Peer Evaluation: Enhancing learning Opportunities and Reducing Marking Effort

Denard Lynch, Bradley Schmid
University of Saskatchewan
denard.lynch@usask.ca, brad.schmid@usask.ca

Abstract: Evaluation of report-based assignments, especially in larger classes, adds a considerable marking load. Even with detailed rubrics, subjectivity may lead to grading variations and inaccuracies. Evaluation of others’ work can also be a very informative and educational experience, improving their skill through exposure to a broader performance range. Involving students in peer evaluation can potentially address both of these issues by reducing marking load, providing alternate (and increased number of) assessments, and by exposing students to a broader spectrum of report skills thus enhancing their own knowledge.

This paper discusses the results of an experiment in peer assessment and whether it can be exploited to reduce marking effort, improve accuracy for report assignment evaluation and improve student skill. The data was gathered from assignments in two different engineering classes: a second year course on safety and environmental stewardship, and a senior course on engineering economics.

For the second-year course, an individual essay assignment was marked by the instructor and two peers. The three evaluations were analyzed to assess the accuracy and assign a grade. For the senior course, a group report on a case study was self and peer evaluated. These evaluations were used to derive a grade for the report directly if the self and peer results were within a prescribed tolerance; other cases were resolved by instructor intervention. The results were analyzed considering the number of outliers, range of scores, and the number of cases which had to be resolved by the instructor.

Parameters considered in assessing the results of the experiment included: the correlation between assessments, the learning opportunities for students, and instructor marking effort required. (preliminary analysis) Results suggest positive gains in reducing effort. Improved accuracy and enhanced student learning are also expected.

Keywords: Student assessment, Assessment tools, summative, grading systems

1. INTRODUCTION

A significant observation often made by instructors is that there is a learning benefit from reviewing the work of students. Providing the opportunity for students to improve their understanding through peer reviews of other’s work may have several pedagogical advantages. The main benefit is described by Lundquist et al in one of their conclusions: “…letting the students apply their knowledge practically and revising this knowledge via a peer-review process gives a deeper understanding of the theory.” [1] [2]. Adding to the incentive to include peer reviews in an engineering educational setting is the reality that, as graduated professionals, engineers are often required to review the work of peers, either as members of a team or as independent reviewers to improve quality [2] [3]. On the resource side, there is the issue of the significant marking resources required to provide grading and feedback on assignments, particularly those involving soft skills like communication and life long learning.

The opportunity and the need for the authors was available in two courses involving relatively large class size and soft skills content. This paper discussed the experiences implementing peer review in two different engineering courses as well as the benefits for the students and the instructors in doing so.

2. THE RESEARCH QUESTION

The easiest question to investigate, from a pedagogical perspective, relates to the potential reduction in marking effort. While the initial implementation requires additional effort, eventual resource savings can be estimated with reasonable certainty and accuracy.

The main condition for reducing marking effort is maintaining marking accuracy and student feedback. While a strictly objective measure may be illusive, a qualified subjective measure will be sought. In the
implementation phase, a comparison with traditional processes can be included (at the expense of additional effort).

The learning benefits for students will also be examined, but primarily based on inference from other work, such as that by Phillips [2] or Topping [5] and anecdotal evidence.

Finally, there is an opportunity to contrast results between junior and senior students.

3. EXPERIMENTAL DESIGN

The setup for the data generation is described in this section. One of the advantages of this experiment is the opportunity to compare the results from both a junior and senior course, and to compare the similarities and differences.

3.1 Junior Course Experimental Setup

The junior course on “Safety and Environmental Stewardship in Electrical and Computer Engineering” is offered in the second year of the program, with an enrollment of 78 students. The assignment selected for this experiment was an individual term assignment on the topic “Technology in Society”, requiring a report of approximately three typewritten pages. The assignment instructions included notice of a voluntary, subsequent peer review exercise and required an online submission of the paper with no student identifying information included. Following submission, the instructor randomly assigned each student two reports to evaluate. The basis for evaluation was the original rubric provided for the original assignment. The voluntary “Peer Review Assignment”, including the rubric, is shown in Appendix 1. Students were asked to grade the reports and to provide one complimentary and one critically constructive comment on the report. Each report was therefore assessed by two peers and the instructor. The final grade for the assignment was determined by the instructor with consideration given to the peer evaluations. The requested comments provided by student reviewers were moderated by the instructor and returned to the authors, anonymously, for additional feedback. All assignments and feedback in this process were masked (blind).

