Comparing Peer-to-Peer Written Comments and Teamwork Peer Evaluations.

Dr. Catherine E. Brawner, Research Triangle Educational Consultants

Catherine E. Brawner is President of Research Triangle Educational Consultants. She received her Ph.D.in Educational Research and Policy Analysis from NC State University in 1996. She also has an MBA from Indiana University (Bloomington) and a bachelor's degree from Duke University. She specializes in evaluation and research in engineering education, computer science education, teacher education, and technology education. Dr. Brawner is a founding member and former treasurer of Research Triangle Park Evaluators, an American Evaluation Association affiliate organization and is a member of the American Educational Research Association and American Evaluation Association, in addition to ASEE. Dr. Brawner is also an Extension Services Consultant for the National Center for Women in Information Technology (NCWIT) and, in that role, advises computer science and engineering departments on diversifying their undergraduate student population. She remains an active researcher, including studying academic policies, gender and ethnicity issues, transfers, and matriculation models with MIDFIELD as well as student veterans in engineering. Her evaluation work includes evaluating teamwork models, statewide pre-college math initiatives, teacher and faculty professional development programs, and S-STEM programs.

Ms. Olivia W. Murch, Purdue University

Senior at Purdue University pursuing a Bachelor of Science degree in Biological, Food Process, Engineering. Currently conducting research under Dr. Ferguson through Engineering Education.

Dr. Daniel M. Ferguson, Purdue University, West Lafayette (College of Engineering)

Daniel M. Ferguson is CATME Managing Director and a research associate at Purdue University. Prior to coming to Purdue he was Assistant Professor of Entrepreneurship at Ohio Northern University. Before assuming that position he was Associate Director of the Inter-Professional Studies Program [IPRO] and Senior Lecturer at Illinois Institute of Technology and involved in research in service learning, assessment processes and interventions aimed at improving learning objective attainment. Prior to his University assignments he was the Founder and CEO of The EDI Group, Ltd. and The EDI Group Canada, Ltd, independent professional services companies specializing in B2B electronic commerce and electronic data interchange. The EDI Group companies conducted syndicated market research, offered educational seminars and conferences and published The Journal of Electronic Commerce. He was also a Vice President at the First National Bank of Chicago [now J.P. Morgan Chase], where he founded and managed the bank's market leading professional Cash Management Consulting Group, initiated the bank's non-credit service product management organization and profit center profitability programs and was instrumental in the breakthrough EDI/EFT payment system implemented by General Motors. Dr. Ferguson is a graduate of Notre Dame, Stanford and Purdue Universities, a special edition editor of the Journal of Engineering Entrepreneurship and a member of Tau Beta Pi.

Dr. Matthew W. Ohland, Purdue University, West Lafayette (College of Engineering)

Matthew W. Ohland is Professor of Engineering Education at Purdue University. He has degrees from Swarthmore College, Rensselaer Polytechnic Institute, and the University of Florida. His research on the longitudinal study of engineering students, team assignment, peer evaluation, and active and collaborative teaching methods has been supported by the National Science Foundation and the Sloan Foundation and his team received Best Paper awards from the Journal of Engineering Education in 2008 and 2011 and from the IEEE Transactions on Education in 2011 and 2015. Dr. Ohland is Chair of the IEEE Curriculum and Pedagogy Committee and an ABET Program Evaluator for ASEE. He was the 2002–2006 President of Tau Beta Pi and is a Fellow of the ASEE, IEEE, and AAAS.

Comparing Peer Ratings of Teamwork Behavior with Peer-to-Peer Written Comments

Abstract

This paper investigates the relationship between peer to peer comments and behavioral ratings of teamwork behavior for first year engineering students using the CATME peer evaluation system. CATME allows team members to rate themselves and each other on five research-based dimensions essential for good team functioning. The five dimensions are: Contributing to the Team's Work (C), Interacting with Teammates (I), Keeping the Team on Track (K), Expecting Quality (E) and Having Related Knowledge, Skills, and Abilities (H). During the spring semester of 2016, students rated themselves and their teammates on these five dimensions and were asked to explain their ratings of themselves and their peers. We find that, in general, the comments were focused primarily on Contributing and secondarily on Having Related Knowledge, Skills, and Abilities; not all five CATME dimensions. However, when detailed comments are given, they often provide additional insights into peer ratings and explanations for the CATME exception codes. These insights into team functional or dysfunctional behavior provide information to the instructor that goes well beyond what can be obtained from the peer ratings alone.

1. Introduction

Working in teams is widely viewed as a key skill for having a successful career. However, effective team behavior does not necessarily come naturally to many students. In engineering education, developing teamwork and communication skills, among other things, are student outcomes necessary for accreditation [1]. Many instructors at four-year colleges use small groups or teams in their courses as a way to help students develop these skills. The development of these skills is enhanced through constructive feedback on teamwork behavior from peers so that students can learn and improve how they perform in teams.

This paper examines the use of written peer-to-peer comments as a tool to enhance the quality of peer evaluation using the Comprehensive Assessment of Team Member Effectiveness (CATME) peer evaluation system. CATME (www.catme.org) is currently used by over 16,000 instructors, across multiple disciplines, in over 2,000 institutions worldwide. CATME is used to develop and support teamwork in higher education. It focuses on creating diverse teams and allowing self and peer evaluations [2]. Teamwork skills in our study are defined and measured as the dimensions of teamwork in the CATME.

