A Preliminary Investigation of the Effectiveness of Peer Ratings in Engineering Design Teams

Richard O. Mines, Jr.[1] *and Joan M. Burtner*[2]

Abstract – This study was undertaken to investigate the effectiveness of peer ratings in engineering design courses. Data on peer ratings, self-ratings, final course grades and minority status were collected for three courses at Mercer University (freshman design, first-semester senior design, and second-semester senior design). Spearman rank correlation coefficients indicate there were significant correlations between final grades and peer ratings for preliminary design reports (PDRs) in the freshman design course and between final grades and peer ratings in the second-semester senior design course. No significant correlation between peer and self-ratings was found in the freshman design course. However, strong correlations exist between peer and self-ratings for both PDRs and critical design reports (CDRs) in senior design courses. In the freshman design course, there was a significant difference in ratings given by non-minority and by minority students for the PDRs; whereas, for CDRs, there was a significant difference in ratings given to non-minority and to minority students. In the senior design courses, there were no significant differences in ratings given by or given to non-minorities and minorities.

Keywords: Correlations, peer ratings, final grades, self-ratings, design teams.

INTRODUCTION

Using collaborative or cooperative learning can result in a learning environment that makes students more confident, retain and acquire information easier, and lead to higher-level thinking skills [4]. Felder and Brent [2] present an excellent overview of successful and unsuccessful techniques that have been tried in cooperative learning. However, this instructional methodology can present some challenges due to teamwork problems such as uneven participation or personality conflicts.

One of the major concerns when collaborative work is assigned is how to evaluate and hold each team member accountable for their role in the project or collaborative effort. Equitable grading is threatened when one or more team members contribute very little to the project. Some faculty members are wary of using collaborative work assignments and feel that team members can agree to give themselves self-inflated ratings or there may be prejudice in assigning a rating. Hatfield and Tester [3] list and briefly describe ten reasons why peer evaluation scores can be distorted. They list the following reasons: 1) reallocation of evaluation points, 2) students not doing an evaluation, 3) difficulty in assessing and quantifying individual contributions, 4) not knowing students, 5) not aware of an individual's contributions, 6) not wanting to hurt someone, 7) bias, 8) personality conflicts, 9) insufficient involvement in the project, and 10) collusion.

At the Mercer University School of Engineering (MUSE), collaborative or cooperative learning is an instructional technique that is used in several courses throughout our general engineering curriculum. Specifically, teams of students work together to accomplish design projects in the following courses: Introduction to Engineering Design (EGR 107), Senior Design Exhibit I (EGR 487/ECE 485), and Senior Design Exhibit II (EGR 488/ECE 486).

[1] Mercer University School of Engineering, Department of Environmental Engineering, 1400 Coleman Avenue, Macon, GA 31207 and mines_ro@mercer.edu

[2] Mercer University School of Engineering, Department of Industrial Engineering and Industrial Management, 1400 Coleman Avenue, Macon, GA 31207 and burtner_j@mercer.edu

Over the past several years, faculty members in the School of Engineering who teach the design courses have been using a peer rating system developed by Professor Rob Brown at the Royal Melbourne Institute of Technology (RMIT) to assist them in evaluating each team and team member. Details on how Professor Brown implemented the peer rating system can be found in [1]. The RMIT form was modified by Professor Rich Felder by adding detailed descriptions to each of the nine categories [5]. The peer rating form that is currently used at MUSE is shown in Figure 1 (included at the end of this paper). Each member of the design team confidentially completes the form performing an assessment of themselves and each team member. The rating system is a qualitative system consisting of nine terms ranging from "excellent" to "no show". Although students do not realize it, the instructor assigns a numerical value to each term. An "excellent rating" equates to 100 points, "very good" equates to 87.5, and each subsequent rating is decreased by 12.5 points until the last rating of "no show" garners a numerical rating of zero. At the end of the semester or for a given assignment, the instructor collects the peer rating forms and calculates the team average rating (TAR) and individual average rating (IAR). A grade adjustment factor is determined by dividing the IAR by the TAR. The instructor then applies the grade adjustment factor to the final grade in the course or for a particular assignment. Therefore, it is possible that individual members of the team can earn a higher or lower grade then the other members depending on the results of the peer reviews.