3.2 Senior Course Experimental Setup

Engineering Economics is a mandatory senior course for all programs, and typically has an enrollment of approximately 150 students per semester. An experimental peer-assessment scheme was conducted on a group term assignment of an economic case study. The assignment required a report approximately 15 pages in length plus appendices to be completed in teams of 4. Each report was self-assessed by each individual team member according to the rubric supplied with the case study prior to submission of the report.

Each student in the course was randomly assigned two papers to review by the course instructor. The instructor ensured that no one received their own report and that only a numeric identifier indicated which report they had received so that the students could not easily determine who had written the report. Assigning two reports to each review resulted in 8 student reviewers per report. This would amount to 300 peer reviews for a class of 150 students.

The student reviewers used the same rubric on each assigned report as they had used for the self-assessment and the original report assignment. Additionally, students were asked to provide anonymous comments about the reports they were reviewing. These comments served as formative feedback to the authors of the reports. The quality of these student peer review comments was worth 5% of the project grade as was assessed by each team collectively on a simple Likert scale of 1 to 5 using the question “On a scale of 1 to 5, with 5 being the most useful, how useful were the comments provided by the reviewer?”

The instructor reviewed the range of peer assessment scores, verbatim comments, and the team reports, to derive a grade. Typically, this would involve simply removing outlier peer evaluations from the calculation of the average score. If the peer assessed grades spanned more than 20% or the standard deviation was larger than 10%, the grades were reviewed by the instructor and outlier grades were removed from the calculated average.

If the self and peer assessments agreed within 5%, the higher of the two was assigned as the final grade for the report. As an incentive to the students to be as objective as possible in their individual self-assessment of their report, bonus grades (2%) were awarded where the self-assessment grade was within 5% of the peer evaluated grade.

4. ANALYSIS & RESULTS

This section describes and discusses the analysis of the data collected and the results from the two courses individually. It will also provide some preliminary comments regarding the comparison of the two courses, although final comments will be included in the Conclusions section.

4.1 Junior Course Analysis

Enrollment in this course was 78 students. From this, 76 assignment reports were received at the time the peer exercise was initiated and assigned to each of two
separate peer reviewers (assignments were random, except that no student was assigned their own paper to review). Of these, 60 received two peer reviews, plus the instructor’s evaluation, and an additional 14 received one peer review plus the instructor’s evaluation.

Table 1 summarizes the basic statistics for the evaluations. The “Peer” column includes all peer reviews (134 in total), and 77 completed by the instructor (“Instructor” column). The “Overall” column gives the basic statistics for all evaluations performed on the subject papers. Although peers were slightly more generous on average (approximately 5%), there is a reasonable visual correlation between peer and instructor ratings.

Table 1: Peer Review Statistics - Junior Course

<table>
<thead>
<tr>
<th></th>
<th>Peer</th>
<th>Instructor</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>80.1%</td>
<td>74.9%</td>
<td>78.2%</td>
</tr>
<tr>
<td>St. Dev.</td>
<td>11.8%</td>
<td>10.1%</td>
<td>8.7%</td>
</tr>
</tbody>
</table>

Determining an acceptable range that can be considered “accurate” for grading purposes followed the author’s previous work on “Speed Marking” [4]. In this case, a difference within two standard deviations will be considered reliable, and an acceptable difference with “soft” assignments.

Examining the difference between different sources of marks also assess the reliability or acceptability of the results. Table 2 shows a summary of comparisons between the first and second peer marks, the first peer mark to the instructor’s mark, the second peer mark to the instructor’s, and finally the peer average to the instructor’s mark. (Note: these are all magnitudes, and out of 100 marks (%)). Finally, the Pearson correlation coefficient, r, between the two data sets is shown in the final column.