CATME includes five common measures of teamwork behavior on which team members are asked to rate themselves and their teammates using a behaviorally anchored rating scale on each dimension. These teamwork dimensions fall into 5 categories:

- Contributing (C) to the Team's Work is being able to add value to a team's work/project. Team members are rated on how well they meet their commitments, do their share of the work, and help their teammates.
- Interacting (I) with Teammates refers to how individuals communicate within their teams. It includes encouraging teammates, communicating ideas clearly, and listening respectfully to others' ideas.
- Keeping the Team on Track (K) is being aware of milestones and deadlines and ensuring that the team is making appropriate progress.
- Expecting Quality (E) is taking steps to ensure that the team meets or exceeds all requirements for project outcomes.
- Having (H) Relevant Knowledge, Skills, or Attributes (KSAs) refers to the base knowledge of individual team members. It means having the required KSAs to solve the problems at hand or being willing to learn the KSAs an individual lacks. [3]

In this work in progress, we explore how peer-to-peer comments inform the ratings that students give to each other on these five CATME dimensions and what might be learned by instructors from the comments. Ideally, when students are asked to explain their ratings, they are better able justify the ratings they give and distinguish between different levels of teammate performance. After a brief literature review on the use of peer ratings in business and educational settings and behaviorally anchored rating scales, we discuss what can be learned from the peer comments that were given by students in a first year engineering (FYE) class. We compare student ratings with the comments they provided as well as our ratings of the same student teamming behavior based on their comments.

2. Literature Review

2.1 Peer evaluation in business settings

Peer evaluations are used in professional work environments, including business and health settings, and in education [4]. In the workplace, peer reviews have been used to assess job performance. One example of peer evaluations with written comments in the workplace is 360-degree feedback which is often used in Fortune 100 companies. With this tool, managers are rated within the company by peers, subordinates, supervisors, and occasionally also by customers. The feedback is written, but generally anonymous [5]. Two factors that affect the quality of 360-degree feedback are relevant content and accountability [6]. Relevant content is important to align all participants on the values of the company and then use these values to rate managers. Accountability, the other factor that affects the quality of feedback, can drive behavioral change [6]. However, accountability alone does not always cause performance improvement. Further intervention from a superior may be necessary [7].

Viswesvaran and colleagues [8] conducted a meta-analysis of feedback given by both peers and supervisors on ten dimensions: overall job performance, productivity, effort, interpersonal competence, administrative competence, quality, job knowledge, leadership, compliance, and acceptance of authority. They found that there was convergence between supervisor and peer ratings on constructs that were relatively easy to rate, such as overall performance, productivity,

and job knowledge; but they found more disagreement between supervisors and peers on constructs that were more difficult to rate such as administrative and interpersonal competence.

2.2 Peer evaluation in classroom settings

From their meta-analysis of feedback literature, Hattie and Timperley [9] suggested a model where effective feedback answers three principal questions: "1) Where am I going?; 2) How am I going?; and 3) Where to next?" Within this model, each question was addressed on four different levels – the task level, process level, self-regulation level, and self level. Effective feedback requires specific goals to which the feedback pertains. The "how am I going" question is best answered with feedback about performance as it relates to a specific standard of behavior and success or failure on a task. Process level feedback gives students cues to let them know if they have the required competence and helps them move toward self regulation and higher confidence and competence in accomplishing the task.

The use of written peer and TA feedback has been used in a classroom setting, among other places, for first year engineering students. Rogers and colleagues [10] investigated how feedback affected teams and project outcomes. They found that students felt that the feedback was of low quality when not enough time and thought was put into the peer review. They believe that peer reviews require focus and time commitment. Additionally, students valued the feedback from TAs more than that of their peers because they felt it had more of an impact on their grades. The students did not believe the peer reviews helped to increase their grades or create change within their teams, but the results of their project indicated otherwise.

2.3 Behaviorally Anchored Rating Scales

Behaviorally Anchored Rating Scales (BARS) are evaluations based on performance, rather than levels of agreement, such as a Likert scale. BARS utilizes multiple dimensions as a feedback tool. The ratings between dimensions do not depend on one another. Thus, a person can be rated very low in one dimension while being rated very highly in another. BARS is most effective when the rating scale is behavior specific. Each number along the rating scale has a specific behavior associated with it. This causes BARS to be job oriented instead of trait oriented [11]. BARS is comparable to other forms of performance evaluation [12].

CATME uses a behaviorally anchored rating scale, the rubric for which is shown in the Appendix. Top performance on the contributing dimension (C5) is described as including "Does more or higher-quality work than expected" whereas expected performance (C3) includes "Completes a fair share of work with acceptable quality" and poor performance (C1) includes "Does not do a fair share of the team's work." The other four dimensions provide similar guidance for behavioral ratings. In the CATME system, raters do not see the numbers to which the behaviors are converted, only the behaviors, and are asked to rate themselves and their peers using these behaviors as a guide.

