POWER OF PEERS: EXPERIENCES USING AN ONLINE PEER ASSESSMENT TOOL TO GRADE STUDENT WORK

Peter M. Ostafichuk and Carol P. Jaeger
University of British Columbia
ostafichuk@mech.ubc.ca

Abstract This paper explores the implementation, outcomes, and student perceptions of the use of an online tool for anonymous peer assessment of student work. Peer assessment, where one student assesses the work of another, provides an opportunity for important skill development, as well as a fully-scalable strategy for rich, timely, and frequent feedback.

In first and third year engineering courses at the University of British Columbia, we have begun using an online peer assessment tool (peerScholar). The tool divides the peer assessment process into three phases: a creation phase where the work is written or uploaded, an assessment phase where students are randomly assigned to assess the work of a set number of their peers, and a review phase where students review the feedback they received, with options to revise their work or assess the quality of feedback received. We have successfully used this tool in two large (n = 750) classes and one moderate-sized (n = 130) class, with a wide range of different types of student work, including letters, technical memoranda, detailed design reports, and video presentations.

Through surveys, student feedback with the tool and the process has been positive. Students at both year levels overwhelmingly recognize the importance of peer assessment—over 90% identified it as an essential skill for an engineer, and over 85% felt opportunities for peer assessment should be embedded in the curriculum. Both groups indicate that they felt the process of reviewing others’ work was beneficial for their own understanding of the material; however, first year students were more likely than third year students to put more effort into their work knowing it would be peer assessed, and that they found the content of the feedback received more helpful to their learning. Student acceptance has been good.

In a third year mechanical design course, three different design assignments were independently assessed by students using peerScholar and by teaching assistants. The outcomes across all measures were encouraging: for each assignment, the students and teaching assistants had similar mean, standard deviation, minimum, and maximum values, as well as reasonable correlation (r = 0.5 overall).

Overall, we consider the adoption of peerScholar a success. Students have been receptive, challenges have been minor, and feedback is more detailed and frequent.

Keywords: peer assessment, online assessment, grading practices

1. INTRODUCTION

Regular, timely feedback is frequently identified as an essential element of effective teaching practice [2],[13]; however, limits on instructor or marker time and availability, and pressures of increasing class sizes, make providing frequent, timely, high-quality feedback difficult. Peer assessment of other students’ work (not to be confused with peer evaluation of behaviours in a group) provides promise to maintain high frequency, rich feedback that is fully scalable with any class size. Keith Topping [16],[18] notes that even though they have less skill at assessment, students can produce equally valid and reliable results to a teacher by virtue of the increased time they have for the task. In addition, the greater volume, frequency, and immediacy of feedback can compensate for quality disadvantages compared to assessment by a teacher or teaching assistant. Even in cases where student and expert assessments are found to differ, it is argued that the benefits of peer assessment outweigh some discrepancy in marks [6],[8],[16].

The topic of reliability and validity of peer assessment has been examined through numerous studies. Falchikov and Goldfinch [7] conducted a meta-analysis of 56 cases comparing peer and teacher evaluations and found good correlation between the two (r = 0.69). In another analysis of 25 studies by Topping [17], 72% reported “acceptably high reliability” and no statistical relationship between student acceptance and reliability was found. Finally, in a three-year study on the use of a rubric in oral presentations, Hafner and Hafner [8] found a valid and reliable results when the rubric was used by untrained student assessors, and a significant 1-to-1 correlation between student and teacher assessments.

Furthermore, the act of assessing others’ work can be a rich learning experience in and of itself since it increases time on task and requires the application of higher level cognitive thinking. In addition, Topping identifies potential cognitive gains in intelligent questioning, misconception identification, reflection, and generalization through peer assessment [18]. Finally, peer assessment affords an opportunity for students to practice and develop
their ability to review technical work prepared by others – an essential skill for most practicing engineers. Sluijsmans found that training through peer assessment tasks leads to the development of the skill to assess others’ work [15].

Two common challenges that remain are how to manage the peer assessment process, particularly in larger classes or when anonymity is preferred, and how to ensure fairness and quality in the assessments. Online peer assessment tools, such as peerScholar [10], Calibrated Peer Review [5], ComPAIR [4] (based on adaptive comparative judgement [12]) and others have emerged as approaches to manage this.

