Peer-Marking and Peer-Feedback for Coding Exercises

Thomas L. Rodgers*

School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL, UK

Abstract: There are many advantages to the use of peer marking to student learning, but when used with code creation it has the added advantage that it mirrors one of the key activities undertaken with code, the reusing of others’ code. Due to the range of roles that chemical engineers can undertake and the increasing use of modelling, it is becoming more important to include skills like this within degree courses. However, if peer-marking is to be used effectively it is important that the marks generated from it are reliable and the feedback generated from peers is useful.

This paper examines the use of peer marking to allow students to develop their use and knowledge of molecular dynamics. Each student is provided with codes to mark and a rubric which guides their marking. The rubric contains information about the code format and using the code for some calculations.

The marks given by the students are consistent with marks checked by the course leader, especially when the student marks are averaged when code is multiply marked. For the averaged mark the effect size between the student mark and the tutor mark is only 0.12 and the Pearson product correlation coefficient is 0.82. Students also received helpful feedback from their peers.

Keywords: peer-marking, coding, assessment.

*Correspondence to: T. L. Rodgers, School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL. e-mail: tom.rogers@manchester.ac.uk

Introduction

Peer assessment has been a technique used in higher education for many years (Weaver and Cotrell 1986) and has shown to be beneficial to help students in the quality of their work and to develop their own skills of reflection (Somervell 1993). In general peer assessment or peer marking is referred to as a student evaluating a fellow student’s work ideally against a pre-defined standard. This mark scheme can either be defined by an expert, generally the class tutor, or in some cases with agreement from the students, for student led assessment.

Many benefits of peer-assessment to the quality of student learning have been discussed in the literature and summarised by Dochy, Segers, and Sluijsmans (1999). These include, but are not limited to, increased student confidence in the ability to perform, increased student reflections on their own behaviour and/or performance, increased student performance on assessments, increased independence of students in learning, and increased student satisfaction.

Nicol, Thomson, and Breslin (2014) discuss the importance of an active role for learners in constructing meaning from received feedback. They suggest that peer-marking allows students to understand that feedback production is recognised as just as valuable for learning as receiving feedback. Peer-marking can provide students with key additional skills. These skills include the ability to engage with and take ownership of evaluation criteria, to make informed judgements about the
quality of the work of others, to formulate and articulate these judgments in written form and, fundamentally, the ability to evaluate and improve one’s own work based on these processes.

Fry (1990) implemented peer-marking and comment that as students spend time marking the scripts and comparing their efforts with the model solution provided, they reinforce the correct solution. They also get to see what the marks are actually awarded for and they can see the marking scheme. An additional advantage is that students are exposed to other students’ solutions and, in attempting to mark the work, they realise the importance of clearly presented work, and alternative methods to solve the problem.

One additional advantage of peer assessment for use with coding is that one of the key activities undertaken by many people in this area is the re-use of existing codes, either as a whole program where understanding must be checked, or as functions or subroutines in a new program which must be checked and corrected, especially when the code is open source (Haefliger, von Krogh, and Spaeth 2008). This means that students need to be able to write new code and use existing code. As chemical engineers can undertake a wide variety of roles and the increasing use of modelling within engineering, it is becoming more important to include these skills in degree courses.

For peer-assessment to be valuable then the marks produced need to be reliable. Falchikov and Goldfinch (2000) carried out a large meta-analysis of peer-assessment reliability and identified some key variables likely to be influential in terms of the agreement between tutor and student assessment. Some of the key variables related to this study which improve reliability are peer assessments should use global judgements based on well understood criteria rather than many individual dimensions, student should be familiar with the marking criteria, and activities based around academic processes rather than professional practice. They also stated that there was no evidence to support the superiority of multiple peer ratings over ratings by individuals and that peer assessment carried out on advanced level courses were no more valid than those conducted on introductory courses.

