Optimum Frequency of Peer Evaluations in Capstone Design Courses

Michel F. Couturier and Guida Bendrich
Department of Chemical Engineering, University of New Brunswick
cout@unb.ca

***Abstract** –Peer evaluations are commonly used in design courses for developmental and evaluative purposes. Peer ratings are however often higher than the instructor ratings and this can induce fears regarding their reliability. This study examined whether it may be possible to increase the agreement between peer and instructor ratings by increasing the frequency of the peer assessments. The premise was that peers may provide less lenient assessments if the impact of single evaluations on the final grade is reduced by increasing the number of evaluations. Increasing the number of peer evaluations in our senior design course from two to six per year did not increase the accuracy of the peer ratings but provided other benefits such as earlier identification of dysfunctional teams, elimination of free riding and more frequent developmental feedback. Peer and instructor ratings can be normalized however to yield similar indicators of the relative performance of teammates. The frequency and timing of peer evaluations are critical to obtain meaningful results and maximize impact on team dynamics.*

Keywords: peer evaluation, teamwork, assessment factor, grade adjustment, developmental feedback

1. INTRODUCTION

Students registered in accredited engineering programs in Canada are expected to develop an ability to work effectively as a member and leader in teams [1]. This skill is mainly developed in design courses where, working in teams of four or more individuals, students must obtain solutions to open-ended design problems. Team work presents challenges however to both students and instructors. Students are concerned that some of their teammates may become free riders who will not do their fair share of the group work, while instructors are concerned about determining individual grades in a context where all tasks are completed in a team setting outside of class. Both challenges can be overcome adequately using peer assessments. The argument for using peer evaluations to determine individual contributions is that team members are better able to observe and evaluate members' contributions than are instructors, who see outcomes but not the behind-the-scene process [2]. Peer evaluations also reduce free-riding by increasing students' accountability to their teammates [3, 4].

For peer assessments to provide meaningful ratings, it is important that students take the rating process seriously and spend time considering each rating [5]. Given that performance is influenced by rewards and rewards for students come primarily in the form of grades [6], it is logical that students will take the peer evaluation process more seriously if it has a great impact on the final grade.

The frequency of peer assessments is also critical. When instructors use only confidential, end-of- course peer evaluations to determine grades, there is not enough time to provide feedback to students and the single assessment may actually undermine team process and performance [6]. When poor team dynamics is encountered, rather than seeking solutions, students may tolerate the undesirable behaviors thinking that they can punish their teammates on the peer evaluation at the end of the course. Peer assessments that provide feedback at both early and multiple points during the course are more effective at reducing free-riding problems and improving team dynamics [3]. Frequent feedback encourages early communication of desirable and undesirable behaviors and provides lagging team members the opportunity to improve.

Peer evaluations used for both developmental and evaluative purposes should be derived from categorical rather than holistic assessment tools [3]. In categorical assessment, students score teammates on a number of individual dimensions whereas a single score that reflects a global judgement is assigned to each team member in holistic assessment [5]. Assessment instruments which contain specific evaluation criteria not only provide team members with more effective feedback but also provide individuals with clearer indications of desirable behavior [3].

In an earlier study [7], we reported that the peer ratings collected at the mid-point and end of our full-year design course were generally higher than the team mentor ratings. Peer evaluations were high unless conflicts existed within the team, and we concluded that peers inflated their ratings to avoid hurting interpersonal relationships. A number of other studies have also reported this phenomenon [5]. In this study, we hypothesized that more accurate individual evaluations could be obtained by increasing the frequency of the peer evaluations. The premise is that peers will provide more accurate assessments of their teammates if the impact of individual evaluations on the final grade is reduced by increasing the number of evaluations. During the 2015-2016 offering of the capstone design course in the Chemical Engineering Department at the University of New Brunswick, we ran peer evaluations at the end of each month for a total of three evaluations in the fall and three more in the winter. In this article, we describe our evaluation process and examine the pros and cons of performing frequent peer evaluations.