In addition to the ratings of team members on their behavior, CATME also allows for comments about team members and their peers to support the ratings. This provides both the relevance and accountability that is called for in best practices in business and educational settings to enhance

team performance and outcomes. This paper discusses the quality of peer feedback and how it informs peer ratings in an educational setting.

3. Methods

The data for this analysis was taken from students enrolled in the second Introduction to Engineering (Engineering 2) course that used team-based learning assignments at a large Midwestern university in the Spring of 2016. This course had 15 sections with a total of 427 teams, each usually having four students. Team members were asked on four occasions to rate themselves and each other using behavioral ratings on the five CATME dimensions and to write comments about their teamwork behavior on the third and fourth of these occasions. They were familiar with CATME, having also been asked to provide ratings, but not peer-to-peer comments, during the first Introduction to Engineering (Engineering 1) course. We used the third rating event – the first one with comments – for this preliminary analysis.

The first analytical filter was to choose teams with four members who had written about themselves and each other in both the third and fourth time periods. Some teams only had three members and drops, absences, and non-compliance reduced the availability of comments for many teams. This reduced the number of teams for analysis to 197. The first team in the first section meeting these criteria was used to calibrate author Brawner's ratings with author Ohland, the CATME Principal Investigator who is considered a subject matter expert. Each comment within a team was coded using the rubric (see Appendix) developed by the CATME project team and provided to the students along with Frame of Reference Training on how to properly rate behaviors. Once the comment codes had been calibrated, she continued to code 3 randomly selected teams per section to yield 46 coded teams, including the calibration team. For this work in progress, author Murch, who had used CATME during her time as an FYE student, coded one randomly selected team from among the three coded teams in each section (15 teams).

Of the 1200 opportunities for agreement, the two coders agreed 77% of the time on both the dimension to be coded and the level (i.e., Contributing dimension, level 3 or C3), including the choice of no code. We agreed on the dimension, but not the level an additional 11% of the time, and of those, 68% were within 1 (e.g., one person gave a level 3 and the other a level 4 on the same dimension). Thus, the coders were in agreement or within 1on a 5 point scale 85% of the time which exceeds the 80% threshhold normally considered acceptable for inter-coder reliability [13, p. 150]. The remaining 12% of the time, one of us, but not the other, assigned a code in a particular dimension to a comment. The final expert codes were then compared with the actual ratings that the team members gave each other to determine if the codes and comments were aligned.

When deciding which comments to present here, we looked not only at the comments themselves, but also our ratings of the comments and the student ratings of themselves and each other. We studied the ratings that the students gave each other and looked for patterns such as two students on a team rating each other low, but other team members high and suggested names for these patterns (e.g., 1 and 4 don't seem to like each other). We then overlayed the CATME exception codes and found both areas of agreement (e.g., "conflict") and areas where our intuitive sense did not align with the CATME algorithms. This helped us to identify teams or

students that were worthy of further exploration. We created matrixes of the comments of the identified teams along the lines advocated by Miles, Huberman, and Saldaña [14] to display the comments that would prove fruitful for further analysis.

We present the actual comments as written by the students, with edits for grammar and spelling, using culturally similar pseudonyms for their names.

4. Findings

4.1 Content of the comments

The Contributing (C) dimension was the easiest for the coders to discern from the comments. Both coded more than 80% of the 240 opportunities on this dimension. A student's contribution can generally be gleaned from even the briefest of comments such as this one:

> Steve completed the algorithm that we chose to go with for Milestone 2. He has also updated it slightly and described what he changed in Milestone 3.

We each gave this comment a C3 (Contributing, level 3) not so much because it specifically addresses the elements in the rubric, but because we inferred that this work constituted completing a fair share of the team's work with acceptable quality. It was clear that the student did *something* to contribute to the team, though not anything particularly special nor anything detrimental. We also knew from the context of the comments as a whole what the assignment was and how much effort might be required to complete it, and thus what would be a reasonable contribution.

Having the KSA's to be an effective team member was the second easiest dimension for us to glean from the comments (e.g., "*Anthony was the one who came up with the algorithm which we chose as the final algorithm. He contributed immensely in improving the algorithm and reducing the run time of the program*" yielded an H5), but we could only do so about half as often as we could discern comments about the contributing dimension. Comments rarely addressed Keeping the Team on Track or Expecting Quality as we could not code either of those dimensions more than 39 times out of 240 comments.

4.2 Level of detail in comments

Of the 1200 possible opportunities to provide a code (i.e., 16 comments per team for 15 teams on five dimensions), the two analysts coded 443 and 407 respectively, indicating that the comments were sufficiently detailed to make a judgment about student behavior on each dimension only about a third of the time. Each team offers 16 x 5 or 80 possibilities for a code, but the most that we could provide based on the comments for any team was 49 while the least was 19. On average, we were able to apply 28 codes per team. We were both able to apply a code for each dimension in only two of the 240 comments that we coded. Two other times, there were 9 codes between us and three other times, there were 8 codes between us.

Our inability to glean ratings from the comments would indicate that students may not have been given instructions to "explain your ratings." In fact, the availability of peer-to-peer comments in

CATME was simply announced in a meeting of Engineering 2 course instructors; they chose how to address it with their students. While the curriculum includes instruction on providing constructive feedback in teams, the connection between that part of the curriculum may or may not have been made explicit at the time peer-to-peer comments were introduced.