As well as the student learning, another potential key benefit to peer assessment, especially with large class sizes is that tutor time for marking can be reduced. Lynch and Schmid (2017) calculated that marking in their study could be reduced by approximately 70%, as the tutor only now needs to review the peer assessment scores and comments to derive the grade. They noted that this time could be reduced through use of more automation if the same evaluation method was used multiple times.

Therefore the key research questions studied during this paper are:

- Is peer-marking reliable and thus suitable for formative assessment
- Are students able to receive the same quality of feedback from peer-marking
- How can peer-marking be implemented
Methodology

Context
Final year undergraduate MEng students in Chemical Engineering and postgraduate MSc students in Advanced Chemical Engineering have the option of taking a module in the second semester on Molecular Modelling for Chemical Engineering. This module covers three key topics Quantum mechanics, statistical mechanics, and molecular dynamics (MD) and how these can be used within chemical engineering. This paper focuses on the MD part of this module. For this part of the module, the assessment is made up of three activities, the creation of and use of a MD code (40%), the marking of the code assignment (10%), and an online test which examines MD theory and the use of the code (50%).

A total of 85 students (32 in 2017, 53 in 2018) undertook the course and thus the marking exercise. Six students did not complete the marking task, (2 in 2017 and 4 in 2018) and one student (in 2018) did not complete the code or the marking. The coding exercise is undertaken in Matlab as although this is an unusual choice for scientific programming of this type due to its speed, it has been selected for this course as it is a very clear language, and it has been used in previous modules by the students, as the intent of this course is not to teach a particular programming language. The three intended learning outcomes (ILOs) for the MD part of the course are:

- Assess force fields, thermostats, and numerical integration algorithms to select optimal choices for molecular dynamics simulations.
- Design optimised, modular code which can be combined with existing functions to produce a molecular dynamics code using Matlab.
- Appraise molecular dynamics code and apply it to problems of chemical engineering interest.

Students are provided with a course handbook, which they follow during the MD sessions. These are undertaken in a self-learn style with support from the tutor and two graduate teaching assistances. All the sessions take place in a computer room. The handbook contains a number of focused activities with build to produce the MD code (See supporting information Figure 1 for an example function). Students are also provided with a tips sheet which provides information on Modularization, Design, Readability, Comments, and Function, which are key aspects included in the marking scheme. There were also some short questions to answer using the code.

After submission of the code assessment students were allocated a randomly selected anonymised script of a peer; one in 2017 and two in 2018. This allocation was preformed automatically using the peer assessment tools on blackboard, the Virtual Learning Environment used by the students on the course. They were asked to mark the code using the same marking rubric as the tutor (See supporting information Figure 2) and to provide 4 or 5 bullet points of additional written feedback on potential improvements to the code. The students don’t receive any training before marking; however, the marking is supported by a drop-in session within a computer cluster, so students can ask for support if needed.

Marking Reliability
To examine the reliability between student and tutor marking of the in-course assessment several statistical tests are undertaken. The similarity in distribution of the tutor and student marks are
checked with Student’s t-distribution (this however doesn’t check the similarity of the mark for each assessment), the similarity between the tutor and student marks is checked with the Pearson product-moment correlation which measures the strength of a linear association between the two variables and the Spearman’s rank correlation which measures how well the relationship between the two variables can be described using a monotonic function.

The reliability is also checked using a Bland-Altman style plot. This plots the differences between the pairs of marks against, in this case, the tutor mark and tells us whether there is any systematic bias (a tendency for the student marking to be higher or lower than the tutor marking). This analysis relies on the distribution of the difference in marks being normal. To test if the mark difference is normally distributed the Lilliefors test is used as the mean and variance are estimated from the sample. The effect size, $d$, is also calculated for the two distributions as in equation (1) (Cooper 1998).

$$d = \frac{\text{Mean}_{\text{student}} - \text{Mean}_{\text{tutor}}}{\left(\frac{\text{Standard Deviation}_{\text{student}} + \text{Standard Deviation}_{\text{tutor}}}{2}\right)}$$

For clarity all statistical probabilities, $p$, are given as a probability of the distributions being the same based on a significance level of 0.05.