At the University of British Columbia, we have access to the above mentioned packages. Seeing the value for such a tool, we selected peerScholar to pilot due to a balance, from our perspective, between simplicity (both for students and faculty) and versatility. We piloted the use of peerScholar in courses at the first and third year levels, starting in 2015. We were aware that Canadian middle- and high-school students held a generally positive view of this software [3], but we wanted to understand acceptance and perceived value in our context and from our engineering students’ perspective. In addition, we intended the software for different year levels and for a variety of different assignment deliverables. We surveyed students at the end of the courses to understand their perceptions of the peer assessment process, fairness when used for course grading, and feelings towards the peerScholar tool in particular. In the case of the third year course, we also had teaching assistants grade the same materials and were able to compare peer grades to teaching assistant grades. The results indicate broad acceptance for the concept of peer assessment, and general acceptance for the specific instances of peer assessment in the courses.

2. CONTEXT

At the University of British Columbia, we have employed peerScholar as a tool to enable anonymous, online peer assessment of various course deliverables. This has been done in first year introduction to engineering courses (APSC 100 and 101), as well as a third year mechanical engineering design course (MECH 326). The sections below outline the course-specific contexts.

2.1. APSC 100 and 101 Courses

PeerScholar was used in two large core course in our common first year engineering program. APSC 100 (Introduction to Engineering I, n = 800) and APSC 101 (Introduction to Engineering II, n = 750) were offered for the first time in the 2015/16 academic year. The courses are each supported by a team of 8 classroom instructors, 7 laboratory/studio instructors, 7 senior graduate students acting as curriculum designers, 10 teaching assistants (TAs), and several instructional support staff members. In the first year of offering of the courses, we had several small written assignments (such as a business letter) that counted for course credit but we lacked the TA resources to fully assess them. As a result, we either had TAs quickly glance at each submission and give full credit if the assignment was submitted and appeared to be at least vaguely on topic, or, later, we gave credit if the student submitted anything. This served the purpose of holding students accountable to hand in something by the assignment deadline, but it provided the students no feedback on their work and it provided us no valid data on the students’ ability. We initially piloted peerScholar in a few assignments in APSC 100 and 101 in 2015-16, and then fully adopted it in both courses for the 2016/17 academic year.

Students have used peerScholar to review and assess two personal letters written on the topic of transitioning to engineering in Term 1, a technical memorandum summarizing on a design project involving CAD and rapid prototyping, a business letter and expression of interest (EOI) for a fictitious remote community water treatment project, and an optional personal statement written as part of application to the student’s preferred second year engineering discipline. The technical memorandum and the EOI are prepared by teams, and all other items are prepared by individuals. All peerScholar reviews were completed individually.

2.2. MECH 326

MECH 326 (Mechanical Design II) is a third year design and analysis course in Mechanical Engineering. It has a practice focus, and is divided into five modules on the topics of fracture, fatigue, shaft design, welding, and finite element analysis. The course is taught using the Team-Based Learning approach [14], and features a semi-constrained design assignment as part of each module. The design assignments are completed in teams, and, as a deliverable, each team prepares a two-part design report consisting of a three-page summary written for a lay audience and a detailed appendix of supporting work suitable for a knowledgeable engineer.

3. EXPERIENCES USING PEERSCHOLAR

Following an introduction to the key elements of the peerScholar platform, the follow sections summarize the types of assignments we used with peerScholar, the considerations we made in applying the software to team-based courses, and some of the challenges we faced.
3.1. peerScholar

peerScholar is an online peer assessment platform (www.pearsoned.ca/highered/peerscholar/). After creation of a peer assessment assignment by an instructor, students are guided through three phases in the assignment:

- Phase 1 – Create: students access peerScholar to view the assignment instructions and then create their solution. As determined in the assignment settings, students can either write their response directly in the peerScholar text editor, or upload their work (with support for video files, audio files, images, or other document types).

- Phase 2 – Assess: students are asked to assess a predetermined number of randomly assigned submissions from their peers. In the setting of the assignment in peerScholar, the instructor determines the number of peer assessments each student is to complete and the type of assessment required. Assessment options include overall comments, inline comments added directly to the work, and rating scales (Likert, stars, or numbers).

- Phase 3 – Reflect/Revise: students are asked to review and reflect on the feedback they have received from their peers in Phase 2. Optionally, the instructor can require students in this phase to assess the quality of the feedback they received, revise and resubmit their work, or reflect and comment on the feedback received.