4.2.1 Lack of detail

Lack of detail in comments reflects a lack of discrimination in students' ratings of themselves and each other. One student received 18 3's of a possible 20 ratings (4 teammates rating 5 dimensions). The comments provided little insight into this unremarkable behavior.

Table 1 – Comments about Gwen Yield Little Information

By Gwen	By Teammate #2	By Teammate #3	By Teammate #4
1. I was responsible for setting up the Google docs, making a code 2. I contributed by adding ideas to the overall project and doing research 3. N/A [3, 3, 3, 3, 3]	Gwen contributes good ideas to the group and is a strong researcher. [3, 3, 3, 3, 3]	Solves a lot of the problem and participates and collaborates with team well [3, 3, 3, 3, 3]	Gwen usually takes care of the word file. She makes sure that all the required information is entered comprehensively and elaborately. She is also usually the first one to show up during group meetings which show that she is very committed. [3, 3, 4, 3, 2]

Note: numbers in brackets are the students' ratings on the C, I, K, E, and H dimensions respectively.

Gwen's gave herself 3's on each dimension. We gave her a C3 for contributing and a "content" rating for stating what she did. This particular comment also represented a pattern that we saw among the comments, specifically that the students stated what they or their teammates completed, what they or their teammates contributed to, and occasionally what they or their teammates may have been assigned but did not do ("3. N/A"). This style would generally be used by only one member of a team, but it was present on five different teams. These comments are activity driven rather than behavioral and are of the form that some Engineering 1 instructors required the semester before. Before peer-to-peer comments were an option in CATME, "confidential comments to instructor" were the only way to provide any comments on team or individual performance. Based on the comments that we read, some students (or perhaps some faculty members) may not have understood the shift in expectations for the content of these comments. What this also shows is that the quality of the comments and the ability to derive information about team behavior from the written comments, depends in some part on the instructions given to the students by instructors when requiring the comments.

Teammates #2 and #3 likewise gave her 3's on each dimension. Based on Teammate #2's comments, we both gave Gwen an H3 for having the required KSA's (due to being a strong researcher); One of us gave her a C3 (because the author so stated) and one gave her an I3 for communicating her ideas. Based on Teammate #3's comments, one of us gave her a C3 (because problem solving and participation are expected) and one gave her an I3 (because she collaborates well).

Teammate #4 was the only one to provide much constructive feedback and to offer any discrimination of the ratings. We were able to code the C and E dimensions from this comment, although we disagreed on the level, with one providing a 4 on both dimensions (due to the adverbs "comprehensively" and "elaborately" and being "very committed") and one a 3 on both dimensions, considering those behaviors to be the expected standard.

4.2.2 Sufficient Detail

When team members are expected to and actually do explain their ratings on each dimension, a much clearer picture of team and team member behavior becomes apparent. In one team, we were able to provide ratings for 48 or 49 dimensions from the comments. It was for this team that we had six of the seven comments that we were able to code 8 or more between us. Arjun contributed four of these six. His comments were both constructive and instructive, although we rated the team members lower than he did based on the comments.

Table 2. Arjun's Detailed Comments about Himself and his Teammates

About Taylor	About Himself	About Ragaav	About Fran
Taylor puts in a lot of effort and gains required skills by searching for the required techniques all over the internet. Expects the team to complete task with acceptable quality. Constructive task-related interaction with teammates can be improved. Has skills but cannot replace other team members. Is not able to help other team members. Is able to finish assigned task with sufficient quality. [4, 3, 3, 4, 4]	I always contribute as much as I can to the project. I believe that the team can do excellent work and I motivate all of us to give the task our best shot. I provide constructive feedback to the team and help them when they have problems in their assigned task. I care that the team does well and I notice things that can potentially derail us. I have extensive knowledge or I gain required skills. I value all the team members and their ideas. I try to make sure we stay focused on the task. [4, 3, 5, 5, 5]	Has required skills and knowledge. Cares that the team completes task with acceptable quality. Able to contribute to task completion. Does not care about doing top-notch quality work. Encourages the team and makes sure everyone knows what to do. Interacts well with teammates. Sometimes chips in with other work. Cannot entirely replace other members. [3, 5, 4, 3, 5]	Fran puts in a lot of effort and gains required skills by searching for the required techniques all over the internet. Expects the team to complete task with acceptable quality. Interacts well with the team and values everyone's opinions. Is able to finish assigned task with sufficient quality. Has skills but cannot replace other team members. Is not able to help other team members. [4, 4, 4, 3, 4]

Note: numbers in brackets are the students' ratings on the C, I, K, E, and H dimensions respectively.

In general, we gave Arjun's teammates lower ratings than he did based solely on the information provided in the comments. We rated Taylor at least a point lower on each dimension. Both of us gave her a C3 and an E3 since the comments indicate that she seemed to do fine, but nothing special. We each gave her an I2 due to the need for improvement in this area. One of us gave her an H3 and one a H2. The lower rating was due to not being able to replace other team members but the 3 was warranted for her willingness to acquire the skills she needed. We gave Ragaav nine 3's and one 2 for Expecting Quality, since he did not care about "doing top-notch quality work." All of the other behaviors, as described, meet expectations, but we did not consider them to be special. Likewise, we both gave Fran 3's for all dimensions except Keeping the Team on Track, which we were unable to code.

