Peer grading reduces instructor's workload without jeopardizing student learning in an undergraduate programming class

Fedor Duzhinr & Amrita Sridhar Narayanan

Nanyang Technological University

*Corresponding Author: fduzhin@ntu.edu.sg

Keywords: *Peer grading; Cooperative learning; STEM teaching; Machine learning in education; Quasi-experiment*

Abstract

In an undergraduate programming class taught at Nanyang Technological University, Singapore, students (*N*=243) were given an opportunity to grade reports submitted by their peers. 10% of all students participated in peer grading and were satisfied with the grade given to them by peers (i.e., this group did not use instructors' resources). 13% participated in peer grading, updated their reports based on peer feedback, and submitted to a course tutor for final grading. We have shown that even though students who participated in peer grading and updated their reports achieved higher scores, but it happened because they were stronger students to begin with. At the same time, scores of students who participated in peer grading and did not re-submit their reports to an instructor were not lower than average scores. Thus peer grading can be recommended in teaching programming classes as a strategy that reduces instructors' workload while not jeopardizing students' learning.

Introduction

Collaborative / cooperative learning is an educational approach that empowers students to take charge of their learning. Requiring higher individual accountability and positive interdependence than traditional learning, this method is sought to be implemented in university education. This is because it is well-suited for learning higher order knowledge (McWhaw & Schnackenberg, 2003). In Asian context, cooperative learning has been shown to reduce perceived difficulty of the subject (Lee et al., 1997).

Very often, researches distinguish between collaborative (Dillenbourg, 1999) and cooperative learning (Sharan, 2002). This distinction is not a focus of our paper. However, it is important to know that most forms of either approach are based on students working in teams. It is understandable since teamwork is a crucial skill for future learners (Koh et al., 2018). We, however, are exploring a different, more free-form approach where students are given a choice whether or not to participate in a collaborative learning activity.

Previous research proves the positive impact of active learning in STEM disciplines, with methods such as Team-Based Learning improving a student's mastery of the content (Freeman et al., 2014). This has been attributed to the structure of collaborative learning where students are required to work more consistently on homework compared to the traditional learning.

The collaborative learning activity that we are interested in is peer grading. In peer grading, students are required to write a report and

then each student grades a certain number of reports written by peers according to a rubric provided by the course instructor. Details vary – peer grading may be formative or summative, may include or not include self-grading etc. Peer grading is collaborative since students learn from each other but it does not involve teamwork.

Research literature on peer grading mainly focuses on three themes. The first theme is validity (Ryan et al., 2007) or reliability (Gopinath, 1999) of peer grading. The main method here is to compare scores given by students to scores given by instructors (validity) or other students (reliability). The main finding is that peer grading is highly consistent with instructor grading when students are given clear grading criteria but biased in the sense that students tend to give lower than deserved score to best peers (Sadler & Good, 2006). Another important finding is that academically strong students are also more accurate assessors (González-Betancor et al., 2019). The second theme is educational value of peer grading (Crossman & Kite, 2012; Baker, 2016). The consensus is that peer grading has some benefits to students – it may help them to learn better quantitatively (e.g., get higher exam scores) or qualitatively (e.g., learn skills or aptitudes that they would not learn otherwise, such as empathy or self-awareness). The third theme is student perception of peer grading (Mulder et al., 2014; Ghahari & Sedaghat, 2018), (Burke Moneypenny et al., 2018). Findings are mixed. Simply put, a lot of students feel anxious about peer grading and yet many are excited about it.

Research literature on peer grading is rich and versatile but, still, gaps exist. One aspect of peer grading that researchers often mention but do not explore further is the fact that it saves instructors' time. This is particularly important in the context of MOOCs (Luo et al., 2014). Our study aims to show that peer grading can save instructors some time without jeopardizing learning and student perception of the course. The novel thing that we implemented is optional participation in peer grading. Optional participation presumably helped to alleviate some of students' worries about peer grading. Another aspect of peer grading or, probably, concern that we have about existing literature is the fact that most existing studies are about business or medicine subjects. Literature on peer grading in STEM disciplines and, in particular, programming is scarce. The authors of (Grov et al., 2017) found certain benefits in peer grading but reported that students are not ready to replace teacher's feedback with peer grading. Our study adopts a more quantitative approach. We attempt to show that, at least sometimes, peer grading can replace teacher's feedback by precisely measuring peer grading and teacher grading.

Methodology

Setup

The word "experiment" is not very accurate because we are using non-experimental data collected in the normal teaching and learning process. Since it was not a deliberate experiment, we will call it "setup".