For each phase, it is possible to set start and stop deadlines with 5-minute resolution. Late submission options are possible (with work received after the initial deadline but before the late deadline marked as “Late”), but a subsequent phase cannot begin until all deadlines in the prior phase have passed. It is possible to have time gaps between phases, for example, to have Phase 1 end on Monday at midnight but Phase 2 not begin until Tuesday at noon.

3.2. Assignment Types

We employed peerScholar for a total of eight assignments in the 2016/17 academic year (five in APSC 100/1, and three in MECH 326). These are summarized in Table 1. Although not included here, we also successfully used peerScholar in the assessment of student video submissions in the 2016 offering of the APSC 101 course.

<table>
<thead>
<tr>
<th>Course</th>
<th>Assignment</th>
<th>Individual or team</th>
</tr>
</thead>
<tbody>
<tr>
<td>APSC 100</td>
<td>Personal letter</td>
<td>Individual</td>
</tr>
<tr>
<td>APSC 100</td>
<td>Technical memo</td>
<td>Team</td>
</tr>
<tr>
<td>APSC 101</td>
<td>Business letter</td>
<td>Individual</td>
</tr>
<tr>
<td>APSC 101</td>
<td>Expression of interest</td>
<td>Team</td>
</tr>
<tr>
<td>APSC 101</td>
<td>Personal statement</td>
<td>Individual (optional)</td>
</tr>
<tr>
<td>MECH 326</td>
<td>Design reports (3)</td>
<td>Team</td>
</tr>
</tbody>
</table>

3.3. Working with Team Projects

In the case of team assignments, the team collaboratively prepared the work, and then each individual team member was responsible for uploading that assignment to peerScholar under their own name. In this way, a team of $N$ students would collectively upload their assignment $N$ separate times. In the peerScholar settings, the number of required reviews, $r$, was specified (we typically used $r$ between one and three). In this way, each team assignment was reviewed $r \times N$ times. With a typical team size of $N = 5$ and a typical number of reviews of $r = 2$ meant that each team’s deliverable was reviewed 10 times on average, and the final peer grade was the composite of 10 unique evaluations. This compares favourably with findings of Paré and Joordens [11], which, for individual assignments, suggested 6 assessments per work was optimal to achieve high validity (more assessments are better) but to avoid grader burnout (fewer assessments are better). In their study, Paré and Joordens found the highest degree of correlation between peer grades and TA graders at 12 peer assessments—for team assignments we are achieving close to that number of assessments while only requiring two assessments per student.

In the review phase, the one to three assessments to be completed by each individual were randomly assigned by peerScholar. Although peerScholar would not assign a student their own work, there was no mechanism to prevent it from assigning a teammate’s submission—that is, an individual could end up evaluating their own team’s work. We notified students this might be the case; we also indicated that we expected a fair an honest evaluation and that we would be looking for evaluation outliers in cases such as this. In addition, with each team assessed through $r \times N = 10$ evaluations, any impacts of an uncorrected outlier evaluation were mitigated.

3.4. Challenges

Overall, the peerScholar platform was well-received by students, and we felt it was highly effective and appropriate in the courses. At the same time, there were several challenges that we needed to work around.

In both courses, we ran into the issue of students missing key deadlines. With three phases (create, assess, reflect/revise), if a student missed the create phase, they were excluded from other aspects of the assignment. We tried to address this with consistent and predictable deadlines (e.g. deadlines were always on Monday’s at 8:00 am, and each phase lasted one week), as well as with regular reminders. In addition, in cases where students were expected to assess the quality of feedback received (in the reflect/revise phase), if all feedback recipients forgot to complete this phase, the original assessor received
a grade of 0 for the quality of their feedback. This had to be identified and corrected through post-processing. This still left questions as to what grade the assessor should receive; options considered included having TAs assess the quality of feedback (which would take time to coordinate and effort to complete), or to automatically assign a grade. The latter option was chosen to simplify and expedite grading, and full marks were awarded to the student for this phase (rather than, say, the class average) in the interest of preserving student goodwill and offering the benefit of the doubt to the assessor.