Arjun's comment about himself indicates that he used the rubric (see Appendix) as a guideline for writing them. He addressed each of the five dimensions and used terminology found in the rubric For example, "**motivating** us to give the task our best shot" and "**caring that the team does well**" (E5). Our ratings of him were closer to his ratings of himself than we were to his ratings of his teammates. One of us gave him 5's on each dimension the other gave him a C5, an E5 and an H5, but only 3's for Interacting and Keeping the Team on Track. The lower ratings were due to discerning that "valuing" teammates' ideas is not the same as soliciting them and "noticing" problems is not the same as "making sure" that teammates stay on task, which are required for 5's.

When asked to explain his ratings, Arjun appears to be frank about the skills he and his teammates bring to the task at hand, but his ratings for everyone except himself seem to be higher than his comments would warrant. Perhaps he is more comfortable explaining his own behavior. There could also be a social compact among team members to rate each other highly (all ratings were 3 or higher) and his ratings were not distinguishably different from the other students about each other. Another possibility was that Arjun misunderstood how the behaviors should be reflected in the ratings.

In contrast, when Fran rated Arjun, her ratings of his behavior were on par with or a little lower than the comments would indicate.

Arjun also has a lot of knowledge concerning concepts relevant to the milestones. He shows up to every meeting on time and ready to expect quality work. Since the last CATME, he has completed the executive function. He has also contributed to making our code more efficient. Arjun has been a huge help in keeping the team on track and laying out our meeting goals before the start of every meeting. [4, 5, 4, 5, 5]

We agreed on ratings of C5, K5, and H5. One of us gave him a 5 and the other a 3 for Expecting quality and one of us gave him a 3 for Interacting for participating fully in team activities.

This team clearly had norms for detailed comments about themselves and each other. Fran's comments about the other teammates were sufficiently detailed for us to code at least 6 items for each team member and Taylor's comments were sufficiently detailed for us to code at least 4 items for each team member. While Ragaav made relatively detailed comments about himself, his comments about the others lacked detail. We infer that this is due to instructor guidance along these lines, although we have no direct evidence to support this assertion. Another explanation would be that Arjun had work or other experience in which he learned how to give constructive feedback which he shared with his team. However, we were able to provide an average of 39 ratings to a team in another section taught by this instructor, second most of the 15 teams, providing further evidence that instructor expectations for student comments matter.

4.3 Explanation of CATME Exception Codes

One feature of CATME is to flag student ratings that are out of line with those of other team members or that indicate a problem with the team. Students are notified if they meet the criteria for the exception codes. The seven exception codes are:

1. Manipulator – tries to skew the curve by rating themselves highly and other team members poorly.
2. Underconfident – rates oneself at least a point lower than teammates rate them.
3. Overconfident – rates oneself at least a point higher than teammates rate them.
4. High Performer – all team members rate the person very highly and higher than other team members
5. Low Performer – all team members rate the person very low and lower than they rate other team members.
6. Cliques – members of a team appear to divide into groups based on how they rated the team members.
7. Conflict – one team member does not appear to get along with the others and has uniformly low ratings relative to the others.

By looking at the student ratings of each other and the student comments as we rated them, we were occasionally able to infer the appropriate CATME exception code without knowing them in advance. In particular, we were able to discern conflict, cliques and overconfidence as it related to low performance. However, because CATME only returns one exception code per student using a heirarchy, we could also infer other types of behavior beyond what CATME flagged based on looking at the ratings together with the comments the students made about each other.

4.3.1 Team Dynamics

The comments for the team shown in Table 3 offer tremendous insight into that team's dynamics, including interpersonal conflict, cliques, and possible sexism or cultural clashes. On this team, CATME flagged conflict between Evan and Habiba as each gave the other low ratings (1's and 2's) on every dimension. Habiba defended her low ratings of Evan by specifically mentioning the behaviors (being bossy, mocking others' ideas, not showing enthusiasm, etc.). Evan, on the other hand, gave Habiba similarly low ratings, but only mentioned the activities that she participated in. We felt that his comment about her warranted a C3 and one of us also thought it deserved an H3. Not only the comment itself, but also his lack of detail, failed to justify his low ratings.

Looking more deeply at the comments and ratings, there appears to be animus between Evan and both women. He gave both of them low ratings while at the same time rating himself and Matthew much more highly. In spite of this, his comments about all of his teammates were essentially the same and we gave each of them a C3. In addition to Habiba's negative comments and ratings, Farah offered negative comments about Evan's communication skills and his disregard for her input. When combined with Habiba's comments, we can infer that the women believed that Evan did not treat them or their ideas with respect. We surmise that there was no conflict flag between Farah and Evan because she gave him mostly 3's with a 2 for interacting. She also gave Matthew all 3's, which would be unremarkable in itself, but she gave herself and Habiba all 4's in her ratings. Similarly, Habiba gave Matthew and Evan all 1's and 2's, but gave herself and Farah 3's or higher. There is little additional insight to be gained from Matthew's ratings, since they were mostly 3's for everyone with some 4's for Evan, however, we agreed that his comment about Farah yielded a C2 and one of us thought that his comment about Habiba warranted a C2 as well as an I2; the other gave it a C3. His comments about Evan were much

more positive than they were about Habiba and Farah – we gave them 3's and 4's for Contributing and Keeping the Team on Track. Taken together, these comments and ratings lead us to speculate that there was a men versus women dynamic on the team or possibly an ethnic clash to the extent that their names are indicative of different cultures. Perhaps it was both.