Other engineering educators also use versions of the RMIT rating form. Kaufman, Felder, and Fuller [5] found statistically significant correlations between peer ratings and average test grades in two sophomore-level chemical engineering course. The authors suggest that more responsible students, i.e., students receiving higher evaluations, are the academically stronger students. The study found no indication of gender bias in the peer rating system; results on racial/ethnic bias were inconclusive. Layton and Ohland used a modified version of the RMIT form in which they listed six team responsibilities. Their results showed no effects related to race/ethnicity. However, they did find significant effects due to gender, with women rating other women much lower than they rate men [6]. Ohland and Finelli [7] studied peer ratings at an institution that requires participation in a cooperative education program. They found no differences in peer ratings due to gender or race/ethnicity. Ohland and Layton [8] assessed the reliability of two different evaluation procedures for performing peer reviews using analysis of variance. The peer evaluation instruments that were evaluated were developed by Brown [1] and Ofori, Pai, and Layton [8]. The results of their study validated both instruments.

The objective of our study was to assess the use of the peer review form in three engineering design classes. One major goal was to determine if there were strong correlations between final course grades and average peer ratings and between final course grades and individual self-assessment ratings. The second major goal was to determine if there were significant differences between the peer ratings of non-minority and minority students.

COURSES EVALUATED

The modified RMIT peer rating system form was used in assessing the following three courses at Mercer University: EGR 107 Introduction to Engineering Design (spring 2012), ECE 485/EGR 487 Senior Design Exhibit I (spring 2011), and ECE 486/EGR 488 Senior Design Exhibit II (fall 2011). Table 1 shows the demographic data for the students enrolled in each course.

Table 1. Demographic Data for Courses in which RMIT Peer Rating System was used.

Course	Number of Students	Number of Men	Number of Women	Number of Non-minorities	Number of Minorities
EGR 107	22	21	1	15	7
EGC 485/EGR 487	19	17	2	10	9
EGC 486/EGR 488	19	17	2	10	9

EGR 107, Introduction to Engineering Design, is a freshman-level, 3 hour credit course in which students are introduced to the design process and must complete two design projects during the semester. At the freshman design level, students were randomly assigned teams. EGC 485/EGR 487 and EGC 486/EGR 488 comprise a two-semester, senior-level, 2 hour credit, design course that all engineering students must complete before graduating. Students select their own projects and team members in the precursor to the design sequence in a course called EGR 480, Introduction to Senior Design. Table 2 reports the composition of the teams in each of the three courses that were evaluated in this study.

Table 2. Team Composition among the Three Courses.

Course	Number of Teams	Mixed Gender	All Non-minority	All Minority	Minority and Non-minority
EGR 107	One, 2-person teams Four, 3-person teams Two, 4-person teams	1	1	1	5
EGC 485/EGR 487	Five, 3-person teams One, 4-person teams	2	1	1	4
EGC 486/EGR 488	Five, 3-person teams One, 4-person teams	2	1	1	4

PEER EVALUATIONS

In EGR 107, students evaluated each other on two occasions during the semester; first for the preliminary design review (PDR) and the second for the critical design review (CDR). Peer evaluations in ECE485/EGR 487 and ECE 486/EGR 488 were conducted at the end of each semester. The modified RMIT form as presented in Figure 1 was used for assessing individual and team performance. As mentioned previously, the instructor applied a numerical value to each of the nine qualitative statements listed on the form. An "excellent" rating equated to a numerical value of 100 with the other ratings receiving a numerical rating in decrements of 12.5 resulting in a numerical value of zero being assigned to the "no show" rating. A spreadsheet was used in calculating the team average rating (TAR) and individual average rating (IAR). A grade adjustment factor was calculated by dividing the IAR by the TAR. This adjustment factor for each student was then multiplied by the PDR and CDR grades in EGR 107 for determining the final grade for each of those components of the student's grade. The final course grades in EGC 485/EGR 487 and EGC 486/EGR 488 were determined by multiplying the student's individual adjustment factor by the student's final average in these courses.

