Evaluating the effect of different teamwork training interventions on the quality of peer evaluations

Daniel M. Ferguson
Engineering Education
Purdue University
West Lafayette, IN USA
dfergus@purdue.edu

Matthew W. Ohland
Engineering Education
Purdue University
West Lafayette, IN USA
ohland@purdue.edu

Chad Lally
Engineering Education
Purdue University
West Lafayette, IN USA
clally@purdue.edu

Hilda Ibriga Somnooma
Statistics
Purdue University
West Lafayette, IN USA
hibriga@purdue.edu

Yuchen Cao
Statistics
Purdue University
West Lafayette, IN USA
cao273@purdue.edu

***Abstract* — This research-to-practice paper describes the results of 12 experiments using different combinations of peer evaluation trainings on the quality of peer evaluations of teamwork skills provided by first year engineering (FYE) students in an introduction to engineering course in a large Midwestern University. Our research question was: Which combination of peer evaluation trainings most significantly improves the quality of peer evaluations in this FYE course as measured by using data collected by the CATME (Comprehensive Assessment of Team Member Effectiveness) system's peer evaluation exercise? The five training interventions combined in the 12 experiments were Frame-of-reference teamwork training, Video-based teamwork training materials, Rater practice, Rater accuracy training and In-class teamwork reflections. The training intervention combinations that included Video based training and Rater accuracy training had the most significant effects on improving rater and target variances or the quality of peer evaluations.**

Keywords— teamwork training, peer evaluations, peer ratings

I. Introduction

An engineer's ability to work in teams is critical to their engineering career and often a significant factor in a corporate hiring process. Recognizing this need most U.S. undergraduate engineering programs use team or project-based courses in their curricula. These programs are also motivated by the ABET teamwork learning requirements [1]. Hundreds of these engineering programs also use a teamwork formation and assessment system, CATME (Comprehensive Assessment of Team Member Effectiveness), that asks students to assess the teamwork behavior of their peers [2].

Teamwork is defined as "cooperative or coordinated effort on the part of a group of persons acting together" [3]. Companies, as diverse as NASA [4], look for teamwork skills when hiring new employees and Chen argues that many students entering into the workplace lack key teamwork skills [5]. In addition teamwork skills training has become more prevalent throughout college programs due to the addition of teamwork accreditation requirements not only in engineering [1], but in business [3] pharmacy and other accreditation bodies and employers recognize the importance of teamwork skills [4, 5].

Our research question was: Which combination of peer evaluation trainings improves the quality of peer evaluations?

CATME is a web-based tool created for academic teamwork environments and is used to help instructors create teams more likely to be effective based on scientific research on team effectiveness and practical considerations. CATME also assists students in giving peer feedback to their team members [6]. CATME is constructed around five behavioral dimensions[7, 8]. These dimensions are defined as follows:

Having (H) relevant KSAs refers to the base knowledge of individual team members.

Contributing (C) to the Team's Work is being able to add value to your team's work/project.

Interacting (I) with teammates refers to the way individuals communicate within their teams.

Keeping (K) the team on track is similar to being a timekeeper.

Expecting (E) quality is taking expectations to the next level and working collaboratively to produce the best possible team outcomes.

Every aspect of the five teamwork dimensions is equally important to team success and a critical element in the peer evaluations [7].

This work is supported by the USA National Science Foundation through grant 1431694.

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

Peer evaluations (PEs) facilitate better learning outcomes in upper level education, encouraging students to continue their engagement with constructive team behavior in future team activities [9]. A behavioral peer evaluation is an evaluation of an individual's contribution to a work activity by their peers, either as students or as professionals in industry [10]. Peer evaluations help to teach individuals how to act in teams and how to evaluate one another's performance. As facilitated by the CATME system, peer evaluations potentially point out behaviors where an individual excels and areas where he/she may need improvement [11, 12].

In this paper we discuss the research population, the types of experimental training interventions, the analysis processes used, and our findings and conclusions.