One of assessment components in the course "Algorithms and Computing III" taught by the first author in 2017 at Nanyang Technological University, Singapore was an individual report on a partially open-ended problem with the objective of modelling an ostrich farm using a stage-structured population model (Lefkovitch, 1965) and finding appropriate parameter values in MATLAB.

While a student may or may not have been able to produce a report that meets all the requirements, it is not too challenging to verify if a given report meets the requirements. For example, one of the requirements was to do all the coding without using explicit loops. Actually doing it is challenging, but verifying if someone else's code does not have explicit loops is easy.

Grading each report takes, on average, about 20 minutes of an instructor's time and is a boring chore. It is tempting to delegate the grading duty to students themselves, especially since grading coding tasks is relatively straightforward. However, there are two main concerns. First, students who participate in peer grading lose the opportunity to learn from the instructor's feedback. Second, students may simply be not comfortable with peer grading.

Group	N	mean	median	SD
Control (TUTOR)	54	76.1	78.5	12.0
Experiment 1 (PG_KEPT)	24	75.9	76.4	9.7
Experiment 2 (PG_CHANGED)	31	80.9	79.0	9.5

Table 1: Aggregated scores in the control group (students who did not do peer grading and were graded by the tutor), experimental group 1 (students who did peer grading and kept the score they were given by peers) and experimental group 2 (students who did peer grading, updated their report based on peers' comments and submitted it to an instructor). To balance group sizes and to mitigate influence of students who failed the report, we removed scores below 50 (there were 21 of them, one participated in peer grading) and we randomly selected 54 out of 187 students who opted out of peer grading to form the control group.

We made peer grading an optional activity. Students who opted for peer grading were required to submit their reports 3 days before the deadline. Each student participating in peer grading had 2 days to grade 3 reports submitted by their peers. Then results of peer grading were released. Students who were satisfied with the score they were given by peers could simply keep it. Students who were not satisfied with the score given by peers had one more day to update their report and to submit it to an instructor.

Each of our students consciously made one of the following choices:

- not participating in peer grading – we randomly selected 54 out of 187 of them to form the control group;
- participating in peer grading and keeping the score given by peers – they formed experimental group 1 after removal of one student who failed the report;
- participating in peer grading, updating their report based on peers' comments and submitting to an instructor for regrading – they formed experimental group 2.

Statistics of final scores for this activity are shown in Table 1.

Ideally, we would like to find out if participation in peer grading had any effect on students' scores. A "good" effect would be observed if students in experimental group 1 (these are students whose reports were not graded by the tutors and hence they did not use the school's resources) did not perform worse than students in the control group. At the same time, we hope that students in experimental group 2 (these are students for whom peer grading was a true feedback loop) performed better than students in the control group.

Specifically, research questions that were are interested in are:

- Did experimental group 1 perform better than they would have had they chosen not to participate in peer grading?
- Did experimental group 2 perform not worse than they would have had they chosen not to participate in peer grading?

Data cleaning

Our non-experimental set-up (data summarized in Table 1) has all the usual disadvantages of a quasi-experiment, such as confounding variables and self-selection bias.

To (at least, partially) control for confounding variables, we measure the initial level of our students by a diagnostic quiz and by their grade point average (GPA). Processed data are summarized in Table 2.

Note that the control group in Table 2 is labelled "TUTOR" for it consists of students whose individual report was graded by the tutor only, experimental group 1 is labelled

	n	mean	SD	median	min	max
PG_CHANGED.quiz	31	44.8	31.7	40.0	0.0	100.0
PG_CHANGED.gpa	31	4.3	0.5	4.3	3.0	5.0
PG_CHANGED.report	31	80.1	9.5	79.0	61.0	96.0
PG_CHANGED.gain_report	31	1.9	8.6	3.5	-15.6	14.3
PG_CHANGED.project	31	87.9	13.1	89.2	62.5	111.1
PG_KEPT.quiz	24	44.8	30.6	55.0	0.0	100.0
PG_KEPT.gpa	24	4.0	0.5	3.8	3.2	5.0
PG_KEPT.report	24	75.9	9.7	76.4	51.3	93.8
PG_KEPT.gain_report	24	-2.0	6.4	-1.3	-19.5	7.6
PG_KEPT.project	24	88.5	12.5	91.5	57.1	105.9
TUTOR.quiz	54	40.5	32.0	40.0	0.0	100.0
TUTOR.gpa	54	4.0	0.6	4.0	2.7	5.0
TUTOR.report	54	76.1	12.0	78.5	51.0	95.0
TUTOR.gain_report	54	-0.2	8.7	0.4	-24.5	15.9
TUTOR.project	54	85.4	10.9	86.5	54.8	105.5

Table 2: Statistics of quiz scores, GPA (grade point average), and individual report in the control group (TUTOR), experimental group 1 (PG_KEPT) and experimental group 2 (PG_CHANGED). Variables "gain_report" and "project" are explained later.

