				0.28 *

Figure 5: Summary of correlations between personal achievements (mini-test, exercise) and group achievements (presentations: mid, final, and open) from 2013 to 2018

tive correlations). The result implies that the student who rated points with larger variances decreased the fairness. The third major correlation was *Avg-Sdev* (positive correlations), but it is natural result in a sense. We focus on the *Avg-Score* and *Sdev-Score* correlations. The former one had negative tendency and the latter had positive tendency. These results show that the high-score students were evaluated seriously. As a result, the average point for the other group decreased and the variance of the point increased. We consider that the significance of negative correlations in *Sdev-Spearman* suggests that many students may have scored carelessly and/or the averaged group point was inappropriate/unfair for scoring. To avoid the careless evaluations, it is possible to introduce a mechanism that can not complete the mutual evaluation without exceeding a certain variance value.

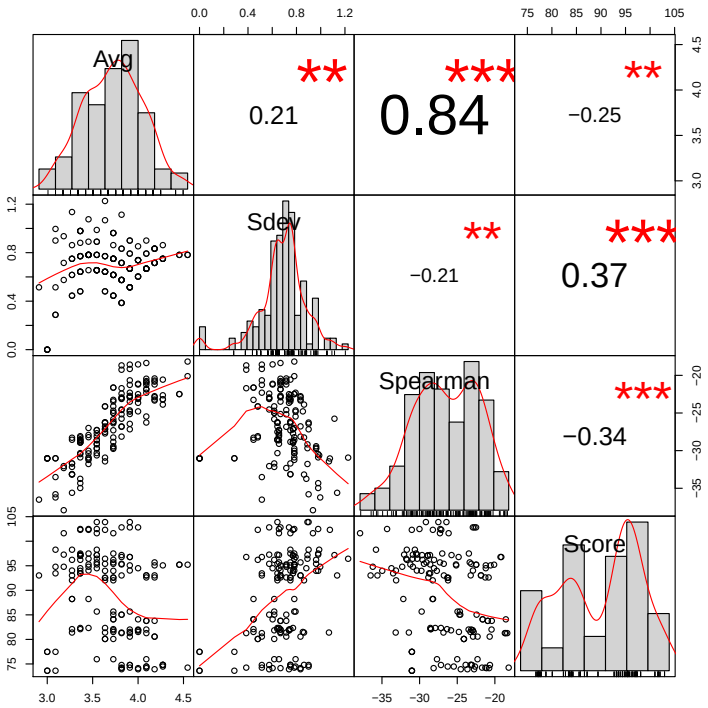

Figure 6: A correlation matrix with students' mutual evaluation stats in 2013

3 Analysis of Peer-Assessment in the Third Perspective: Effect of Feedback Comments

The third perspective is to investigate trends in the amount of feedback comments entered during peer assessment, peer assessment scores, and student grades. When conducting mutual evaluation of group presentations, it is common to include free descriptions as part of the feedback to students. However, practical research on the amount of feedback comments, the scores of mutual evaluations, and students' awareness of mutual evaluation has hardly been conducted. Generally, writing comments puts a greater burden on students than giving marks, but students are expected to make a thorough evaluation accordingly. Therefore, we investigated the difference in mutual evaluation activities between cases where comment writing was compulsory and cases where it was not compulsory.

3.1 Target Lectures

In order to compare the length of comments, the following two lectures were targeted.

1. Basic Experiments in Information and Communication Engineering (experimental lecture, compulsory subject, 2020)

2. Algorithm and Data Structure (regular lecture, elective subject, From 2020 to 2022)

Both lectures were aimed at second-year university students. In the first experimental lecture, students work on five themes for two weeks each. After that, the students prepare the presentation on the last theme and present it in the final week by group. The number of

	Avg-Sdev	Avg-Spearman	Avg-Score	Sdev-Spearman	Sdev-Score	Spearman-Score
2013mid		0.18 *		-0.50 ***		
2013final	0.21 **	0.84 ***	-0.25 **	-0.21 **	0.37 ***	-0.34 ***
2014mid	0.24 **		-0.22 **	-0.61 ***	0.14 +	
2014final		0.81 ***	-0.34 ***	-0.30 **	0.25 *	-0.40 ***
2015mid	0.19 +	0.82 ***	-0.20 *	-0.19 +		-0.20 *
2015final	0.35 ***	0.79 ***				
2016mid	0.20 *	0.70 ***		-0.29 **	0.38 ***	
2016final				-0.63 ***		
2017mid	0.50 ***	0.83 ***				
2017final	0.30 **		-0.20 *	-0.34 ***		
2018mid	0.47 ***	0.86 ***	0.21 *		0.35 ***	
2018final	0.51 ***	0.76 ***			0.21 *	-0.17 +