DATA REDUCTION AND ANALYSIS

Non-parametric statistical analyses were performed on the peer ratings. Ranked Spearman correlation analyses were performed between average student peer rating and final grade received by the student, and between self-rating and final grade in each of the courses. Wilcoxon tests were performed on the peer ratings. A p-value of ≤ 0.05 was considered statistically significant. An Excel spreadsheet was used to organize the peer ratings and for computing the averages. The non-parametric statistical analyses, ranked Spearman correlation coefficients, and Wilcoxon tests were performed in Minitab.

Correlations between Peering Ratings and Final Grades

We were interested in determining if a strong correlation existed between peer ratings and final grades in a course. Table 3 shows the results of the Spearman Rank correlation analyses between final course grades and peer- and self-evaluations.

Table 3. Spearman Rank Correlation Coefficients between Final Grades and Evaluations.

Course	Comparison	n	α	r_s Calculated	r_s Table	Sign. (Y/N)
EGR 107	Final grade and self-evaluation PDR	22	0.05	-0.084	0.359	N
EGR 107	Final grade and peer-evaluation PDR	22	0.05	0.573	0.359	Y
EGR 107	Final grade and self-evaluation CDR	18	0.05	-0.054	0.399	N
EGR 107	Final grade and peer-evaluation CDR	17	0.05	0.295	0.412	N
EGC 485/EGR 487	Final grade and self-evaluation	19	0.05	0.320	0.388	N
EGC 485/EGR 487	Final grade and peer-evaluation	19	0.05	0.195	0.388	N
EGC 486/EGR 488	Final grade and self-evaluation	19	0.05	0.282	0.388	N
EGC 486/EGR 488	Final grade and peer-evaluation	19	0.05	0.599	0.388	Y

Based on the Spearman rank correlation coefficient statistical analyses, there was a significant correlation between final grades and peer evaluations for the PDR in the freshman design course (EGR 107). The correlation between final grades and peer evaluations was also significant for the second-semester senior design course (EGC 486/EGR 488). There were no statistically significant correlations between the final grades and self-evaluations for any of the three courses included in this study.

Correlations between Peer and Self-Evaluations

Spearman rank correlation coefficients were calculated between the peer reviews and self-evaluations in the three courses. Table 4 shows the results of the correlations from EGR 107 and the results from EGC 485/EGR 487 and EGC 486/EGR 488.

Table 4. Spearman Rank Correlation Coefficients between Peer Reviews and Self-Evaluations.

Course	Comparison	n	α	r_s Calculated	r_s Table	Sign. (Y/N)
EGR 107	Peer review and self-evaluation PDR	20	0.05	0.003	0.377	N
EGR 107	Peer review and self-evaluation CDR	17	0.05	0.359	0.412	N
EGC 485/EGR 487	Peer review and self-evaluation PDR	19	0.05	0.465	0.388	Y
EGC 486/EGR 488	Peer review and self-evaluation CDR	19	0.05	0.435	0.388	Y

Statistically, there was no significant correlation between peer reviews and self-reviews for both the PDR and CDR assignments in EGR 107. However, at the 0.05 level of significance, there were strong correlations between peer reviews and self-reviews for both the PDR and CDR for the senior design courses.

Presence of Hitchhikers

The term "Hitchhiker" is given to those students who are not responsible team members. They allow the other members to perform their duties hoping that they will receive the same grade as their peers. Students who received less than a "Satisfactory" (translates to a numerical score of 75) average peer rating were categorized a" Hitchhiker". Although most low peer ratings are associated with those students who do not contribute to the welfare of the team, some receive low ratings due to personality conflicts, racial and/or gender prejudice, or reactions to students who try to dominate the team.