II. Research Methods

Research Population

All 16 sections of First-Year Engineering (FYE) students at a major Midwestern university in the fall of 2016 were involved in one of the 12 training experiments as described in Table 1. FYE students at this institution in fall 2016 numbered over 1,600 including 23% women and 6% minorities. Each FYE section was composed of up to 120 students. All 16 FYE sections completed the identical syllabus. Individual results from peer evaluations were a component of the individual's FYE course grade in all 16 FYE sections. Twelve different instructors were the instructors of record for the experimental and control FYE sections.

Experimental Training Interventions

The five different experimental trainings included both in-class and outside-class activities. They were introduced by section instructors as appropriate and supported as required by each section's teaching team. Teaching teams were a graduate teaching assistant, five undergraduate peer teachers and the instructor. The five peer evaluation training activities were:

- ***Frame-of-reference (FOR) teamwork training*** materials which explain and provide practice in the peer evaluation rating schema of each CATME teamwork dimension.
- ***Video-based teamwork training*** materials which contain demonstrations of teamwork behavior for each CATME dimension, using scenes from popular movies, and in-class exercises.
- ***Rater practice training*** is an on-line simulation that presents a written scenario describing the teamwork behaviors of three students. Students identify which behaviors described for a hypothetical student are related to each teamwork dimension and then rate each hypothetical student on all five CATME dimensions.
- ***Rater accuracy training*** presents three problematic peer evaluation rating patterns. Students engage in discussions as a team in class as to what could cause those rating patterns and the consequences of rating in that manner
- ***Feedback exercises[FE]*** is a discussion of CATME feedback, but one that does not explicitly attempt to promote more accurate rating.
- ***In-class reflections*** uses written statements describing student behaviors. Students as a team discuss how to classify and rate these word descriptions of student behavior in a peer evaluation

Experimental Design

The five Experimental Training Interventions were combined in the 10 experimental combinations described in Table I with the remaining FYE sections serving as the control group. No peer evaluation training activities were scheduled for the control sections. This pattern of interventions in the experiment was designed to separate out the effects of the interventions or combinations of interventions that we most wanted to test. We were prompted in this experimental design by preliminary experimental results gathered in the summer of 2016 [13] and results from a predecessor study in the Fall of 2015 [14].

TABLE I. Experimental Peer Evaluation Training Plan

Training	Abbreviation
FOR + feedback exercise	F + FE
FOR + feedback exercise + rater practice training	F + FE+ RP
FOR + rater accuracy training	F + RA
FOR + rater accuracy training + rater practice	F + RA + RP
In-class reflection + feedback exercise	TR + FE
In-class reflection + feedback exercise + rater practice training	TR + FE + RP
In-class reflection + rater accuracy training	TR + RA
In-class reflection + rater accuracy training + rater practice training	TR + RA + RP
Video-based training + feedback exercise	V + FE
Video-based training + feedback exercise + rater practice training	V + FE + RP

Experimental Procedure

FYE students were required to do three CATME Peer Evaluations (PEs) of their teammates per semester.

These PEs were scheduled concurrently with the completion of three major team project milestones. The FYE team projects are also designed to require substantial team interactions.

CATME peer evaluations are conducted online and students receive peer evaluation feedback in an online form. Peer evaluation feedback is shown as pointers to word descriptions of behaviors as shown in Figure 1 for the CATME dimension: Contributing. No numbers are provided to students and behavior descriptions appropriate to improving their average ratings are also cited. For each of the five CATME peer rating dimensions students see their own ratings of themselves, their average ratings by their teammates and the average ratings for all team members as shown in Figure 1.

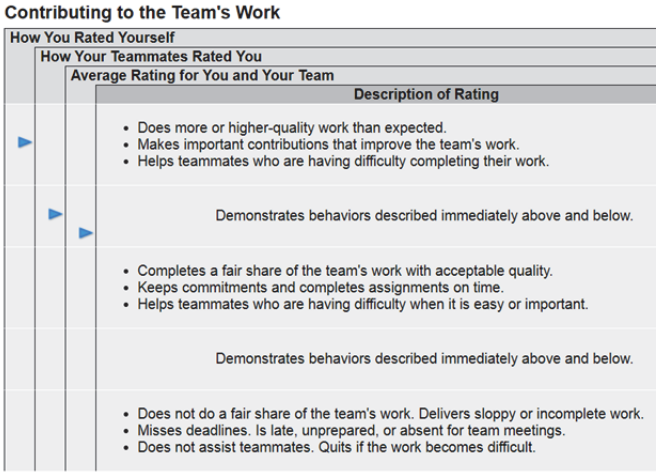

Figure 1: Sample feedback screen from the CATME Peer Evaluation tool

Students do not see their actual numeric ratings by their teammates as all ratings are held confidential, although the numeric rating data is provided to instructors for review as shown in Figure 2.