Figure 7: Summary of correlations with students' mutual evaluation stats

students in one group is 4. Usually, the presentation of the experiment is held face-to-face, but due to the influence of Covid-19, it was held online in 2020. At that time a mutual evaluation was carried out with a maximum of 5 points with a web interface (like Figure 3 but there were no comment text fields for each presenter). Students were instructed to write descriptive comments on other groups' presentations.

The second lecture is a conventional regular lecture. At the end of this lecture, students are expected to summarize what they have learned, develop a program that utilizes what they have learned, and make presentations on the content and objectives in groups of two or three students. We also prepared web interface for the mutual evaluation with 7-point Likert scale, with feedback comment input areas for each group.

3.2 Result

Figure 8 and Figure 9 show pair-plots of forced and optional comment entry, respectively. *Score* was the personal achievement score on the lecture except the *Presen*, *Presen* was the score of peer evaluation for the presentation, and *AvgTxtLen* and *SdevTxtLen* indicated the average and standard deviation of the feedback comment text length. Also, *AvgEval* and *SdevEval* showed the average and standard deviation of points given to other groups as mutual evaluation. In general, it can be said that the higher the *SdevEval*, the student evaluated other groups more seriously.

From the Figure 8, the correlation between *SdevTxtLen* and *SdevEval* was 0.23. This indicates that the standard deviation of the feedback comment text length is larger for students who earnestly conduct mutual evaluations. For the standard deviation *SdevTxtLen* to be large, it is necessary to include a certain number of long comments, so it can be said that the student entered feedback comments seriously.[1] The correlation between *SdevEval* and *Score* was also 0.20. It can be said that the students who earnestly evaluated their work also worked earnestly on their experimental reports.

The distribution of *AvgTxtLen* by comparing the hisograms of *AvgTxtLen* in Figure 8 and Figure 9. In the regular lecture with arbitrary comment input, the students entered less or short feedback comments. Similarly, from the difference in the distribution of *AvgEval*, it can be said that there were many students who scored relatively high in mutual evaluation where comments were optional. This phenomenon is also reflected in the skewed distribution of *Presen*. From the Figure 9, the correlation between *AvgTxtLen* and *AvgEval* was

[1] Students with a high peer rating comment average but a small standard deviation might enter similar/uniform comments.