"PG_KEPT" since they participated in peer grading and kept the score they received from peers, and experimental group 2 is labelled "PG_CHANGED" since they participated in peer grading and changed their report based on peer reviews.

Comparing raw scores

Let us compare raw scores for the individual report in three groups, control ("TUTOR"), experimental group 1 ("PG_KEPT") and experimental group 2 ("PG_CHANGED").

We tested four hypotheses and reported results in Table 3.

The difference in mean report scores between groups "PG_CHANGED" (experimental group 2) and "TUTOR" (control group) is statistically significant. However, a superficial look at Table 2 provides an explanation – students in "PG_CHANGED" are, on average, stronger than students in "TUTOR" as they have a higher mean quiz score and a higher grade point average. At the same time, while students in "PG_KEPT" had about the same mean final score as students in "TUTOR", the former group had a higher average quiz score. This may be an indication that stronger students in "PG_KEPT" did not perform as well as they could have, i.e., participation in peer grading may have been detrimental for them. A more careful analysis is in the next section.

Comparing report gains

Following (Duzhin and Gustafsson, 2018), we have constructed a mathematical model that predicts students' report score from their quiz and GPA scores. Let:

$$M_1 = 0.25\max(\text{gpa}^3,\text{quiz}) + 0.8\sqrt{\text{quiz}} - 27.2\text{gpa} - \frac{382.73}{\text{gpa}} + 260.66,$$

$$M_2 = 0.75\sqrt{\text{quiz}} - 45.91\ \text{gpa} + 3.62\ \text{gpa}\sqrt[3]{\max(\text{gpa}^3,\text{quiz})} - \frac{419.6}{\text{gpa}} + 303.22,$$

$$M_3 = 115.02 - \frac{234.27}{\max(\text{gpa}, \sqrt[3]{\text{quiz}}) \cdot \sqrt[9]{2\text{gpa} + \text{quiz} + 10}},$$

$$M_4 = 116.86 - \frac{218.62}{\max(2, \text{gpa}, \sqrt[3]{\text{quiz}}) \cdot \sqrt[12]{2\text{gpa} + \text{quiz} + 5.79}}.$$

Null hypothesis	Test	p-value	Reject at p=0.05	Cohen's d (95% CI)
Mean report score in "PG_CHANGED" is the same as in "TUTOR"	t-test	0.04	reject	0.43 [-0.02, 0.88]
Mean report score in "PG_KEPT" is the same as in "TUTOR"	t-test	0.92	accept	0.02 [-0.46, 0.50]
Median report score in "PG_CHANGED" is the same as in "TUTOR"	u-test	0.12	accept	NA
Median report score in "PG_KEPT" is the same as in "TUTOR"	u-test	0.64	accept	NA

Table 3: Statistical significance of comparing raw final scores

The model's prediction is then the median of M_1, M_2, M_3 and M_4. It may look complicated, especially roots of degree 9 and 12 (the former ranges between 1.3 and 1.7, the latter between 1.2 and 1.5), but it is still an explicit formula and it can be interpreted. For instance, it is clear that the predicted report score is an increasing function of both gpa and quiz, which makes perfect sense.

The model was inferred from data by the method called symbolic regression (Vladislavleva et al., 2008). The R^2 of the model is 0.44.

The residual of this predictive model, or *gain*, i.e., the actual report score minus the predicted score, is interpreted as a measure of learning in the course. The gain results from observed treats, e.g., choice of learning activities, and unobserved treats, e.g., rapport with the instructor. Our argument that the reader may or may not agree with is that unobserved treats are random noise, i.e., they equally randomly affect all the three groups.

Report gains in the three groups are shown in Figure 1. There does appear to be a desired effect that gains in "PG_CHANGED" are higher than in "TUTOR". At the same time, we seem to have an undesired effect that gains in "PG_KEPT" are lower than in "TUTOR". However, the seemingly poor average performance of "PG_KEPT" is due to two statistical outliers. These statistical outliers are students who underperformed the model's prediction by a large margin and the fact that both of them are in "PG_KEPT" may be a mere co-incidence. The difference in gains in the three groups is not statistically significant.