2. ASSESSMENT METHOD

2.1 Course Structure

A collaborative approach has been used for teaching the full-year capstone design course *ChE 4225 Chemical Plant Design* in the Chemical Engineering Department at the University of New Brunswick since 2010 [8]. Design projects undertaken in the course are sponsored by outside clients. Two teams of four or five students are assigned to each project. The two teams assigned to each project work independently and are co-mentored by a faculty member and a practicing engineer. Funding for the industrial co-mentors is provided by the NSERC-UNB Chair for Collaborative Engineering Design Education held by one of the co-authors (M. Couturier).

The two course coordinators (M. Couturier and G. Bendrich) ensure that all groups are progressing at the same rate and are evaluated against the same criteria by setting eight evenly-spaced milestones (four per term) that all teams must complete. The outline for each milestone defines the tasks that need to be accomplished and set the marking scheme that will be used by the co-mentors to evaluate the deliverables.

Design review meetings take place with the clients in November and February. These meetings start with an oral presentation by the students to update their client on their progress and conclude with questions and feedback from the client. The design work of the students is also showcased at our annual Engineering Design Symposium held in early April.

2.2 Individual Grades

The breakdown of the final grade of the students is provided in Table 1. The final report has the largest weighting because it is the main product of the design exercise and the only document given to the client. Each milestone report is only worth 5% of the overall mark (Table 1).

Table 1: Breakdown of the final grade

Deliverable	Value
1. Final report	45%
2. First 7 milestones	35%
3. 3 presentations	20%

Students are individually evaluated during the presentations and their team grades on milestones are adjusted for individual performances using the following equation.

$$\text{Individual Mark} = \text{CF} \times \text{Group Mark} \qquad (1)$$
$$\text{on Milestone} \qquad \qquad \text{on Milestone}$$

The correction factor CF, which can be greater or smaller than unity, takes into account the teammate and co-mentor evaluations and the participation of the student in the course.

$$\text{CF} = 0.2*\text{CF}_{\text{participation}} + 0.2*\text{CF}_{\text{co-mentor}} + 0.6*\text{CF}_{\text{peer}} \qquad (2)$$

The active participation of every student is required to make the course *Chemical Plant Design* a positive experience for everyone. The students are expected to attend the weekly two-hour lectures (attendance is recorded) and to submit the following information:
- monthly logbooks which list the main tasks accomplished by each student and the time spent on each task,
- self and peer evaluations at the end of each month.

If the student misses no more than 30% of the lectures per term and submits all required information, $\text{CF}_{\text{participation}}$ is assigned a value of unity; otherwise it takes on a value of zero.

The peer evaluations collected at the end of each month are used to calculate the assessment factor CF_{peer} as follows,

$$\text{CF}_{\text{peer}} = \frac{\text{Student's evaluation by peers in \%}}{\text{Average of peer evaluations for the group in \%}} \qquad (3)$$

and the co-mentor evaluations collected at the end of each term are used to calculate $\text{CF}_{\text{co-mentor}}$ using

$$CF_{co\text{-}mentor} = \frac{\text{Student's evaluation by co-mentors in \%}}{\text{Avg of co-mentor evaluations for the group in \%}} \quad (4)$$

The correction factor CF is calculated at the end of the fall term and used to determine the individual marks for milestones 1 to 4. The assessment factor CF_{peer} appearing in Eq. 2 is then the average of the values obtained at the end of September, October and November. The correction factor CF is again calculated at the end of the winter term and used to obtain individual marks for milestones 5 to 7. The assessment factor CF_{peer} used in this second calculation is the average of the values obtained at the end of January, February and March. The average of the fall and winter correction factors is then used to determine the individual marks of the students on the final report using the following formula.

$$\text{Individual Mark on Final Report} = \frac{(CF_{fall} + CF_{winter})}{2} * \text{Group Mark on Final Report} \quad (5)$$