In EGR 107, two students received average peer rating scores below 75 on the PDR and only one on the CDR. Therefore, it appears that only three students appeared to be "Hitchhikers" during the EGR 107 course. There may have been a few more, however, all students did not complete the peer evaluations and in some cases, they did not evaluate themselves.

For the senior design courses, there was a greater incidence in the number of students that received average peer rating scores below 75. During the first semester of senior design (EGC 485/EGR 487), five of the nineteen students received unfavorable ratings. At least two of the three-person design teams had an individual that was added to their team at the last minute and perhaps it was not a good fit for the team. During the second semester of senior design (EGC 486/EGR 488), only two of nineteen students received average peer ratings below 75. During senior design, the instructor serves as the project manager who meets with the team at regular intervals throughout the semester. One of the authors (Mines) especially worked with the two teams that had difficulty during the first semester to resolve problems with their teams. Peer ratings for these two groups improved but still resulted in a member from each team being given a low rating. The high incidence of hitchhikers may be due to the fact that the two senior design courses included in this study were the off-cycle sections. In discussions with other professors who teach senior design, there does seem to be a disparity among the design teams that are formed during the off-cycle offering of the senior design sequence. This in part is a result of some students who transfer in from other colleges and are off-cycle and other students are off-cycle due to repeating core engineering courses.

Peer Review and Self-Assessment Comparisons

A comparison of peer and self-assessment of the PDR and CDR in each course was made to determine if students rated themselves differently at each stage. Table 5 shows the results of the Wilcoxon tests.

Table 5. PDR/CDR Comparisons.

Course	Comparison	n	α	p-value
EGR 107	PDR Peer versus CDR Peer	17	0.05	0.6175
EGR 107	PDR Self versus CDR Self	17	0.05	0.3096
EGC 485/EGR 487 EGC 486/EGR 488	PDR Peer versus CDR Peer	19	0.05	0.1444
EGC 485/EGR 487 EGC 486/EGR 488	PDR Self versus CDR Self	19	0.05	0.2933

Based on the data presented in Table 5, there were no significant differences between the ratings performed in the courses between the PDR and CDR phases. This suggests that students consistently evaluated themselves when given a second opportunity.

Effects of Gender and Ethnicity on Ratings

The effects of gender on the peer and self-evaluations could not be ascertained since there was only one female in the first author's EGR 107 section during the spring semester and only two females were enrolled in EGC 485/EGR 487 and EGC 486/EGR 488.

Wilcoxon rank sum tests were performed to determine the effects of ethnicity on student ratings. The results of the Wilcoxon rank sum tests are reported in Tables 6 - 9. The only statistically significant results in Table 6 indicated that minority students gave higher ratings to team members than did non-minority students when evaluating the PDR. Only the average rating given by minorities to non-minorities could be determined since there was no more than one minority in each of the groups. This is with exception to a two-person minority group whose members did not complete the peer review.

Table 6. Wilcoxon Rank Sum Tests Results for Evaluation for EGR 107 PDR.

Average rating given	n	Rating	p-value
By non-minorities By minorities	36 10	84.38 96.25	0.04
To non-minorities To minorities	39 12	86.54 87.50	0.74
By non-minorities to non-minorities By non-minorities to minorities	24 12	82.81 87.50	0.80
By minorities to non-minorities By minorities to minorities	12 0	94.79 0	---

Table 7. Wilcoxon Rank Sum Tests Results for Evaluation for EGR 107 CDR.