An additional challenge we identified involved the limitations imposed by the assessment structure. Most of our assessments were rubric-based, and while we could re-create the criteria and scoring scales of our rubrics, we could not easily add descriptors directly to peerScholar. Students had to rely on a .pdf attachment of the rubric with descriptors to refer to during their evaluations. This was partially addressed by using the star-scale ranking system in peerScholar and adding the same star graphics to our .pdf rubric attachment as a visual way to more clearly map each descriptor to each rating.

When we started using the software, we found we needed to create a number of test student accounts and test assignments to be able to see the student view of the platform. The instructor preview was slightly different than the student working view. Once we had mapped out the relationships between the two views, this was no longer an issue.

Finally, during a particularly busy time of term, with simultaneous assessments active for the APSC 100 and MECH 326 courses, there was a string of system outages at both the UBC side and peerScholar host side (at the University of Toronto) which resulted in some student frustration and numerous extensions to assignments over about one week. Since that time, the system has been stable for us and has functioned without incident or interruption.

4. RESULTS

For each course, feedback on peerScholar was solicited from students with informal online surveys after the end of the course. In total, there were 54 survey responses (7% of cohort) in APSC 100/101, and 42 responses in MECH 326 (32% of cohort). The perceived value of peer assessment, feedback on the experiences in the courses, and feedback on peerScholar are summarized below, followed a statistical comparison of student and TA grader assessments.

4.1. Perceived Value of Peer Assessment

The first set of survey questions asked about the perceived value in the peer assessment process using a five-point Likert response scale (strongly agree, mildly agree, neither agree nor disagree, mildly disagree, strongly disagree). The prompts were:

- Q1: The ability to quickly and accurately review technical work prepared by other engineers is an essential skill,
- Q2: Part of an engineering degree should include opportunities to review technical work prepared by others,
- Q3: I see myself one day taking on the role of manager or leader of an engineering team, and
- Q4: Including peer assessment activities in [course name] was appropriate.

The resulting data is shown in Fig. 1 with APSC 100/101 as the top set of four bars and MECH 326 as the bottom set.

![Fig. 1. Student perception of peer assessment (top: APSC 100/101, bottom: MECH 326)](image)

In both courses, there was broad perceived value in peer assessment, with almost all students mildly or strongly indicating it was an essential skill (Q1, over 90% agreement in each course) and that opportunities should be given in an undergraduate degree to exercise it (Q2, over 85% agreement). This broad support for peer assessment in general did not translate directly to support for the peer assessment activities in the specific courses (Q4); still, over 65% in each course agreed that peer assessment was appropriate for the course. No statistically significant correlation was observed between whether someone saw themselves as a future leader (Q4) and any other prompt.
4.2. Feedback on Peer Assessment

The second block of survey questions asked about the experience of engaging in the peer assessment process using the same five-point Likert scale. The prompts were:

- Q5: The effort I and/or my teammates put into the MECH 326 assignments increased knowing they would be peer reviewed,
- Q6: The process of reviewing and assessing the work of my peers helped deepen my learning of the assignment topics,
- Q7: The feedback I received from my peers helped to deepen my learning of the assignment topics, and
- Q8: Overall, the use of peer grading mixed with TA grading for the MECH 326 assignments is fair.

The resulting data is shown in Fig. 2, again with APSC 100/101 as the top set of four bars and MECH 326 as the bottom set.

Since the third year student design reports were also graded by TAs, an additional question was added to the survey for those students. Specifically, once a student’s response to the question of fairness (Q8) was recorded, if they did not strongly agree that the process of using peer grades in the course was fair (i.e. they responded mildly agree to strongly disagree on Q8), a comparison of average grades from peer markers to TA markers was shown. Notably, students were notified that peer grades were 1.1% higher than TA grades on average, and the two sets of grades were positively correlated ($r = 0.50$). The students were then asked to rate their perceived fairness of the use of peer assessment in the course on a four point scale, with the results shown in Table 2. These indicate that the overall perception of fairness increased knowing that grade differences were small between peers and teaching assistants. In addition, at worst, a small group of students (6%) felt the use of peer assessment was “somewhat unfair,” and no students felt it was “very unfair.”

<table>
<thead>
<tr>
<th>Category</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very fair</td>
<td>23%</td>
</tr>
<tr>
<td>Somewhat fair</td>
<td>71%</td>
</tr>
<tr>
<td>Somewhat unfair</td>
<td>6%</td>
</tr>
<tr>
<td>Very unfair</td>
<td>0%</td>
</tr>
</tbody>
</table>

This increased perception in fairness knowing the student marks were comparable to the TA (expert) marks, is consistent with the increase in credibility suggested by Paré and Joordens in their study of peerScholar [10].