2.3 Peer and Co-mentor Evaluations

Each team member evaluated his/her teammates at the end of each month using a survey developed in Microsoft Excel. The survey used in November and March measured the sixteen team member dimensions listed in Table 2. The dimensions are divided into five general areas as suggested by Ohland et al. [1] and attempt to measure the quantity and quality of individual work as well as the teamwork skills of students. A shorter survey comprising ten dimensions was used at the end of September, October, January and February by omitting the second and third dimensions listed under areas 4, 5 and 6. Each dimension was evaluated using a pull-down rubric which contained four levels from poor to excellent. The levels were assigned scores of 0% (fails), 50% (marginal), 75% (meets expectations), and 100% (exceeds expectations) for averaging purposes. The survey was sent to the students by email and the completed surveys were returned to the instructors by the same method. The scores provided by peers for each dimension were averaged and the overall average of all dimensions assessed for each student was used in Eq. 3 to calculate CF_{peer}.

After the peer evaluations were completed at the end of November and March, the co-mentors were also asked to complete an evaluation of the students under their responsibility. Their survey was similar in structure to the one used by the students but measured fewer team member qualities or behaviors as shown in Table 3. The dimensions which require information that only peers are able to observe were not evaluated by co-mentors. Information used by co-mentors to assess individual contributions included the average peer assessment results, discussions with students at weekly team meetings, the logbooks submitted monthly by each student, and the milestone reports. The scores provided by the two co-mentors were averaged for each student and used in Eq. 4 to calculate $CF_{co\text{-}mentor}$.

In addition to being used for adjusting team grades, the peer and co-mentor evaluations were also shared with the students as formative feedback on their performance. To maintain the anonymity of the evaluators, only the aggregated peer and co-mentor results for each dimension were given to the students.

Table 2: Dimensions assessed by peers

Area	Dimension
1. Contribution to team work	• interest in work • initiative • quantity of work
2. Interaction with teammates	• dependability • interpersonal behavior • conflict resolution
3. Leadership	• planning & organizing • setting goals • leadership qualities
4. Quality expectations	• quality of work • creativity • judgement
5. Skills	• integration of prior learning • written communication • oral communication • problem solving skills

Table 3: Dimensions assessed by co-mentors

Area	Dimension
1. Contribution to team work	• attitude toward project • quantity of work
2. Interaction with teammates	• dependability and punctuality • team work
3. Leadership	• leadership qualities
4. Quality expectations	• quality of work • creativity
5. Skills	• problem solving skills • written communication • technical knowledge

3. RESULTS AND DISCUSSION

The results of the surveys performed during the academic year 2015-2016 are presented in this section. A

total of 53 students were evaluated. The students were divided into 13 teams (one team of 5 students and twelve teams of 4 students). Each student was therefore evaluated by two co-mentors and 3 or 4 peers. Seven faculty members and 7 practicing engineers acted as co-mentors.

3.1 Peer Ratings for Developmental Feedback

Peer evaluations were performed at the end of each month whereas milestone reports were due every three weeks. We purposely set different due dates for the milestone reports and peer evaluations to ensure that the evaluations would not be rushed. We also reduced the length of the peer survey from sixteen to ten questions in September, October, January and February to avoid a decrease in student motivation caused by 'evaluation fatigue'. The longer peer survey was only used in the months of November and March when students were also being evaluated by their co-mentors. The long survey provided a greater overlap between the dimensions assessed by peers and co-mentors.