Table 3: Comment Matrix for a Team with Conflict

About→ By ↓	Matthew	Evan	Habiba	Farah
Matthew	I believe that this project overall has helped me learn more about MATLAB and coding than anything else previously did this semester [3, 3, 3, 3, 3]	Evan is a very good leader for our group. He is always on top of things and finishes his work whenever we have any. [3, 4, 4, 4, 3]	Habiba is often very hard to communicate with when we are trying to organize meetings. She has attended almost all of them however and has done almost all the parts of the assignments assigned to her. [3, 3, 3, 3, 3]	Farah responds to almost all messages we have about meeting however she was a little late to tell she couldn't come to some meetings. She completed her parts of the project we have had due up until now. [3, 3, 3, 3, 3]
Evan	Matthew completed the regression code and the SSE code for M3 as well as revised the codes for M4. He contributed to the executive function and several other parts of M3. Matthew completed all work assigned to him on time. [4, 3, 3, 3, 5]	I wrote several parts of our code, including the algorithm, the refined algorithm, and the code to evaluate arbitrary numbers in the algorithm. Furthermore, I wrote a majority of the improvement information for M3. I contributed to the regression code and the executive function. I completed all tasks that were assigned to me on time [5, 3, 4, 4, 5]	Habiba help to write the executive function. She also contributed to some of the graphs that we used to analyze our solution in M3. [2, 2, 1, 2, 2] [conflict]	Farah help to created the graphs that we used in our evaluation of our code as part of M3. She also helped to write the executive function for M3. [2, 2, 1, 2, 2]
Habiba	Completes his work but does not contribute to team discussions not willing to work with all members of the team and does not care if the team meets its goal or not does not show interest in the work but finishes the work assigned to him [2, 2, 1, 1, 2]	Does not communicate well with other team members. Bosses and does not want to discuss ideas with other team members. Does not show respect to others ideas and mocks them. Not willing to share ideas with other team mates or help them. Didn't show any enthusiasm and didn't care about the quality of the work in M4. Not excited about meeting with the team and doing work does not ac-	Does fair share of the work and finish the work assigned to me. Communicates with others and accepts other ideas. Trying to resolve team issues. Respects others ideas and willing to work with any team member. Completed my work in milestone 3 and 4 was responsible for the technical report in milestone 4. [3, 5, 4, 3, 3]	Willing to work with other team members. Cares about meeting with other people to finish the work. No problems arise when working with her. Finishes the work assigned to her before deadlines cares about completing assignments and getting all the points possible willing to discuss ideas and respects other team members working with her is productive and

		cept comments from other team members does not accept feed-back not willing to change his work even of there are mistakes [2, 1, 1, 1, 2] [conflict]		there is no tension unlike working worh other members [3, 3, 3, 3, 3]
Farah	Matthew is probably the only person in the team that I have a very little communication with. I believe that he is good at completing the tasks and giving his knowledge. However, I rated him a little low on the areas about communication because I almost have no communication with him. Other than that he finishes the tasks on time and he usually attends to most of the meetings [3, 3, 3, 3, 3]	Evan is unfortunately the only person in the team that I have nega-tive communication with. We don't get along very well and that is affecting our com-munication a lot. I feel like he doesn't pay at-tention to my contribu-tions, my ideas or my questions. That is why I rated him low in some of the categories. Other than that, he is good at finishing the tasks on time and being to the meetings on time. [3, 2, 3, 3, 3]	I also feel like Habiba has improved a lot. She is focused during the class and finishes the tasks on time. She is very helpful about explaining the milestones of the project if I miss something or if I don't understand. She pays attention to my ideas and questions and respects them which is the biggest reason I gave her good ratings. [4, 4, 4, 4, 4]	I think I've improved myself since the last evaluation. I was on time for the meetings and did all of the parts of the tasks that were assigned to me. I be-lieve I am trying my best to be beneficial for my team. I am still having conflicts with some of the team members which will probably affect the ratings given to me. Unfortunately, I think some of the team members won't realize the improvements about me. [4, 4, 4, 4, 4]

Notes: pairs of comments share the same color (e.g., #2 and #3's comments about each other are in red). Self comments are in gray. Numbers in brackets are the students' ratings on the C, I, K, E, and H dimensions respectively. CATME exception codes, if indicated, follow the ratings.

4.3.2 Overconfidence

One exception code that the comments inform well is "overconfident." Colin's teammates gave him generally low ratings (1s and 2s) and their comments supported their ratings as shown in Table 4.