				Rater 1				
Student Name	Student ID	Team ID	Rater #	C	I	K	E	H
Brown, Carson	1010101	1	1	5	5	5	5	5
Miller, Chase	1010111	1	2	5	5	5	5	5
smith, addison	1	1	3	5	5	5	5	5
williams, ali	2	1	4	5	5	5	5	5

Figure 2: Ratings from sample student for a sample CATME Peer Evaluation

Data Preparation

For the analysis of the peer evaluation data a standard deviation of each student's ratings of themselves and each team member across the five CATME dimensions was calculated. For a team of 4 students and the 5 CATME dimensions this is a matrix of 20 data points for Carson Brown as shown in Figure 2 (rater 1=Carson Brown and Carson is not a real student).

Each FYE section had 30 teams and the standard deviations for each student of their ratings of team members across all 5 CATME dimensions were calculated. This standard deviation is referred to as the dispersion factor for the CATME dimensions for that student[[15]]. Figure 2 shows the ratings of one student for themselves (Brown Carson) and their other team members (rows 2, 3 and 4). The resulting standard deviation matrix for all students across all the experimental sections and the control sections was placed in an analysis matrix. A higher standard deviation means a student is differentiating the ratings across the five CATME dimensions. This calculation procedure was repeated for each of the three FYE peer evaluations for all 16 sections totaling approximately 1600 students and 400 teams. Similar calculations were done to create a mean absolute score for each student and the mean scores were placed in a matrix of mean scores for each student in both samples.

Data Analysis

An improvement in the quality of a peer rating in this paper is first defined as having a larger dispersion (measured by standard deviation) in the ratings across teamwork dimensions [16]. A larger dispersion means that an individual is differentiating their perception of a teammate's contribution on CATME teamwork dimensions. While it is clearly possible that an individual exhibits the same level of performance across all five CATME dimensions, it is not likely, and similarly it is unlikely that there are no differences in teamwork performance among all team members on any of the five CATME dimensions as illustrated in Figure 2 [16, 17].

For each of the 10 experimental design sections their dispersion matrixes were then compared to the control section dispersion matrix for each PE (1, 2, 3) or to a prior PE using a repeated measure ANOVA. The purpose of this analysis was to determine if any significant change in dispersion had occurred as a result of any of the 10 training interventions.

Second, we examined self and peer ratings for convergence in ratings. Convergence analysis examines the difference between self and peer ratings. A higher quality rating is one where a rater gives themselves a rating that is similar to the rating given them by their teammates Further, if the two ratings converge over multiple PEs, individuals were either changing their perceptions of team work behavior and/or modifying their teamwork behavior, with both types of changes viewed as positive teamwork training or experience results. Similar to dispersion, we compared the self and

peer ratings standard deviation matrices for each of the 10 experimental sections with the control sections or to a prior PE using a repeated measure ANOVA.

Third, we examined the variances in self and peer ratings using the Social Relations Model (SRM) [18-20], SRM identifies changes in different variance components:

1. how individuals perceive others (Rater effect), **Rater Effect** measures how consistently an individual rates his/her teammates. A **larger** Rater Effect would mean that on average an individual gives all his/her teammates the same rating. (unreliable rating)
2. how individuals are perceived by others (Target Effect), **Target Effect** measures how consistently an individual is rated by his/her teammates. A **larger** Target Effect would mean that on average an individual receives similar ratings from all his/her teammates. (reliable rating)
3. the extent to which the ratings give and received by each pair are related to each other (Relationship Effect). **Relationship Effect** measures the uniqueness of relationships after the rater and target variance have already been accounted for. A **larger** Relationship Effect would indicate that ratings are based on personal interaction inside or outside of team activities.