The average ratings provided by the peers of each student were analyzed by determining the percentage of students who fell into the four evaluation categories (fail, marginal, meets expectations and exceeds expectations) for each dimension and then averaging the percentages over all dimensions. The results thereby obtained are shown in Fig. 1 for each month. Only a small percentage of students received fail and marginal ratings. Most students received ratings above marginal and the percentage of students receiving a rating of 'exceeds expectations' increased with time. This suggests that students used the feedback provided by their peers to improve their evaluations. The large weight associated with peer evaluations in the calculation of individual marks (see Eq. 2) provided a large incentive to improve on individual weaknesses and discouraged free riding.
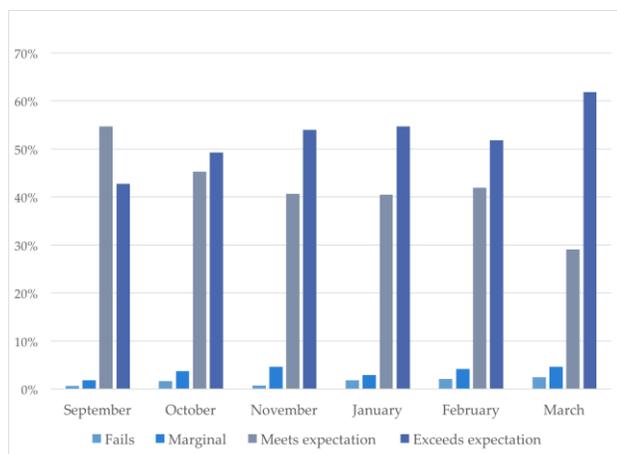
By performing peer evaluations early in the design process and more frequently than in the previous academic year, we were able to identify dysfunctional teams more quickly and better monitor the progress of poor performers. The more frequent peer evaluations also helped the teams move more quickly through the team development stages of forming, storming, norming, and performing defined by Tuckman [9]. The feedback provided by peers promoted team discussions about unacceptable behavior and was used by students to improve weaknesses. Brooks and Ammons [4] also concluded that an evaluation which provides feedback on specific criteria at both early and multiple points during a group project can have a positive influence on a group's ability to work well together.

3.2 Peer Ratings for Evaluative Purposes

Although increasing the frequency of peer evaluations had a positive impact on team dynamics, it did not improve the accuracy of peer evaluations. This is demonstrated in Fig. 2 where the average of the two co-mentor evaluations obtained for each student in November is plotted versus the average of the peer evaluations for the same students. The fact that the majority of the data points are below the 45 degree line indicates that peer evaluations are generally higher than co-mentor evaluations. The same trend was also observed in March as can be seen in Fig. 3 and is no different than the trend observed in our previous study [7] where we performed peer evaluations at the midpoint and end of the course only. Therefore, although increasing the number of peer evaluations reduces the impact of individual evaluations on the final grade, it does not lead to better agreement between peer and co-mentor assessments. Our initial hypothesis is therefore not validated. Peer evaluations were higher than co-mentor evaluations for all dimensions evaluated except for quality of work, dependability and interpersonal behavior where more similar ratings were obtained.
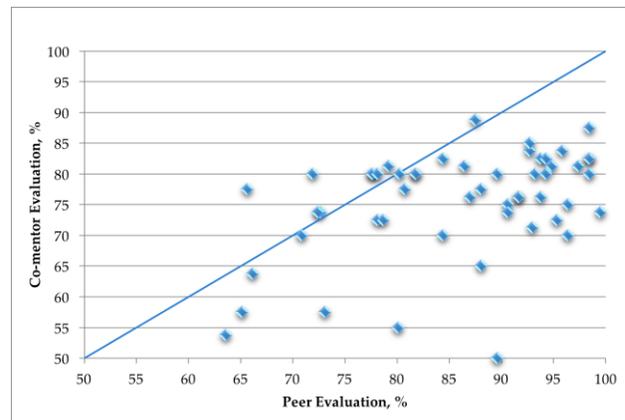

Fig. 1. Monthly avg. individual evaluations

Fig. 2. Co-mentor vs. peer evaluations, November
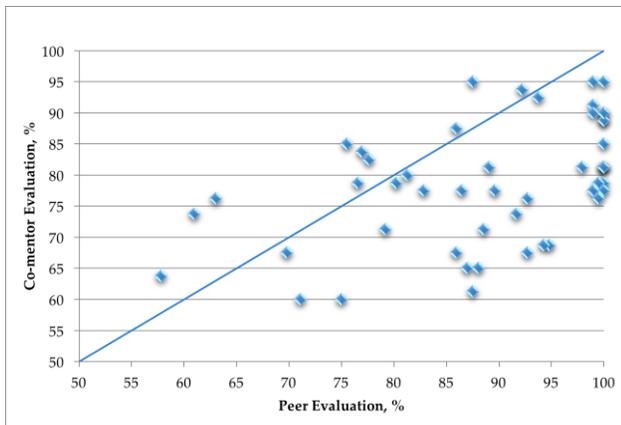
Fig. 3. Co-mentor vs. peer evaluations, March