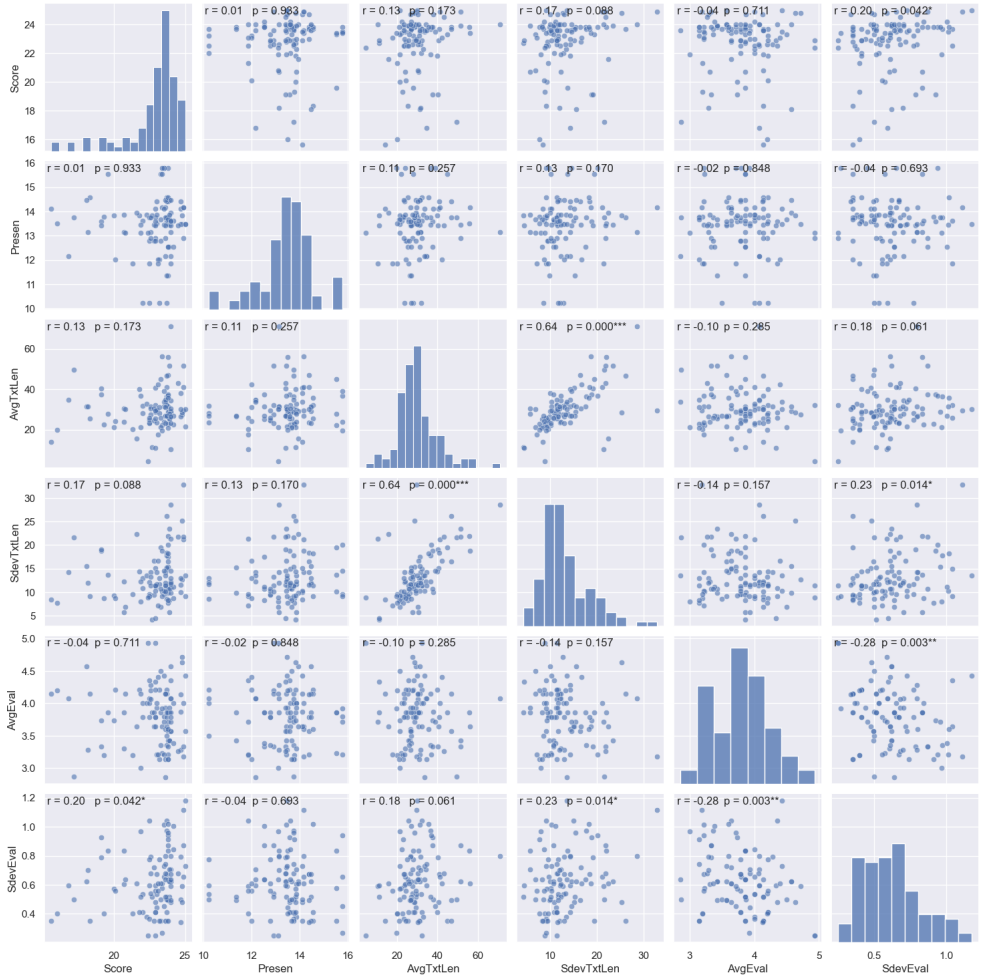

Figure 8: Pair-plot of the experimental lecture with mandatory comment input

0.26. Therefore, the students who input many comments tended to give higher evaluation scores to other groups. This phenomenon did not appear in comment-forced mutual evaluations. This is probably because the experimental lecture is a compulsory subject and the other is an elective subject.

From this research, we can conclude that students are more likely to be rigorous in their mutual evaluations if comments are required. However, since differences in subject characteristics are also possible, it is necessary to continue to investigate.

4 Conclusions

In this study, we analyzed relationships between student performance and mutual evaluation activities. Firstly, we analyzed correlations between personal achievements (mini-test, exercise) and group achievements (mid-term, final, and open presentations) from 2013 to 2018. From the correlations, we confirmed that the correlations between oral presentations (mid-term and final) and demonstration (open) were high. We also revealed the positive correlations between personal skills and presentations.

After that, we assessed correlation of students' mutual evaluation stats (average and