III. Findings

Table II displays the only experiments showing positive dispersion ANOVA results from the training experiments. The significant changes in mean dispersion were the V+RA section between PE1 and PE2 ($p<0.01$). The V+RA+RP section also showed significant changes in dispersion but it only lasted for one peer evaluation (PE1 to PE2):

TABLE II. Sample of Dispersion Analysis Results

Experiment	PE	Experiment	PE	Mean Dispersion	p value exp to cntrl
V + RA	2	V + RA	1	-0.1041	0.0096
V + RA	3	V + RA	2	-0.00345	1.0000
V + RA + RP	2	V + RA + RP	1	-0.09564	0.0410
V + RA + RP	3	V + RA + RP	2	-0.02451	1.0000

Table III shows a small sample of the convergence analysis results. Overall the standard deviation of self and peer ratings started out converged ($p > .05$) and stayed converged. Mean ratings were also converged compared to control sections in all but 4 experimental sections for all PEs. However the sections with V+RA and V+RA+RP experiments moved to convergence in the 2nd PE.

Table IV shows a sample of the results of the SRM analysis. We see in Table IV significant and sustained improvements in Rater, Target and Relationship Variance.: In particular, it appears that adding RA, RP or PE to V will produce significant and sustained improvements in target and rater variance, that is, better quality peer feedback.

TABLE III. Sample of Convergence Results

Experimental Sample	Peer Review Time	Difference Self to Peer std dev	p	Difference average	p
V + RA	1	-0.03457	0.21	0.1220	<0.01
V + RA	2	-0.00312	0.91	0.07388	0.12
V + RA	3	-0.03763	0.17	0.1234	<0.01
V + RA + RP	1	-0.04868	0.09	0.1135	0.02
V + RA + RP	2	-0.03716	0.20	0.06644	0.19
V + RA + RP	3	-0.02382	0.41	0.01644	0.75

TABLE IV. Sample of SRM Results

Sample	V+PE			V+RA+RP			V+RA		
Peer Evaluation	PE1	PE2	PE3	PE1	PE2	PE3	PE1	PE2	PE3
Rater Variance	65%	35%	34%	56%	22%	40%	59%	56%	42%
Target Variance	7%	21%	23%	14%	35%	21%	16%	7%	30%
Relationship Variance	28%	44%	43%	30%	43%	39%	25%	38%	28%
Notes	13 teams, 52 individuals	10 teams, 40 individuals	25 teams, 100 individuals	33 teams, 132 individuals	25 teams, 100 individuals	45 teams, 180 individuals	47 teams, 188 individuals	27 teams, 108 individuals	45 teams, 180 individuals

IV. Conclusion

Our principal conclusion is that CATME Video trainings seem to have the most significant effects on improving the quality of peer ratings for FYE teams. However, Frame of Reference training and Team Reflections in class or Rater Accuracy discussions also had some positive effects on peer evaluation quality.

Some limitations, however, must be noted. In these experiments, all participants were first-year engineering students. Different results may be obtained with senior capstone STEM students or other non-STEM majors who complete peer evaluations. More than 10 instructors were involved in these experiments, so fidelity of implementation is a concern. Other instructors may experience results that differ from our research findings.

V. Further Research

We plan to extend these training experiments to capstone courses and other STEM disciplines. In these peer ratings training experiments we are looking to identify training strategies that best provide increased dispersion, improved convergence or improvements in the SRM Target and Rater variances in order to improve the constructive feedback received by students [21-23].

References

[1] *2010-2011 Criteria for Accrediting Applied Science Programs,* Applied Science Accreditation Commission, 2009.

[2] CATME SMARTER Teamwork. *info.catme.org*.

[3] AACSB International, "Accreditation Standards," ed: AACSB International, 2013.

[4] Calloway, " A Report on Recruiters' Perceptions of Undergraduate Business Schools and Students," School of Business and Accountancy of Wake Forest University2004.