Table 4: Comments about Colin Reveal His Overconfidence

Teammate #1	Teammate #2	Teammate #3	Colin
Colin is my paired pro-gramming partner. He usually attempts to help me program, however he usually does not take an initiative to take a role in the team. I feel like if he took more of an initiative when paired programming or working on a milestone, we would be a more effi-cient team. Otherwise he is a good teammate. He	Colin doesn't do anything ever. He's either on his phone or just blankly staring at us. [1,2,1,1,2]	Colin is apathetic and unproductive. He will occasionally contribute to a google doc but often contributions are minimal or must be supplemented by [Teammate 2] or me. [1,2,1,2,2]	This class was very hard for me because of my heavy workload and because I struggled in picking up on MATLAB. I would go to office hours and try but I just wasn't good at using the program. [3,4,3,3,3] [over]

always tries to research solutions to issues we encounter. [2,2,2,3,2]			

Note: numbers in brackets are the students' ratings on the C, I, K, E, and H dimensions respectively. CATME exception codes, if indicated, follow the ratings.

Interestingly, even Colin, through his comments, indicated that he was struggling with the material. Yet, he gave himself at 3 for Having the KSA's required to be a good team member. One of us gave him a 2 for trying to gain the skills by going to office hours and the other gave him a 1 for his lack of success. Clearly, this would be a case where an intervention on the part of a faculty member might be required, both to help Colin with his skills and to help him gain a realistic perspective on his performance as a teammate.

5. Discussion and Conclusions

Peer feedback through ratings and comments can be a very powerful tool for both assessing and improving teamwork behavior. When done constructively, team members have the opportunity to learn where they are doing well and where they can improve. Yet as we saw in this study, high quality peer comments are a rare thing among first year engineering students. Fewer than half of the 427 teams that semester met the criteria simply to qualify for further analysis, that is, they did not have all four team members provide written feedback about each other on both occasions. To be sure, there were undoubtedly acceptable reasons in a number of these cases, including odd numbers of students in a section leading to 3 person teams, excused absences, and the inevitable drops by the later part of the semester. Even taking these reasons into account, however, there were likely many more cases of simple non-compliance as no grade was associated specifically with completing comments in this Engineering 2 course.

We can infer from the lack of detail in many comments as well as comments that centered only on the Contributing dimension, that students simply lack an understanding of how to provide constructive comments related to all five of the CATME dimensions. We know that students were provided with Frame of Reference Training early in the semester to calibrate their actual ratings to expected behaviors as represented by the rubric. However, no such training was provided at the time peer to peer comments were added to the CATME system and expected of students. An announcement was made to Engineering 2 instructors about the availability of peer to peer comments and they could then introduce it as they wished. It seems likely that at least one instructor set clear standards for high quality comments that were followed by Arjun's team. Having reviewed hundreds of comments, we know that this level of quality was out of the ordinary throughout the different sections and must have been taught. We suggest that instructors learn from this one how to effectively introduce peer to peer feedback, although knowing exactly what she did is out of the scope of this paper.

Where good quality written comments are provided, they generally are congruent with peer ratings and also very informative about individual dysfunctionality (e.g., overconfidence) or dysfunctional team behaviors (e.g., cliques). High quality comments can shed light on dysfunctional team behavior that might not be evident from the ratings alone as evidenced by the team in Table 3. The men's comments did not generally justify their ratings, but the women's comments were clear and to the point about the behaviors exhibited by the men toward them in

that team. By looking at the comments and the ratings together in matrix form, the patterns became clearer and showed not only the conflict between Habiba and Evan flagged by CATME, but also the antipathy between the men and women on that team. A team that exhibited these behaviors should warrant intervention on the part of the instructor.

The value of written peer to peer feedback to students and instructors is only as good as the constructiveness of the comments themselves. Students, particularly those in their first year in college, cannot be expected to know how to provide constructive comments. They must be taught. We suggest that instructors consider the model presented by Hattie and Timperley [9] as they instruct their students on how to provide feedback that is effective in helping team members know "how they are going." We recommend, therefore, that when using peer to peer comments in CATME, instructors teach constructive feedback skills and expect students to exhibit them in the comments that they provide. If such training is provided to the students, further research should be done to determine if the training was effective for eliciting constructive feedback from students.

Limitations

Even with a rubric, getting exact agreement on short snippets of comments is difficult. While the absence of a code is relatively easy to discern (we did so nearly 700 times), coding dimensions and levels from short statements sometimes requires inference. For instance, if a person lists a lot of activities that they completed, they haven't mentioned behaviors, per se (what the ratings are based on); however, they have given evidence of making a contribution to the team. That is why the Contributing dimension was the most often coded, but also why we sometimes disagreed on the level of contribution. We also recognize that the absence of so many student comments suggests a bias in our data that is hard to resolve or avoid.

References

[1] ABET, "Accreditation policy and procedure manual," 23 November 2016. [Online]. Available: http://www.abet.org/wp-content/uploads/2016/12/A001-17-18-Accreditation-Policy-and-Procedure-Manual-11-29-16.pdf. [Accessed 16 March 2018].

[2] M. Loughry, M. Ohland and D. Woehr, "Assessing Teamwork Skills for Assurance of Learning Using CATME Team Tools," *Journal of Marketing Education* , vol. 36, pp. 5-19, 2013.