Average rating given	n	Rating	p-value
By non-minorities By minorities	31 9	88.71 94.44	0.36
To non-minorities To minorities	30 9	92.50 80.56	0.05
By non-minorities to non-minorities By non-minorities to minorities	21 9	90.28 80.56	0.09
By minorities to non-minorities By minorities to minorities	9 0	94.44 0	---

In Table 7, the only statistically significant result was that higher ratings were given to non-minorities than to minorities in EGR 107 when evaluating the CDR assignment. As with the evaluation of the PDR, only the average rating given by minorities to non-minorities could be calculated since there was only one minority in each group with the exception as was noted above.

Table 8. Wilcoxon Rank Sum Tests Results for Evaluation for First Semester Senior Design -EGC 485/EGR 487

Average rating given	n	Rating	p-value
By non-minorities	22	79.55	0.13
By minorities	21	84.52	
To non-minorities	22	85.80	0.12
To minorities	20	77.50	
By non-minorities to non-minorities	12	83.33	0.14
By non-minorities to minorities	10	75.00	
By minorities to non-minorities	10	87.50	0.60
By minorities to minorities	10	80.00	

Table 9. Wilcoxon Rank Sum Tests Results for Evaluation for Second Semester Senior Design -EGC 486/EGR 488

Average rating given	n	Rating	p-value
By non-minorities	22	85.23	0.80
By minorities	21	87.50	
To non-minorities	22	89.20	0.18
To minorities	20	81.25	
By non-minorities to non-minorities	12	90.63	0.14
By non-minorities to minorities	10	78.75	
By minorities to non-minorities	10	91.25	0.31
By minorities to minorities	10	83.75	

As shown in Table 8, the Wilcoxon tests indicate there were no statistically significant differences in average ratings of the PDR in EGC 485/EGR 487 based on minority status. Similarly, as reported in Table 9, there were no statistically significant differences in average ratings of the CDR in EGC 486/EGR 488 with respect to minority status.

SUMMARY AND CONCLUSIONS

In this study, the application of a peer rating system developed by Professor Rob Brown at the Royal Melbourne Institute of Technology and modified by Professor Rich Felder was assessed in one freshman-level and two senior-level engineering design courses. Spearman rank correlation coefficients and Wilcoxon tests were performed on the data collected. The major conclusions that can be drawn from the data are listed below.

- Based on the Spearman rank correlation coefficient statistical analyses, there were significant correlations between final grades and peer evaluations for the PDR in EGR 107 and between final grades and peer evaluations in the second-semester senior design (EGC 486/EGR 488) courses.

- Statistically, there was no significant correlation between peer reviews and self-review both for the PDR and CDR assignments in EGR 107. However, there were strong correlations between peer reviews and self-reviews for both the PDR and CDR for the senior design courses at a 0.05 level of significance.

- At least three students in EGR 107 and five students in EGC 485/EGR 487 appear to have been "Hitchhikers" since they received unfavorable ratings below 75.

- A comparison of peer and self-assessment of the PDR and CDR in each course indicated that students did not evaluate themselves differently given a second opportunity.

- In EGR 107, there was a significant difference in the ratings given *by* non-minority and those given *by* minority students for the Preliminary Design Report; whereas, for the EGR 107 Critical Design Report, there was a significant difference in ratings given *to* non-minority and those given *to* minority students.

- In both senior design courses, there were no significant differences in the ratings given by or given to non-minorities and minorities.

In our current study, the results are different from those of Ohland and Finelli [7] who found no differences in peer ratings based on race/ethnicity. However the relatively small sample size in the Ohland and Finelli study as well as in our current study suggests that further investigation is necessary before drawing strong conclusions. Using an expanded version of the peer rating form, Layton and Ohland [6] observed no effects relating to race/ethnicity. It is possible that the use of Layton and Ohland's expanded form at our institution might result in a more valid measure of individual performance in cooperative teams. We plan to pursue this alternative in future sections of our freshman and senior design courses.

REFERENCES

[1] Brown, R.W. Autorating: Getting Individual Marks from Team Marks and Enhancing Teamwork, 1995 Frontiers in Education Conference Proceedings. Pittsburgh, IEE/ASEE, November 1995.