Table 2: Evaluation Difference Summary

<table>
<thead>
<tr>
<th></th>
<th>Avg</th>
<th>StDev</th>
<th>Min</th>
<th>Max</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1-P2</td>
<td>10.6</td>
<td>9.6</td>
<td>0</td>
<td>58</td>
<td>0.31</td>
</tr>
<tr>
<td>P1-Inst</td>
<td>10.2</td>
<td>7.8</td>
<td>0</td>
<td>36</td>
<td>0.51</td>
</tr>
<tr>
<td>P2-Inst</td>
<td>10.0</td>
<td>8.2</td>
<td>0</td>
<td>33</td>
<td>0.44</td>
</tr>
<tr>
<td>PAvg-Inst</td>
<td>8.7</td>
<td>6.3</td>
<td>1</td>
<td>28</td>
<td>0.59</td>
</tr>
</tbody>
</table>

It is clear that there is more variation in marks between the two peer evaluations than between either peer evaluation and the instructor evaluation, and the best correlation is between the average of the peer marks and the instructor mark. Figure 1 shows the difference between the adjusted peer average mark and the instructor mark.. Using the two standard deviation guideline for acceptability (±~9%), 61.3%, about two thirds, fall within this range.

Finally, moderation of requested feedback was trivial, and in the author’s opinion, generally provided insightful and valuable feedback.

4.2 Senior course Analysis

Parameters considered in assessing the results of the experiment included: the correlation between assessments, the learning opportunities for students, and instructor marking effort required.

For the semester course being analyzed, there were a total of 273 valid peer reviews. Some reviews were incomplete or not submitted. Of these 273 reviews, a total of 11 outliers were removed based on the range of grades submitted for each report (>20%) or the standard deviation of these grades (>10%). The individual self-assessment results were compared to the peer assessment results after outliers had been removed. The difference between an individual student’s self-assessment and the average of their peer assessment is shown in Figure 2.

For 74% of the students, the difference between their own self-assessment and the average assessment of their peers differed by less than 10%. This indicates that for the large majority of students, the grade they expect to receive on the assignment is within a letter grade of the actual grade they receive.

In further comparing the self-assessment to peer assessment grades it was found that 71% of students graded themselves higher than their peers graded them, but only 21% of students submitted self-assessments that were more than 10% higher than their average peer assessment grade.
While Figure 2 suggests there could be a reasonable correlation between the self-assessment and the peer assessment, a closer look at the data is presented by cross plotting the self-assessment and average peer assessment values (Figure 3). This results in a low correlation coefficient (Pearson Correlation R=0.02) indicating very little correlation between the self-assessment values and the average peer assessment values.

Comparing the average of the individual self-assessments for each team improved the correlation somewhat (Figure 4). Again, there’s appears to be a slight, but not significant, trend for reports assessed lower by peers to be overestimated by the authors, even as a group.

If these higher marks could be explained by the Dunning-Kruger effect [7], where less-skilled students tend to overestimate their capability, one would expect to see a tendency for plotted pairs to cluster to the lower-right of the r=1 correlation line in Figure 3. However, the scatter plots shows only a slight indication that peer-assessed lower grades were “overestimated” in the self-assessments.

A total of 35 reports were submitted and analyzed for the course which in previous years would mean approximately 50 to 60 hours of instructor marking effort and provide only a limited amount of feedback. The total amount of time required to process the self-assessments, peer assessments, and evaluate the student’s reports grades amounted to approximately 16 hours. This time could be reduced through use of more computer software automation if the same evaluation method was used multiple times.

Although each report team had the option of requesting an instructor review of their report because they felt the grade given was not accurate, none of the 35 teams requested an instructor review. Had requests been received, this would have amounted to individual paper reviews that would have been expected to take 1 to 1.5 hours per report to complete.

Feedback from students regarding the peer evaluation was mostly positive. Most frequently the students commented that they enjoyed seeing how other groups of students approached the problem. Additionally many commented that the reports were hard to follow and that they could have been written much more clearly and
5. CONCLUSIONS

The objectives of this work were to: reduce marking effort while maintaining marking accuracy, provide formative feedback and learning by peer evaluators, and contrast the results between junior and senior course.

Based on instructor effort, marking effort could be reduced by approximately 70% (senior class; 16hrs/55hrs) by appropriate implementation of student peer reviews. An estimate of saving in the junior course would be assumed comparable, although for this implementation, the instructor marked all reports.