4.3. Assessment of peerScholar

The final three core questions asked about experiences specifically relating to peerScholar in the courses:

- Q9: peerScholar was intuitive and easy to use,
- Q10: It was time-consuming and/or burdensome to use peerScholar in the course, and
- Q11: Overall, the assignment review process through peerScholar contributed positively to my learning.

The resulting data is shown below in Fig. 3. First year students were more likely to report that the software was easy to use (Q9), but also that the process was more burdensome (Q10). Roughly half of the students in each group indicated that peerScholar contributed positively to their learning (Q11). This compares to results reported by Basnet et al. [1] where approximately 60% of students indicated that peer assessment is a useful learning tool.
4.4. Student and TA Grade Comparison

In the MECH 326 course, three assignments were independently peer-marked and TA-marked. The grading criteria for each were similar, although students had four criteria and TAs had five (as shown in Table 3). A different TA (and only one) marked each assignment. Each student was assigned to grade two assignments; however, due to some students forgetting to submit their team’s assignment (average of 1% of the class did this per assignment), some students starting but forgetting to finish their assessments (2%/assignment), and some not beginning their assessments (16%/assignment), there were roughly 8.5 evaluations per team’s assignment on average (less than the 10.4 per team that would be seen with full participation).

A comparison of the grading data by assignment is given below in Table 3. The rank ordering of grades was the same between students and TAs (i.e. Assignment 1 received the lowest grades from both groups and Assignment 3 received the highest). There was more spread in the TA assignment averages compared to the student averages, which could represent reluctance on the part of students to give more extreme grades, more variability in the grading of TAs, or an averaging effect by having many student graders. The correlation coefficient between the two groups of markers ranged from 0.19 to 0.57 (with Assignment 3 and the composite grade showing a statistically significant correlation). While we would have liked to see stronger correlations for Assignments 1 and 2, we reminded by through the literature that there is value in the act of assessment and in the feedback, even without perfect agreement between student and TA assessors.

Table 4: MECH 326 assignment grading criteria

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Student grade</th>
<th>TA grade</th>
<th>Pearson r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80.0%</td>
<td>74.6%</td>
<td>0.34 (p=0.09)</td>
</tr>
<tr>
<td>2</td>
<td>82.0%</td>
<td>81.7%</td>
<td>0.19 (p=0.36)</td>
</tr>
<tr>
<td>3</td>
<td>86.9%</td>
<td>89.5%</td>
<td>0.57 (p&lt;0.01)</td>
</tr>
<tr>
<td>Composite</td>
<td>81.9%</td>
<td>83.0%</td>
<td>0.50 (p&lt;0.01)</td>
</tr>
</tbody>
</table>

5. DISCUSSION AND CONCLUSIONS

We consider our piloting of peerScholar to be successful. We used it in large first and third year courses, and for multiple types of deliverables, including design reports, letters, technical memoranda, and more. peerScholar is an online tool with three phases: creation of the assignment, assessment of peers, and then review and revision. We used all three phases in our pilots.

Students strongly recognized the need to develop peer assessment skills and practice them while in university, and they identified the use of peer assessment as being appropriate in the respective courses. Compared to third year students, first year students were more likely to report putting additional effort into their work knowing it would be peer reviewed, and they reported finding the feedback they received as being more useful towards their learning. Both first and third year students found learning value in the act of reviewing others’ work.

In comparing student and TA grades, we found good overall agreement in terms of grades (including averages, standard deviations, minimums, and maximums), and reasonable agreement in terms of correlations (r = 0.5 overall). Students perceived increased fairness in the peer assessment process once they were informed of the similarity between peer and TA grades.

In addition, the anonymous nature of the software tool addresses our concerns of students favouring friends (or
penalizing adversaries) in assessments. Furthermore, the ability in the review phase to require students to evaluate the quality of feedback received enhances grader accountability while providing a mechanism for students to identify unfair grading. We only encountered minor issues when using the software, and plan to continue and potentially expand its use in future.

Acknowledgements

We wish to thank the University of British Columbia Teaching and Learning Enhancement Fund (TLEF) for funding the work to redevelop our first year engineering curriculum. References

References