**Student Opinion**

After the online test part of the assessment was completed (i.e. all the assessment tasks completed for the MD part of the course) the students were invited to complete an online survey about the course. This survey contained several questions related to aspects of the MD part of the course, but the ones relevant to this study, and thus examined here are, “The feedback that I received on my work was helpful” (which referred to feedback from the tutor), “The learning outcomes were achieved”, and “The tutorials were helpful in understanding the module content”\(^1\). All responses were recorded on a Likert scale from 1 to 5, with 1 being “strongly disagree” and 5 being “strongly agree”. For data from Likert scales normality is not checked as the student t-test provides good statistical representation regardless of distribution and number of samples (Norman 2010, Sullivan and Artino 2013), thus is used here if needed. In 2017 this received 15 responses and in 2018 received 20 responses.

Via a separate survey students were also asked “The feedback that I received on my work was helpful” relating to the feedback given from other students, again on a Likert scale from 1 to 5. For 2018 they rated each marker separately. In 2017 this received 14 responses and in 2018 received 18 responses (i.e. 36 ratings).

**Results**

**Reliability of peer-marking**

Figure 1 shows the marks given by the tutor and students for each of the 78 codes received (29 for 2017 and 49 for 2018) and includes a line of equality (i.e. agreement between marks). Students were

---

\(^1\) The tutorials were where the coding activity and peer-marking was undertaken.
slightly more likely to be the generous markers. Students marks were on average 1.034 times tutor marks (95% confidence interval 0.995 to 1.074) in 2017 and 1.014 times tutor marks (95% confidence interval 0.979 to 1.044) in 2018. However, there is no statistic difference between the student marks as the tutor marks with a t-test p value for similarity between the student and tutor mark distributions are 0.57 for 2017 and 0.54 for 2018 with the Pearson product correlation coefficient of 0.66 (p = 0.99) for 2017 and 0.82 (p = 1) for 2018.

There is one clear outlier from the 2017 marks received from the student markers (37.8% from the student versus 66.7% from the tutor). This was investigated and it was clear from the feedback provided by the student marker that they couldn’t get the code provided to run and some files were missing. However, this was not the case during the tutor marking as the code ran and all files were present. It is not clear if this problem was caused by marker error or a software error in delivery of the files to the marker. It is therefore important that issues like this are checked for; however, the mark difference is clearly visible and the individual student marker feedback rating received for this marker was one of the ones (see Figure 3(a)), so would trigger a checking from the tutor anyway.

In 2018, as mentioned, each code was marked by two students, whereas in 2017 each code was only marked by one student. The average mark of these two marks was then attributed as the student mark (in Figure 1). Table 1 shows the comparison of the average difference and the standard deviation of the difference between the student and tutor mark. The most important value for the reliability is the standard deviation. When each code is marked by one student the standard deviation is larger than when the two student marks are averaged. This suggests that not only is the reliability of the mark increased when two student marks are averaged, but students also get the opportunity to examine two assessments rather than one.

To test the difference between the tutor marks and the student marks a Bland-Altman plot can be used. This relies on the distribution of the difference in marks being normal. To test if the mark difference is normally distributed the Lilliefors test is used as the mean and variance are estimated from the sample, at a significance level of 0.05 with 49 values the critical value is 0.126, the max deviation based on grouping into ranges of 2 marks is 0.066, which gives a p value of 1 for the distribution being normal. Figure 2(a) shows the Bland-Altman style plot for the difference between the student and tutor mark plotted against the mark received by the student marker (note each student marks two reports which are plotted as individual data points). This allows us to see that the distribution of the mark difference is consistent against the received mark (this can be thought of as the student’s understanding of the assessment). The dotted lines are the upper and lower limits, which are set at 1.96 times the standard deviation of the mark difference. There is potentially an indication that students who receive a lower mark (< 50%) also gave a mark lower than the tutor; however, this is not statistically significant based on the number of results.