The accuracy (reliability) of the final marks can be inferred from the number of requests for instructor reviews (none in these trials) and the conclusions from previous work [2] [3], which also suggest the advantages for improving student skill and knowledge. Also, even with relatively low correlations between self-peer or instructor-peer assessments, the result was generally acceptable by students and useful by the instructors.

Anecdotal review of formative comments from peer reviewers suggests that students were serious about the task and made honest and insightful suggestions in both junior and senior sources. In addition, participation rates suggest, but do not prove, that voluntariness was not a significant factor, further supporting the observation that the process was considered useful by students.

There was no direct comparison available between junior and senior courses on self versus peer assessment, but comparing instructor-versus-peer (junior course) and self-versus-peer (senior course) statistics, peer assessment was higher than instructor assessment (junior course) and lower than self assessment (senior course).

6. FUTURE WORK

Possibilities for future work to realize resource savings and improve student benefits related to the original questions being investigated.

Investigation of computer-based automation for the process may lead to improved efficiency and the ability to provide timely and frequent formative feedback from peers. Training and calibration of peer assessors, similar to that reported by Phillips [2] could enhance the accuracy of grades as well as improve student skills.

To further improve confidence in enhanced learning opportunities for students and improve efficiencies for instructors, an integrated approach at developing and integrating peer reviews as part of the course strategy could be analyzed.

7. ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the University of Saskatchewan for providing the means and opportunity for its instructors to explore and enhance their pedagogical skill and knowledge and contribute to the body of knowledge in this area. The authors would also like to thank Dr. Sean Maw and the other members of the Innovative Teaching and Research in Engineering Education Group at the University of Saskatchewan for their support and suggestions. Finally, the author thanks the students in the subject courses for their patience and cooperation related to this research.

8. WORKS CITED

9. APPENDIX A1: ASSIGNMENT RUBRIC (JUNIOR COURSE)

Assignment: Review the two assigned “Technology in Society” reports using the rubric below as a guide (original rubric for that assignment).

Due Date: Nov 14, 2016, or thereabouts. I expect it should only take a half hour or so, try to fit it in as soon as you can.

Deliverables:
An evaluation for each report consisting of:
1. A grade out of 100 points. There are 4 rubric categories. It would be helpful, if you can, to break the mark down for each. E.g.: “7/11/47/15 = 80/100” would mean 7/10 for “Selection of Topic”; 11/20 for “Research” etc.
2. Also, make the following two comments on each of the reports:
   i. One thing I liked about this report was......
   ii. One thing I felt would help improve the report was...

You can show the result and your comments right in the assignment submission area by typing it in as a text submission in the appropriate area. You are expected to do a serious, professional evaluation and make meaningful, constructive comments. While only I will see your input, I may share appropriately made comments with the authors for their benefit. Your evaluation may also be considered in the instructor evaluation and final grade for the original assignment.

Benefits to the student:
1. Peer-assessment provides the student with the opportunity to critically examine other’s work, reflect on their own performance and use the opportunity to improve. When students enter the work force they will be required to have their work reviewed by peers and review the work of others.
2. Providing ‘reverse’ peer assessment provides feedback to the reviewer about the quality of the review provided.

Report Rubric (from original assignment):
The following guidelines will be used in grading this assignment:
1. Selection of topic
   o 10 Clear justification of relevant topic selection related to assignment objectives
   o 5 Some suggestion that relevance was considered in topic selection
   o 0 No indication that suitability of topic was considered

2. Research (related to number and quality of resources)
   o 20 A number (>=4) of relevant sources were used to research topic and were clearly evident in report
   o 10 2 – 3 relevant sources were used to research topic and were evident in report
   o 0 Report did not utilize relevant sources to any noticeable degree

3. Analysis and interpretation
   o 50 Clear evidence of analysis or interpretation related to the interaction with, or effect of, technology on society considering specifically the focus of this assignment.
   o 30 Some instances of evidence of analysis or interpretation
   o 0 Very limited or no evidence of analysis or interpretation

4. Formatting, grammar and spelling (including proper citations)
   o 20 Well formatted/organized to accommodate reader; only very minor or no grammatical or spelling errors; very readable.
   o 10 Some grammatical or spelling errors, but did not substantially detract from readability
   o 0 Noticeable formatting, grammatical or spelling errors that significantly affected readability