[5] Elrick L. , "The Importance of Teamwork Skills in Work & School," Rasmussen College2015.

[6] R. A. Layton, M. L. Loughry, M. W. Ohland, and G. D. Ricco, "Design and validation of a web-based system for assigning members to teams using instructor-specified criteria," *Advances in Engineering Education,* vol. 2 pp. 1-28, 2010.

[7] M. W. Ohland, M. L. Loughry, D. J. Woehr, C. J. Finelli, L. G. Bullard, R. M. Felder, R. A. Layton, H. R. Pomeranz, and D. 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, pp. 609-630, 2012.

[8] M. L. Loughry, Ohland M.W. , and D. D. Moore, "Development of a Theory-Based Assessment of Team Member Effectiveness," *Educational and Psychological Measurement,* vol. 67, pp. 505-524, 2007.

[9] E. J. Thomas, "Improving teamwork in healthcare: current approaches and the path forward," *BMJ quality & safety,* 2011.

[10] Dictionary.com Unabridged. (2016, April 05, 2016). *peer review (http://www.dictionary.com/browse/frame-of-reference ed.)*. Available: http://www.dictionary.com/browse/peer-review?s=t

[11] M. L. Loughry, M. L. Ohland, and D. J. Woehr, "Assessing teamwork skills for assurance of learning using CATME Team Tools," *Journal of Marketing Education,* vol. 36, pp. 5-19, 2014.

[12] M. W. Ohland, R. A. Layton, M. L. Loughry, and A. G. Yuhasz, "Effects of behavioral anchors on peer evaluation reliability," *Journal of Engineering Education,* vol. 94, pp. 319-326, 2005.

[13] B. Natalia, C. Lally, Y. Cao, and D. Ferguson, "Evaluation of Training in the CATME Peer Evaluation Schema," presented at the The Purdue Undergraduate Research Symposium, West Lafayette,IN, 2018.

[14] Ferguson Daniel, Chad Lally, Somnooma Hilda, Murch Olivia, and Ohland Matthew W, "Using Frame-of-Reference Training to Improve the Dispersion of Peer Ratings in Teams," in *Frontiers in Education*, Eire PA, October 2016.

[15] D. C. Howell, *Statistical Methods for Psychology*: Cengage Learning, 2012.

[16] T. Poling, D. J. Woehr, L. M. Arciniega, and A. Gorman, "The impact of personality and value diversity on team performance.," in *Annual Meeting for the Society for Industrial and Organizational Psychology*, Dallas, TX, 2006.

[17] C. J. Resick, M. W. Dickson, J. K. Mitchelson, L. K. Allison, and M. A. Clark, "Team composition, cognition, and effectiveness: Examining mental model similarity and accuracy," *Group Dynamics: Theory, Research, and Practice,* vol. 14, pp. 174-191, 2010.

[18] Cohen J. , *Statistical power analyses for the social sciences*. Hillsdale, NJ: Lawrence Erlbauni Associates, 1988.

[19] M. D. Back and D. A. Kenny, "The social relations model: How to understand dyadic processes," *Social and Personality Psychology Compass,* vol. 4, pp. 855-870, 2010.

[20] Loignon Andrew, Woehr David J., Shumski Thomas Jane, Misty Loughry, Ohland Matthew W., and Ferguson Daniel, "Facilitating Peer Evaluation in Team Contexts: The Impact of Frame-of-Reference Rater Training," *Academy of Management Learning & Education,* vol. 16, pp. 562-579, 2017.

[21] H. E. Tinsley and D. J. Weiss, "Interrater reliability and agreement of subjective judgments," *Journal of Counseling Psychology,* vol. 22, p. 358, 1975.

[22] H. Shi, Ferguson D.M., Beagley J., and Huyck M., "Improving inter-rater reliability used to measure learning outcomes," presented at the 38th IEEE/ASEE Frontiers in Education Conference, Saratoga Springs, NY, 2008.

[23] Ferguson Daniel M., Govekar M., and Stype A., "Quality and Consistency in Idea Pitch, Research Proposal and Business Plan Competition Judging," presented at the ASEE Annual Conference Louisville, KY, 2010.