[2] Felder, R. M. and Brent, R. Effective Strategies for Cooperative Learning, *J. Cooperative & Collaborative College Teaching*, 2001, 10(2), 69-75.

[3] Hatfield, J.M. and Tester, J. T. Assessing Individual Performance with a Team Using Peer Evaluations, Proceedings 2004 American Society of Engineering Education Annual Conference & Exposition, Session 1725, 2004.

[4] Johnson, D.W., Johnson, R. T., and Smith, K. A., Active *Learning: Cooperative in the College Classroom*, Edina, MN. Interaction Book Co., 1998.

[5] Kaufman, D. B., Felder, R. M., and Fuller, H. Peer Ratings in Cooperative Learning Teams, Proceedings of the 1999 Annual ASEE Meeting, ASEE, June 1999

[6] Layton, R.A. and Ohland, M.W. Peer Ratings Revisited: Focus on Teamwork, Not Ability, Proc. American Society of Engineering Education, Albuquerque, June 2001.

[7] Ohland, M.W. and Finelli, C.J. Peer Evaluation in a Mandatory Cooperative Education Environment, Proc. American Society of Engineering Education, Albuquerque, June 2001.

[8] Ohland, M.W. and Layton, R.A. Comparing the Reliability of Two Peer Evaluations Instruments, Proc. American Society of Engineering Education, St. Louis, June 2000.

Richard O. Mines, Jr.

Dr. Mines is Director of MSE/MS Programs and Professor of Environmental Engineering at Mercer University. He is a graduate of the Virginia Military Institute with a BS in Civil Engineering, a Masters in Civil Engineering from the University of Virginia, and a PhD in Civil Engineering from Virginia Tech. Dr. Mines has over six years of

experience with CH_2M Hill and BLACK & VEATCH consultants and twenty-five years of teaching experience. He is a registered Professional Engineer in New Mexico. Dr. Mines has authored/co-authored over 100 technical and educational papers. He is the primary author of *Introduction to Environmental Engineering* published by Prentice-Hall. His research interests lie in water and wastewater treatment, modeling of bionutrient removal systems, and enhancing learning in the classroom. Dr. Mines is an active member of ASEE and a Fellow in ASCE.

Joan M. Burtner

Dr. Joan Burtner is an associate professor of Industrial Engineering and Industrial Management at Mercer University. She is a Certified Quality Engineer and a senior member of the American Society for Quality and the Institute of Industrial Engineers. She teaches courses in statistics, statistical quality control, quality management, quality engineering, reliability, and healthcare performance improvement.

EGR 107 Introduction to Engineering Design
Spring 2012 Competition Project
Peer Rating of Team Members

Please print the names of all of your team members, INCLUDING YOURSELF, and rate the degree to which each member fulfilled his/her responsibilities in completing course assignments during the semester. The possible ratings are as follows:

Excellent	Consistently went above and beyond carried more than his/her fair share of the load.
Very Good	Consistently did what he/she was supposed to do, very well-prepared and cooperative.
Satisfactory	Usually did what he/she was supposed to do, acceptably prepared and cooperative.
Ordinary	Often did what he/she was supposed to do, minimally prepared and cooperative.
Marginal	Sometimes failed to show up or complete assignments, rarely prepared.
Deficient	Often failed to show up or complete assignments, rarely prepared.
Unsatisfactory	Consistently failed to show up or complete assignments, unprepared.
Superficial	Practically no participation.
No Show	No participation at all.

These ratings should reflect each individual's level of participation and effort and sense of responsibility, not his or her academic ability.

Print Name of Team Member	**Rating**

I understand that Mercer University has an honor code and that cheating or dishonesty of any kind is unacceptable. Furthermore, I certify that this submittal is a fair evaluation of the effort and participation of each individual team member including my own.

Signed: _____

Print Name:

COMMENTS:

Figure 1. Typical Peer Review Evaluation Form.