Figure 2(b) shows the Bland-Altman style plot for the averaged 2018 students marks compared to the tutor marks. In this case it can be seen that by averaging the wo student marks the range of differences are reduced, reflected by the standard deviations in Table 1, thus the reliability of the

---

2 Max deviation is 0.071 based on ranges of 1 mark, and 0.074 based on ranges of 5 marks, both have a p value of 1.
3 1.96 times the standard deviation is the limit for a significance level of 0.05.
marks are increased. It is also clear that the distribution of mark differences is consistent across the whole range of marks, there is no significant bias at high or low marks.

**Student Opinion**

Data on the student opinion of the quality of the feedback produced was measured using a Likert scale from 1 to 5, with a score of 5 being the best. This was undertaken for both 2017 and 2018 with no statistical difference between the two different years. Therefore the combined results for the two years are given in Figure 3. The average score for the student generated feedback was 3.9 with a standard deviation of 0.96 and that for the tutor provided feedback was 4.5 with a standard deviation of 0.69. The tutor provided feedback scored statistically higher than the student generated feedback (student t-test p value for similarity between the 2 distributions is 0.008). Although students were given some guidance on the feedback to provide it is however not surprising that the score was lower as they haven’t carried out such an activity before. It is positive that 80% of the student agreed that the student generated feedback was helpful (either a 4 or 5). It has previously been reported that peer marking, especially the feedback, requires skills in summarising and evaluation which students do not always have (Williams 1992). It should also be noted here that none of the students in 2018 (with 2 sets of feedback) rated both sets a low score, so they always felt they received some helpful feedback – even if it was just one set from their peers.

As well as the quality of the student feedback produced, it is also useful to know if the students found the activity useful. Figure 4(a) presents the student response to the statement “The learning outcomes were achieved” (the same as listed in the Methodology section of this paper) from a Likert scale from 1 “strongly disagree” to 5 “strongly agree”. The average score given was 4.2 with a standard deviation of 1.06. Figure 4(b) presents the student response to the statement “The tutorials were helpful in understanding the module content”, using the same Likert scale. The average score given was 4.2 with a standard deviation of 1.11.

This data provided evidence that the students value the assessment and peer-marking activity as they feel that it achieves the learning outcomes for the course and helps them to understand the module content.

**Implementation**

In this study the peer-marking was implemented alongside marking by the tutor. This means that students receive both the feedback of the tutor and their peers. The success of this implementation was likely due to the fact that the marking was randomly allocated and all the assessments and marks were kept anonymous (but not to the tutor). The marking drop in session also allowed students with any problems to seek help.

An alternative to this would be to provide the peer-marking of the assessment as formative feedback. The students can then look at the comments received and improve their assessment before a final submission which is only marked by the tutor.

This is not; however, the only potential implementation. Due to the check on the marks, with double marking, this means that the amount of marking performed by the tutor could be reduced if they only mark the assessments with a difference between the peer marks, or if a student complains about a mark. In the 2018 cohort this would mean that 10 out of the 49 reports would have to checked by the tutor, as they were over 10% different in marks between the student markers, but
there were no complaints about the mark received. This would be approximately an 80% reduction in the marking time by the tutor. As the allocation of marking was performed automatically using the peer assessment system on Blackboard and the assessment and mark submission is collection via Blackboard, this takes no longer than it would to set up if the tutor was marking the assessments. In this instance it would be important to make sure that students were undertaking the activity properly and providing sensible marks.