[3] M. Loughry, M. Ohland and D. Moore, "Development of a Theory-Based Assessment of Team Member Effectiveness," *Educational and Psychological Measurement,* 2007.

[4] M. Ohland, M. L. Loughry, D. Woehr, L. Bullard, R. Felder, C. Finelli, R. Layton, H. Pomeranz and D. Schmucker, "The Comprehensive Assesment 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.

[5] L. Atwater and D. Waldman, "Accountability in 360 Degree Feedback," *Military and Intelligence Database Collection,* p. 96+, 1998.

[6] D. Bracken and D. Rose, "When Does 360-Degree Feedback Create Behavior Change? And How Would We Know It When It Does?," *J Bus Psychol,* pp. 183-192, 2011.

[7] R. Parrigin, "Accountablility and Professional Development: Use of the 360-Degree Feedback Appraisal," Order No. 3401294, Ann Arbor, 2009.

[8] C. Viswesvaran , F. Schmidt and D. Ones, "The Moderating Influence of Job Performance Dimensions on Convergence of Supervisory and Peer Ratings of Job Performance: Unconfounding Construct-Level Convergence and Rating Difficulty," *Journal of Applied Psychology* , p. 345–354 , 2 87.

[9] J. Hattie and H. Timperley, "The power of feedback," *Review of Educational Research,* vol. 77, no. 1, pp. 21-112, 2007.

[10] K. Rodgers, A. Horvath, H. Jung, A. Fry and H. Diefes-Dux , "Students' Perceptions of and Responses to Teaching Assistant and Peer Feedback," *Interdisciplinary Journal of Problem Based Learning,* vol. 9, no. 2, 2014.

[11] D. Schwab, H. Heneman and T. DeCotlls, "Behaviorally Anchored Rating Scales: A Review of the Literature," *Personal Psychology,* pp. 549-562, 1975.

[12] R. Jacobs, D. Kafry and S. Zedeck, "Expectations of Behaviorally Anchored Rating Scales," *Personal Psychology,* pp. 595-640, 1980.

[13] J. R. Fraenkel and N. E. Wallen, How to Design and Evaluate Research in Education, New York: McGraw-Hill, 1993.

[14] M. B. Miles, A. M. Huberman and J. Saldana, Qualitative Data Analysis, 3rd ed., Los Angeles: Sage, 2014.

Appendix
CATME Teamwork Rating Scale

Score	Contributing to Team's Work	Interacting with Teammates	Keeping the Team on Track	Expecting Quality	Having Related Knowledge, Skills, and Abilities
5	• Does more or higher-quality work than expected. • Makes important contributions that improve the team's work. • Helps teammates who are having difficulty completing their work.	• Asks for and shows an interest in teammates' ideas and contributions. • Makes sure teammates stay informed and understand each other. • Provides encouragement or enthusiasm to the team. • Asks teammates for feedback and uses their suggestions to improve.	• Watches conditions affecting the team and monitors the team's progress. • Makes sure that teammates are making appropriate progress. • Gives teammates specific, timely, and constructive feedback.	• Motivates the team to do excellent work. • Cares that the team does outstanding work, even if there is no additional reward. • Believes that the team can do excellent work.	• Demonstrates the knowledge, skills, and abilities to do excellent work. • Acquires new knowledge or skills to improve the team's performance. • Able to perform the role of any team member if necessary.
4	Demonstrates behaviors described immediately above and below.				
3	• Completes a fair share of the team's work with acceptable quality. • Keeps commitments and completes assignments on time. • Helps teammates who are having difficulty when it is easy or important.	• Listens to teammates and respects their contributions. • Communicates clearly. Shares information with teammates. • Participates fully in team activities. • Respects and responds to feedback from teammates.	• Notices changes that influence the team's success. • Knows what everyone on the team should be doing and notices problems. • Alerts teammates or suggests solutions when the team's success is threatened.	• Encourages the team to do good work that meets all requirements. • Wants the team to perform well enough to earn all available rewards. • Believes that the team can fully meet its responsibilities.	• Demonstrates sufficient knowledge, skills, and abilities to contribute to the team's work. • Acquires knowledge or skills as needed to meet requirements. • Able to perform some of the tasks normally done by other team members.
2	Demonstrates behaviors described immediately above and below.				
1	• Does not do a fair share of the team's work. Delivers sloppy or incomplete work. • Misses deadlines. Is late, unprepared, or absent for team meetings. • Does not assist teammates. Quits if the work becomes difficult.	• Interrupts, ignores, bosses, or makes fun of teammates. • Takes actions that affect teammates without their input. Does not share information. • Complains, makes excuses, or does not interact with teammates. • Is defensive. Will not accept help or advice from teammates.	• Is unaware of whether the team is meeting its goals. • Does not pay attention to teammates' progress. • Avoids discussing team problems, even when they are obvious.	• Satisfied even if the team does not meet assigned standards. • Wants the team to avoid work, even if it hurts the team. • Doubts that the team can meet its requirements.	• Missing basic qualifications needed to be a member of the team. • Unable or unwilling to develop knowledge or skills to contribute to the team. • Unable to perform any of the duties of other team members.