Fig. 5. CF$_{co-mentor}$ versus CF$_{peer}$, March

Despite the lack of agreement between the peer and co-mentor assessments, we can use peer ratings to determine individual grades because we are concerned not by the absolute values of the ratings but by their relative values. The ratings are used to arrive at relative assessment factors (Eqs. 3 and 4) which are then combined (Eq. 2) and multiplied with the group project grade (Eqs. 1 or 5) to determine each individual's grade. In the calculation of the assessment factors, the ratings are normalized by dividing the average assessment for each student by the team average. This normalization process attenuates the biases that might operate when peers and co-mentors perform their evaluations. As can be seen in Figs. 4 and 5, there is a good correlation between the co-mentor and peer assessment factors, CF$_{co-mentor}$ and CF$_{peer}$, obtained in November and March, respectively. Similar results were obtained in our previous study [7].

performance level in the rubric is nonetheless subjective and open to biases. Peers and co-mentors use different reference groups and observations when making their evaluations. Furthermore, peers are being asked to evaluate individuals who are in turn also evaluating them. Therefore, unless there are conflicts within their team, peers tend to be lenient in their evaluations to avoid hurting interpersonal relationships. They therefore submit more generous ratings than co-mentors but are nonetheless differentiating between their teammates. As a result, once the bias caused by the differing anchor points is removed via normalization, peer and co-mentor ratings produce similar assessment factors. In other words, although the peer and co-mentor ratings are different in absolute terms, they generate similar indicators of the relative performance of individuals in the same team.
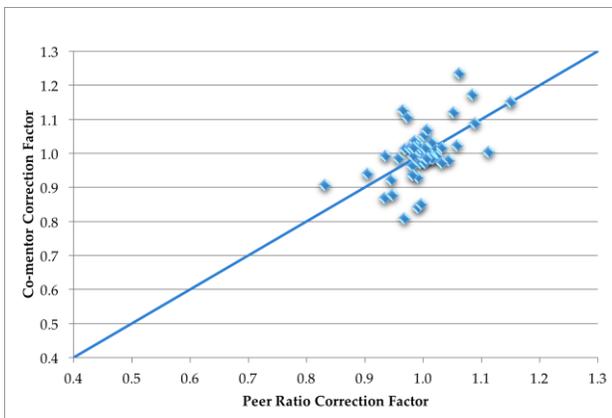
Table 4: Rubric for the dimension 'quality of work'

Description	Ranking	Value
Excellent productivity	Exceeds expectations	100%
Good productivity	Meets expectations	75%
Satisfactory productivity	Marginal	50%
Unsatisfactory productivity	Fails	0%

Fig. 4. CF$_{co-mentor}$ versus CF$_{peer}$, November

Differences between peer and co-mentor ratings likely exist because the two groups use different anchor points for what they consider satisfactory behavior in their scale of evaluation.   The rubrics used in this study were designed to minimize this type of bias (see sample rubric in Table 4) but the selection of the appropriate