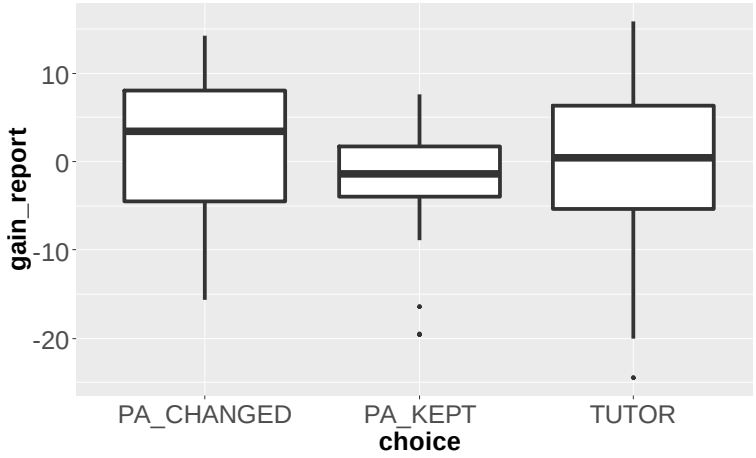

Figure 1: Distribution of report gain depending on self-selection.

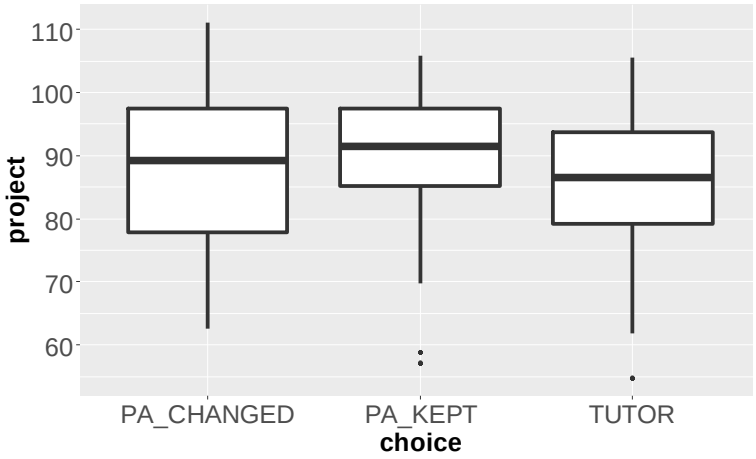

Figure 2: Distribution of final project scores depending on self-selection

Following learning activity

The true indicator of how effective a teaching approach is is its effect on further students' performance. It means that the right way to measure educational value of peer grading is looking not at the report scores that were obtained through peer grading but on students' scores for some follow-up activity. The main objective of peer grading in our course was to prepare students for a team project by giving them prompt and targeted feedback on their writing and programming skills. The score for individual contribution to the team project was the main component of the total course mark.

The comparison in final project scores across the three groups is shown in Figure **2**. Students who had opted for tutor-graded report achieved, on average, lower scores for the final project than students who had participated in peer grading. However, the difference in final project scores is not statistically significant (t-tests and u-tests all have large p-values not even worth mentioning here). Besides, whatever small difference that there is can be explained by the fact that students who had opted for tutor-graded reports are academically slightly weaker than the two groups that had chosen peer grading.

Results and discussion

Participation in peer grading seems to have had a mild positive impact on students who chose to do it and to update their report based on peer reviews. At the same time, participation in peer grading seems to have an even milder negative impact on students who chose to do it and to keep the peer

https://doi.org/10.29311/ndtps.v0i15.3466

score. Finally, participation in peer grading seems to have a very mild positive impact on a follow-up high stakes learning activity. However, all these positive and negative effects are not only small, but statistically insignificant.

On a larger perspective, this suggests that peer grading is similar to teacher grading on a university level. This may be explained by the fact that most students perceive the report as yet another test rather than as a learning opportunity. Such students would probably not learn much from an instructor's feedback anyway – when the test is over, it is too late for feedback. Of course, there are students who take feedback from an instructor seriously. But these are just strong students and they would learn much from peer feedback because they are serious about it too. In either case, peer grading seems to be a cheap way to scale-up a course that would otherwise require a lot of manpower – students' learning will not suffer.

Thus, in order to cut down instructor's workload and to free instructors' resources for more meaningful activities, this paper proves that peer grading can instead be implemented. However, the negligible difference brings into question the effectiveness of this form of collaborative learning: why do the students not learn more with an increase in accountability? The short duration of the peer-grading implemented in our study does not provide any substantial evidence to answer this question. As such, it would be of interest to study whether the impact of peer grading is proportional to the duration and weightage of their accountability.