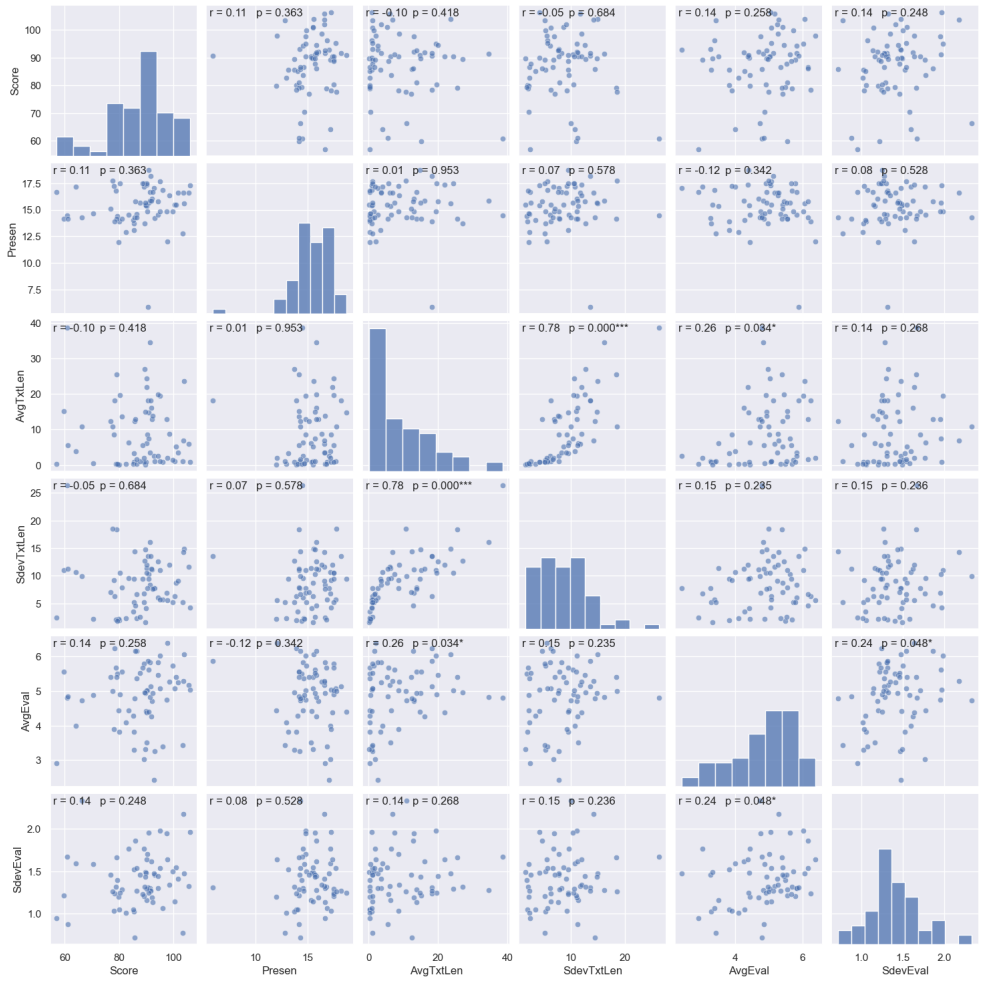

Figure 9: Pair-plot of the regular lecture with arbitrary comment input

standard deviation of points), rank correlations, and the total score. The most significant tendency was the high correlation between average points and spearman rank correlations. However, the significance of negative correlations between the standard deviation and the spearman rank correlation may suggest inappropriate/unfair mutual evaluations by students.

Finally, we compared correlations of mutual evaluation stats and the personal/group scores with different instructions on feedback comments. We confirmed that students with a larger standard deviation of the comment length are more serious about mutual evaluation. Also, students are more likely to be rigorous in their mutual evaluations if comments are required and mandatory.

Since it would be a burden on students to force all groups to make comments, we would like to explore a method of rigorously conducting mutual evaluations while reducing the burden on students.

Acknowledgments

The part of this research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research JP19K03056 and JP22K12319.

References

[1] Motoki Miura. Analysis of student performance and mutual evaluation activity in creative project-based learning using LEGO mindstorms. In *The 41st Annual Conference of Japan Creativity Society, Proceedings of International Session*, pages 17–20, September 2019.

[2] Keith Topping. Peer assessment between students in colleges and universities. *Review of educational Research*, 68(3):249–276, 1998.

[3] Yumeno Shiba, Haruna Umegaki, and Toshiharu Sugawara. Fair assessment of group work by mutual evaluation with irresponsible and collusive students using trust networks. In *PRIMA 2015: Principles and Practice of Multi-Agent Systems*, pages 528–537, Cham, 2015. Springer International Publishing.

[4] Fabiane Barreto Vavassori Benitti. Exploring the educational potential of robotics in schools: A systematic review. *Computers & Education*, 58(3):978–988, 2012.

[5] Alexander Behrens, Linus Atorf, Robert Schwann, Bernd Neumann, Rainer Schnitzler, Johannes Balle, Thomas Herold, Aulis Telle, Tobias G Noll, Kay Hameyer, et al. Matlab meets lego mindstorms—a freshman introduction course into practical engineering. *IEEE Transactions on Education*, 53(2):306–317, 2009.

[6] Carmen Fernández Panadero, Julio Villena Román, and Carlos Delgado Kloos. Impact of learning experiences using lego mindstorms® in engineering courses. In *IEEE EDUCON 2010 Conference*, pages 503–512. IEEE, 2010.

[7] Ciarán Mc Goldrick and Meriel Huggard. Peer learning with lego mindstorms. In *34th Annual Frontiers in Education, 2004. FIE 2004.*, pages S2F–24. IEEE, 2004.

[8] Mike Carbonaro, Marion Rex, and Joan Chambers. Using LEGO robotics in a project-based learning environment. *The Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 6(1):55–70, 2004.