3.3 Optimum Frequency of Peer Evaluations

It is a challenge to assign fair grades to individual students in a capstone design course like ChE 4225 where all deliverables are completed in a team setting outside of class. When we first started co-instructing ChE 4225 in 2010, we gave every team member the same grade on milestone reports. We quickly realized that this was an unsatisfactory method of assigning individual grades because it rewarded and even encouraged free riding. We

therefore introduced a peer survey at the end of the year and adjusted final report grades using a multiplier similar to the one proposed in section 2.2. The administration of the original paper survey was however time consuming and we replaced the end-of-year paper survey by the current electronic survey in 2013. The electronic survey has greatly reduced the time required to collect and analyze the rating data and has made it possible to run the survey more often. We ran the peer survey twice in 2014-15 and six times in 2015-16. The benefits provided by the additional peer evaluations performed in 2015-16 have outweighed the disadvantages. Although more frequent peer assessments have not led to more accurate peer ratings, they have eliminated free riding, have helped students progress more quickly through team development stages and have allowed us to identify dysfunctional teams earlier. We doubt that increasing the frequency of peer assessments further will bring additional benefits and fear that it may lead to 'evaluation fatigue'. We therefore conclude that the current frequency is likely optimal for the course ChE 4225.

The timing of the assessments is also critical. It would not be productive to ask students to perform peer assessments during the stressful period preceding the due date of a major deliverable. Currently, students are submitting their evaluations about one week after the submission of each major team deliverable. This is a less stressful period to reflect on the performance of their teammates. It is also an opportune time to analyze the performance of their team during the execution of the previous milestone.

4. CONCLUSIONS

Increasing the frequency of peer evaluations had a positive impact on team dynamics but did not improve the accuracy of peer evaluations. Peers submit more generous ratings than co-mentors but are nonetheless differentiating between their teammates. As a result, although the peer and co-mentor ratings are different in absolute terms, they produce similar indicators of the relative performance of individuals on the same team once normalized.

The frequency of peer evaluations in the design course ChE 4225 at UNB was gradually increased over several years and found to provide maximum benefits when performed once per month. The timing of the evaluations is also critical and is set to occur about one week after the submission of each major deliverable.

Acknowledgements

References

[1] Canadian Engineering Accreditation Board, *Accreditation Criteria and Procedures*. Ottawa, ON: Engineers Canada, 2015, 115 pp. {ISSN: 1708-8054}

[2] Matthew W. Ohland, Misty L. Loughry, David J. Woehr, Lisa G. Gullard, Richard M. Felder, Cynthia J. Finelli, Richarch A. Layton, Hal R. Pomeranz, and Douglas G. Schmucker, "The Comprehensive Assessment of Team Member Effectiveness: Development of a Behaviorally Anchored Rating Scale for Self- and Peer Evaluation," *Academy of Management Learning & Education*, vol. 11, no. 4, pp. 609-630, 2012.

[3] Larry K. Michaelson, Arletta B. Knight, and L. Dee Fink, *Team-based Learning: A Transformative Use of Small Groups*. Westport, CT: Praeger Publishers, 2002, 288 pp. {ISBN: 0-89789-863-X}

[4] Charles M. Brooks and Janice L. Ammons, "Free riding in group projects and the effects of timing, frequency, and specificity of criteria in peer assessments," *Journal of Education for Business*, vol. 78, no. 5, pp. 268-272, 2003.

[5] Barry A. Friedman, Pamela L. Cox, and Larry E. Maher, "An expectancy theory motivation approach to peer assessment," *Journal of Management Education*, vol. 32, no. 5, pp. 580-612, 2008.

[6] Donald R. Bacon, Kim A. Stewart and William S. Silver, "Lessons from the best and worst student team experiences: how a teacher can make the difference," *Journal of Management Education*, vol. 23, no. 5, pp. 467-488, 1999.

[7] Michel F. Couturier and Guida Bendrich, "Evaluation of individual members in engineering design teams," in *Proc. CEEA Canadian Engineering Education Conf., CEEA15*, Hamilton, ON; (May 31-June 3 2013), 2013.

[8] Michel F. Couturier, and Guida Bendrich, "Collaborative Approach for Teaching Chemical Process Design," in *Proc. CEEA Canadian Engineering Education Conf., CEEA13*, (Montreal, PQ; 17-20 June 2013), 2013.

[9] Bruce W. Tuckman, "Developmental sequence in small groups," *Psychological Bulletin*, vol. 63, no. 6, pp. 384-389, 1965.