10. APPENDIX A2: ASSIGNMENT RUBRIC (SENIOR COURSE)

1. Writing Style, Clarity, Spelling, Grammar, Proper use of technical terms, Readability, Font size, etc
   10: Professional, well written, third person, concise wording, no spelling or grammar mistakes, excellent use of technical terms, references and citations used appropriately
   8: Generally well written, some explanations too long or too short, very few (1-2) spelling or grammar mistakes, good use of technical terms, included references
   6: Acceptable writing, some explanations vague, few spelling and grammar mistakes, reasonable use of technical terms, missing references
   4: Writing is weak and hard to understand, many spelling and grammar mistakes, poor use of technical terms, missing references, difficult to read
   2: Very poor writing, explanations are ambiguous or absent, many spelling and grammar mistakes, poor use of technical terms, difficult to read, poor font size or formatting
   0: Extremely poor, unacceptable

2. Report Organization
   10: Includes technical report sections: Summary, table of contents, list of figures, list of tables, Introduction,
8: Most assumptions clearly stated.
6: Some assumptions stated, lacks some clarity
4: Few assumptions stated and/or lacking clarity
2: Assumptions not stated and/or a considerable lack of clarity
0: Very poor or unacceptable assumptions made.

6. Alternative Solutions
15: Multiple options were considered and appropriately explained
12: Some alternatives developed but lacked clarity in their definition
9: Alternatives were identified but were incomplete
6: Only one alternative considered, but clearly explained
3: Poor assessment of alternatives
0: Solution not explained

7. Evaluation of Alternatives
20: Evaluation included relevant economic analysis. Methodology is clearly explained. Use of NPW, sensitivity, and other relevant analyses. Identified alternatives considered.
16: Evaluation included relevant economic analyses. Methodology explained. Use of NPW analyses and sensitivity analyses. Identified alternatives considered.
12: Evaluation included relevant economic analyses. Methodology not well explained. Use of NPW analyses only. Not all alternatives evaluated.
8: Poor evaluation. Did not considered the time value of money, methodology poorly explained, did not complete a NPW analysis. Only one alternative analyzed.
4: Very poor evaluation. Economic concepts not considered.
0: Evaluation method, considerations, results are not understandable, are missing, or do not include the consideration of any economic principles.

8. Project Cost Estimates
15: Estimates of costs well justified.
12: Method of estimating cost are reasonable.
9: Estimates of cost documented but unclear.
3: Estimates of cost not documented and appear to be of poor quality.
0: Estimates of cost violate methods taught in class.

background, methodology, discussion, conclusions and recommendations
8: Generally complete, minor items missing or additional unnecessary items included
6: Report requires modification to meet the expectation of a professional report
4: Appropriate sections not included, the document lacks a logical structure, items missing
2: Report is very poorly organized, many sections missing. Report requires major modification.
0: Report lacks any organization, many sections missing, incomplete. Extremely poor. Unacceptable.

3. Use of Tables, Figures, and Appendices
10: Tables and figures used appropriately to assist in the understanding of the report. Appendices include additional information referenced in the report and spreadsheet calculations used in the report. Calculations well documented to be understood.
8: Includes relevant information and calculations in the appendix. May be missing minor items or contains some additional information that is not required. Calculations documented to be understood.
6: Includes most relevant information. Missing some calculations or reference information. Calculations not well documented.
4: Contains too much or too little information to support the main report. Calculations not documented.
2: Missing major items from the appendices. Information provided is not understandable or related to the report.
0: No appendices. No calculations. Profuse amounts of unrelated information provided.

4. Use of Information
10: Used a variety of quality information. Information sourced from government publications, journal publications, and other reputable sources.
8: Used a variety of information. Information obtained from some quality publications. Significant portion of the information used was from less credible sources (eg. News articles, blogs, non-government websites)
6: Used a variety of information. Some quality publications and sources used. Many biased or unreliable sources used.
4: Used few sources of information. Used mostly low quality internet resources.
2: Used very few sources of information. Used exclusively low quality internet resources. No government or technical resources used.
0: Used no sources of information or used only one sources of poor quality information.

5. Documentation of Assumption
10: All assumptions clearly stated.