Conclusions
One of the key advantages of peer marking for coding is that students get experience of looking at existing code (and correcting it) as well as creating their own. As students have just tried to write the same code they are familiar with what is needed and get to see alternative options. This means that in this case the peer marking is very successful with a good similarity between the marks given by the students to each other and the marks given by the tutor. Having each assessment marked by 2 other students and averaged improved the reliability to the tutor mark over a single student mark. The feedback provided by the peer marking exercise also seemed to be valued by the students, although not quite as much as the feedback from the tutor.

In this study there is a clear advantage of two students markers per assessment compared to one. This is firstly in the improvement of reliability in the average mark, secondly that if there is a large difference in the peer markers marks then this assessment can be flagged to be checked by the tutor, and thirdly that students always received one set of peer feedback that they score highly.

References
Table 1. Similarity metrics between the tutor mark, $T$, and the student mark, $S$.

<table>
<thead>
<tr>
<th></th>
<th>2017</th>
<th>2018</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual</td>
<td>Average</td>
<td></td>
</tr>
<tr>
<td>Number of Marks</td>
<td>29</td>
<td>98</td>
<td>49</td>
</tr>
<tr>
<td>Average Difference ($T - S$)</td>
<td>-1.60</td>
<td>-1.28</td>
<td>-1.28</td>
</tr>
<tr>
<td>Standard Error (Average)</td>
<td>1.31</td>
<td>1.07</td>
<td>1.13</td>
</tr>
<tr>
<td>Standard Deviation of Difference</td>
<td>9.01</td>
<td>10.03</td>
<td>7.65</td>
</tr>
<tr>
<td>Effect size, $d$</td>
<td>0.15</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Standard Error (Effect size)</td>
<td>0.26</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>Student's $t$-distribution</td>
<td>0.57 ($p = 0.57$)</td>
<td>0.66 ($p = 0.51$)</td>
<td>0.61 ($p = 0.54$)</td>
</tr>
<tr>
<td>Pearson product correlation</td>
<td>0.66 ($p = 0.99$)</td>
<td>0.71 ($p = 1.00$)</td>
<td>0.82 ($p = 1.00$)</td>
</tr>
<tr>
<td>Spearman's rank correlation</td>
<td>0.69 ($p = 1.00$)</td>
<td>0.65 ($p = 1.00$)</td>
<td>0.79 ($p = 1.00$)</td>
</tr>
</tbody>
</table>
Figure 1. Parity plot of tutor and student marks [for 2017, n = 29 (1 student enrolled on the course did not complete the coursework and 2 students did not complete the marking exercise); 2018, n = 49 (4 students enrolled on the course did not complete the coursework)]. The dashed lines represent tutor mark ±10. $p$ value for similarity between the student and tutor mark distributions is 0.57 for 2017 and 0.54 for 2018. Pearson product correlation coefficient 0.66 for 2017 and 0.82 for 2018.
Figure 2. Bland-Altman style plots of (a) tutor mark minus individual student mark against the students own assessment mark, and (b) tutor mark minus average student mark against the tutor mark, for 2018 [n = 49 (4 students enrolled on the course did not complete the coursework)]. In both cases the horizontal line is the average difference, -1.28, and the horizontal dashed lines are 1.96 standard deviations from the mean, see Table 1. The error bars on the horizontal lines are one standard error of the calculation.
Figure 3. Students responses to the question “The feedback that I received on my work was helpful” on a Likert scale of 1 “strongly disagree” to 5 “strongly agree” for (a) the feedback received from student markers, average 3.9 standard deviation 0.96 (2017 n = 14, 2018 n = 18) and (b) the feedback received from the tutor, average 4.5 standard deviation 0.69 (2017 n = 15, 2018 n = 20). $p$ value for similarity between the 2 distributions is 0.008.
Figure 4. Students responses (2017 n = 15, 2018 n = 20) on a Likert scale of 1 “strongly disagree” to 5 “strongly agree” to the questions (a) "The learning outcomes were achieved", average 4.2 standard deviation 1.06 and (b) "The tutorials were helpful in understanding the module content", average 4.2 standard deviation 1.11.