The potential issue with our approach stems from the fact that a lot of students are not comfortable with peer grading, which is indicated by a small proportion of our students who opted for it. An important direction for further research is therefore developing scaffolding techniques to alleviate students' anxiety about peer grading. But in any case, we strongly believe that students will be more comfortable with peer grading if peer grading it optional.

References

Baker, K.M. (2016) *Peer review as a strategy for improving students' writing process*. Active Learning in Higher Education 17: 179-192. DOI: 10.1177/1469787416654794

Burke Moneypenny, D., Evans, M. & Kraha, A. (2018) *Student Perceptions of and Attitudes toward Peer Review*. American Journal of Distance Education 32: 236-247. DOI:10.1080/08923647.2018.1509425

Crossman, J.M. & Kite, S.L. (2012) *Facilitating improved writing among students through directed peer review*. Active Learning in Higher Education 13: 219-229. DOI: 10.1177/1469787412452980

Dillenbourg, P. (1999) *Collaborative learning: Cognitive and computational approaches.* advances in learning and instruction series: ERIC.

Duzhin, F. & Gustafsson, A. (2018) *Machine Learning-Based App for Self-Evaluation of Teacher-Specific Instructional Style and Tools*. Education Sciences 8: 7. DOI: 10.3390/educsci8010007

Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H. & Wenderoth, M.P. (2014) *Active learning increases student performance in science, engineering, and mathematics*. Proceedings of the National Academy of Sciences 111: 8410-8415. DOI: 10.1073/pnas.1319030111

Ghahari, S. & Sedaghat, M. (2018) *Optimal feedback structure and interactional pattern in formative peer practices: Students' beliefs*. System 74: 9-20. DOI: 10.1016/j.system.2018.02.003

González-Betancor, S.M., Bolívar-Cruz, A. & Verano-Tacoronte, D. (2019) *Self-assessment accuracy in higher education: The influence of gender and performance of university students*. Active Learning in Higher Education 20: 101-114. DOI: 10.1177/1469787417735604

Gopinath, C. (1999) *Alternatives to instructor assessment of class participation*. Journal of

Education for Business 75: 10-14. DOI: 10.1080/08832329909598983

Grov, G., Hamdan, M., Kumar, S., Maarek, M., McGregor, L., Shaikh, T., Wells, J.B. & Zantout, H. (2017) *Transition from Passive Learner to Critical Evaluator through Peer-Testing of Programming Artefacts*. New Directions in the Teaching of Physical Sciences. DOI: 10.29311/ndtps.v0i12.2398

Koh, E., Hong, H. and Tan, JP-L. (2018) *Formatively assessing teamwork in technology-enabled twenty-first century classrooms: exploratory findings of a teamwork awareness programme in Singapore*. Asia Pacific Journal of Education 38: 129-144. DOI: 10.1080/02188791.2018.1423952

Lee, CK-E,, Lim, T-K. & Ng, M. (1997) *Affective outcomes of cooperative learning in social studies*. DOI: 10.1080/02188799708547744

Lefkovitch, L. (1965) *The study of population growth in organisms grouped by stages.* Biometrics: 1-18. DOI: 10.2307/2528348

Luo, H., Robinson, A. & Park, J-Y. (2014) *Peer grading in a MOOC: Reliability, validity, and perceived effects*. Online Learning Journal 18.

McWhaw, K. & Schnackenberg, H. (2003) *From co-operation to collaboration: Helping students become collaborative learners. Cooperative Learning*. Routledge, 79-96.

Mulder, R.A., Pearce, J.M. & Baik, C. (2014) *Peer review in higher education: Student perceptions before and after participation.* Active Learning in Higher Education 15: 157-171. DOI: 10.1177/1469787414527391

Ryan, G.J., Marshall, L.L., Porter, K., et al. (2007) *Peer, professor and self-evaluation of class participation*. Active Learning in Higher Education 8: 49-61. DOI: 10.1177/1469787407074049

Sadler, P.M. & Good, E. (2006) *The impact of self-and peer-grading on student learning*. Educational assessment 11: 1-31. DOI: 10.1207/s15326977ea1101_1

Sharan, S. (2002) *Differentiating methods of cooperative learning in research and practice*. Asia Pacific Journal of Education 22: 106-116. DOI: 10.1080/0218879020220111

Vladislavleva, E.J., Smits, G.F. & Den Hertog, D. (2008) *Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming*. IEEE Transactions on Evolutionary Computation 13: 333-349. DOI: 10.1109/TEVC.2008.926486